DOI: 10.1111/epp.12884

EPPO STANDARD ON DIAGNOSTICS

PM 7/151 (1) Considerations for the use of high throughput sequencing in plant health diagnostics[1]

Specific scope: This Standard describes elements to take into consideration for the use of high throughput sequencing (HTS) tests, including validation, quality control measures and interpretation and reporting of results to ensure HTS test results are robust and accurate, have biological significance in a phytosanitary context, and are implemented in a harmonized way. This Standard applies to all plant pest groups and HTS technologies.

This Standard should be used in conjunction with PM 7/76 *Use of EPPO diagnostic protocols*.

Specific approval and amendment: Approved in 2022–09.

Authors and contributors are given in the Acknowledgements section.

1 | INTRODUCTION

High-throughput sequencing (HTS), also known as next generation sequencing (NGS) or deep sequencing, is one of the most significant advances in molecular diagnostics since the advent of the PCR methods in the early 1980s. HTS can potentially detect the nucleic acids of any organism present in a sample without any a priori knowledge of the sample's phytosanitary status (Hadidi et al., 2016; Massart et al., 2014). HTS can be used for targeted detection of regulated pests and can also help identifying pests causing novel diseases or diseases of unknown aetiology that might be a potential threat to plant health (Aritua et al., 2015; Barba et al., 2014; Malapi-Wight et al., 2016; Maliogka et al., 2018). As described previously (Olmos et al., 2018), HTS technologies open new possibilities and opportunities in routine diagnostics for (a) understanding the status of a pest in a region through surveillance programmes, (b) certifying nuclear stock and plant propagation material, (c) (post-entry) quarantine testing to prevent the introduction of pests into a country or area, and (d) monitoring of imported commodities for new potential risks. In HTS, the target organism(s) can be one or more variants, species, genera, families or groups of organisms (e.g. bacteria, fungi, viruses) that are being tested as individual specimens or isolates or for a range of matrices (e.g. plant, soil, water). In any case, the scope of the HTS test should be defined according to EPPO Standard PM 7/98 *Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity* (EPPO, 2021a).

Two different applications of HTS are used to detect and/or identify plant pests: amplicon sequencing (also called targeted sequencing or specific sequencing) and shotgun sequencing of nucleic acids (also called random sequencing). For amplicon sequencing, specific standardized genetic marker(s) (called barcodes) are amplified (mainly by PCR, although recent protocols used rolling circle amplification or LAMP) and sequenced. Barcode regions can be used for the identification of the organisms present in a sample at a certain taxonomic level. Barcodes have been proposed and described in EPPO Standard PM 7/129 *DNA barcoding as an identification tool for a number of regulated pests* (EPPO, 2021b) for arthropods, bacteria, fungi, nematodes, oomycetes, invasive plants and phytoplasmas by classical Sanger sequencing. Some of these barcodes have been successfully used in metabarcoding (Ahmed et al., 2019; Dormontt et al., 2018; Nilsson et al., 2019; Ritter et al., 2019; Tremblay et al., 2018). Given the high sequence diversity within plant viruses, no generic plant virus barcodes are available although conserved motifs within specific virus genera that allow virus identification have been identified. Shotgun sequencing consists of the random sequencing of any nucleic acid present in a sample, whatever its origin (e.g. pest, endophytic micro- and macroorganisms, host). Using shotgun sequencing can help to recover the whole genome of specific pests e.g. *Xylella fastidiosa* (Simpson et al., 2000) or *Pyrenochaeta lycopersici* (Dal Molin et al., 2018).

A recommendation on 'Preparing the use of high-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes' was adopted by the Commission on Phytosanitary Measures governing body of the International Plant Protection Convention (IPPC) in 2019. This recommendation encourages the development of best-practice operational guidelines covering result and quality control measures for HTS that

[1]Use of HTS in plant pest diagnostics is a new developing area, consequently the standard will be revised in 2024 based on experience following its use in laboratories until this date.

'ensure HTS data outputs are robust and accurate, have biological significance in a phytosanitary context, and are implemented in a harmonized way, test validation and quality assurance' (FAO, 2019). In addition, it highlights the need of validating HTS tests.

In line with this recommendation, the present Standard which was developed based on an outcome of the VALITEST project (Lebas & Massart, 2020; Trontin et al., 2021), describes specific elements to take into consideration for the use of HTS tests in laboratories. Specific challenges related to the laboratory procedures and bioinformatic analyses of the HTS test, to validation, to quality controls and to the interpretation and reporting of the results are explained and addressed when possible. These considerations are irrespective of the chemistry, instrumentation, and software, and apply to any plant pest in any matrix to allow flexibility within this fast-evolving technology. In this Standard, the HTS process has been divided into eight distinct steps (Figure 1; see also Section 3.2). For each of these steps a range of procedures has been developed and published and further improvements are expected in the future.

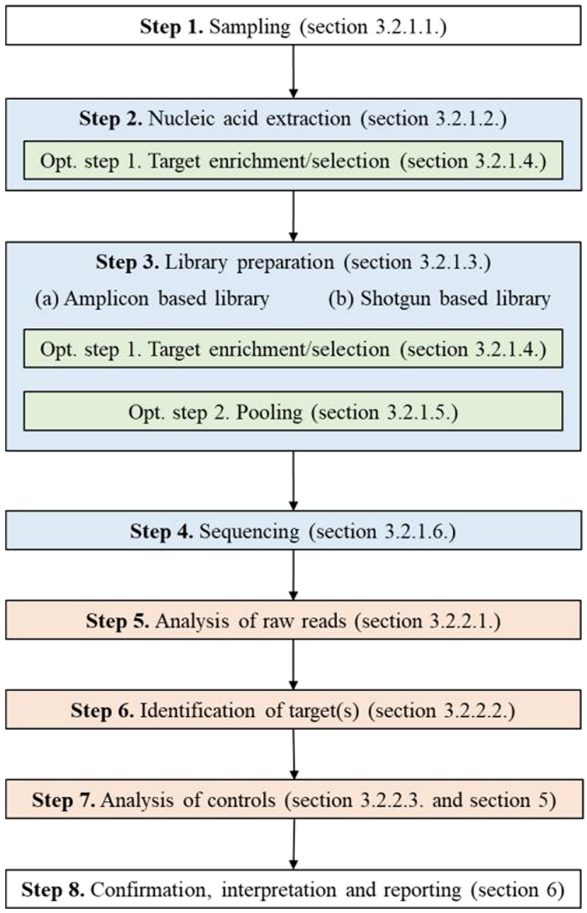

FIGURE 1 Scheme representing the eight main steps of the HTS tests as described in these guidelines. The laboratory steps are highlighted in blue and the bioinformatic steps in orange. Two optional steps are included in green: the target enrichment or selection and the pooling of samples.

2 | DEFINITIONS

*Only definitions specifically relevant to this Standard are included. Other definitions are included in PM 7/76 Use of EPPO diagnostic protocols (*EPPO*, 2018).*

Adapter: platform-specific oligonucleotide sequences attached onto target DNA molecules during library preparation that bind to, or are otherwise recognized by, the sequencing flow cell.

Amplicon sequencing: HTS test based on PCR amplification, such as metabarcoding. The PCR primers are usually designed to broadly amplify a specific genome region for a range of target organisms (e.g. bacteria, fungi, plants, viruses, insects, nematodes) and should be able to generate sequences from as many species as possible within this range.

Annotation: information describing properties and features of a sequence region; sequence annotation can be either taxonomic (e.g. giving a taxonomic rank) or functional (e.g. identifying functional element like coding region, intron, promoter, micro RNA (miRNA), long non-coding RNA (lncRNA), transposon, repeated sequences) depending on the intended use of the HTS test.

Amplicon sequence variants: single DNA sequences recovered from amplicon sequencing, following the removal of erroneous sequences generated during PCR and sequencing.

Background reads removal: a sub-step of the bioinformatic component of the HTS process in which non-target sequences are completely or partially excluded from the dataset. Also called background depletion or subtraction, reference subtraction or negative selection.

Background reads: sequences not related to the targets. These sequences may be for example (part of) host sequences and its associated microbiome, phage sequences, environmental contaminants sequences (e.g. bacteria commonly found in the air, on plant surfaces, in reagents).

Base quality scores: indicates the probability that a base is called incorrectly. Each base in a read is assigned a quality score by a Phred-like algorithm. A quality score of 10 means there is a 1/10 chance that the base call is incorrect (90% probability to be accurate); a score of 20 means there is a 1/100 chance that the base call is incorrect (99% probability to be accurate) and a score of 30 means there is a 1/1000 chance that the base call is incorrect (99.9% probability to be accurate; Illumina, 2022). Also called Phred quality score.

Bioinformatic pipeline: a suite of several pieces of software that usually follow each other in order to conduct the complete bioinformatic analyses.

Clustering: a bioinformatic operation (used in metabarcording and metagenomics) in which reads with

related sequences (e.g. similar genomic features, identical or homologous gene or protein) are grouped together.

Contiguous sequences: assembly of overlapping reads that together form a consensus region of DNA/RNA. Also called contigs.

De novo *assembly*: a computational process in which the HTS-generated reads are assembled into longer, continuous sequences, and sometimes (near) complete sequences without using a reference sequence (see definition).

Denoising: a bioinformatic operation (specific to metabarcording) in which reads with artefacts introduced during PCR amplification and sequencing (noisy sequences, e.g. nucleotide substitutions, length variation) are removed or corrected in order to preserve the highest quality reads.

Duplicated reads: identical reads generated during a sequencing run. Also called duplex reads.

Genome completeness: proportion of obtained sequence compared to the (near) complete genome. Also called completeness of the sequence, genome length coverage, horizontal coverage.

Index: a short sequence of oligonucleotides added during the library preparation when sequencing several samples in parallel. It is unique to each sample and allows the assignment of the generated sequences to the corresponding samples.

Index-hopping: a known phenomenon that causes incorrect assignment of reads to libraries from the expected index to a different index (in the multiplexed pool). Can also be called barcode[2] bleeding, barcode[2]-hopping, barcode[2] misassignment, cross-talk, index misassignment, index switching, miss-tag.

Library preparation: laboratory preparation of nucleic acids to make them compatible to the sequencing platform. There are two main ways to prepare the libraries for plant pest detection: the shotgun sequencing (also called random sequencing of nucleic acids) and the targeted sequencing of PCR products, also called amplicon sequencing (e.g. metabarcoding).

Metabarcoding: amplification and sequencing at high throughput of specific standardized genetic marker(s) that allows the simultaneous identification of many taxa within a single sample.

Metagenomics: study of genetic material recovered directly from environmental samples, typically untargeted (i.e. by shotgun sequencing). In case RNA is sequenced instead of DNA, it is often called metatranscriptomics.

N_{50}: 'the length of the smallest contig such that 50% of the sum of all contigs is contained in contigs of size N_{50} or larger' (Castro & Ng, 2017).

Operational Taxonomic Unit (*OTU*): OTU is a cluster of sequences based on their sequence similarities.

[2]Note that in that context, the barcode is referring to the index and not to a short standardized genetic marker used for species identification.

Poorly characterized organism: a known organism for which there are no existing tests for diagnosis.

Quasispecies (viruses): a population of closely related viral genomes of a virus within a host, and which act as a unit of selection.

Read: inferred sequence of nucleotides corresponding to a DNA fragment, resulting from a high-throughput sequencing experiment.

Read depth: number of aligned reads covering a specific nucleotide position. Also called vertical coverage, depth of coverage, coverage read depth, or coverage fold. Mean (or average) read depth is often calculated as follows: for a given contig, the mean read depth is the number of reads mapping that contig multiplied by the read lengths and divided by the contig length.

Reference mapping: a computational process in which sequences (reads or contigs) are compared to an existing reference sequence (or backbone sequence). Can also be called reference assembly or reference-based mapping.

Reference sequence: sequences [partial or (nearly) complete genome, gene] used to map or annotate the reads or contigs.

Scaffold: created by joining contigs together using additional information (introducing arbitrary N letters) about the relative position and orientation of the contigs in the genome (Jung et al., 2019).

Sequencing run: single use of a sequencing machine to sequence one or several libraries.

Shotgun sequencing: random sequencing of any DNA or RNA molecule present in a sample, whatever its origin: for example, pest, endophytic micro- and macroorganisms, host (e.g. plant). Also called random sequencing.

Strand bias: on a single genome position, a strand bias occurs when the proportion of reads from a forward sequence and from its corresponding pairs deviates from the expected result of an equal likelihood of sequencing the plus and minus strands.

Tagmentation: illumina defined this as the 'step included in shotgun library preparation which involves the transposon cleaving and tagging of the double-stranded DNA with a universal overhang'.

Trimming: removal of nucleotides at one or both extremities of reads. These nucleotides usually correspond to low quality nucleotides or to nucleotides added to the sample DNA (e.g. primers, adapters, indexes). The aim is to either to remove nucleotides not of interest or to keep reads and nucleotides of appropriate quality for further analysis.

U_{50}: 'the length of the smallest contig such that 50% of the sum of all unique, target-specific contigs is contained in contigs of size U_{50} or larger' (Castro & Ng, 2017).

Uncharacterized organism: an organism that has never been detected/identified before (e.g. new species or new variant/strain, organism new to science not yet taxonomically characterized) or whose biological properties (e.g. host range, transmission, symptomatology) are not known.

Unexpected organism: a known plant pest unexpectedly found in a new host.

Variants: single nucleotide polymorphism (SNP), insertion or deletion of nucleotides, integration or deletion of genes (structural variants) or homologous recombination observed in a sequence compared to reference sequence target(s).

3 | TECHNICAL REQUIREMENTS TO PERFORM HTS TESTS

The EPPO Standard PM 7/84 *Basic requirements for quality management in plant pest diagnostic laboratories* (EPPO, 2021c) describes basic requirements for quality management in plant pest diagnosis. The EPPO Standard PM 7/98 (EPPO, 2021a) further describes specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. The general management and technical requirements described in those Standards also apply to laboratories that intend to implement HTS tests. This Standard focuses on the requirements that are either specific to HTS tests or for which specific aspects should be considered when implementing an HTS test.

