https://doi.org/10.1094/PHYTOFR-03-22-0030-FI

PhytoFrontiers™
An Open Access Journal from The American Phytopathological Society

Research

Validation of High-Throughput Sequencing as Virus Indexing Test for *Musa* Germplasm: Performance Criteria Evaluation and Contamination Monitoring Using an Alien Control

Wei Rong[1] | Johan Rollin[1,2] | Marwa Hanafi[1] | Nicolas Roux[3] | Sebastien Massart[1,†] |

[1] Laboratory of Plant Pathology – TERRA - Gembloux Agro-Bio Tech – University of Liège, 5030 Gembloux, Belgium

[2] DNAVision, 6041 Gosselies, Belgium

[3] Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier, France

[†] Corresponding author: S. Massart; sebastien.massart@uliege.be

Accepted for publication 3 October 2022.

W. Rong and J. Rollin contributed equally.

Funding
The work has been supported by the NGS cross-center project from the CGIAR Fund, and in particular by the Germplasm Health Unit (GHU) of the CGIAR Genebank Platform, and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 813542T (INEXTVIR).

e-Xtra: Supplementary material is available online.

The author(s) declare no conflict of interest.

Abstract

High-throughput sequencing (HTS) technologies have brought tremendous improvements in the ability to detect plant viruses and have great potential for application in virus routine diagnostics. The performance criteria of an HTS test need therefore to be estimated and compared with traditional virus indexing tests before it can be used in routine diagnostics. In this study, 78 *Musa* accessions previously indexed for viruses by molecular tests and/or electron microscopy were tested individually or in pools using an HTS protocol based on total RNA sequencing. The analytical sensitivity of HTS and RT-PCR was also compared by independent testing on serial dilutions of RNA extracts. In total, 136 libraries were sequenced in five batches, and the sequences were analyzed for virus detection. The external alien control, a wheat sample infected by barley yellow dwarf virus, monitored the contamination burden and determined an adaptative detection threshold. Overall, the HTS test displayed a better analytical sensitivity than the RT-PCR and a better inclusivity than the classical indexing protocol, as distant isolates and new viral species were only detected by the HTS test. The repeatability and reproducibility of virus detection were both 100%, although differences in number of sequencing reads per virus were observed between replicates. The diagnostic sensitivity was very high, but false positive results were observed. Finally, the results also underlined the need for expert judgement in the interpretation of the results. In conclusion, the HTS test with an alien control and completed by expert evaluation fulfilled the criteria of the virus indexing protocol for *Musa* germplasm.

Keywords: contaminant, detection, diagnostics, high-throughput sequencing, *Musa*, performance criteria, virus

Plant viruses are a major threat for crop production and food security worldwide (Calil and Fontes 2017). The ability to provide a fast, cost-effective, and reliable diagnostic test for any given plant virus infection is a key parameter to control these ubiquitous pathogens and to support efficient plant quarantine or certification programs (Kumar et al. 2021; Massart et al. 2014; Soltani et al. 2021). From the first discovery of a plant virus in 1898, the ability to detect plant viruses has followed technological evolution over time (Maclot et al. 2020). In the first part of the 20th century, the detection of plant viruses relied on symptomatology on natural host and indicator plants (bioassays), biochemistry, and electron microscopy (Boonham et al. 2014; De Clerck et al. 2017). Bioassays have several drawbacks, including asymptomatic responses in investigated plants, reduced number of indicator results due to bud grafting failure, and relatively long testing period before release (Al Rwahnih et al. 2015; Soltani et al. 2021). Later, the detection of plant viruses was steadily improved by the development of targeted serological and molecular tests (Boonham et al. 2014). These tests are generally sensitive and rapid, but they require a priori knowledge of the targeted viruses and can lack inclusivity for viruses with high genetic diversity (Al Rwahnih et al. 2015; Maree et al. 2018).

The development of high-throughput sequencing (HTS) technologies represents a promising method for universal virus detection (Massart et al. 2014). The sequencing machine will produce millions to billions of sequences, which will be called reads in this publication to differentiate them from assembled sequences or genomes. Since their first application for plant virus detection, HTS technologies have become easier to perform, at both the laboratory and bioinformatics levels, and cheaper. Their use is therefore increasing for research purposes and, progressively, their adoption is envisioned for virus diagnostics (Olmos et al. 2018). HTS has been used as part of routine virus diagnostic workflows in several diagnostic laboratories to identify novel viruses from plant hosts (Adams et al. 2018), as well as complementary to conventional methods to inform diagnostic workflows in the identification of well-characterized pathogens (Adams et al. 2014) or new pathogens following initial detection using targeted generic tests (Fox et al. 2018).

Transferring an HTS-based detection test from research toward routine application in diagnostics is a challenging task. A workflow allowing this transfer was drafted by an international consortium of plant pest diagnostics specialists (Massart et al. 2022). First, one or several protocols that fit the purpose of the test should be selected and evaluated within the laboratory. Once a protocol is defined and well described, its performance characteristics should be evaluated through a validation scheme. Several performance criteria are proposed in international guidelines such as EPPO standard PM7/98: analytical sensitivity, analytical specificity (including inclusivity and exclusivity), selectivity, repeatability, and reproducibility. The diagnostic sensitivity, corresponding to the ratio between the number of infected samples that tested positive and the total number of infected samples, and the diagnostic specificity, corresponding to the ratio between the number of healthy samples that tested negative and the total number of healthy samples, are both key performance criteria for a diagnostic test. The evaluation of performance characteristics of various HTS tests, and their comparison with classical tests based on targeted detection and/or bioassays, has already been carried out for several crops, including grapevine, *Prunus*, *Malus*, *Pyrus*, *Citrus*, and ornamentals (Al Rwahnih et al. 2015; Bester et al. 2021; Di Gaspero et al. 2022; Gauthier et al. 2022; Rott et al. 2017; Soltani et al. 2021).

An HTS test can be divided into eight steps, among which laboratory (sampling, nucleic acids extraction, library preparation, sequencing) and bioinformatics (analysis of raw sequencing reads, identification of targets, analysis of controls) steps are completed by the last step of confirmation, interpretation, and reporting of the results (Lebas et al. 2022). It is mandatory to have a fixed and well-described HTS protocol before starting the validation. Indeed, whatever the protocol used (total RNA sequencing, small RNA sequencing, dsRNA sequencing), several scientific publications have underlined that each step of an HTS test can impact its performance: sampled tissue (Di Gaspero et al. 2022; Malapi-Wight et al. 2021), RNA extraction (Bester et al. 2021), library protocol (Bester et al. 2021; Di Gaspero et al. 2022; Maachi et al. 2021; Pecman et al. 2017), sequencing technology (Bester et al. 2021), sequencing service provider (Gauthier et al. 2022), number of generated sequences per sample (Gauthier et al. 2022; Pecman et al. 2017; Visser et al. 2016), bioinformatics analysis (Bester et al. 2021; Gaafar et al. 2021; Galan et al. 2016; Gauthier et al. 2022; Massart et al. 2019; Tamisier et al. 2021), and the diagnostic laboratory performing the test (Gaafar et al. 2021).

In side-by-side comparisons published in the literature, the inclusivity of HTS tests was identical to or better than the other compared tests (molecular, immunological, or bioassays) (Al Rwahnih et al. 2015; Hanafi et al. 2020; Velasco and Padilla 2021). On the other hand, the exclusivity of HTS tests is intrinsically high but depends on the number of reads generated for the detected virus (and the genome coverage of the virus), the quality of these reads, as well as the bioinformatic procedure (e.g., the software, the parameters, and the database used) at each step of the bioinformatics analysis. The repeatability and reproducibility of HTS tests have been evaluated in the literature and were generally very high, with up to 100% repeatability and reproducibility for virus detection (Bester et al. 2021; Di Gaspero et al. 2022; Soltani et al. 2021).

A key performance criterion is the analytical sensitivity and the corresponding limit of detection of the HTS test. When compared with reverse transcription PCR (RT-PCR), the limit of detection of the HTS test for virus detection was often improved, for example, by 10× for Potato Virus Y (Santala and Valkonen 2018). The analytical sensitivity of an HTS test is theoretically very low, as a single read from a target can be potentially identified by an appropriate bioinformatics pipeline. Nevertheless, the analytical sensitivity is limited by the cross-contamination level between samples (Massart et al. 2022). Although cross-contamination has long been known to occur during the detection of plant viruses by HTS, very few reports have analyzed it extensively, and none used controls to monitor it in routine settings. Cross-contamination between samples was previously estimated to range between 5 and 10% (Sinha et al. 2017) and between 0.2 and 6% (Costello et al. 2018) of the reads. Although improvements were gained at specific steps, for example by improved washing between sequencing runs or by alternating sequencing indexes between runs, contaminations can still potentially occur at any step of the process. High proportions could cause false positive detections that complicate the data interpretation. As a consequence, the limit of detection should consider the cross-contamination level between samples to minimize false positive rate while maintaining an analytical sensitivity (limit of detection) and diagnostic sensitivity (limiting number of false negatives due to very low level of infection) that fit the purpose of the test. This issue and its impact on false positive rate have been discussed recently for plant viruses based on reference samples with characterized virus infection (Gauthier et al. 2022).

In this context, it is recommended to monitor the cross-contamination burden when using HTS tests. For this purpose, a new type of external control should be used during the validation and in routine diagnostics: the alien control. For plant virus diagnostics, an alien control corresponds to a plant sample previously sequenced and containing one or several plant viruses (called alien viruses) that should not be present in the test tissues. Therefore, the detection of reads from an alien virus in a sample or another non-host control can be unequivocally considered a cross-contamination from the alien control (Massart et al. 2022). The alien control should be preferred over the negative control to monitor the cross-contamination. Indeed, it has the same role as a negative control for viruses infecting the tested plants as any read from a host virus can be considered cross-contamination (controlling the reads exchange from any sample to a single sample). In addition, the alien control monitors the cross-contamination from a single sample to any other sample by detecting alien virus reads in all the samples. The monitoring of cross-contamination is therefore greatly improved.

This paper reports the validation of the main performance characteristics of an HTS test, based on the HTS of ribosomal depleted RNA, to detect viruses infecting *Musa* germplasms. It worth mentioning that most edible bananas belong to the genus *Musa* with a genome predominantly originating from *M. acuminata* (A genome) and/or *M. balbisiana* (B genome). They can be diploid, triploid, or tetraploid and comprise solely A genomes or have combinations of A and B genomes. B genomes have an important specificity as they almost universally carry sequences of one or more banana streak virus (BSV) species within their chromosomes; some of these integrant sequences can be activated and can trigger an infection with viral particle of BSV.

The performance criteria, that is, analytical sensitivity, analytical specificity, repeatability, reproducibility, diagnostic sensitivity, and diagnostic specificity (focusing on false positives), were evaluated considering the cross-contamination burden, monitored for the first time in plant pest diagnostics by an alien control, and the biology of the tested viruses, with the BSV particularly.

MATERIALS AND METHODS

Plant materials

The list of plants included in the validation experiment is provided in Supplementary Table S1. A total of 78 distinct *Musa* accessions were used, and they were provided by the Bioversity International Musa Germplasm Transit Center (ITC, Belgium), ANSES (La Réunion, France), CIRAD (Montpellier, France), and the University of Queensland (Australia). The plants were previously tested following the standard virus detection protocols (Geering et al. 2000; Thomas et al. 2015), and the following nine virus species were detected in at least one sample: banana mild mosaic virus (BanMMV), banana bract mosaic virus (BBrMV), banana bunchy top virus (BBTV), cucumber mosaic virus (CMV), banana streak OL virus (BSOLV), banana streak CA virus (BSCAV), banana streak GF virus (BSGFV), banana streak IM virus (BSIMV), and banana streak MY virus (BSMYV). These plants were either sequenced individually or in a pool (corresponding to a single sequencing library) in five different batches.

For each sequencing batch, at least one pool of five positive controls was sequenced together with the samples. These positive controls were plants maintained for more than a decade in the greenhouse of Gembloux Agro-Bio Tech (Liège University, Belgium) and used as positive controls during routine virus detection tests based on molecular tests and electron microscopy (De Clerck et al. 2017). The positive control plants from batch 1 were pooled using the same quantity of tissue from each plant before the RNA extraction for batch 1 (e.g., using 20 mg per plant instead of 100 mg). For batches 3, 4, and 5, a different proportion for each virus was used (BBTV : BanMMV : BSV [either BSMYV or BSOLV] : BBrMV : CMV : Healthy plants = 6:4:4:2:1:5). This adaptation of proportion was carried out to standardize read numbers above a theoretical limit of detection of the test (as BBTV tends to provide fewer reads and CMV a lot). The analytical sensitivity was also evaluated using serial dilutions of four pools of five plants (see hereunder), always using the same proportion per sample (20 mg).

For each sequencing batch, at least one alien control (wheat plant infected with barley yellow dwarf virus [BYDV]) was processed in parallel with the samples as an external control. The goal of an alien control is to monitor the cross-contamination level among the samples in a single batch (Massart et al. 2022). Leaves of wheat plants infected with BYDV-PAS and BYDV-PAV and maintained in the greenhouse of Gembloux Agro-Bio Tech were sampled via the same process as for banana plants. The concentration of virus reads in the sequence dataset from such a sample was observed previously as very high in the growing conditions of the greenhouse (Tamisier et al., *unpublished data*). Importantly, *Musa* species are not a host for BYDV, so the presence of reads from BYDV in any *Musa* sample reveals a cross-sample contamination.

In total, the sequenced samples corresponded to 78 different plants on which 702 (78 × 9) individual predictions could be made with the nine virus species mentioned above.

Sampling and nucleic acids extraction

For the accessions grown in a greenhouse, one third of the uppermost fully expanded leaves were sampled from four individual plants for each accession, and 100 mg of tissue (random punches of 0.6 cm diameter circles) was used for nucleic acids extraction. For samples received from collaborative laboratories, 100 mg of tissue (random punches of 6 mm diameter) and 10 mg of tissue (from randomly selected leaf parts) were sampled from fresh or lyophilized leaves, respectively (weights used were adapted to maintain homogeneity between fresh and dry samples). The same process was carried out for the alien control.

Total RNA was extracted from the sampled leaves using the RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands) according to the manufacturer's instructions and DNase treated with DNase I, Amplification Grade (Invitrogen, CA, United States) according to the manufacturer's instructions. The RNA samples were aliquoted and stored at −80°C before sending for sequencing or confirmation analysis.

For the confirmation study on BSV detection, genomic DNA was extracted from the sampled leaves using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions.

PCR-based protocols

The five abovementioned banana viruses were detected by PCR and RT-PCR using cDNA generated from extracted total RNA, respectively. The sequences of the primers used are listed in Supplementary Table S3. For targeted RT-PCR, cDNA was synthesized from 1 μg of total RNA with SuperScript III Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions, and random hexamers (Invitrogen) were used. PCRs were performed in 25-μl reactions. The mix used for BanMMV, BBrMV, BBTV, and CMV contained 2 μl of cDNA, 1 μl of forward primer (25 μmol l^{-1}), 1 μl of reverse primer (25 μmol l^{-1}), 5 μl of 5X PCR buffer, 1 μl of 50 mM $MgCl_2$, 0.5 μl of 10 mM dNTP mixture, 0.5 μl of Mango *Taq* DNA Polymerase (Bioline, London, United Kingdom), and 14 µl of sterilized and distilled water. The temperature cycles used for each virus were:

BanMMV: 94°C for 1 min, followed by 35 cycles of 15 s at 94°C, 20 s at 60°C, and 20 s at 72°C; final elongation step at 72°C for 3 min.

BBTV, BBrMV, and CMV: 94°C for 1 min, followed by 35 cycles of 20 s at 94°C, 20 s at 60°C, and 40 s at 72°C; final elongation step at 72°C for 3 min.

BSV detection (from cDNA): 2 µl of cDNA, 2.5 µl of primer mixtures (see primer list in Supplementary Table S3; the final concentration for each primer was 1 µmol l^{-1} and 2 µmol l^{-1} for BSMYV detection primers), 5 µl of 5X PCR buffer, 0.75 µl of 50 mM $MgCl_2$, 0.5 µl of 10 mM dNTP mixture, 0.5 µl of Mango *Taq* DNA Polymerase (Bioline), and 13.75 µl of sterilized and distilled water. The PCR and RT-PCR programs used for BSV were as follows. The PCR cycle consisted of a pre-incubation step at 94°C for 30 s, followed by 35 cycles of 15 s at 94°C, 30 s at 60°C, and 1 min at 72°C and a final elongation step at 72°C for 10 min. To confirm the HTS detection of BSMYV and BSIMV previously undetected by virus indexing, the same mastermix and temperature cycles were used, with the only change being in the reverse primer sequence (Supplementary Table S3).