3.1 | General technical requirements

3.1.1 | IT equipment

The implementation of HTS requires significant investment in information technology and bioinformatics (Olmos et al., 2018). Large files, from a few megabytes to several gigabytes per sample, are generated and need to be stored and properly backed-up on remote internal and/or external servers for a duration of time that meets customer and legal requirements [see PM 7/77 *Documentation and reporting on a diagnosis*, EPPO (2019)].

To build the relevant IT infrastructure for the storage, analysis and transfer of data, the laboratory should consider:

- the expected number of samples,
- the sequencing capacity of the technology of interest (i.e. the maximum amount of gigabases theoretically produced per year with platforms operating at full use),
- the volume of data per sample depending on the files (i.e. raw generated reads, intermediate data files and final results) that need to be kept,
- the legal or commercial obligations related to data protection and privacy, especially when dealing with official testing and quarantine pests,
- the maintenance and data back-up,
- that a fast, stable and secured network is needed to ensure the integrity and the time needed for the data transfer,
- machines with appropriate computational power and operating system environment (e.g. Windows, MacOS, Linux) needed to run the bioinformatic pipeline.

Laboratories that do not have extensive data analysis capabilities can outsource the bioinformatic data analysis to external facilities or rent the computational power and storage space on commercially available computer clusters.

3.1.2 | Managing contamination

High-throughput sequencing tests are more prone to reveal contamination than other molecular tests because of their ability to detect nucleic acids from any organism. In addition, the multiple handling steps and the use of more reagents in the sample preparation process may introduce additional sources of contamination. Contamination can occur at different steps of the process in the laboratory due to poor sample handling or laboratory surface, reagent or equipment contamination (Asplund et al., 2019; Champlot et al., 2010; Dickins et al., 2014; Gaafar & Ziebell, 2020; Galan et al., 2016; Rosseel et al., 2014; Salter et al., 2014).

For example, contamination between successive runs of a sequencing machine, called carry-over contamination has often been observed (Quail et al., 2014). Technical advances partially overcome some of those contamination issues (see Section 3.2.1.5). In addition, contamination can occur when multiplexing several samples in a single sequencing experiment, i.e. the cross-contamination between prepared nucleic acids due to traces of other samples or index-hopping between samples (see Section 3.2.1.5; Buschmann et al., 2014).

The EPPO Standard PM 7/98 (EPPO, 2021a) provides guidance on how to avoid contamination with specific requirements for molecular laboratories and specific guidelines for monitoring contamination.

Although precautions are taken, some contamination can still occur. Therefore, the level of contamination should be monitored during the entire process from sampling to the analysis of data using relevant controls (see Section 5.2.1) and should be taken in consideration during interpretation of the results. It should be noted that the identification of contamination in the sequencing datasets is not yet standardized and many scientific, technical and bioinformatic developments are expected in the near future to improve this.

3.1.3 | Reference material (including reference sequence databases)

Reference material should be used for the validation of HTS tests and to monitor the performance of HTS tests. The production of biological reference material

should follow the EPPO Standard PM 7/147 *Guidelines for the production of biological reference material* (EPPO, 2021d).

Reference materials specific to HTS tests may be used, for example, artificial reference materials such as synthesized DNA/RNA (External RNA Controls Consortium 2005) or sequence datasets obtained from biological reference material (Brinkmann et al., 2019; Budowle et al., 2014; Massart et al., 2019; Trimme et al., 2015) or artificial reference datasets containing known target(s) (Tamisier et al., 2021). Such reference datasets should be stored and maintained properly and preferably be publicly available.

The selection of (an) appropriate sequence database(s) is important for the correct taxonomic, structural and/or functional assignment of the sequences obtained in an HTS test (see Section 3.2.2.2).

Sequence databases can be incomplete or contain errors. In addition, their content is constantly evolving because of scientific discoveries and changes in taxonomy. The use of inappropriate sequence database(s) can lead to incorrect results e.g. false negative results or misidentification of the detected organism (Brinkmann et al., 2019; Massart et al., 2019; Nilsson et al., 2019; Piper et al., 2019).

Sequence databases can either be publicly available (see Table 1) or can be developed and maintained by the laboratory. The choice of a specific database depends on the intended use of the HTS test. For example, when the focus of the HTS test is on a limited range of known pests, a curated database can be created with sequences of high quality that are accurately annotated and not redundant. When available, this database should also ideally include sequences obtained from documented reference material (for example, vouchered specimens). However, when searching for uncharacterized or unexpected organisms, a more extensive and less curated database might provide a much better chance for their discovery than a well curated database with a limited number of entries (Lambert et al., 2018; Piper et al., 2019).

Sequence databases should be evaluated for their ability to identify at least some of the expected target(s) of the HTS tests. The information on the sequence database used should be documented and include the version number, the date of download, the download source. It is important to note that the content of downloaded database might change, and sequences or associated metadata might not be retrievable at a later date. The laboratory needs to make sure that for each new release, the database is still fit for purpose. For example, the laboratory needs to make sure that (a subset of) target organisms are still part of the databases.

Note that ensuring that the sequencing data obtained by different laboratories employing HTS are shared with the entire diagnostic community (e.g. in different databases) will contribute to a reliable and better identification of pathogens threatening plant health.

3.2 | Requirements specific to particular steps of the HTS tests

All the steps listed in this section should be described and standardized into a standard operating procedure (SOP).

TABLE 1 Examples of sequence databases (in alphabetical order).

Name of the database	Description/type of sequences	Url
BOLD	Barcode of DNA species	http://www.boldsystems.org/
EMBL-EBI	Wide range of sequences including plant pests	https://www.ebi.ac.uk/services
ENA	Wide range of sequences including plant pests	https://www.ebi.ac.uk/ena/browser/home
EPPO-Q-bank	Curated sequences of plant pests	https://qbank.eppo.int/
EzBioCloud	Bacteria and archaeal sequences	https://www.ezbiocloud.net/
GenBank	Wide range of sequences including plant pests	www.ncbi.nlm.nih.gov
Genome Taxonomy Database	Bacteria and archaeal sequences	https://gtdb.ecogenomic.org/
GreenGenes	Bacteria and archaeal sequences	https://greengenes.secondgenome.com/
InsectBase	Insects	http://v2.insect-genome.com/
JGI	Fungi	https://genome.jgi.doe.gov/portal/
NEMBASE4	Nematode sequences	http://www.nematodes.org/nembase4/
NemaGene	Nematode sequences	http://nematode.net/NemaGene/
SILVA	Ribosomal RNA sequence data	https://www.arb-silva.de/
UNITE	Eukaryotic nuclear ribosomal ITS regions	https://unite.ut.ee/
WormBase	Nematode sequences	https://wormbase.org/#012-34-5

3.2.1 | Laboratory component

3.2.1.1 | Sampling and sample handling

See EPPO Standards PM 7/84 (EPPO, 2021c) and PM 7/98 (EPPO, 2021a).

3.2.1.2 | Nucleic acid extraction

The selection of the extraction method depends on the type of nucleic acid (e.g. DNA or RNA) of the target(s), the fragment size required by the sequencing platform and the type of matrix from which the nucleic acids are extracted (e.g. seed, leaf, stem, purified cultures, soil, water, insects). In most cases, nucleic acid extraction protocols for PCR or real-time PCR purposes are suitable for HTS tests, particularly for amplicon-based HTS tests. However, some library preparation protocols (e.g. used with long-read HTS technologies) require a higher nucleic acid integrity and minimal concentration.

3.2.1.3 | Library preparation

The selection of the protocol for library preparation depends on the HTS test used.

For shotgun sequencing, several protocols are available, often provided as kits with all the reagents included. The appropriate protocol should be chosen depending on the sequencing technology, technical criteria (e.g. minimum required quantity and integrity of the extracted nucleic acid and expected proportion of target nucleic acid), the time needed, the staff required, the costs of reagents and consumables.

For amplicon sequencing which usually relies on a PCR step, special care should be taken for the selection of primers to ensure the target organisms will be amplified (Scibetta et al., 2018). A high-fidelity polymerase should preferably be used to minimize errors due to the miss-incorporation of nucleotides (Budowle et al., 2014; McInerney et al., 2014). The number of PCR cycles should be selected to ensure the PCR is still in the exponential phase (e.g. usually 25–30 cycles for a quantitative metabarcoding test).

3.2.1.4 | Target enrichment for shotgun sequencing

When the amount of (a) target(s) in a sample is expected to be very low compared to background sequences (e.g. host or non-target organisms present in the sample) the nucleic acid extraction protocol may include a target enrichment or selection step to improve the analytical sensitivity of the HTS test. For example, in water samples, the enrichment of the target(s) was found to be essential for the detection of some viruses (Mehle et al., 2018).

The selection of the enrichment protocol depends on the target genome (e.g. ssRNA, dsRNA, total RNA, circular DNA for viruses), its physical properties (e.g. viroid naked RNA, encapsidated viral RNA/DNA, DNA of bacteria and fungi protected by a cell wall), the matrix (e.g. plant material, soil, water). Three examples of protocols that can improve the analytical sensitivity of the HTS test for plant samples (Adams & Fox, 2016) are:

- viral particle enrichment by ultracentrifugation before nucleic acid extraction,
- depletion of ribosomal RNA (rRNA) from total RNA or
- enrichment of dsRNA by cellulose affinity chromatography with or without additional nuclease treatment.

Rolling circle amplification is also frequently used as an enrichment procedure when targeting DNA viruses with circular genomes (Johne et al., 2009).

The enrichment of target nucleic acids can also be carried out during library preparation. It can be based on a size selection or on the use of specific oligonucleotides to either eliminate non-target nucleic acids (such as ribosomal RNA in plant samples) or to specifically select the target nucleic acids. For example, it has been shown that the removal of plant ribosomal RNA by specific oligonucleotides resulted in a 10-fold enrichment of viral sequences (Adams & Fox, 2016).

3.2.1.5 | Pooling of libraries

Several libraries (samples) can be pooled together to reduce sequencing costs. During library preparation, nucleic acids extracted from each sample are tagged with an index so that each obtained sequence can be traced back to the original sample (Budowle et al., 2014; Piper et al., 2019).

The process of pooling increases the risk of misassignment of reads to a sample due to indexes cross contamination during library preparation and sequencing (i.e. index-hopping) or inter-run contamination when identical indexes are used in successive runs (Galan et al., 2016; Kircher et al., 2011; van der Valk et al., 2018). This risk is increased when high sequencing depths are obtained with pooled libraries (Budowle et al., 2014; Massart et al., 2019). Depending on the sequencing technology, index misassignments can also occur at the demultiplexing step due to sequencing errors on indexes. Finally, the creation of chimeric sequences due for example to the ligation of free adapters can also result in the misassignment of reads to a sample (Wright & Vetsigian, 2016).

When pooling samples, solutions to limit the misassignment of reads should be considered. On the Illumina platform, sample misassignments can be reduced by using dual indexes (Kircher et al., 2011) and almost abolished by using unique dual indexes (MacConaill et al., 2018). Another option is to use indexes that are sufficiently long and different, so that their identification is robust and tolerates several sequencing errors. Pooling libraries just prior to sequencing or adding a step to remove free adapters can also reduce these misassignment issues. The sequences of sets of indexes included in each run should be recorded for traceability purposes and to plan the succession of sequencing runs properly.

Pooling also requires that the amount of nucleic acid of each library in the pool is normalized in order to minimize the pooling bias resulting in uneven numbers of sequences between samples (Hébrant et al., 2018). The laboratory should be aware of the risk associated with pooling and demonstrate that the pooling strategy used, does not affect the test performance, as generating a lower amount of reads per sample can limit the analytical sensitivity, and pooling several libraries can trigger more contamination. The number of pooled samples depends on the desired read depth of the targets to be sequenced and should be optimized to ensure that the HTS test meets the criteria of its intended use (Hébrant et al., 2018).

3.2.1.6 | Sequencing platforms and methods

The laboratory should use a sequencing platform and length of reads best suited for the intended use of the HTS test and taking into account the following points:

- the total number of samples to sequence and minimal number of reads per sample (considering variability in the number of reads per sample),
- required test turn-around time (e.g. urgent testing for imported perishable material),
- total number of generated reads per sequencing run: it should be compared to the requested reads per sample and the number of samples in order to determine if a complete or partial sequencing run is required, which has an impact on the turn-around time,
- multiplexing capacity of the platform: is it compatible with the expected number of samples per batch?
- read length and type (e.g. single, paired, mate-pair): the choice for these two parameters will depend on the HTS test used, short single reads are appropriate for sRNA sequencing whereas amplicon sequencing might need the longest reads, provided the error rate is acceptable,
- error rate and type of error: the error rate varies between the sequencing platforms and between runs. It can be critical for some HTS technologies, such as for amplicon sequencing where a small number of errors in the sequence can modify its annotation, or for shotgun sequencing when SNPs are important,
- availability of bioinformatic support, laboratory resources and technical expertise and level of manufacturer technical support in order to solve (re-) occurring problems quickly,
- the downstream bioinformatic analyses (which depends on the number of reads, their length, their quality and accuracy; Budowle et al., 2014; Jennings et al., 2017).

A cost study may be carried out taking into account the previous criteria and also (i) the three main expenses involved in the operation of a sequencing machine: purchase, running (reagents and consumables, e.g. the cost per sequence) and maintenance; (ii) the personnel time and expertise needed to run and maintain the machine (Rehm et al., 2013). These considerations can be important for a decision to invest in a desktop or stand-alone sequencer or to outsource the sequencing step.

Sequencing platforms are regularly updated, and the laboratory should closely monitor these updates and evaluate their potential impact on the test results.

3.2.2 | Bioinformatic component

The bioinformatic component of an HTS test consists of using a combination of software to analyse the raw data. The results generated by the bioinformatic pipeline depend on the (version of the) different type of software used, the parameters and the thresholds applied, as well as the accuracy and completeness of the sequence database(s) used for sequence comparison (see Section 3.1.3). The impact of the bioinformatic pipeline on the identification of target(s) has been shown in a test performance study (Massart et al., 2019).

Many bioinformatic pipelines have been developed that can operate either on a Linux system, statistical programmes, web interface, as well as commercial packages or user-friendly open-source software. A current general trend is to simplify the use and the parameterization of these tools, making them usable without extended bioinformatic knowledge or, sometimes, as a 'one-click' solution. For such simplified pipelines, it is paramount that the personnel using them is competent and understands how the pipelines work to use them appropriately according to the data and the goal of the analysis.

The laboratory should keep track of software versions and updates/upgrades with algorithms and parameter settings and keep records of changes to the underlying operating systems which might affect how pipelines and tools perform (e.g. integrate a Log system to track all versions in the bioinformatic pipeline).

The bioinformatic component of the HTS test can be divided in three steps with several sub-steps which are described below. It should be noted that the result of each sub-step of the bioinformatic analyses depends on the selected parameters and metrics of the previous sub-step(s). For example, the selected minimal quality score of trimming from the first step of the bioinformatic analyses (see Section 3.2.2.1) can impact the quality of the reads assembly from the second step of the bioinformatic analyses (see Section 3.2.2.2).

3.2.2.1 | Analysis of raw reads

The analysis of raw reads consists of different sub-steps (Figure 2) detailed below. These sub-steps may not all be relevant and may be performed in a different order depending on the HTS test (e.g. single species vs metagenomics, short vs long reads, single-end vs paired–end reads).

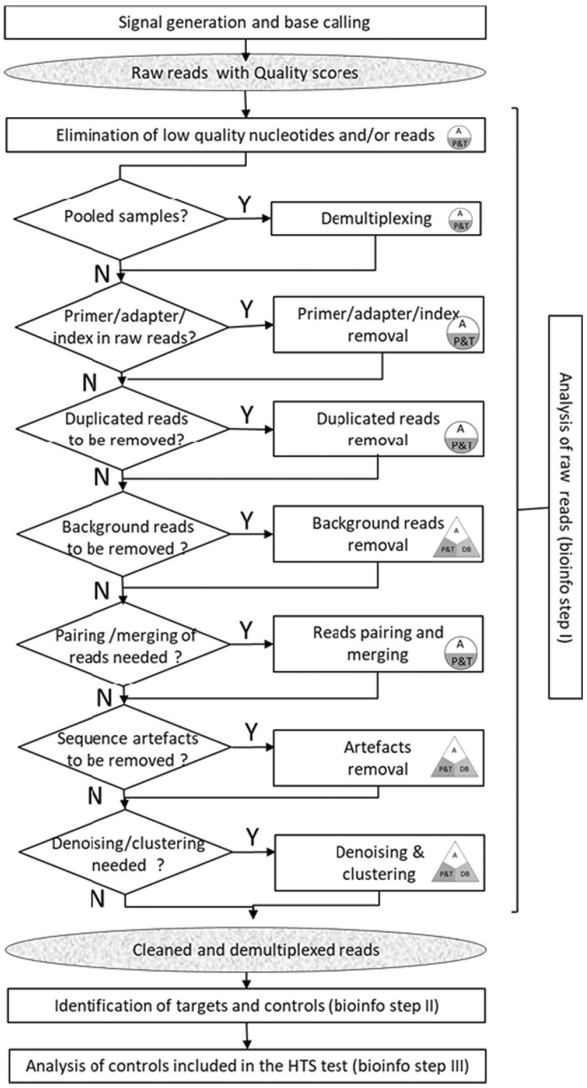

FIGURE 2 Overview of the step 'analysis of raw reads'. The order of sub-steps can be modified, depending on the bioinformatics pipeline that is used (for example the elimination of low-quality reads and nucleotides can be carried out at any time during the process). The rectangles correspond to operations/calculation while the grey ellipses correspond to file(s) containing sequence information and generated by the analysis. The triangles and circles represent the influence on the generated results of bioinformatics triad and duet respectively. Meaning of acronyms: A, Algorithm; DB, Database; P&T, Parameters and thresholds. More information about each of the steps can be found in the text.

The sub-steps needed, and their order should be defined during test development with their parameters and corresponding quality metrics and thresholds (Hébrant et al., 2018; Weiss et al., 2013). When implementing the HTS test in routine diagnostic activities, if the thresholds are not met, a decision concerning whether to repeat part of the HTS test or to proceed with it should be made and documented.

The first step of the bioinformatic analyses is to check the overall quality of the sequencing dataset by looking at the metadata produced during the sequencing run (e.g. cluster densities, quality profiles, number of and size of reads) and the specification metrics. These run quality metrics are platform-dependent, and the most relevant ones should be determined during development with the setting of (a) minimal threshold(s) (Hébrant et al., 2018). The run quality metrics and associated thresholds may be adjusted during validation. Note that the analysis of the run quality metrics can be carried out after the trimming of primers, adapters, and indexes.

Then, the quality of the raw reads should also be checked. The objective of quality filtering is to retain sequences of appropriate quality for the next steps of the bioinformatic analyses (Budowle et al., 2014; Hébrant et al., 2018; Weiss et al., 2013). Nucleotides or reads whose quality does not meet an established threshold should be removed, when relevant together with its pair. The reads quality is checked using base quality scores (for example, Phred quality score) which can vary depending on the sequencing platform. A minimal threshold of the base quality scores should be defined during development and validated during the validation procedure. The choice of an optimal threshold for read trimming is always a trade-off between sequence loss and dataset quality (Del Fabbro et al., 2013). This score is logarithmically related to the base calling error probability which is used to measure the quality of the identification of each nucleotide by the sequencing platform (Lambert et al., 2018).

Other sub-steps may have to be considered depending on the HTS test:

- *Demultiplexing*: If several libraries were pooled for sequencing, the reads are assigned in silico to their respective samples of origin by cross-checking the index sequences associated with each read (Budowle et al., 2014; Hébrant et al., 2018). The laboratory should be aware of the mismatch tolerance used, so that the tolerance of index errors does not cause misassignment of the reads. It is also possible to search for index sequences that have not been used in the sequencing run to estimate and filter cross-contamination that may have occurred during the indexing or sequencing steps (i.e. inter-run contamination; Galan et al., 2016; Kircher et al., 2011; van der Valk et al., 2018). Misassignment can occur during this step due to errors in index sequencing and inappropriate bioinformatic parameters (for example mismatch tolerance).
- *Primer, adapter and indexes removal* (also known as clipping, trimming): primers, adapters and indexes that are included in the sequence of the reads generated should be removed before continuing the bioinformatic analyses (Davis et al., 2013; Hébrant et al., 2018). The removal of indexes is usually done during the demultiplexing step (see above).
- *Duplicated reads removal*: Duplicated reads originate from the same amplified fragments. Their characteristics are common coordinates (e.g. the same start and

end coordinates after mapping), same sequencing direction (or mapped strand) and identical sequences. The presence of duplicated reads depends on the initial sequence complexity of extracted nucleic acids, the library preparation procedure and the sequencing technology. They can be generated during a fragmentation or tagmentation step or by an amplification-based technology (Hébrant et al., 2018; Maliogka et al., 2018). A dataset containing lots of duplicated reads might also be the result of a failed library preparation, where too little input material was available. The high abundance of duplicated reads can limit the analytical sensitivity of the HTS test as they can compete with low abundance targets. It is therefore recommended to evaluate the proportion of duplicated reads during the quality control stage of data analyses. Excess duplicated reads can be removed by using read normalization tools to facilitate the downstream analysis. The elimination of duplicated reads depends on the protocol and is not required in protocols that use the number of reads to estimate the relative abundance of a target like amplicon sequencing for metabarcoding.

- *Background reads removal*: Some sequences not related to the target(s), called background reads (e.g. host sequences, ribosomal sequences, phage sequences, environmental contaminant sequences), can be removed to facilitate the search of target(s) sequences and to reduce the risk of reporting incorrect results (Lambert et al., 2018). These reads are mainly observed with shotgun sequencing, and their presence also depends on the nucleic acid extraction procedure used (e.g. total nucleic acid extraction vs. target enrichment or selection). They can be removed by reference subtraction (i.e. host genome reads or host rRNA reads removal). The host control and/or no template control can be used to find the background reads (see Table 2). The removal of background reads can be particularly important when the target(s) is/are present in low concentration (Baizan-Edge et al., 2019). Caution should be taken when dealing with organisms that are capable of being completely or partially integrated into their host genome because they may be removed during this process (e.g. pararetroviruses in plants, bacteriophages in bacteria; Hohn et al., 2008, Sharma et al., 2017, Massart et al., 2019). In addition, there may be a risk of removing target reads during this process when high sequence identities exist between the host and target or if the reference genomes used for the removal of background reads contain themselves contaminant target reads. The quality of the host reference genome used for background reads removal is hence very important. Some (typically lower quality) reference genomes in databases can contain contaminant sequences (from endophytes), or are incomplete and their annotations are still in progress.
- *Reads pairing and merging*: In paired-end sequencing, the DNA fragment is sequenced from both ends (sense and antisense sequencing). Depending on the intended use of the HTS test, it may be useful to merge both reads of a single DNA fragment, if they overlap. For some sequencing technology, such as Illumina, the quality of the sequence tends to diminish towards the end of the reads (Kwon et al., 2014; Lambert et al., 2018). The pairing of reads can increase the overall quality and the length of the sequences. The laboratory should define the parameters (e.g. number of allowed mismatches) for merging the two sequences. In any case, it is important to retain all the reads for downstream analyses.
- *Artefact removal*: Amplicon sequencing can generate chimeric sequences where the first part of a read comes from a target organism while the other part comes from another target organism as a result of an amplicon accidentally acting as a primer during PCR. Similarly, whole genome amplification techniques such as multiple displacement amplification (MDA) or emerging single cell sequencing techniques which are commonly used within low-input library preparation protocols for shotgun sequencing can produce chimeric sequences (Lasken & Stockwell, 2007; Quince et al., 2011). It is important to monitor and remove these sequences using appropriate tools before the target identification (Anslan et al., 2018; Lu et al., 2019; Quince et al., 2011).
- *Denoising/clustering (specific to metabarcoding)*: PCR and sequencing errors inherent to amplicon sequencing introduce noise through the generation of high numbers of unique amplicons differing from the original sequences by one or more nucleotides. As a consequence, spurious results can be generated, and data analysis can become more complex. Within metabarcoding analyses sequencing reads are commonly clustered in representative bins called Operational Taxonomic Units (OTUs) using a nucleotide similarity threshold that ideally broadly approximates species boundaries (Mahé et al., 2015). Nevertheless, the optimal selection of a threshold can vary across taxa and can result in over-clustering (putting different species together in one cluster) or under-clustering (splitting one species over different clusters; Anslan et al., 2018; Quince et al., 2011).

  Alternatively, denoising algorithms have been developed. They do not cluster the sequences based on their similarity but resolve erroneous sequences by assuming that erroneous sequences will be closely related and will show a similar occurrence pattern than an authentic 'parent' haplotype while showing lower abundances and/or lower quality scores (Laehnemann et al., 2016; Yang et al., 2012). After read correction, this denoising process produces amplicon sequence variants (ASVs) or exact sequence variants (ESVs) that are taxonomically identified.

3.2.2.2 | *Identification of target(s)*

The second step of the bioinformatic analyses aims to identify the target(s) in the datasets.

Accurate identification of the target(s) bioinformatically is important to avoid false positive (incorrect taxonomic position, gene annotation or variant detection) or false negative (absence of identification) results.

The identification of target(s) consists of different sub-steps (Figure 3) detailed below. These sub-steps may not all be relevant and may be performed in a different order depending on the HTS test. The sub-steps needed, and their order should be defined during test development along with their parameters and their corresponding quality metrics and thresholds (Budowle et al., 2014; Hébrant et al., 2018). When implementing the HTS test in routine diagnostic activities, if the thresholds are not met, a decision concerning whether to repeat the part of the HTS test or to proceed should be made and documented.

The optional sub-steps of the second step of the bioinformatic analyses are:

- *Direct annotation of individual reads*: The reads can be annotated at taxonomic or functional levels without any assembly, clustering or mapping. The specificity of the annotation process will depend on the length of the sequences and on the database(s) used (see taxonomic position and functional assignment sub-steps).
- De novo *assembly* (also called contiguous assembly, reads assembly): The reads from a shotgun sequencing library can be assembled de novo to create longer

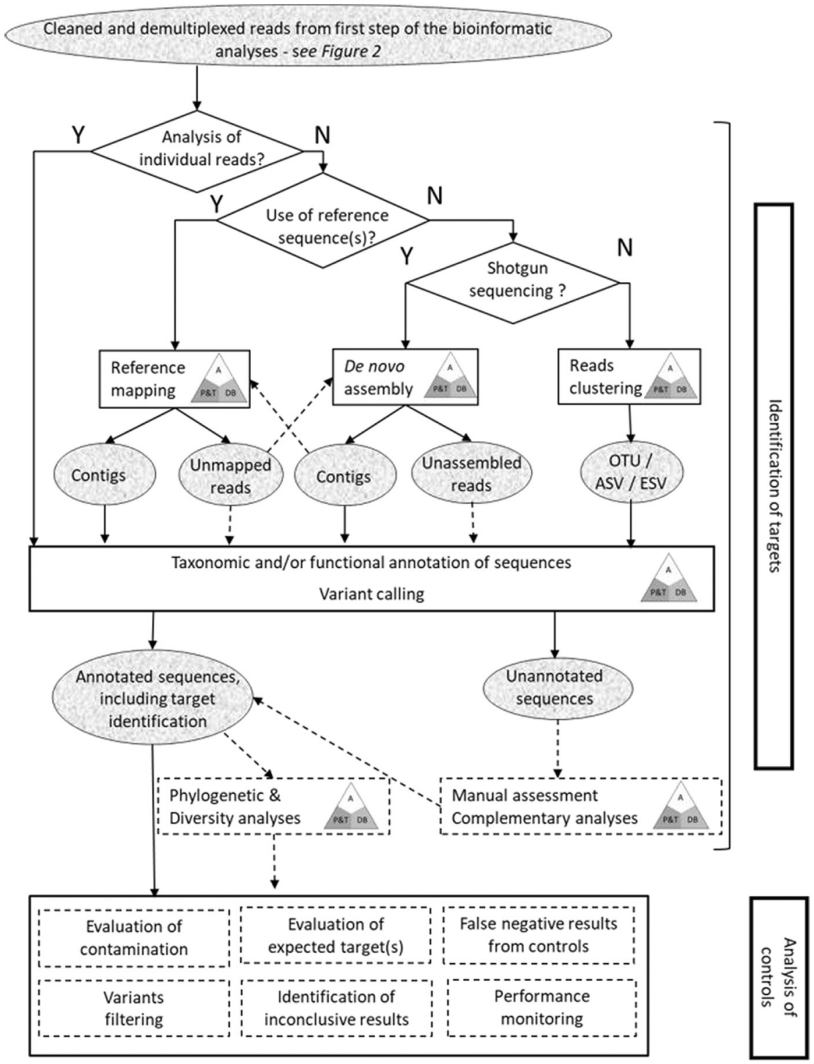

FIGURE 3 Overview of the steps: Identification of target(s) and analysis of controls. The selection of the sub-steps depends on the bioinformatic pipeline that is used. Dashed arrows are alternative steps. The rectangles correspond to operations/calculation while the grey ellipses correspond to file(s) containing sequence information and are generated by the analysis. The triangle represents the influence of bioinformatics triad corresponding to the algorithm(s) (A), its (their) parameters and thresholds (P&T) and the sequence database(s) (DB), on the generated results. Meaning of acronyms: ASV: Amplicon sequence variants; ESV: Exact sequence variants; OTU: Operational taxonomic units. More information about each of the steps can be found in the text.