To evaluate the presence of BSV and BanMMV viral particles, previously described protocols based on immunocapture followed by (RT-)PCR were applied (De Clerck et al. 2017).

The obtained PCR products were separated on a 1.0% agarose gel in TAE 1× stained with GelRed Nucleic Acid Gel Stain (Biotium, Fremont, CA, United States).

HTS of total RNA extracts

The library preparation and sequencing were carried out at the Interdisciplinary Center of Biomedical Research of Liège University (GIGA, Liège, Belgium). Before sequencing, the RNA concentration and RNA integrity number (RIN) of each sample were analyzed (qPCR, using the KAPA kit). Stranded Total RNA Library Prep Human/Mouse/Rat (Illumina, San Diego, CA, United States; hereafter referred as the "old kit") or TruSeq Stranded Total RNA Library Prep Plant (Illumina; hereafter referred as the "new kit") were used for batches 1-2 and 3-5, respectively. As ribosomal depletion step was not included in the old kit; Ribo-Zero Plant Leaf Kit (Illumina) was used before applying the old kit. The prepared libraries were quantified, pooled and paired-end sequenced on a NextSeq 500 sequencer (2 × 150 bp read length for batches 1, 2, 3, 4) and NovaSeq 6000 for batch 5 from Illumina and operated by the GIGA facilities of Liège University. Sequencing batches were processed on the following dates: April 2018 (batch 1), October 2018 (batch 2), January 2019 (batch 3), September 2019 (batch 4), and November 2020 (batch 5).

For small RNA sequencing, a single batch was prepared and sequenced. The library preparation and sequencing were carried out in the GIGA facilities using the SMARTer smRNA-Seq Kit (Clontech – Takara Bio). The sequencing of 1 × 50 nt was carried out on NextSeq 500.

Bioinformatics analyses

The obtained sequence reads (FastQ file format) were quality controlled on both ends using a minimal nucleotide Phred quality score of 25 and a minimal length of 35 bp with BBDuk (Kechin et al. 2017) (v38.37). Then, the trimmed reads were merged, and duplicated merged reads were removed using Dedupe (Bushnell et al. 2017) (v38.37) with kmer seed 31.

All the cleaned reads (merged and unmerged) were mapped to the custom-built database with 820 whole genomes of BanMMV, BBrMV, BBTV, BSV, CMV, and BYDV, which were available from NCBI (accessed 12/12/2020; Supplementary Table S4). Except for BanMMV, all the other tested viruses had more than one reference genome available, and all the cleaned reads were mapped to these available references at the same time. The mapping tool from GeneiousPrime (2020.0.5, Biomatters) was used as the high-sensitivity mapping method. We used the default parameters associated with the "Low sensitivity/Fastest" profile allowing two iterations and manually setting a 20% mismatch and a maximum of three nucleotide gaps allowed. Those parameters were used for all samples except for small RNA on which adapted parameters (two iterations, 10% mismatch, and maximum three nucleotide gaps allowed) were used. No reads with multiple best matches between different viruses were observed. The BSV species detection was considered separately for the test performance characteristics.

De novo assembly with rnaSPAdes (Bushmanova et al. 2019) (v3.13.0) to obtain contig and tBlastX with RefSeq (November 2020) analysis were performed to check the presence of a new or known unexpected virus (not targeted by the PCR test). For a new virus detected, we combined tBlastX with the NCBI nonredundant database (November 2020) and pairwise alignment (GeneiousPrime, 2020.0.5, Biomatters) to identify the closest reference.

Evaluation of analytical sensitivity

The analytical sensitivity between RT-PCR and HTS was compared using four pooled samples from five plants each infected by one viral species: BanMMV, BSOLV, BBTV, BBrMV, or CMV (Supplementary Table S2). RNA extractions were carried out individually for each plant and pooled together using the same quantity of RNA from each extract. Each pool represented a 5× dilution for each virus. The pools were further diluted to reach 100-, 1,000-, and 10,000-fold dilutions. The pools were diluted in RNA extracted from virus-free plants. All the RNA pools were aliquoted and immediately stored at −80°C.

Evaluation of repeatability and reproducibility

Eight independent samples (including positive, negative, and alien controls) were analyzed in replicates from different samplings of plant tissue from the same plant at the same time. The number of replicates analyzed in parallel ranged from 2 to 5 and is indicated by "r" in the sample code column of Supplementary Table S1.

The reproducibility of the HTS test over time was also evaluated on the positive control mix at each sequencing batch (five independent sequences starting from plant material, although in different proportions in batches 3 to 5 compared with batch 2). In addition, three samples among the replicates were tested in two different sequencing batches (ITC1543, Sample J, Sample EM4). The samples tested for reproducibility over time always corresponded to new plant tissue samples. The impact of the library preparation kit was also evaluated due to the disruption of the first kit used during this validation by using the last reagents available for three samples.

Bioinformatics analyses to investigate false positive detections

To investigate false positive detections, single-nucleotide polymorphism (SNP) profiling was done, and comparative deduplication was performed between supposed contaminating and contaminated samples. When the reads abundance allowed it, SNPs were detected using the Geneious (Prime 2020.0.5, Biomatters) "find variant" functionality (only variants with more than 25% frequency were kept). Then, the SNP distribution was compared between samples. The second strategy aimed to evaluate the number of identical reads between contaminating and contaminated samples. It was performed using a comparative deduplication strategy. The mapped unique reads from samples with a putative false positive detection and from samples positives for the targeted virus were grouped into a single pool (using "Group sequences into a list"). Then, the deduplication tool Dedupe V38.37 (Bushnell et al. 2017) (from BBMap) was used with the parameters kmer seed length, maximum edit, and maximum substitutions set as "31," "0," and "0," respectively.

Calculation of diagnostic sensitivity

The diagnostic sensitivity was calculated according to recent recommendations for the statistical analysis of validation datasets (Massart et al. 2022). The known formula was used (true positive/(true positive + false negative)) but without considering diluted samples or alien controls. However, for analytical sensitivity calculation, diluted samples were used. The occurrence of false positives was monitored and investigated.

Alien controls were used to evaluate the impact of threshold determination on the balance between diagnostic sensitivity and false positive occurrence.

RESULTS

Generated sequencing data

In total, 136 libraries were sequenced by HTS from individual or pooled total RNA in five separate HTS runs, each comprising 19 to 38 libraries (Supplementary Table S1). These libraries originated from 78 different *Musa* accessions, nine positive controls corresponding to the same mix of plants (same for all batches except for batch 1), seven negative controls (virus-free *Musa* accessions), 11 alien controls, and 18 pooled samples (containing *Musa* infected with BanMMV, BSOLV, BBrMV, BBTV, and CMV at various dilution levels). The data (about 2 billion reads in total) generated by each independent sequencing batch is summarized in Table 1 and publicly available (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA777477).

Whatever the batch, the minimal and maximal number of reads per sequencing library were 4.7 and 36.8 M, respectively. The average numbers of reads generated per sequencing batch ranged from 8 to 12 M, except batch 4, with 26 M. In addition, the average numbers of contigs generated after de novo assembly for each library ranged from 52,949 to 102,211. Small RNA from 27 samples were sequenced in a single batch. The minimal and maximal number of reads per sequencing library were 8 and 11 M, respectively, with an average of 9.5 M.

Comparison between small RNA and total RNA sequencing

The summary tables (Supplementary Table S1A and B) show that small RNA sequencing and total RNA sequencing protocols generated a similar number of reads per sample, with an average of 9.5 and 7.9 M, respectively. The minimal number of reads was 8.3 and 4.7 M for small RNA and total RNA, respectively. Nevertheless, the proportion of viral reads in samples sequenced by the small RNA protocol ($n = 101{,}168$ reads) was much lower compared with the total RNA sequencing protocol ($n = 396{,}574$ reads). The relatively lower number of reads could limit the analytical sensitivity of the protocol. In addition, 11 virus detections were achieved with 10 or fewer reads, and one false negative result was observed. Based on these results, the total RNA sequencing protocol was selected for further validation of its performance criteria.