sequences, called contiguous sequences (or contigs; Brinkmann et al., 2019). The reads are assembled when they present similar sequences on a portion or on the totality of their length. The reads assembly can be complex when they are short (like for small RNA sequencing; Massart et al., 2019). The parameters to consider for reads assembly such as, the percentage of identity between reads, the minimum overlap, the minimal length of contigs, the k-mer length or bubble size, depend on the type of algorithm used. For the genome sequencing of isolates of cellular organisms, the quality of assembly in contigs can be evaluated, for example by summarizing the length of the contigs using N50 or U50 values (Castro & Ng, 2017) or by comparing the contigs with related genomes and/or genes, using CheckM or BUSCO (Parks et al., 2014; Seppey et al., 2019).

- *Reference mapping* (also called reference assembly) for selected target(s): The reads can be directly mapped against the available targets reference sequence(s), which can be partial or full genome(s) (Budowle et al., 2014; Hébrant et al., 2018; Roy et al., 2018). Several reference sequences can be used for each target in order to take into account genetic variability (Massart et al., 2019) and improve the number of mapped reads and the annotation quality.

  The mapping parameters such as, number of mismatches or gaps allowed, or minimal percentage of identity, are critical to avoid incorrect results. If the mapping parameters are set too low, non-specific mapping to another species can happen, while too stringent mapping parameters can result in the poor mapping of reads from a distant isolate (Roy et al., 2018; Weiss et al., 2013). Important mapping results metrics include genome completeness, average read depth, distribution of reads on the reference sequence and percentage of identity with reference sequence(s). Their relevance depends on the technology used (e.g. PCR amplified targets will result in greater read depth; Asplund et al., 2019; Weiss et al., 2013).

  A combination of reference mapping and de novo assembly can be required to increase the likelihood of identifying target(s) present in low concentration (Maliogka et al., 2018). The ordering of contigs along a genome (i.e. scaffolding) can improve downstream analyses like the taxonomic and functional annotation (Sahlin et al., 2016) or de novo assembled (meta) genome contiguity.
- *Taxonomic position for pest identification*: when using reference mapping, the taxonomic position can be obtained from the annotation of the reference sequences but there can be a risk of misassignment (reads belonging to another species are mapped on the reference used) and the contigs generated from reads assembly might need to be further annotated independently. For individual reads, clustered reads and de novo contigs, the taxonomic position should be determined using the latest taxonomic information, including up to date sequence-based demarcation criteria (see box below for a list of current databases) and appropriate sequence databases and software (see Section 3.1.3). Searches for similarities with sequences in the reference database can be performed from assembled contigs or individual reads using dedicated tools (e.g. ANI, AODP, BLAST, DIAMOND, EDNA, Mash, Kraken, KAIJU) and provide indications on the taxonomic position and the closest organisms, most often with a confidence threshold (Lambert et al., 2018; Maliogka et al., 2018; Massart et al., 2019). These similarity searches use algorithms analysing alignment, k-mer, signature short motifs, and are continuously evolving (Budowle et al., 2014; Lefebvre et al., 2019; Rott et al., 2017; Ye et al., 2019). Simply taking the top hit in a BLAST search can lead to incorrect conclusions [see also PM 7/129 (EPPO, 2021b)]. In addition to sequence similarity searches, some taxonomic classifiers, such as RDP classifier, QIIME or SYNTAX, also take into account other similar sequences in the reference database and provide a confidence score using approaches such as bootstrapping. The level of certainty of the similarity searches should always be retained and mentioned (e.g. *e*-value) together with the tool and database (including the version) used. Expert judgement may be needed to evaluate the result of a taxonomic position (Massart et al., 2017; Matthijs et al., 2016). This is particularly challenging when dealing with uncharacterized organisms or with a sequence identity close to the threshold of species demarcation. When it is

The IPPC diagnostic protocols usually consider the following species lists for the latest taxonomy information:

International Committee on Taxonomy of Viruses (ICTV), https://talk.ictvonline.org/
International Committee on Systematics of Prokaryotes (ISCP), http://www.the-icsp.org/
International Commission on the Taxonomy of Fungi (ICTF), https://www.fungaltaxonomy.org/
Committee on Taxonomy of Plant Pathogenic Bacteria -International Society for Plant Pathology, https://isppweb.org/about_tppb.asp
Remark: some recently discovered species might not be listed in the official taxonomy list although they may be described in the literature and published in genome databases.

possible to retrieve the whole genome of a target, through shotgun sequencing, genome completeness and read depth can support the result of a taxonomic annotation (i.e. the more complete the genome is, the more reliable the taxonomic position). Additional analyses such as phylogenetic analysis may also be required. For amplicon sequencing, the resolution of the taxonomic assignment of the OTUs depends on different factors with the chosen barcode, the completeness of the reference database and the taxonomic position algorithm as the main ones. Currently used barcodes are relatively short (a few hundred nucleotides), and hence can provide only a limited taxonomic resolution. Classification methods such as naive Bayesian classifiers, lowest common ancestor-based methods, or phylogenetic placement methods are more reliable, but often also more conservative, hence not always leading to a satisfactory species-level classification. These limitations are inherent to amplicon sequencing or to the annotation of individual reads from shotgun sequencing and should be considered and explored in silico during the test selection and development, to verify whether the barcode is suited to detect the target organism(s) at a satisfactory taxonomic level.

- *Functional assignment*: The determination of the (potential) function of genes, the (prediction of) genomic features related to pathogenicity, resistance to antibiotics or to pesticides, proof of irradiation of live insects (provoking nucleotide mutations) intercepted at a border or any other sequence feature that may be of importance to plant health (Davis et al., 2016; Leifert et al., 2013; Zheng et al., 2015) may be useful/required depending on the intended use of the HTS test.
- *Recovering the whole genome of pests*: Obtaining the complete (or nearly complete) genome sequence may be needed to validate the taxa identified. Obtaining the (near) complete genome sequence for viruses/viroids is relatively easy because of their small genome sizes. The ability to recover a (near) complete genome becomes more complex with bacteria, phytoplasmas, and eukaryotic pests. When a (near) complete genome is needed, an iterative combination of reference mapping and de novo assembly with varying parameters can be carried out. Alternatively, a combination of sequencing strategies such as short and long read sequencing can assist in obtaining the (near) complete genome.
- *Variant calling*: Variants can consist of single nucleotide polymorphisms (SNP), insertion and deletion of nucleotides or the integration/deletion of entire genes compared to a reference sequence or to the consensus contigs generated (for example, the quasispecies complex of haplotypes for a virus isolate). The number of variants identified on a pest genome compared to a reference sequence can be used to evaluate if a new species (for viruses) or divergent isolate (for bacteria) has been identified and if the used reference is appropriate. To identify those variants accurately, longer reads can be used and if possible, retrieved from several samples and the associated metadata such as mapping quality, base-calling quality and strand bias should be checked (Gargis et al., 2015; Roy et al., 2018; Weiss et al., 2013). Replicates from the same sample can be processed in parallel to verify that the variant is identified in all datasets.
- *Unused high-quality reads*: A number of reads that have passed all the quality checks may still not be assembled, mapped or annotated after the bioinformatic analyses. These reads, called unused reads or unmapped reads, can be gathered as a separate output during analysis and their number or proportion calculated. Depending on the purpose of the test and the algorithms used, these reads can be discarded or re-analysed using other algorithms in order to validate the absence of target sequences or of unexpected organisms among them. Some individual sequences or some contigs, may still not be annotated during the second step of bioinformatic analyses. These unannotated sequences are sometimes referred to as 'dark matter'. Periodic re-analysis can be carried out to see if progress in strategies, algorithms or in knowledge of organisms allows a progress in annotation of such 'dark matter'.

3.2.2.3 | Analysis of controls included in the HTS test

The third and last step of the bioinformatic analyses is to verify that all the controls included in the HTS run performed as expected. This step is important to identify potential false positive (e.g. resulting from contamination) and/or negative results (e.g. resulting from the inhibition of enzymatic reactions, sample degradation or the generation of few sequences).

To detect false positive and/or negative results, different controls can be included at different stages of the HTS test. The type of controls that can be included are provided in Table 2 (see Section 5.2.1). All controls should be checked and should meet their respective acceptance criteria. The origin of incorrect results should be investigated and addressed and the decision on whether to repeat (parts of) the HTS test should be documented.

The analysis of controls may consist of different substeps that may not all be relevant, depending on the HTS protocol and the controls used (see Section 5). The substeps needed, and their order should be defined during test development along with their corresponding quality metrics and thresholds (Budowle et al., 2014; Hébrant et al., 2018). If thresholds are not met, the decision on whether to repeat (parts of) the HTS test should be documented and the reason for the failure of the control(s) should be investigated.

These sub-steps are:

- *Performance monitoring*: The performance of HTS tests may be checked routinely by including appropriate controls (see Section 5.2.1). For example, for HTS tests used for the detection of quarantine pests, a positive control close to the limit of detection should be included in each sequencing run and the results monitored over time.
- *Evaluation of contamination*: To check for contamination that may occur during a HTS test, positive, negative and alien controls (see Section 5.2.1) can be used at different stages of the test (Table 2).
- *Evaluation of expected target(s)*: The detection of expected target(s) can be carried out using positive and alien controls (see Section 5.2.1). These targets should all be detected according to the specified metrics (for example: genome completeness, number of generated sequences/reads, read depth and percentage of identity compared with their reference sequences).
- *False negative results from controls*: False negative results can be expected when one of the targets from the positive control(s) (see Section 5.2.1) is not detected in the sequence data. The result metrics for reference mapping (see Section 3.2.2.2) such as genome completeness, read depth and percentage of identity compared with reference sequences are important for filtering false negative results (Asplund et al., 2019; Weiss et al., 2013).
- *Variant filtering*: Variants generated during the HTS test due to sequencing errors, polymerase errors or reverse transcriptase errors, should be flagged and taken into account (Hébrant et al., 2018; Roy et al., 2018).
- *Inconclusive results*: If there are some issues with the controls of a sequencing run, for example when a quality metric is just below the defined threshold, the origin of the issue should be investigated and addressed (e.g. a reference sequence data set can be used to check that the bioinformatic pipeline performs as expected). The HTS test may need to be repeated or confirmatory tests other than HTS may be required to ascertain the HTS results. Whatever the decision, it should be documented as part of quality assurance [PM 7/77, EPPO (2019)].

4 | VALIDATION AND VERIFICATION OF HTS TEST

High-throughput sequencing tests should be validated or verified according to the EPPO standard PM 7/98 (EPPO, 2021a)

High throughput sequencing tests (e.g. all the steps described in Section 3) should be validated with reference biological material. When relevant, the bioinformatic pipeline can be validated/verified independently from the laboratory part of the test using sequence datasets obtained from biological reference material or artificial reference datasets containing known target(s) (Brinkmann et al., 2019; Budowle et al., 2014; Massart et al., 2019; Tamisier et al., 2021; Trimme et al., 2015).

Because HTS tests target a broad range of organisms, it is not possible to validate them for all possible combinations of organism, host or matrix. The validation of the HTS test should focus on key representatives of the targets/pests and use samples that mimic the concentration and composition of real samples expected to be tested.

Sections 4.1 and 4.2 describe specific considerations that should be taken into account for the evaluation of analytical sensitivity and specificity but all performance criteria described in the EPPO Standard PM 7/98 (EPPO, 2021a) should be evaluated if relevant.

4.1 | Specific considerations for analytical sensitivity

Theoretically, an HTS test can detect a very low amount of an organism as a single read from a target could be identified by an appropriate bioinformatics pipeline. In practice, many factors influence the ability of an HTS test to detect a target and determining the analytical sensitivity of an HTS test is particularly challenging.

Firstly, the analytical sensitivity of a HTS step depends on the proper execution of all the steps performed in the laboratory before sequencing. Then, the number of reads generated per sample also influence the analytical sensitivity of an HTS test. This number may vary in each sequencing run due to the variation of the number of reads per run. In addition, it varies between samples when several samples are pooled together, and this variability can increase with the level of pooling (see 3.2.1.5). The probability of detecting a target rises with the increase in the number of reads generated per sample (Massart et al., 2019; Pecman et al., 2017; Visser et al., 2016). However, increasing this number also increases the probability of detecting contaminants (H. Ziebell, JKI, pers. comms., 2020). Therefore, the optimal number of reads generated per sample and the minimal number of reads per sample, should be defined to reach an analytical sensitivity that fits the intended use of the test. This level has been established in literature using reference samples (Pecman et al., 2017), by comparing the results of dilution series of samples containing the relevant range of targets with those of PCR-based tests (Santala & Valkonen, 2018). Other metrics should be considered, such as sequence duplication levels in case of shotgun sequencing. A sample sequenced by shotgun sequencing can have many reads, but the diversity of the sequenced molecules could be low due to a poor library preparation. The minimal number of reads per sample is determined during test development and can be re-evaluated during validation by the bioinformatic

analysis. The generated reads for a sample can be rarefied by randomly selecting part of them (Gaafar & Ziebell, 2020; Pecman et al., 2017). This rarefaction will generate subsamples of reads corresponding to variable lower sequencing depths. The bioinformatic analyses of all these subsamples will identify the sequencing depth(s) at which a target is no longer detected.

The organisms that are present in a sample, in particular when at very high concentrations (e.g. in the case of co-infections), can also affect the ability of the HTS test to detect a target (Maclot et al., 2020). This situation has been observed for both shotgun sequencing (Rolland et al., 2017) and amplicon sequencing (Chandelier et al., 2020). For amplicon sequencing, this may be due to e.g. competition for the primers in the PCR reaction, or differences in copy number of the targeted region between target organisms. However, the competition between targets cannot be anticipated for all the combinations of targets tested. To mitigate this risk, the validation can include reference samples with different proportions/quantities of the targets, some very abundant while others at very low level. Such series of controls have been recently used for amplicon sequencing to survey the presence of fungal species in spore traps (Chandelier et al., 2020).

Furthermore, the DNA extraction step may also influence analytical sensitivity. For example, some nematode species were difficult to detect by amplicon sequencing because of poor cell lysis. As a result, contaminants were more easily picked up and amplified than the target (Waeyenberge et al., 2019).

Finally, the analytical sensitivity of an HTS test is limited by the contamination level between samples that can vary between sample batches and runs and, within a batch or a run, between target organisms. The analytical sensitivity will depend on the contamination threshold fixed for the run or the batch. In addition, it will be influenced by the presence of other samples containing the same target within the run or the batch. For example, 10 reads of a target have been detected in a sample. If there is at least one other sample in the batch with a very high abundance of this target (for example 500 000 reads), there is a risk of cross-contamination from this sample. If this target is not detected in any other sample from the batch, the 10 reads are more likely to represent a true infection at very low level.

4.2 | Specific considerations for analytical specificity

The analytical specificity of an HTS test depends on the strategy used to generate the sequencing library, the genetic variability of the target organisms, the software and parameters used for the bioinformatic analyses and the reference sequence database(s) (see section 3.1.3). The desired taxonomic resolution (e.g. genes, isolates/strains, pathovars, *formae speciales*, species, genera or families relevant to plant health) should be determined when describing the scope and the intended use of the test.

The taxonomic resolution of a region can vary. For example, a genomic region may allow distinction between all the species in one genus but may be unable to allow distinction between the species of another genus because of the lack of divergence between these species in that genomic region. For amplicon sequencing, the ability of the target regions to allow differentiation between closely related species can be at least partially evaluated theoretically by analysing all the sequences of the target regions available in databases, taking into account the intended use of the HTS test. If sequence similarities exist between organisms that could potentially be present in the samples and might interfere with pest detection and identification, those organisms should be included in validation. The analytical specificity can be evaluated using artificial reads datasets with known pest composition or of positive controls containing a mix of targets whose presence has been confirmed by different methods. Ideally, the concentration of the target(s) should reflect as much as possible the concentration in real samples that will be tested.

When applying a shotgun sequencing protocol to a sample composed of multiple organisms, the analytical specificity might depend on the number of sequences generated from each organism, the percentage of the genome covered, the genomic regions that have been sequenced (conserved or specific) and their read depth. As for analytical sensitivity, the taxonomic resolution of a shotgun sequencing will depend on the number of sequencing reads per sample and the appropriate target coverage to achieve the intended taxonomic resolution. Sufficient and reproducible sequence coverage and quality needs to be obtained and a minimal number of generated sequences per sample needs to be clearly stated during the development and/or validation phases.

For bacteria, determination of the analytical specificity can be complicated when applying HTS on a complex sample (not a single colony). This is because of their genome size and the presence of commensal bacterial species which may be related to the pathogenic ones present in the samples. For many bacteria, the appropriate discrimination between family, genus, species or strains may rely on a few specific genes, from which sequences need to be obtained. In cases where no specific genes are available for the identification to species level, the full genome should be obtained, which can be complicated for target organisms that are not isolated. Currently, the sequence databases are not yet representative of the diversity of bacterial species and the limited availability of genome sequences for bacteria will hamper their identification and may lead to the false positive detection of a related species whose genome is in the database(s). The use of a curated databases such as the Genome Taxonomy Database (https://gtdb.ecogenomic.org/) is encouraged.

For fungi, oomycetes, protists, nematodes, arthropod pests, invasive plants or weeds, the determination of the analytical specificity is even more difficult than for bacteria. This is because they have larger genomes than bacteria and also because of the limited availability of genomic sequences in current sequence databases. In 2020, it was estimated that only a very small proportion of fungal DNA is described in databases with about 1% of fungal species having DNA sequences annotated. In addition, low-quality reference genomes or sequences can be contaminated by microbial sequences (some fungi host bacterial cells, for example *Paenibacillus* spp. can live inside fungi) which could interfere with the calculation of the analytical specificity.

For viruses, determination of the analytical specificity can be better achieved because of their small genomes that can be fully sequenced and of the higher sequence divergence that exists between species. However, the sequence variability of the envelopes or coat proteins of viral species is sometimes close to the species demarcation threshold. This can be an issue for establishing the limit between divergent isolates and closely related species (for example, four molecular discrimination criteria exist for the family *Betaflexiviridae*: nucleotides and amino acids percentage for the coat protein and the replication polymerase genes). Therefore, wherever possible, the full genome of viruses should be sequenced or at least, several genomic regions should be sequenced although some uncertainties can remain, and the demarcation criteria could be met only partially.

5 | ENSURING THE VALIDITY OF HTS TEST RESULTS

Recommendations for ensuring the validity of test results listed in EPPO Standard PM 7/98 (2021a) are valid for HTS tests and should be followed.

Quality metrics (e.g. see sections 3.2.2.1 and 3.2.2.2) should be monitored for each sequencing run, routinely collected and compared to those of an optimal validated run. Any significant deviation should be investigated which may require the test to be repeated (for example when one of the targets from the positive controls is not detected in the sequence data). Such data checks can also help when investigating the source of the problem in an underperforming test (Hébrant et al., 2018).

5.1 | Risk analysis for ensuring the validity of results

The risks associated with running HTS tests should be identified before their use as diagnostics tests [see PM 7/98, (EPPO, 2021a)]. Given the complexity of a HTS test, it is particularly recommended to assess holistically the factors influencing the results. The severity of their impact should be estimated, and appropriate measures should be implemented to reduce, minimize or when possible, eliminate the risk (Hébrant et al., 2018; Jennings et al., 2017).

The risk analysis will be the basis for establishing the critical parameters and quality checks of the HTS test

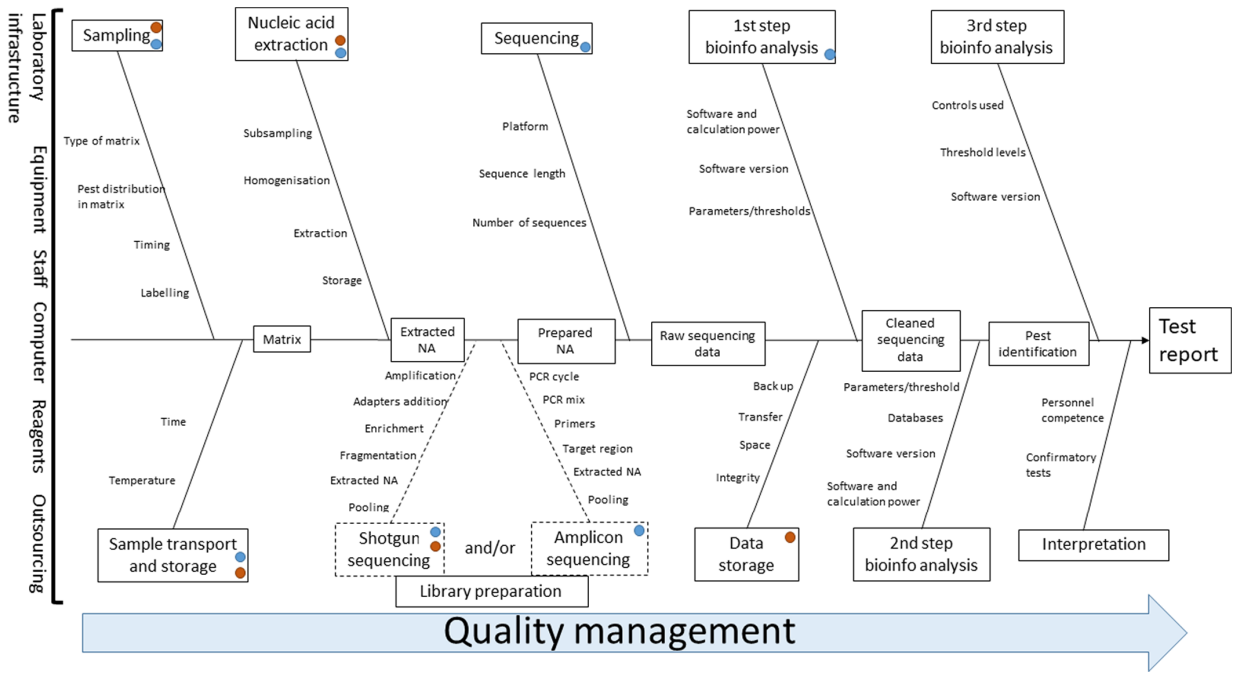

FIGURE 4 Ishikawa diagram adapted from Mehle et al. (2014) representing the cause and effect of each component of HTS tests. Acronyms: NA: Nucleic acids, bioinfo: Bioinformatics, PCR: Polymerase chain reaction. Risk of degradation or loss of integrity marked with red circle, risk of contamination marked with blue circle. Dashed boxes/lines: Possible types of library preparation.

for routine use. The thresholds, acceptable range and proper interpretation, should be defined in the procedure used during routine analysis and be used for continuous monitoring of the performance through time.

The risks can be analysed through the methodologies described in EPPO Standard PM 7/98 (2021a). An example, an Ishikawa diagram, adapted from Mehle et al. (2014), is proposed in Figure 4.

The risk analysis should be regularly updated (e.g. change in the level of and/or type of risks) depending on the results of the quality checks obtained during the development and validation of the HTS test and during the routine use of the HTS test. The risk analysis should be documented.

A non-exhaustive list of risks associated with HTS, and the corresponding metrics/controls can be found in supporting information and presents examples developed in the framework of VALITEST.

5.2 | Internal and external quality checks

5.2.1 | First line controls

Different types of first line controls are required for HTS tests. Table 2 provides a description of first line controls that can be used in HTS tests and their application in relation to the main steps of the HTS process. The purpose of each type of control in an HTS context is detailed below. It should be emphasized that each step of the HTS test should be monitored during each run.

Regardless of the type of control, the absence of targets (i.e. no targets in negative controls, absence of other targets in positive and alien controls) or the presence and abundance of target(s) (i.e. positive and alien controls) should be known unequivocally in each control and should be stable over time. The known abundance of target(s) in controls is also important for the determination of a quantitative threshold for contamination (see Section 3.1.2). This threshold can be an absolute number of reads and/or can be calculated as a relative proportion of reads from the alien targets in the samples and the positive controls. For example, 100 reads of the alien target have been detected as a contamination in a sample or another control. The relative level of contamination will be different if, within the run, the number of reads of this target in the alien control is 1000 (meaning 10% contamination) or 10 000 000 (meaning 0.001% contamination).

5.2.2.1 | Positive controls

The positive controls are external controls used to monitor the execution of the test and the correct detection of targets.

A positive control will usually contain a small but representative fraction of the possible targets because of the broad range of targets an HTS test could detect. It can be prepared as a mix of individual positive controls. It is recommended to use positive controls for which at least some target concentrations are close to the limit of detection to ensure that low-levels of target can be detected. The low concentration of those targets/controls also limits the risk of contaminating other samples. Positive controls can also be used to monitor contamination. The detection of an unexpected target in the positive control [in addition to the expected target(s)] may be a signal of contamination from another sample that can be confirmed with the percentage of nucleotide identity of the potential source of contamination.

5.2.2.2 | Negative controls

The negative controls of HTS tests are used to monitor contamination. Detection of target(s) in the negative controls indicates that contamination has occurred during the HTS test.

A very low amount of contamination by target sequences will often be present in the data generated from negative controls. Contamination can be more prevalent in amplicon sequencing because the amplification of traces of contaminant DNA will be very efficient in the absence of other DNA in the sample. This phenomenon leads to the risk of overestimating the contamination as compared to a sample or control in which trace contaminant DNA is extremely low in sample DNA. For this reason, the use of positive and/or alien controls containing a DNA quantity similar to the analysed samples may allow a better estimation of contamination in this specific case.

5.2.2.3 | Alien controls

A third group of controls, called alien controls, can be used in HTS tests. The alien control is used to monitor the detection of an alien target (role of positive control) and to check for cross contamination between samples (role of a negative control).

An alien control corresponds to a matrix containing a target (called alien target) which belongs to the same group as the target organism(s) but cannot be present in the samples to be tested. This alien target can be a pest or not. For example, an alien control can be a bacterial or fungal strain from a species or genus restricted to an ecological niche that is not related to the analysed matrix (e.g. extremophile species with plant samples or spore trapping). For insects or plants, a species restricted to temperate climates could be used as an alien control when analysing tropical crops or environments using insect or pollen traps and vice versa. For viruses, a *Phaseolus vulgaris* (cultivar: Black Turtle) infected with the cryptic viruses *Phaseolus vulgaris alphaendornavirus* 1 and 2 (PvEV1 and PvEV2; Kesanakurti et al., 2016) can be used as alien control when analysing viruses infecting potato or banana samples. In this case, the detection of PvEV1 and PvEV2 sequences in the analysed potato or banana samples would indicate that cross contamination has occurred.