Monitoring cross-contamination burden with the alien control

As *Musa* species are not a host for BYDV, the cross-sample contamination between the alien control and any other samples or control can be monitored by identifying and counting BYDV reads in each sequencing dataset. The results are summarized in Table 2.

Table 2 underlines several trends in the cross-contamination level according to the analysis of the detection of BYDV reads in *Musa* samples. First, the cross-contamination level is highly variable depending on the sequencing batch. Indeed, the ratio between reads generated in the alien control(s) and the cross-contaminating (BYDV) reads ranged between 190× and 7,269× in the different batches. In addition, the number of *Musa* samples with cross-contaminating (BYDV) reads is also variable, from less than 10% ($n = 2$ in batch 1) to more than 90% ($n = 15$ in batch 2). It is worth mentioning that the lowest number of BYDV reads was observed with the first batch processed, whereas in batch 2, most samples had fewer than 20 cross-contaminating reads. In addition, within a batch, the number of cross-contaminating reads detected in a sample could vary significantly, ranging from 0 (five samples) to 288 within batch 3. The high ratio of batch 3 is in fact mainly due to two samples with 184 and 288 BYDV reads. Without these two samples, the ratio dropped to 668×. Therefore, the cross-contamination detected could be quite constant between samples (batch 2) or concentrated in a few samples (batch 3).

Determining the adaptative detection/contamination threshold and impact on performance criteria

The cross-contamination burden monitored with BYDV reads, and its observed variability between and within sequencing batches, raised the fundamental question of fixing a threshold for detection of the viruses infecting the *Musa* samples. As reference samples were used in this study, it was possible to evaluate the

TABLE 1

General information on the high-throughput sequencing data generated by five independent sequencing batches

Parameter	Sequencing batch 1 total RNA	Sequencing batch 2 total RNA	Sequencing batch 3 total RNA	Sequencing batch 4 total RNA	Sequencing batch 5 total RNA	Sequencing batch 6 small RNA
Number of samples sequenced	27	19	27	38	25	31
Average number of reads generated	8,026,387	10,177,611	9,626,664	26,868,110	12,072,128	9,446,662
Maximum number of reads generated	10,340,172	11,379,224	12,309,622	36,861,696	15,597,574	11,285,911
Minimum number of reads generated	4,678,326	6,888,958	7,043,406	20,469,598	9,889,872	7,965,444
Average number of contigs generated	52,949	141,862	91,092	102,211	54,746	NA

impact of several detection thresholds on the false positive (when a virus was detected while it corresponded to cross-contamination) and false negative (when a virus infection was missed because it was considered cross-contamination due to the low number of viral reads) rates. We tested two simple contamination thresholds to evaluate their impact on the accuracy, false positive rate, and false negative rate compared with the absence of any threshold. The first threshold was 10 reads, which has been proposed empirically in the literature (Bloom et al. 2021; Soltani et al. 2021; Strong et al. 2014). Nevertheless, the fixation of a unique threshold did not consider the variability of the cross-contamination burden shown by the alien control. The second threshold is based on the cross-contamination observed from the alien control and is therefore variable between batches. Different metrics could be tested (average number of reads, average or maximum ratio between contaminants and alien reads, etc.) but, after preliminary evaluation (data not shown), a conservative threshold corresponding to the maximum of cross-contaminating reads (here from BYDV) observed in a sample for each batch was selected. This variable threshold considered the inter-batch variability observed and was considered "conservative" because the maximum level was selected. The alien threshold for virus detection was therefore 7 reads for batch 1, 75 reads for batch 2, 288 reads for batch 3, 44 reads for batch 4, and 69 reads for batch 5 (Table 2).

It is important to mention that an unusually high abundance of BSV reads was detected in many samples indexed negative for BSV species, probably due to the integration of episomal BSV genomes in the B genome of *Musa* species. Consequently, the accuracy, false positive rate, and false negative rate were presented independently for BSV species.

Table 3 shows that, whatever the filter, the false positive rate for BSV species was particularly high, confirming the necessity to address this viral complex separately. The absence of a threshold allows for the detection of all the viruses present in the samples at the expense of a high false positive rate (up to 92% for BSV and 71% for the other viruses). These results confirmed the overall cross-contamination burden observed with the alien control.

Despite recent recommendations for the statistical analysis of validation datasets for plant pests (Massart et al. 2022), the serially diluted samples (up to 10,000× dilution) were kept in this preliminary analysis (explaining the higher false negative rate for batch 4). As a result, even with dilutions included, the alien threshold presented a low false positive rate (2%), showing that even with a low virus concentration or number of reads in a sample, no additional misidentification was observed. In fact, the two false positives for this batch corresponded to two BBrMVs, which are discussed later. The application of the "10 reads" threshold improved the accuracy considerably (ranging between 84 and 93% for the viruses other than BSV species) with up to 10-fold reduction of the false positive rate while still ranging between 3 and 16%. Interestingly, the false negative rate did not improve for batches 2, 3, and 5, but it reached 2% for batch 1. The threshold based on the alien control provided overall the best accuracy (always equal to or higher than the "10 reads" threshold) with, again, a marked decrease in the false negative rate. Nevertheless, it caused more false negative results in batches 2 (3%) and 3 (4%), but no false negative was observed for batches 1 and 5. For further analysis, the alien-based threshold was kept because it showed the highest accuracy.

Additionally, a background cross-contamination of *Musa* virus reads was observed in the alien controls, with 1 read in batch 3, 29 reads in batch 1 (4 alien controls), and 387 reads in batch 2 (3 alien controls). No *Musa* virus read was observed in the alien control of batch 4.

BSV genome integrated in *Musa* B genome limits the HTS test performance

Table 3 suggests a high rate of false positive results for BSV species. As stated earlier, BSV sequences integrated in the genome can be activated and can trigger an infection with a viral particle of BSV. So far, the complete genomes of five species have been reported in the *Musa* B genome: BSOLV, BSGFV, BSVNV, BSMYV, and BSIMV, among which BSOLV, BSIMV, and BSGFV have proven to be activatable, whereas partial sequences of BSV are integrated (Chabannes et al. 2021). The factors triggering gene transcription and/or spontaneous infection by viral particles and the plant recovery are not yet well understood and vary between accessions, growth stage of the plant, environmental conditions, and possibly other unknown factors. This means that reads from integrated BSV could be potentially detected in the absence of viral particles. To analyze the false negative rate in more depth, the detected BSV reads were summarized depending on the *Musa* genome and the presence of viral particles (Table 4).

For the plants without BSV particles, a marked difference was observed between accessions containing a B genome (whatever their ploidy and genome combination) and accessions with only an A genome. Indeed, BSV reads were detected in more than 90% of the samples (31 of 34 datasets) with a B genome, with an average number of 224 reads. For 12 samples, the number of BSV reads was above the alien threshold of the corresponding batch. In contrast, 100 times fewer BSV reads were detected in the samples without a B genome, with an average number of three reads. Most of the samples with an A genome had only one or two BSV reads (contamination background), whereas one sample had 17 reads and the other one 44 reads (both under the alien threshold of the respective batches). To independently confirm the results, six *Musa* samples with a B genome but without a BSV particle detected, 1 *Musa* sample with only an A genome, and one *Musa* sample with a B genome and infected with BSOLV particles were further investigated. All those samples went through both PCR and immunocapture PCR (IC-PCR) detection. The comparison of PCR and IC-PCR results of the selected samples is shown in Figure 1. According to the results, BSMYV, BSOLV, and BSIMV could be detected by PCR from the genomic DNA of selected samples

TABLE 2

Summary of the barley yellow dwarf virus (BYDV) reads detected in the sequencing datasets for each batch of sequencing

Parameter	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Total number of BYDV reads in alien controls	87,224	259,736	125,616	61,204	102,714
Total number of cross-contaminating BYDV reads in banana sample	12	300	660	93	79
Ratio of cross-contaminating reads	1/7,269	1/866	1/190	1/658	1/1,300
Number of samples cross-contaminated	2	15	21	8	3
Number of samples without cross-contaminating reads	21	1	5	15	34
Maximum number of cross-contaminating reads in a single sample	7	75	288	44	69
Average number of cross-contaminating reads per sample	6	20	27	11	26

with a B genome but tested negative according to IC-PCR, whereas no band was found by PCR or IC-PCR from the sample (ITC1833) with only an A genome, and a band was observed by PCR and IC-PCR for the sample infected by BSV particles.