TABLE 2 Description of first line controls[a] that may be used in HTS tests and their application in relation to the main steps of the HTS process

	Negative controls	Positive controls	Alien controls	Internal controls
Aim	To monitor contamination	To monitor contamination To ensure the detection of specific targets To ensure that low-levels of target can be detected when used at low concentration	To monitor contamination To ensure the detection of specific targets when used at high concentration To ensure that low-levels of target can be detected when used at low concentration	To ensure that low-levels of target can be detected
Description	Same matrix of the analysed samples but free of the target(s) (host control) or extraction buffer (NIC), or molecular grade water (NAC; no template controls)	Same matrix and range of target(s) expected to be detected in the analysed samples and processed alongside the samples and preferably at low concentration (naturally infected or spiked)	A target not expected to be found in the analysed samples (i.e. alien target) and processed alongside the samples and not expected to be detected in the samples to be tested when used at high concentration and/or expected to be detected only in the alien control when used at low concentration	Non target nucleic acids not related to the sample targets naturally present (e.g. plant genes) or known target spiked at low concentration in the samples (e.g. synthetic nucleic acids, known target not expected to be found in the samples to be tested)
Analysis	Absence of target(s) Contamination target(s) below a set threshold[b,c]	Presence of positive control targets Contamination targets below a set threshold[c]	Absence of the alien target in the analysed samples (when used at high concentration)[c] Contamination targets below a set threshold[c] Presence of expected alien target in the alien control (when used at low concentration)	Presence of internal control in each sample

HTS steps	Negative controls	Positive controls	Alien controls	Internal controls
Sampling and nucleic acids extraction	NIC: matrix without target(s), if not available, extraction buffer	PIC: matrix containing target(s) from a single or pooled individual(s)	Matrix containing alien target processed alongside samples	Not applicable as included in the analysed samples
Library preparation	Nucleic acids previously extracted from a NIC during another HTS test (that can be used as a NAC) Molecular grade water to verify the absence of contamination[b,d]	Nucleic acids extracted from a PIC during another HTS test (that can be used as a PAC)	Nucleic acids previously extracted from an alien control during another HTS test	Spiked nucleic acids to be analysed with non-target nucleic acids of natural, synthetic origin or known target not expected to be found in the samples to be analysed
Sequencing	Previously prepared libraries from the respective controls can be sequenced for specific monitoring of sequencing DNA sequence of the positive controls designed by the HTS technology manufacturer, present in the sequencing reagents			
Bioinformatic analysis	Raw sequencing data generated during previous HTS tests from respective controls or artificially generated data can be used to specifically monitor the bioinformatic analysis			

[a]Abbreviations of first line controls: NAC, negative amplification control; NIC, negative isolation control; PAC, positive amplification control; PIC, positive isolation control.

[b]The absence of target sequences is practically nearly not possible in a negative control.

[c]If an unexpected target is detected in any control or an alien target is detected in the samples, their presence should be quantified and compared with the controls and samples infected by the target.

[d]For shotgun sequencing, the same matrix as the analysed samples but free of the target(s) is preferred over molecular grade water as negative control.

The alien control should contain a high concentration of the alien target [for example a plant with a high virus concentration (the higher the better), purified viruses or a pure isolate of bacteria or fungi]. A high concentration of the alien target allows a better detection and quantification of alien contamination in the analysed samples. The number and/or proportion of the alien target sequences in the samples can be analysed (e.g. maximum, average, standard deviation, distribution) and compared to the number and proportion of alien target sequences in the alien control (relative quantification of contamination). If the alien control is also used as a positive control, at least another alien target should be present at a low concentration in addition to the alien target at high concentration to ensure that low-levels of target can be detected.

As the composition of the alien control is known, the presence of an unexpected target in the generated sequence data from the alien control would also indicate a potential contamination from a sample or another control (when) used in the HTS test (Galan et al., 2016).

5.2.2.4 | Internal positive controls

As an alternative or in addition to positive and alien controls, internal positive controls (IPC) may also be used in an HTS test.

Internal positive controls can correspond to sequences that are expected to be always present in the nucleic acids extracted from the sample (endogenous nucleic acids), for example a plant gene (e.g. nad5 gene, 18S gene, COI) constitutively expressed when analysing RNA shotgun sequencing data from plants to identify pests. Ideally, the selected sequences should be present at a stable and low level in the analysed matrix but above the level of detection to ensure that low-levels of target can be detected.

Alternatively, each sample may be spiked with synthetic nucleic acids or a known target not expected to be found in the samples to be analysed (this target could therefore be another alien control). An advantage of using synthetic nucleic acids is that they are more readily quantifiable than total nucleic acids. The spiked material should be easily and unambiguously detected by the HTS test. It should be spiked at a low concentration (ideally close to the detection level) to ensure that low-levels of target can be detected and to avoid masking the targets present in the sample. For example, black bean tissue containing an endornavirus has been used to spike grapevine samples to monitor the sensitivity of the test and set a threshold for the presence or absence of the target (Kesanakurti et al., 2016).

In metabarcoding, synthetic 16S rRNA gene spike-in controls have been used to aid in sample tracking and to detect and quantify cross-contamination that may have occurred during the laboratory processes. A distinct spike-in or mixtures of spike-ins were added in low concentration(s) in each sample before starting the DNA extraction (Tourlousse et al., 2018). Similarly, synthetic ITS spike-in controls (mock communities) were used in metabarcoding of forestry fungi. These synthetic controls proved to be useful for monitoring index-hoping and parameterizing the bioinformatic pipelines (Palmer et al., 2018).

In shotgun sequencing, a synthetic community of artificial microbial genomes called sequins (standing for sequencing spike-ins) mimicking the microbial community of the real samples, can be added to environmental DNA samples prior to library preparation. This enables the measurement and mitigation of technical variation (e.g. library preparation protocols) that can influence sequencing. Synthetic RNA spike-ins sets have also been used on zebrafish total RNA extracts for monitoring size-selection of RNA and for sample-to-sample normalization of RNA in small RNA sequencing. This improves the technical reproducibility of the test (Locati et al., 2015) but such an approach has not yet been evaluated in plant pest diagnostics.

Internal sequencing controls designed by the HTS technology manufacturers, are available for some sequencing platforms. The manufacturer's instructions should be followed when using these controls. For example, for the Illumina technology, the PhiX phage is used to monitor the sequencing run and is included in sequencing reagents and is always spiked in any sequencing reaction. Its genome sequence is known, and it is therefore used to automatically evaluate the accuracy of sequencing (e.g. the proportion of sequencing errors). Similarly, Oxford Nanopore Technologies have a control sequence that can be spiked in.

Commercialized spike-in controls are now becoming available. For example, a common set of external RNA controls called ERCC RNA spike-in mix, has been developed by the External RNA Controls Consortium (ERCC; ERCC, 2005) for RNA analysis, including gene expression profiling and whole transcriptome surveying. This control has been used routinely in some plant health diagnostic laboratories.

5.2.2 | Replicates

Biological and/or technical replicates may be used to validate the results although the costs can be prohibitive for example for library preparation. Technical or biological replicates could be more affordable for amplicon sequencing due to the lower costs per sample.

Additive processing (i.e. pooling the replicates) can be useful to overcome sampling stochasticity and controlling for false-negative results, while restrictive processing (i.e. only retaining sequences present in several replicates) effectively controls for cross-contamination. To balance the merits of both approaches, it may be best to include a minimum number of technical or biological replicates to allow a majority-rules approach (e.g. 2 or

3 replicates count as a detection; Piper et al., 2019). The processing of replicates could be systematic for only a few samples or the controls and would be limited by their costs.

5.2.3 | Second and third line controls and performance monitoring

Recommendations from PM 7/98 (EPPO, 2021a) should be followed.

6 | CONFIRMATION, BIOLOGICAL INTERPRETATION AND REPORTING

6.1 | Confirmation of the detection and identification of the pest(s)

The need to confirm the detection and identification of a pest depends on the context of the analysis and on the type of organism identified. For regulated pests (i.e. quarantine pest or regulated non-quarantine pest) the results should be confirmed for the critical cases described in EPPO Standard PM 7/76 (EPPO, 2018). When HTS is used as an identification test, confirmation may not be required, and the laboratory should document this decision. Sometimes it may not be possible to confirm the detection and identification of a pest in a sample. In such case(s), the laboratory should document the results and its decision for quality assurance purposes and in case further work should be conducted.

The identity of any uncharacterized organism with potential risks to plant health should also be confirmed and should be documented. For example, two isolates of *Xanthomonas* sp. causing rice bacterial grain rot disease were identified by HTS as close to *Xanthomonas sontii*, a species that is usually considered to be a harmless endophyte. Further testing and Koch's postulates were required, given that it was an unlikely candidate for causing disease (Mirghasempour et al., 2020).

6.2 | Interpretation of the biological relevance of the identified target(s)

High-throughput sequencing data do not provide any information on the biological relevance of the sequences identified, whether they correspond to a pathogenic organism with its associated risks and also whether the detected nucleic acids come from living organisms. For example, detected viral sequences may correspond to a bona fide virus infecting other organisms associated with the sample, including bacteria, fungi or arthropods (Al Rwahnih et al., 2011; Marzano & Domier, 2016) or to viral sequences integrated in the plant genome (Baizan-Edge et al., 2019; Brinkmann et al., 2019; Massart et al., 2017, 2019). Bacterial, fungal or viral sequences attributed to a pest species might be originating from closely related species that are not pathogenic but living as endophytes without causing any harm under the specific environmental conditions. Therefore, the analysis of the biological relevance of the target(s) identified by an HTS test is important for evaluating the potential risk the detected organism(s) would pose to plant health. It applies mainly to poorly characterized and uncharacterized organisms and, in some cases, to known plant pests unexpectedly found in a new host (called unexpected organisms). However, the biological characterization may take time or may not be possible for various reasons (e.g. lack of human and/or financial resources) or be carried out by another laboratory.

Relevant scientific expertise is essential to biologically interpret HTS results and their implications, in particular in the case of the identification of a target at a low concentration, a poorly characterized organism or an uncharacterized organism, and of viral sequences that might result from integration in the host genome (Brinkmann et al., 2019; Massart et al., 2019).

The extent of the biological characterization depends on the potential risk the detected organism(s) would pose to plant health (Massart et al., 2017) Decisions related to the biological characterization of the identified organism(s) should be documented.

The interpretation of the biological significance should cover some or all of the following items, depending on the context of the analysis. The information should be documented.

- *Sample information*: The following sample metadata can be used to support biological interpretation: information about the nature of the material (i.e. host identity to species and, whenever possible, to cultivar level and part of the plant sampled), the precise description of symptoms (if any) and time of appearance (if available), the time of sampling, the geographical origin of the sample, the presence of other relevant organisms and any other information relevant for the biological interpretation of the HTS results (e.g. estimation of the extent of infestation, hosts destined for import or export, size of the consignment; PM 7/77, EPPO, 2019; Massart et al., 2017).
- *Taxonomic information*: The (provisional) taxonomic position of a sequence can provide some information on its biochemical properties (e.g. bacteria belonging to a taxonomic group that have specific biochemical properties) and/or morphological characteristics (e.g. insects and nematodes) and even its biology. For example, for plant viruses, the taxonomic position can give an indication of the putative host range and its potential pathogenicity to these hosts, the modes of horizontal and/or vertical transmission, including

the identification of candidate vectors (Massart et al., 2017). However, these properties should be confirmed.
- *Genome information*: When relevant and wherever possible, identification of putative genes and the prediction of relevant gene products and functions should be determined (Budowle et al., 2014). This is particularly important when uncharacterized organisms are detected, as it would allow differentiation between known pathogenic and non-pathogenic organisms or strains (Zaluga et al., 2014). For example, virulence genes were found in three bacterial species consistently detected in the necrotic stem lesion of acute oak decline disease (Denman et al., 2018).
- *Confirmation of the results*: See Section 6.1.
- *Causation/aetiology*: Evidence of disease association is important when dealing with diseases potentially caused by several organisms (Adams et al., 2014; Denman et al., 2018; Lamichhane & Venturi, 2015). Some complex diseases may also be influenced by abiotic factors such as temperature, moisture, stage of host development (Denman et al., 2018).
  Fulfilling Koch's postulates, where one pathogen causes one disease, can be impractical following HTS results and does not apply to diseases caused by several organisms and abiotic factors. A combination of different approaches may be more effective (Adams et al., 2014; Denman et al., 2018). In particular, Fox (2020) proposed a systematic integrated approach for plant virology, utilizing epidemiological observations and supported by statistical analysis. The proposed approach may possibly be extended, with some modification, to other plant health disciplines.
- *Viability of the organisms*: Determining the viability of an organism may be required. Recommendations provided in EPPO Standard PM 7/76 (EPPO, 2018) should be followed.

6.3 | Reporting

6.3.1 | General recommendations

Reporting of the diagnostic results should follow the recommendations of the EPPO Standard PM 7/77 (EPPO, 2019). The reporting of HTS test results should be accompanied with a statement giving an expert judgement and with other confirmatory tests results, when needed. This is particularly important for the reporting of uncharacterized organisms.

The laboratory should also have a procedure on reporting to the NPPO the finding of any uncharacterized or unexpected organisms with a potential risk to plant health. Information to consider in the report to NPPO includes (if relevant):

- relationship with other organisms in the same taxon (e.g. closely related to an economically important pest).
- relationship with its host (e.g. mycovirus, insect virus).
- potential risk of causing damage to its host.
- potential risk for other hosts (economically and/or ecologically important).
- potential risk of spreading.
- location risk (e.g. horticultural area versus isolated area).
- viability of the organism (e.g. bacteria alive or dead, virus sequence integrated in plant genome leading to replicative form).
- possible influence of abiotic factors.
- presence of other organisms in the same host (e.g. symbiotic or antagonistic effect).
- recommendation for re-sampling/re-testing or other extended analyses.

Comment: organisms relevant to human health or animal health may be detected with HTS, it is consequently good practice to have mechanisms in place to inform the relevant authorities.

6.3.2 | Inconclusive results

Inconclusive results may be obtained with an HTS test (Boukari et al., 2020). Recommendations of the EPPO Standard PM 7/76 (EPPO, 2018) regarding the reporting of such results should be followed. The sources of uncertainty in an HTS test can be that the level of the pest is close to the limit of detection, it is present only in a single technical replicate out of several, the quality of the sample is poor, it is difficult to distinguish between episomal and integrated viruses, the lack of completeness of the databases, the limitations of the barcode used.

6.3.3 | Additional remarks and disclaimers

The laboratory should include in the report additional remarks and disclaimers related to any limitation in the HTS test (for example, the impossibility to distinguish viable and non-pests) and in the performance analysis of the sample (Hébrant et al., 2018; Weiss et al., 2013). Indeed, the HTS test results depend on the algorithms and sequence databases used. If confirmatory tests have been carried out (such as bioassay or viability tests), some limitations of the HTS test may not be relevant.

The HTS test results may be affected by the quality of the sample received. In this case, the report may state that the results apply to the sample as received (ISO 17025, 2017).

7 | FEEDBACK ON THIS STANDARD

If you have any feedback concerning this Diagnostic Standard, please contact diagnostics@eppo.int.

8 | PROTOCOL REVISION

An annual review process is in place to identify the need for revision of diagnostic Standards. Standards identified as needing revision are marked as such on the EPPO website.

When errata and corrigenda are in press, this will also be marked on the website.

ACKNOWLEDGEMENTS

This Standard was originally prepared based on the VALITEST deliverable D 2.2 by Vazquez-Iglesias I (FERA, UK), Santala J (Finnish Food Safety Authority Research Department, FI), Vossenberg B (NVWA, NL), Gaafar Y (JKI, DE), Massart S (U. Liege, BE). It was reviewed by the Panel on Diagnostics and Quality Assurance. The writing of VALITEST deliverable D 2.2 involved many international experts and lead to the publication of two scientific articles (Lebas et al., 2022; Massart et al., submitted).

REFERENCES

Adams IP, Skelton A, Macarthur R, Hodges T, Hinds H, Flint L, Nath PD, Boonham N, Fox A (2014) *Carrot yellow leaf virus* is associated with carrot internal necrosis. *PLoS ONE* 9 (11): e109125.

Adams IP, Fox A (2016) Diagnosis of plant viruses using next-generation sequencing and metagenomics analysis. *In*: Wang A., Zhou X. (eds.), *Current research topics in plant virology*, Springer International Publishing, Switzerland, pp. 323–335.

Ahmed M, Back MA, Prior T, Karssen G, Lawson R, Adams I, Sapp M (2019) Metabarcoding of soil nematodes: the importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). *Metabarcoding and Metagenomics*, 3: 77–99.

Al Rwahnih M, Daubert S, Úrbez-Torres JR (2011) Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Archives of Virology*, 156: 397–403.

Aritua V, Musoni A, Kabeja A, Butare L, Mukamuhirwa F, Gahakwa D, Kato F, Abang MM, Buruchara R, Sapp M, Harrison J, Studholme D.J, Smith J (2015) The draft genome sequence of *Xanthomonas* species strain Nyagatare, isolated from diseased bean in Rwanda. *FEMS Microbiology Letters*, 362 (4): 1–4.

Anslan S, Nilsson RH, Wurzbacher C, Baldrian P, Tedersoo L, Bahram M (2018) Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycoKeys*, 39: 29.

Asplund M, Kjartansdóttir K.R, Mollerup S, Vinner L, Fridholm H, Herrera JAR, Friis-Nielsen J, Hansen TA, Jensen RH, Nielsen IB, Richter S.R, Rey-Iglesia A, Matey-Hernandez ML, Alquezar-Planas DE, Olsen PVS, Sicheritz-Pontén T, Willerslev E, Lund O, Brunak S, Mourier T, Nielsen LP, Izarzugaza JMG, Hansen AJ. (2019) Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clinical Microbiology and Infection*, 25 (10): 1277-1285.

Baizan-Edge A, Cock P, MacFarlane S, McGavin W, Torrance L, Jones S. (2019) *Kodoja*: a workflow for virus detection in plants using *k-mer* analysis of RNA-sequencing data. *Journal of General Virology*, 100: 533–542.

Barba M, Czosnek H, Hadidi A. (2014) Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6: 106–136.

Boukari W, Mollov D.S, Wei C, Tang L, Grinstead S.C, Tahir N, Mulandesa E, Hincappie M, Beiriger R, Rott P. (2020) Screening for sugarcane yellow leaf virus in sorghum in Florida revealed its occurrence in mixed infections with sugarcane mosaic virus and a new marafivirus. *Crop Protection*, 139: 105373.

Brinkmann A, Andrusch A, Belka A, Wylezich C, Höper D, Pohlmann A, Nordahl Petersen T, Lucas P, Blanchard Y, Papa A, Melidou A, Oude Munnink BB, Matthijnssens J, Deboutte W, Ellis R.J, Hansmann F, Baumgärtner W, van der Vries E, Osterhaus A, Camma C, Mangone I, Lorusso A, Maracci M, Nunes A, Pinto M, Borges V, Kroneman A, Schmitz D, Corman VM, Drosten C, Jones TC, Hendriksen RS, Aarestrup FM, Koopmans M, Beer M, Nitsche A. (2019) Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated *in silico* high-throughput sequencing datasets. *Journal of Clinical Microbiology*, 57 (8): e00466-19.

Budowle B, Donnell ND, Bielecka-Oder A, Colwell RR, Corbett C.R, Fletcher J, Forsman M, Kadavy DR, Markotic A, Morse SA, Murch RS, Sajantila A, Schmedes SE, Ternus KL, Turner SD, Minot S. (2014) Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics*, 5: 9.

Buschmann T, Zhang R, Brash DE, Bystrykh LV. (2014). Enhancing the detection of barcoded reads in high throughput DNA sequencing data by controlling the false discovery rate. *BMC Bioinformatics*, 15: 264.

Castro CJ & Ng TFF. (2017). U_{50}: A new metric for measuring assembly output based on non-overlapping, target-specific contigs. *Journal of Computational Biology*, 24 (11): 1071–1080.

Champlot S, Berthelot C, Pruvost M, Bennett EA, Grance T, Geigl EM. (2010) An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS ONE*, 5 (9): e13042.

Chandelier A, Hulin J, Martin S, Debode F, Massart S. (2020) Comparison of real-time PCR and metabarcoding methods as tools for the detection of airborne inoculum of forest fungal pathogens. *Phytopathology*, 111(3):570–581.

Dal Molin A, Minio A, Griggio F, Delledonne M, Infantino A, Aragona M. (2018) The genome assembly of the fungal pathogen *Pyrenochaeta lycopersici* from single-molecule real-time sequencing sheds new light on its biological complexity. *PLoS ONE*, 13 (7): e0200217.

Davis M.P.A, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright A.J. (2013) Kraken: a set of tool for quality control and analysis of high-throughput sequence data. *Methods*, 63: 41–49.

Davis J.J, Boisvert S, Brettin T, Kenyon R.W, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam A.R, Will R, Xia F, Stevens R. (2016) Antimicrobial resistance prediction in PATRIC and RAST. *Scientific Reports*, 6: 27030.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi F.M. (2013) An extensive of read trimming effects on Illumina NGS data analysis. *PLoS ONE*, 8 (12): e85024.

Denman S, Doonan J, Ransom-Jones E, Broberg M, Plummer S, Kirk S, Scarlett K, Griffiths A.R, Kaczmarek M, Forster J, Peace A, Golyshin P.N, Hassard F, Brown N, Kenny J.G, Mcdonald J.E. (2018) Microbiome and infectivity studies reveal complex polyspecies tree disease in Acute oak decline. *The ISME Journal*, 12: 386–399.

Dickins B, Rebolledo-Jaramillo B, Su M.S.-W, Paul I.M, Blankenberg D, Stoler N, Makova K.D, Nekrutenko A. (2014) Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach. *Biotechniques*, 56 (3): 134–141.

Dormontt E.E, Van Dijk K.J, Bell K.L, Biffin E, Breed M.F, Byrne M, Caddy-Retalic S, Encinas-Viso F, Nevill P.G, Shapcott A, Young J.M. (2018) Advancing DNA barcoding and metabarcoding applications for plants requires systematic analysis of herbarium collections—an Australian perspective. *Frontiers in Ecology and Evolution*, 6: 134.

EPPO (2018) PM 7/76 (5) Use of EPPO diagnostic standards. *Bulletin OEPP/EPPO Bulletin*, 48 (3): 373–377.

EPPO (2019) PM 7/77 (3) Documentation and reporting on a diagnosis. *Bulletin OEPP/EPPO Bulletin*, 49 (3): 527–529.

EPPO (2021a) PM 7/98 (5) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *Bulletin OEPP/EPPO Bulletin*, 51: 468–498.

EPPO (2021b) PM 7/129 (2) DNA barcoding as an identification tool for a number of regulated pests. *Bulletin OEPP/EPPO Bulletin*, 51: 100–143.

EPPO (2021c) PM 7/84 (3) Basic requirements for quality management in plant pest diagnostic laboratories. *Bulletin OEPP/EPPO Bulletin*, 51: 457–467.

EPPO (2021d) PM 7/147 (1) Guidelines for the production of biological reference material. *Bulletin OEPP/EPPO Bulletin*, 51: 499–506.

ERCC (2005) Proposed methods for testing and selecting the ERCC external RNA controls. *BMC genomics*, 6: 150.

FAO (2019) Preparing to use high-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes. Commission on Phytosanitary Measures Recommendation, No 8, https://www.ippc.int/en/publications/87199/.

Fox, A. (2020), Reconsidering causal association in plant virology. *Plant Pathol*, 69: 956–961.

Gaafar Y.Z.A, Ziebell H. (2020) Comparative study on three viral enrichment approaches based on RNA extraction for plant virus/viroid detection using high-throughput sequencing. *PLoS ONE*, 15 (8): e0237951.

Galan M, Razzauti M, Bard E, Bernard M, Brouat C, Charbonnel N, Dehne-Garcia A, Loiseau A, Tatard C, Tamisier L, Vayssier-Taussat M, Vignes H. (2016) 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. *Clinical Science and Epidemiology*, 1 (4): e00032-16.

Gargis A.S, Kalman L, Bick D.P, da Silva C, Dimmock D.P, Funke B.H, Gowrisankar S, Hegde M.R, Kulkarni E, Mason C.E, Nagarajan R, Voelkerding K.V, Worthey D.A, Aziz N, Barnes J, Bennett S.F, Bisht H, Church F.M, Dimitrova Z, Gargis S.R, Hafez N, Hambuch T, Hyland F.C.L, Lunna R.A, MacCannell D, Mann T, McCluskey M.R, McDaniel T.K, Ganova-Raeva L.M, Rehm H.L, Reid J, Campo D.S, Resnick R.B, Ridge P.G, Salit M.L, Skums P, Wong L.-J.C, Zehnbauer B.A, Zook J.M, Lubin I.M. (2015) Good laboratory practice for clinical next-generation sequencing informatics pipelines. *National Biotechnologies*, 33 (7): 689–693.

Hadidi A, Flores R, Candresse T, Barba M. (2016) Next-generation sequencing and genome editing in plant virology. *Frontiers in Microbiology*, 7: 1325.

Hébrant A, Froyen G, Maes B, Salgado R, Le Mercier M, D'Haene N, De Keersmaecker S, Claes K, Van der Meulen J, Aftimos P, Van Houdt J, Cuppens K, Vanneste K, Dequeker E, Van Dooren S, Van Huysse J, Nollet F, van Laere S, Denys B, Ghislain V, Van Campenhout C, Van den Bulcke M. (2018) The Belgian next generation sequencing guidelines for haematological and solid tumours. *The Belgian Journal of Medical Oncology*, 11 (2): 56–67.

Hohn T, Richert-Pöggeler K.R, Staginnus C, Harper G, Schwarzacher T, Teo C.H, Teycheney P.Y, Iskra-Caruana M.L, Hull R. (2008) Evolution of integrated plant viruses. *In*: *Plant Virus Evolution*, Roossinck M.J. (Ed.), Springer-Verlag, Berlin, Germany, pp. 53–81.

Illumina (2022) https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html (last access: 2022 Feb)

ISO 17025 (2017) General requirements for the competence of testing and calibration laboratories. International organization for standardization, Geneva, Switzerland.

Jennings L.J, Arcila M.E, Corless C, Kamel-Reid S, Lubin I.M, Pfeifer J, Temple-Smolkin R.L, Voelerding K.V, Nikiforova M.N. (2017) Guidelines for validation of next-generation sequencing – based oncology panels, A joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. *The Journal of Molecular Diagnostics*, 19 (3): 341–365.

Johne R, Müller H, Rector A, van Ranst M, Stevens H. (2009) Rolling-circle amplification of viral DNA genomes using phi29 polymerase. *Trends in Microbiology*, 17 (5): 205–211.

Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P. (2019) Tools and strategies for long-read sequencing and *de novo* assembly of plant genomes. *Trends in Plant Sciences*, 24 (8): 700–724.

Kesanakurti P, Belton M, Saeed H, Rast H, Boyes I, Rott M. (2016) Screening for plant viruses by next generation sequencing using a modified double strand RNA extraction protocol with an internal amplification control. *Journal of Virological Methods*, 236: 35–50.

Kircher M, Sawyer S, Meyer M. (2011) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40 (1): e3, doi: 10.1093/nar/gkr771.

Kwon S, Lee B, Yoon S. (2014) CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. *BMC Bioinformatics*, 15 (suppl. 9): S10.

Lamichhane J.R, Venturi V. (2015) Synergisms between microbial pathogens in plant disease complexes: a growing trend. *Frontiers in plant science*, 6: 385.

Laehnemann D, Borkhardt A, McHardy A.C. (2016) Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in bioinformatics*, 17 (1): 154–179.

Lambert C, Braxton C, Charlebois R.L, Deyati A, Duncan P, La Neve F, Malicki H.D, Ribrioux S, Rozelle D.K, Michaels B, Sun W, Yang Z, Khan A.S. (2018) Considerations for optimisation of high-throughput sequencing bioinformatic pipelines for virus detection. *Viruses*, 10: 528.

Lasken R.S, Stockwell T.B. (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology*, 7: 19.