On average, there were 30× more BSV reads in the presence of viral particles of BSV compared with plants with a B genome but without particles. Nevertheless, a clear threshold could not be set to distinguish the categories. Indeed, the maximum number of reads (840) of negative accessions with a B genome was higher than the number of BSV reads generated for seven plants infected with BSV particles.

Higher inclusivity at the isolate and species level for the HTS test

The HTS test detected a BSIMV infection in accession ITC1843 and a BSMYV infection in accession ITC1599, although these samples were indexed as negative for the respective viruses. To investigate these divergent results, the primers used during virus PCR detection were aligned with the whole genome sequence generated. Six and two mismatches were found for the BSIMV and BSMYV sequences, respectively. Based on the alignment of these new genome sequences and on the sequences available on the NCBI database for each species (accessed on 24/10/2019, with two and seven whole genome sequences available for BSIMV and BSMYV, respectively), new primers were designed (Supplementary Table S3). These primers were tested on both accessions and compared with the former primers following the IC-PCR protocol to validate the presence of viral particles as BSMYV can be integrated in the *Musa* B genome. A positive result was only obtained with the newly designed primers (Fig. 1), confirming the higher inclusivity of HTS testing.

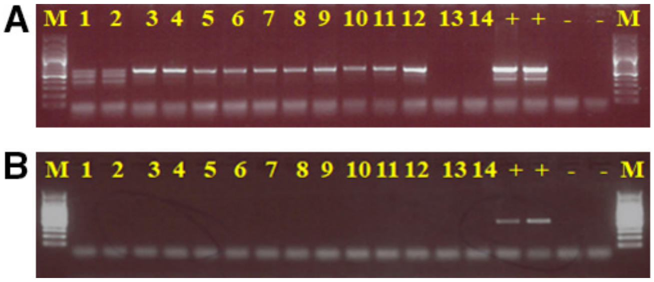

A
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 + + - - M
B
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 + + - - M

FIGURE 1

A, PCR and **B,** immunocapture PCR results of banana streak virus (BSV) species detection (banana streak OL virus [BSOLV], banana streak MY virus [BSMYV], banana streak IM virus [BSIMV]) on selected samples. 1,2: ITC1607 'Birbutia' (ABB), BSV uninfected; 3,4: ITC0148 'Isansi' (AAB), BSV uninfected; 5,6: ITC1498 'Libanga Dark Green' (AAB), BSV uninfected; 7,8: ITC1565 'Halahala' (AA), BSV uninfected; 9,10: ITC0146 'Mushaba' (AAB), BSV uninfected; 11,12: ITC1852 'Amagaba' (AAB), BSV uninfected; 13,14: ITC1833 'ShweNi' (AAA), BSV uninfected; +: ITC1867 'Atili' (AAB), BSOLV infected in duplicate; -: ddH_2O as templates; M: GeneRuler 100 bp DNA ladder from Thermo Scientific. When three bands are observed, they correspond to the detection of BSOLV (700 bp), BSMYV (589 bp), and BSIMV (400 bp).

The BLAST annotation of de novo contigs allowed for the detection of other virus species compared with the targeted molecular test and the immuno-capture electron microscopy protocol (De Clerck et al. 2017). More specifically, contigs presenting homologies with ampelovirus species were detected in accessions ITC1845, ITC1872, and one of the pooled samples. One contig per sample corresponded to nearly complete genomes of

TABLE 4

Summary of detection of banana streak virus (BSV) reads depending on the presence of B genome and the detection of BSV particles by immunocapture PCR

Musa genome	All genotypes	B genome present	B genome absent
Detection of viral particles	Yes	No	No
Total number of BSV reads	218,906	7,647	73
Number of samples with BSV reads	33	31	12
Number of samples without BSV reads	0	3	11
Average number of reads	6,633	224	3
Maximum number of reads	66,741	840	40
Minimum number of reads	144	1	0

TABLE 3

Evaluation of the impact of two alien thresholds on the accuracy, false positive rate, and false negative rate per batch[a]

		Non-BSV[b] virus predication			BSV virus predication		
Sample	Filter	Accuracy	False positive	False negative	Accuracy	False positive	False negative
Batch 1	No filter	49%	51%	0%	14%	86%	0%
	10 reads	93%	5%	2%	57%	43%	0%
	Alien control	93%	5%	2%	52%	48%	0%
Batch 2	No filter	29%	71%	0%	8%	92%	0%
	10 reads	84%	16%	0%	51%	49%	0%
	Alien control	94%	3%	3%	89%	11%	0%
Batch 3	No filter	43%	57%	0%	13%	87%	0%
	10 reads	94%	6%	0%	48%	52%	0%
	Alien control	92%	4%	4%	98%	2%	0%
Batch 4 (with dilution)	No filter	86%	11%	2%	17%	83%	0%
	10 reads	88%	3%	9%	48%	48%	4%
	Alien control	77%	2%	20%	72%	21%	7%
Batch 5	No filter	80%	20%	0%	11%	89%	0%
	10 reads	90%	10%	0%	51%	49%	0%
	Alien control	100%	0%	0%	81%	19%	0%

[a] For each batch, the first line represents the high-throughput sequencing read base detection with no minimum reads required, the second with 10 reads required, and the third with a number depending on the alien (highest number of contaminated alien virus [barley yellow dwarf virus] reads from alien the control).
[b] BSV = banana streak virus.

ampeloviruses. Pairwise comparison of these contigs with reference sequences of other species belonging to family *Closteroviridae* confirmed that these contigs belonged to the genus *Ampelovirus* but were divergent from known species. The highest similarity at the amino acid level was 85% for one of the contigs with sugarcane mild mosaic virus (SCMMV, GenBank accession No. MN116751). These results will be detailed and discussed in a future publication. In addition, two contigs presented homologies with RNA1 and RNA2 of crinivirus species. They were detected in accession ITC1905 and corresponded to nearly complete genomes, with the highest genome identity of 63% and 66% with RNA1 and RNA2 of lettuce chlorosis virus (GenBank FJ380118 and FJ380119). An in-depth analysis of these contigs and their presence in *Musa* germplasm will be presented in another publication (Rong et al., *unpublished data*). The presence of these contigs related to crinivirus or ampelovirus species was confirmed by independent RT-PCR (Rong et al., *unpublished data*).

Diagnostic sensitivity of the HTS test

The alien control datasets and the datasets from diluted samples were not included in the in-depth analysis of diagnostic sensitivity (DSE), as recommended in the recent guidelines for statistical analysis of validation datasets (Massart et al. 2022). Therefore, a total of 111 datasets were included in the analysis, corresponding to a total number of 999 detection events. Among the 111 datasets, 58 were generated from plants without viruses, whereas 53 came from plants infected by at least one virus species (and up to five). According to PCR detection results, a total of 120 virus detections should occur within these 53 datasets.

Using the alien threshold, 115 true positives were detected, representing a DSE of 96%. Considering only the nonintegrated viruses (CMV, BBTV, BanMMV ,and BBrMV), 81 virus detections should be observed (among 444 events). Seventy-six true positives were detected (DSE = 94%). When analyzing the DSE per viral species, the DSE dropped to 75% for BBTV. This high false negative rate is likely due to low abundance of BBTV reads in the different datasets from infected samples, with an average of 380 reads and a maximum of 2,397 reads per dataset. Of note, the positive samples with BBTV corresponded to a 5× dilution of the concentration due to the sample pooling. The alien threshold was fixed based on the abundance of reads in the alien controls that was between 10× and 72× higher than the maximum number of BBTV reads. Therefore, even if the number of BBTV reads of several positive datasets is below the alien threshold, it is unlikely that they come from a cross-contamination. They could therefore be interpreted as positive results during expert review of the data.

Repeatability and reproducibility of the HTS test

Considering the problem of the low BBTV reads number, four samples with a number of reads under the alien threshold were considered positive for BBTV in this analysis (as a consequence of the expert analysis). The repeatability of virus detection is 100%. The reproducibility of virus detection from the same plants between sequencing batches was also 100%.