Lebas B & Massart S (2020) VALITEST deliverable D2.2 "Best practice" guidelines for validation and routine use of non-targeted techniques in diagnostic setting which could serve as a basis for a new EPPO Standard – Zenodo link. https://zenodo.org/record/7113284#.YzGl2XZBy70

Lebas B, Adams I, Al Rwahnih M, Baeyen S, Bilodeau GJ & Blouin AG et al. (2022) Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: A step-by-step description. *EPPO Bulletin*, 52, 394–418.

Lefebvre M, Theil S, Ma Y, Candresse T. (2019) The VirAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OUT) assignation for virome sequencing data. *Phytobiomes Journal* 3:4, 256–259.

Leifert W.R, Glatz R.V, Siddiqui M.S, Collins S.R, Taylor P.W, Fenech M. (2013). Development of a test to detect and quantify irradiation damage in fruit flies. *Final Report for Horticulture Australia Ltd,* Project VG09160.

Locati M.D, Terpstra I, de Leeuw W.C, Kuzak M, Rauwerda H, Ensink W.A, van Leeuwen S, Nehrdich U, Spaink H.P, Jonker M.J, Breit T.M, Dekker R.J. (2015) Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. *Nucleic Acids Research*, 43 (14): e89.

Lu N, Li J, Bi C, Guo J, Tao Y, Luan K, Tu J, Lu Z. (2019) ChimeraMiner: An Improved Chimeric Read Detection Pipeline

and Its Application in Single Cell Sequencing. *International Journal of Molecular Sciences*, 20 (8): 1953.

MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, Light M, Lai K, Jarosz M, McNeill MS, Ducar MD, Meyerson M, Thorner AR. (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, 19: 30.

Maclot F, Candresse T, Filloux D, Malmstrom CM, Roumagnac P, van der Vlugt R, Massart S. (2020) Illuminating an ecological blackbox: Using High Throughput Sequencing to characterize the plant virome across scales. *Frontiers in Microbiology*, 11: 578064.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3: e1420.

Malapi-Wight M, Salgado-Salazar C, Demers J, Clement DL, Rane K, Crouch JA. (2016). Sarcococca Blight: use of whole genome sequencing for fungal plant disease diagnosis. *Plant Disease*, 100 (6): 1093–1100.

Maliogka VI, Minafra A, Saldarelli P, Ruiz-García AB, Glasa M, Katis N, Olmos A. (2018) Recent advances on detection and characterization of fruit tree viruses using high-throughput sequencing technologies. *Viruses*, 10: 436.

Marzano SL, Domier LL. (2016) Novel mycoviruses discovered from metatranscriptomics survey of soybean phyllosphere phytobiomes. *Virus Res.* 2;213:332–342.

Massart S, Adams I, Al Rwahnih M, Baeyen S, Bilodeau GJ, Blouin A, Boonham N, Candresse T, Chandelier A, De Jonghe K, Fox A, Gaafar YZA, Gentit P, Haegeman A, Ho W, Hurtado-Gonzales, Jonkers W, Kreuze J, Kutnjak D, … Lebas BSM (2022) Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. *Peer Community in Infections*. https://doi.org/10.5281/zenodo.7142136

Massart S, Olmos A, Jijaki H, Candresse T (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostic. *Virus Research*, 188 (8): 90–96.

Massart S, Candresse T, Gil J, Lacomme C, Predajna L, Ravnikar M, Reynard J.-S, Rumbou A, Saldarelli P, Škorić D, Vainio EJ, Valkonen JPT, Vanderschuren H, Varveri C, Wetzel T (2017) A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in Microbiology*, 8: 45.

Massart S, Chiumenti M, De Jonghe K, Glover R, Haegeman A, Koloniuk I, Komínek P, Kreuze J, Kutnjak D, Lotos L, Maclot F, Maliogka V, Maree HJ, Olivier T, Olmos A, Pooggin MM, Reynard J.-S, Ruiz-García AB, Safarova D, Schneeberger PHH, Sela N, Turco S, Vainio EJ, Varallyay E, Verdin E, Westenberg M, Brostaux Y, Candresse T (2019) Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. *Phytopathology*, 109: 488–497.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, Scheffer H, Bauer P (2016) Guidelines for diagnostic next-generation sequencing. *European Journal of Human Genetics*, 24: 2–5.

McInerney P, Adams P, Hadi MZ (2014) Error rate comparison during polymerase chain reaction by DNA polymerase. *Molecular Biology International*, 2014: 287430.

Mehle N, Dreo T, Jeffries C, Ravnikar M (2014) Descriptive assessment of uncertainties of qualitative real-time PCR for detection of plant pathogens and quality performance monitoring. *Bulletin OEPP/EPPO Bulletin*, 44 (3): 502–509.

Mehle N, Gutiérrez-Aguirre I, Kutnjak D, Ravnikar M (2018) Water-mediated transmission of plant, animal, and human viruses. *Advances in Virus Research*, 101: 85–128.

Mirghasempour SA, Huang S, Studholme DJ, Brady CL (2020) A grain rot of rice in Iran caused by a *Xanthomonas* strain closely related to *X. sacchari*. *Plant Disease*, 104 (6):1581–1583.

Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L (2019) Mycobiome diversity: high-throughput sequencing and identification of fungi. *Microbiology*, 17: 95–109.

Olmos A, Boonham N, Candresse T, Gentit P, Giovani B, Kutnjak D, Liefting L, Maree HJ, Minafra A, Moreira A, Nakhla MK, Petter F, Ravnikar M, Rodoni B, Roenhorst JW, Rott M, Ruiz-García AB, Santala J, Stancanelli G, van der Vlugt R, Varveri C, Westenberg M, Wetzel T, Ziebell H, Massart S. (2018) High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. *Bulletin OEPP/EPPO Bulletin*, 48 (2): 219–224.

Palmer JM, Jusino MA, Banik MT, Lindner DL (2018) Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, 6: e4925.

Parks DH, Imelfort M, Skennerton C.T, Hugenholtz P, Tyson GW (2014) Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25: 1043–1055.

Pecman A, Kutnjak D, Gutiérrez-Aguirre I, Adams I, Fox A, Boonham N, Ravnikar M (2017) Detection and discovery of plant viruses and viroids: comparions of two approaches. *Frontiers in Microbiology*, 8: 1998.

Piper AM, Batovska J, Cogan N.O.I, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ (2019) Prospects and challenges if implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8: 1–22.

Quail MA, Smith M, Jackson D, Leonard S, Skelly T, Swerdlow HP, Gu Y, Ellis P (2014) SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics*, 15: 110.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, 12 (1): 38.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hedge MR, Lyon Y (2013) ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine*, 15 (9): 733–747.

Ritter CD, Häggqvist S, Karlsson D, Sääksjärvi IE, Muasya AM, Nilsson RH, Antonelli A (2019). Biodiversity assessments in the 21st century: the potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. *Genome*, 62 (3): 147–159.

Rolland M, Villemot J, Marais A, Theil S, Faure C, Cadot V, Valade R, Vitry C, Rabenstein F, Candresse T (2017) Classical and next generation sequencing approaches unravel Bymovirus diversity in Barley crops in France. *PLOS One*, 12 (11): e0188495.

Rosseel T, Pardon B, De Clercq K, Ozhelvaci O, Van Borm S (2014) False-positive results in metagenomics virus discovery: a strong case for follow-up diagnosis. *Transboundary and Emerging Diseases*, 61: 293–299.

Rott M, Xiang Y, Boyes I, Belton M, Saeed H, Kesanakurti P, Hayes S, Lawrence T, Birch C, Bhagwat B, Rast H (2017) Application of next generation sequencing for diagnostic testing of tree fruit viruses and viroids. *Plant Disease*, 101: 1489–1499.

Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB (2018) Standards and guidelines for validating next-generation sequencing bioinformatic pipelines, a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *The Journal of Molecular Diagnostics*, 20 (1): 4–27.

Sahlin K, Chikhi R, Arvestad L (2016) Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics*, 32 (13): 1925–1932.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and

laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12: 87.

Santala J, Valkonen JPT (2018) Sensitivity of small RNA-based detection of plant viruses. *Frontiers in Microbiology*, 9: 939.

Scibetta S, Schena L, Abdelfattah A, Pangallo S, Cacciola S (2018) Selection and experimental evaluation of universal primers to study the fungal microbiome of higher plants. *Phytobiomes*, 2 (4): 225–236.

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. *In*: Kollmar (eds.) Gene Prediction. *Methods in Molecular Biology*, Humana, New York, vol. 1962: 227–245.

Sharma S, Chatterjee S, Datta S, Prasad R, Dubey D, Prasad RK, Vairale MG (2017) Bacteriophages and its applications: an overview. *Folia Microbiology*, 62: 17–55.

Simpson AJG, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LMC, Araya JE, Baia GS, Baptista CS, Barros MH, Bonaccorsi ED, Bordin S, Bové JM, Briones MRS, Bueno MRP, Camargo AA, Camargo LEA, Carraro DM, Carrer H, Colauto NB, Colombo C, Costa FF, Costa MCR, Costa-Neto CM, Coutinho LL, Cristofani M, Dias-Neto E, Docena C, El-Dorry H, Facincani AP, Ferreira AJS, Ferreira VCA, Ferro JA, Fraga JS, França SC, Franco MC, Frohme M, Furlan LR, Garnier M, Goldman GH, Goldman MHS, Gomes SL, Gruber A, Ho PL, Hoheisel JD, Junqueira ML, Kemper EL, Kitajima JP, Krieger JE, Kuramae EE, Laigret F, Lambais MR, Leite LCC, Lemos EGM, Lemos MVF, Lopes SA, Lopes CR, Machado JA, Machado MA, Madeira AMBN, Madeira HMF, Marino CL, Marques MV, Martins EAL, Martins EMF, Matsukuma AY, Menck CFM, Miracca EC, Miyaki CY, Monteiro-Vitorello CB, Moon DH, Nagai MA, Nascimento ALTO, Netto LES, Nhani Jr A, Nobrega FG, Nunes LR, Oliveira MA, de Oliveira MC, de Oliveira RC, Palmieri DA, Paris A, Peixoto BR, Pereira GAG, Pereira Jr HA, Pesquero JB, Quaggio RB, Roberto PG, Rodrigues V, Rosa AJ de M, de Rosa VE Jr, de Sá RG, Santelli RV, Sawasaki HE, da Silva ACR, da Silva AM, da Silva FR, Silva WA Jr, da Silveira JF, Silvestri MLZ, Siqueira WJ, de Souza AA, de Souza AP, Terenzi MF, Truffi D, Tsai SM, Tsuhako MH, Vallada H, Van Sluys MA, Verjovski-Almeida S, Vettore AL, Zago MA, Zatz M, Meidanis J, Setubal JC (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* 406, 151–159.

Tamisier L, Haegeman A, Foucart Y, Fouillien N, Al Rwahnih M, Buzkan N, Candresse T, Chiumenti M, De Jonghe K, Lefebvre M, Margaria P, Reynard JS, Stevens K, Kutnjak D, Massart S (2021) Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. *Peer Community Journal*, 1: e53.

Tourlousse DM, Ohashi A, Sekiguchi Y (2018) Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls. *Scientific Reports*, 8: 9095.

Tremblay ED, Duceppe MO, Bérubé J.A, Kimoto T, Lemieux C, Bilodeau GJ. (2018) Screening for exotic forest pathogens to increase survey capacity using metagenomics. *Phytopathology*, 108 (12): 1509–1521.

Trimme RE, Rand H, Shumway M, Trees EK, Simmons M, Agarwala R, Davis S, Tillman GE, Defibaugh-Chavez S, Carleton HA, Klimke WA, Katz LS (2015) Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*, 5: e3893.

Trontin C, Agstner B, Altenbach D, Anthoine G, Bagińska H, Brittain I, Chabirand A, Chappé AM, Dahlin P, Dreo T, Freye-Minks C, Gianinazzi C, Harrison C, Jones G, Luigi M, Massart S, Mehle N, Mezzalama M, Mouaziz H, Petter F, Ravnikar M, Raaymakers T.M, Renvoisé JP, Rolland M, Santos Paiva M, Seddas S, van der Vlugt R. and Vučurović A (2021) VALITEST: Validation of diagnostic tests to support plant health. *Bulletin OEPP/EPPO Bulletin*, 51: 198–206.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. (2018) Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources*, 00: 1–11.

Visser M, Bester R, Burger JT, Maree HJ (2016) Next-generation sequencing for virus detection: covering all the bases. *Virology Journal*, 13: 85.

Waeyenberge L, de Sutter N, Viaene N, Haegeman A (2019). New insights into nematode DNA-metabarcoding as revealed by the characterization of artificial and spiked nematode communities. *Diversity*, 11: 52.

Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Bruggenwirth HT, Lekanne Derez RH, Mook O, Ruivenkamp CAL, van der Stoep N (2013) Sequencing applications in genome diagnostics: A national collaborative study of Dutch genome diagnostic laboratories. *Human Mutation*, 34 (10): 1313–1321.

Wright ES, Vetsigian KH (2016). Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC genomics*, 17: 876.

Yang X, Chockalingam SP, Aluru S (2012) A survey of error-correction methods for next-generation sequencing. *Briefings in bioinformatics*, 14 (1): 56–66.

Ye SH, Siddle KJ, Park DJ, Sabeti PC (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178: 779–794.

Zaluga J, Stragier P, Baeyen S, Haegman A, Van Vaerenbergh J, Maes M, De Vos P (2014) Comparative genome analysis of pathogenic and non-pathogenic *Clavibacter* strains reveals adaptations to their lifestyle. *BMC Genomics*, 15: 392.

Zheng Z, Hou Y, Cai Y, Zhang Y, Li Y, Zhou M (2015) Whole-genome sequencing reveals that mutations in myosin-5 confer resistance to the fungicide phenamacril in *Fusarium graminearum*. *Scientific reports*, 5: 8248.

SUPPORTING INFORMATION

Additional supporting information can be found online in the html version of this article.

How to cite this article: (2022) PM 7/151 (1) Considerations for the use of high throughput sequencing in plant health diagnostics. *EPPO Bulletin*, 52, 619–642. Available from: https://doi.org/10.1111/epp.12884