The reproducibility of the library preparation kit was also evaluated as, during the experiments, the first kit used (Stranded Total RNA library Prep Human/Mouse/Rat, the "old kit") was no longer produced by the provider. For batch 3, another kit (TruSeq Stranded Total RNA Library Prep Plant, the "new kit") was used. A small-scale comparison between old and new kits (Supplementary Table S1) was performed. The same viruses were detected between both kits (100% reproducibility), with minor differences in the number of reads mapped to each virus.

Although 100% repeatability was obtained for virus detection, a marked variation in read number was observed between biological replicates (e.g., from different samplings from the sample tissue at the same time). For example, with a similar number of reads generated, the number of reads per virus in the four replicates of the positive control varied from 835 to 79,496 for CMV or from 22 to 437 for BanMMV. A similar variation was observed when evaluating the reproducibility. The sequencing of the same sample (sample J) generated 1,519× more BBrMV reads with batch 3 compared with batch 2 (Supplementary Table S1A).

Evaluation of analytical sensitivity for HTS and RT-PCR

As the concentration of viruses can be heterogeneous in banana plants, the comparison of analytical sensitivity between different tests must be carried out from the same extract. Therefore, the HTS analytical sensitivity was compared with RT-PCR performance starting from the same RNA extracts.

The limit of detection by targeted RT-PCR was variable depending on the virus species as BanMMV could only be detected by RT-PCR with 5× dilution, whereas CMV could still be detected even after 1,000× dilution. This observation cannot be generalized as the virus concentration can be variable depending on the genotype and physiological stage of the plant, as well as the virus replication stage. Additionally, BBrMV was detected with up to 5× dilution for pool C, 100× dilution for pools B and D, and 1,000× dilution for pool A. The total RNA samples with all the dilution levels from pools A to D were tested by HTS. All the results are summarized in Figure 2.

First, the impact of any detection threshold on the analytical sensitivity was important. Indeed, without a threshold, all the viruses are detected, even at 10,000× dilution, in some cases with a few reads. As previously shown, this high analytical sensitivity was nevertheless counterbalanced by a high number of false positive detections due to the cross-contamination burden. As expected, the limit of detection varied depending on the threshold and the virus species as the threshold setting impacted BBTV and BSOLV much more than CMV (with many more reads generated). Therefore, as for the evaluation of other performance characteristics of the HTS test, the comparison with RT-PCR was based on the alien threshold. The limit of detection of HTS tests was between 10× and 100× better for BanMMV, equal to or 10× better for BBTV, equal to or 100× better for BBrMV, and equal to or 10× better for CMV. For BSOLV, the limit of detection of the HTS test was equal or even lower. Overall, the HTS test presented an analytical sensitivity equal to or improved compared with the RT-PCR test on the same RNA extracts.

Investigating unexpected results

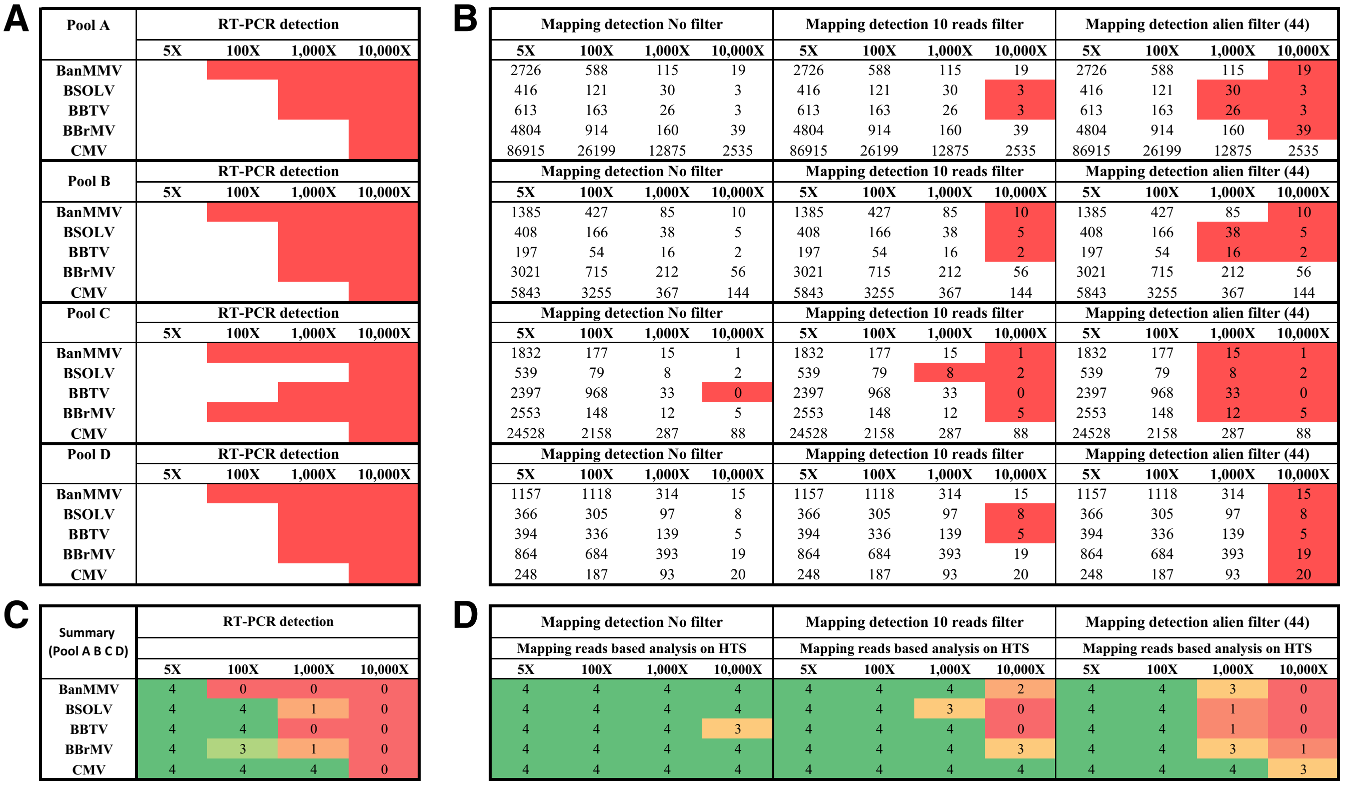
The DSE reached 100% for the other viruses, except for BanMMV. Indeed, one dataset (ITC1536) did not generate BanMMV reads, although it was indexed as positive. For further verification, the preserved lyophilized leaf and the extracted total RNA used for the HTS was tested for BanMMV using IC-RT-PCR and RT-PCR, respectively. BanMMV was not detected from extracted total RNA, confirming the absence of BanMMV RNA in the extract, but a positive result was obtained using IC-RT-PCR. The most probable origin of this result is the heterogeneous distribution of BanMMV inside the plant tissues and its absence in the sampled tissue.

An unexpected detection of CMV in samples ITC1498 and 1565 (batch 3) was observed as the accessions tested negative during PCR, but 845 and 686 reads were detected, respectively. Mapping of CMV reads on a reference genome presented a genome coverage (all 3 RNA fragments) of 99.4% with 17.6 mean depth and 97.8% with 14.7 mean depth for samples ITC1498 and ITS1565, respectively. When analyzing the sequence of the primers used during PCR and the reads generated, a single mismatch was observed. The CMV contigs obtained for this dataset and the ones generated from other datasets with CMV infection for this batch of samples (two positive controls and the sample BBrMV2 No. 208) were compared, but they were different, with many SNPs between them, suggesting that this detection did not come from cross-contamination between these samples. As the RNA extract was not available, additional sequencing from a new sample prepared from the same plant (sampled from lyophilized tissue) showed no trace of CMV (0 reads observed on approximately 9 M reads per sample). We later investigated the case by analyzing the viral reads presence in samples coming from different projects (RNA extracted in another laboratory) but with the library prepared and sequenced in the same batch by the sequencing provider. Importantly, two samples from an independent project presented a high quantity of CMV reads: On a total of 12 M (10 M) reads, there were 8 M (5 M) reads, corresponding to 708,736 (1 M) unique reads mapped on the CMV genome. The SNP profile of the alignment was compared with the SNP profile of the aligned reads from ITC1498 and ITC1565. Nearly all (>99%) the SNPs detected in the ITC1565 (1498) were also found in external sample 1 (2) suggesting two specific cross-contamination events. The SNP list observed after mapping the four datasets on the RNA3 reference sequence is detailed in Supplementary Table S5. The rate of contamination from the original sample was low as it corresponded to 0.01 and 0.02% for ITC1565 and 1498, respectively.

For batch 4, 199 and 121 reads of BBrMV were observed in datasets from ITC1854 and ITC1858, respectively, even though both accessions were PCR-negative for BBrMV. After mapping, the genome coverage observed was 72% with 3 mean depth and 61% with 1.7 mean depth for ITC1854 and ITC1858 respectively. The primer sequences were not covered by the sequencing reads, and the presence of mismatches could not be analyzed. As the genome coverage and depth were low, an SNP profiling analysis could not be done; instead, a deduplication analysis was performed by comparing these 199 and 121 reads with the 38,666 reads generated from the most abundant BBrMV sample. First, the 320 reads were deduplicated, leading to 249 unique reads (79%). Later, the duplication analysis of the 38,666 reads from BBrMV-positive samples of batch 4 generated 32,425 unique reads (84%). None of the 249 unique reads was 100% identical to any of the 32,425 unique reads from infected samples, suggesting that cross-contamination between these samples did not occur. Another HTS test was carried out from new samples taken from lyophilized tissue of these two accessions, but no BBrMV reads were detected in any of the samples (approximately 11 M per sample, *unpresented data*). No clear conclusion could be reached on the origin of the detection of BBrMV reads for these samples.

DISCUSSION

HTS of ribosomal depleted total RNA or small RNA are the two most popular protocols applied for plant virus detection. When comparing the performance characteristics of both protocols,



A

Pool A	RT-PCR detection			
	5X	100X	1,000X	10,000X
BanMMV		🟥	🟥	🟥
BSOLV			🟥	🟥
BBTV			🟥	🟥
BBrMV				🟥
CMV				🟥
Pool B	RT-PCR detection			
	5X	100X	1,000X	10,000X
BanMMV		🟥	🟥	🟥
BSOLV			🟥	🟥
BBTV			🟥	🟥
BBrMV			🟥	🟥
CMV				🟥
Pool C	RT-PCR detection			
	5X	100X	1,000X	10,000X
BanMMV		🟥	🟥	🟥
BSOLV				🟥
BBTV			🟥	🟥
BBrMV		🟥	🟥	🟥
CMV				🟥
Pool D	RT-PCR detection			
	5X	100X	1,000X	10,000X
BanMMV		🟥	🟥	🟥
BSOLV			🟥	🟥
BBTV			🟥	🟥
BBrMV			🟥	🟥
CMV				🟥

B

Mapping detection No filter				Mapping detection 10 reads filter				Mapping detection alien filter (44)			
5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X
2726	588	115	19	2726	588	115	19	2726	588	115	🟥 19
416	121	30	3	416	121	30	🟥 3	416	121	🟥 30	🟥 3
613	163	26	3	613	163	26	🟥 3	613	163	🟥 26	🟥 3
4804	914	160	39	4804	914	160	39	4804	914	160	🟥 39
86915	26199	12875	2535	86915	26199	12875	2535	86915	26199	12875	2535
Mapping detection No filter				Mapping detection 10 reads filter				Mapping detection alien filter (44)			
5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X
1385	427	85	10	1385	427	85	🟥 10	1385	427	85	🟥 10
408	166	38	5	408	166	38	🟥 5	408	166	🟥 38	🟥 5
197	54	16	2	197	54	16	🟥 2	197	54	🟥 16	🟥 2
3021	715	212	56	3021	715	212	56	3021	715	212	56
5843	3255	367	144	5843	3255	367	144	5843	3255	367	144
Mapping detection No filter				Mapping detection 10 reads filter				Mapping detection alien filter (44)			
5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X
1832	177	15	1	1832	177	15	🟥 1	1832	177	🟥 15	🟥 1
539	79	8	2	539	79	🟥 8	🟥 2	539	79	🟥 8	🟥 2
2397	968	33	🟥 0	2397	968	33	🟥 0	2397	968	🟥 33	🟥 0
2553	148	12	5	2553	148	12	🟥 5	2553	148	🟥 12	🟥 5
24528	2158	287	88	24528	2158	287	88	24528	2158	287	88
Mapping detection No filter				Mapping detection 10 reads filter				Mapping detection alien filter (44)			
5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X
1157	1118	314	15	1157	1118	314	15	1157	1118	314	🟥 15
366	305	97	8	366	305	97	🟥 8	366	305	97	🟥 8
394	336	139	5	394	336	139	🟥 5	394	336	139	🟥 5
864	684	393	19	864	684	393	19	864	684	393	🟥 19
248	187	93	20	248	187	93	20	248	187	93	🟥 20

C

Summary (Pool A B C D)	RT-PCR detection			
	5X	100X	1,000X	10,000X
BanMMV	4	0	0	0
BSOLV	4	4	1	0
BBTV	4	4	0	0
BBrMV	4	3	1	0
CMV	4	4	4	0

D

Mapping detection No filter				Mapping detection 10 reads filter				Mapping detection alien filter (44)			
Mapping reads based analysis on HTS				Mapping reads based analysis on HTS				Mapping reads based analysis on HTS			
5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X	5X	100X	1,000X	10,000X
4	4	4	4	4	4	4	2	4	4	3	0
4	4	4	4	4	4	3	0	4	4	1	0
4	4	4	3	4	4	4	0	4	4	1	0
4	4	4	4	4	4	4	3	4	4	3	1
4	4	4	4	4	4	4	4	4	4	4	3

FIGURE 2

Comparison of analytical sensitivity between RT-PCR and high-throughput sequencing (HTS). **A,** Virus detection in diluted pooled samples by PCR (non-detection highlighted in red). **B,** Virus status by mapping HTS reads from diluted pooled samples on all tested virus reference genomes (if the read number is below the filter, samples are highlighted in red). **C,** Number of sample pools in which a virus was detected by PCR depending on dilution. **D,** Number of sample pools in which a virus was detected by mapping (HTS) with all reference genomes, depending on dilution. In C and D, samples in which the virus could not be detected are highlighted in different shades of red according to the number of pools in which the detection worked.

contrasting results have been observed as the overall performance of small RNA was better in a study on various commodities (Gauthier et al. 2022), whereas the total RNA protocol was better with other commodities, such as tomato, lemon tree, or grapevine, in several studies (Di Gaspero et al. 2022; Pecman et al. 2017; Visser et al. 2016). In our study, the ribodepleted total RNA protocol generated more viral reads and was therefore selected. The plant host, its physiological status, and the viruses to be detected will influence the performance of either protocol. It is therefore advised to evaluate which protocol fits the purpose of the test better and, if needed, to carry out a preliminary comparison of protocols as done in this study.

If a PCR-based result is considered positive or negative depending on the visualization of a band on a gel or a Ct value, an HTS-based result should be considered positive if the number of reads of the virus is above the detection threshold. The determination of the detection threshold, differentiating between infection and any potential cross-contamination, is therefore a cornerstone of the data analysis.

In this context, our results also underline the utility of an alien control to evaluate the cross-contamination burden between samples and to adjust the detection threshold performance metric. Cross-contaminations by alien virus reads followed a random pattern with marked variation between batches for the number of samples with alien virus reads, as well as the total number of cross-contaminating alien reads for each sample or their maximal reads abundance. Within batches, the cross-contamination burden was also variable, with the number of alien virus reads ranging from 0 to 288 reads per sample in batch 3, whereas the maximum number in batch 1 was seven reads. The presence of a background level of cross-contaminating reads was also observed on grapevine as, in 46 cases of samples negative for one viral species, between 1 and 10 viral reads were observed (Soltani et al. 2021). Putative cross-contaminating reads were also observed in another study, causing a false discovery rate of up to 11% (Gauthier et al. 2022). Therefore, the determination of a threshold for detection, considering possible cross-contamination, is a way to limit the false positive rate of an HTS test. The threshold of 10 reads has often been proposed in the literature (Bloom et al. 2021; Soltani et al. 2021; Strong et al. 2014), but it does not consider the possible variability in cross-contamination burden. With a threshold of 10 reads, a high false discovery rate (21%) was observed on grapevine (Soltani et al. 2021), but an in-depth analysis of the results showed that 40% of the false positives corresponded to samples with fewer than 30 reads for the detected virus. Therefore, we recommend that the threshold for plant virus detection should be adapted to each sequencing batch using the alien control. We proposed here a conservative threshold, corresponding to the highest number of cross-contaminating alien reads per sample for each batch, but this conservative threshold caused false negative detection for viruses always at low abundance, such as BBTV. We tested (Rong et al., *unpublished data*) other thresholds based on the alien control, such as average number of contaminating reads per sample for each batch or the ratio between contaminating virus alien reads in the samples divided by the virus alien reads in the alien control. These thresholds significantly raised the number of false positives and reduced the accuracy. These results underline that further research is needed to optimize the determination of detection thresholds, for example, by considering ratio of normalized viral reads abundance of a virus species between samples (Gauthier et al. 2022) that could be adapted for an alien virus. For example, the maximum number of reads in a sample was 2,397 for BBTV, 288,675 for BBrMV, and 86,915 for CMV. Therefore, detecting a few reads of BBTV in a sample is less likely a cross-contamination event than for BBrMV or CMV when 36× or 120× more reads have been generated in at least one sample. An alien control can be an efficient alternative to the use of several positive and negative controls at several steps of the HTS test to monitor cross-contamination as it can generate an adaptative threshold considering the cross-contamination burden observed for each batch. Nevertheless, the alien control as used in this study still presents some limitations that need to be discussed. Indeed, it is not able to determine at which step the exchange occurred. The addition of different alien controls at different steps of the process (such as sample tissue, RNA extract, and/or prepared library) would allow for an improved cross-contamination analysis at the expense of higher cost as more samples would need to be processed. An alien control is also not able to identify one-time cross-contamination events occurring between two samples during the test (including with samples from other projects), as demonstrated for CMV. In addition, our threshold (based on maximal cross-contaminating alien reads number) might not be ideal for viral species that typically present a low abundance of reads, such as BBTV. Nevertheless, our alien threshold was selected to minimize false positive results. Indeed, the determination of the threshold is not easy as it will always involve a trade-off between the ability to detect low-level infection (raising the true positive ratio) and baseline cross-contamination (raising the false positive ratio). The most appropriate threshold should therefore be determined during the validation process for each HTS test in the laboratory so the threshold can optimize the downstream interpretation of the results and ensure that the HTS still fits the purpose of the diagnostics.

To complement the alien threshold, an expert analysis was needed to identify one-time cross-contamination between samples, considering the other samples from the project and the viral read abundance of the species in these samples. We have shown for CMV that it is also important to consider any other sample from other projects processed at the same time as the studied samples because they might also be the origin of cross-contamination. The analysis of the SNP profile between samples, as well as a duplication analysis between the hypothetical sample of origin and the potential cross-contaminated samples, can improve the identification of one-time cross-contaminations and should be carried out to investigate dubious results.

Even with the conservative threshold used in this study and expert analysis of the results, two unconfirmed detections of BBrMV remained unclear as the confirmation tests were all negative for these viruses. The sample inversion is unlikely as these two samples were the only ones infected by BanMMV in the sequencing batch, and BanMMV reads were detected. On the other hand, the current geographical spread of BBrMV is restricted to Asia (Thomas et al. 2015), and its presence on samples from Congo is unexpected, although not impossible.

In conclusion, we recommend the use of an adaptative threshold based on an alien virus control read detection in the samples of interest, with an expert review of the data considering the relative abundance of viral reads in samples for each viral species. The determination of such a threshold should be evaluated during validation and will correspond to a trade-off between the diagnostic sensitivity (improving when the threshold diminishes, and calculated at 100% without a threshold) and the diagnostic specificity/false discovery rate (worsening when the threshold diminishes). The determination of our adaptative threshold based on the alien viral reads was the cornerstone for the downstream evaluation of performance criteria and, in routine diagnostics, will be a key factor for discriminating viral infection from cross-contamination.

In this study, the HTS test presented higher inclusivity, analytical sensitivity, and diagnostic sensitivity than RT-PCR or IC-RT-

PCR. A previous report on banana viruses underlined the ability of the HTS test to detect distant isolates of BanMMV that were not amplified by existing primers (Hanafi et al. 2020). In this publication, the HTS test detected isolates of two BSV species and triggered the development of new primers able to detect them. This again underlined the usefulness of HTS data to improve the inclusivity of targeted PCR-based protocols as suggested earlier (Adams et al. 2018). Another report on virus detection from other banana samples demonstrated the higher inclusivity of the HTS test as a new virus infecting *Musa* plant was detected only by the HTS test (Hanafi et al. 2022). In this report, at least two new viral species were detected by the HTS test and are currently characterized. Similar cases of new viral species discovery when evaluating the HTS test on a broad range of samples have been reported previously (Pecman et al. 2017; Rott et al. 2017). The application of the HTS test on banana in vitro plants (Hanafi et al. 2022) also revealed a better diagnostic sensitivity of the HTS test compared with RT-PCR carried out on the same RNA extract, most probably caused by its deeper analytical sensitivity as demonstrated in this publication.

It is also important to mention that the HTS test performed poorly (high false positive rate) for detecting BSV species, most particularly the BSV species whose genome is integrated in the *Musa* B genome (BSOLV, BSMYV, BSIMV, BSGFV, and BSVNV). This biological constraint makes it necessary to complement the HTS test with an IC-PCR-targeted test (detecting viral particles) when viral reads from integrated BSV species are detected, particularly in samples containing the B genome. In addition, BSV sequences from hypothetical species of clade 2 (no viral particle has been detected for them so far) have been detected in the A genome, but there is no proof of transcription nor production of viral particles (Chabannes et al. 2021). Therefore, it will be important to specifically identify the BSV species detected in the A genome to evaluate the risk of presence of viral particles. More globally, other virus species, such as badnavirus or geminivirus, are known to be integrated into the genome of their plant host as a complete or partial sequence. If these sequences can be transcribed, it must be considered when evaluating the performance criteria of an HTS test for virus detection with those plant species, underlining again the importance of knowledge on the biology of detected viruses when analyzing the results.

The repeatability and reproducibility of virus detection were 100%, confirming the excellent results observed during previous evaluations on grapevine and *Citrus* spp. (Bester et al. 2021; Di Gaspero et al. 2022; Soltani et al. 2021). Behind this 100% value, a huge variability in reads number was observed between replicates within and between different batches. This variability was confirmed by the evaluation of the number of reads observed per sample for the dilution series. One of the probable origins of this phenomenon is the heterogeneity of virus distribution in *Musa* plants. This heterogeneity is high as complete virus indexing of an accession requires testing a pool of at least four plants with midribs and limbs from the three last leaves to minimize the risk of false negatives (Thomas et al. 2015).

The minimal sequencing depth for appropriate virus detection has been evaluated in several publications through subsampling reads (also called read rarefaction). For example, one million reads were considered appropriate for some viruses, although this number was not applicable to all viruses (Gauthier et al. 2022; Visser et al. 2016). The reads subsampling and its normalization between samples were also proposed to minimize the detection of cross-contamination events, limiting the false positive detections (Gauthier et al. 2022), although it can reduce the genome coverage and depth. The use of the alien threshold follows the same objective without the need for subsampling. Indeed, subsampling will have no effect because it also reduces proportionally the alien threshold for detection by rarefying the number of cross-contaminating alien virus reads in the samples.

In conclusion, our report demonstrates the usefulness of an alien control to monitor cross-contamination burden, although specific cross-contamination events remain difficult to trace back. For routine virus detection of *Musa* germplasm, the use of the HTS test and the alien threshold, completed by an expert analysis of the results, fit the purpose of viral detection. In addition, the HTS test must be complemented by a targeted IC-PCR for BSV if BSV reads are detected in datasets generated from germplasm containing the B genome to confirm the presence of viral particles.

ACKNOWLEDGMENTS

We thank Angelo Locicero for technical support and Gladys Rufflard for administrative support. Delphine Masse (ANSES, La Réunion, France), Kathy Crew and John Thomas (Queensland Alliance for Agriculture and Food Innovation, Brisbane, Australia), and Mathieu Chabannes and Marilyne Caruana (CIRAD, Montpellier, France) are also acknowledged for kindly providing reference samples. We also thank John Thomas for his suggestions after carefully reading the manuscript.

LITERATURE CITED

Adams, I. P., Fox, A., Boonham, N., Massart, S., de Jonghe, K., Tahzima, R., Foucart, Y., Peusens, G., Beliën, T., Massart, S., and de Jonghe, K. 2018. The impact of high throughput sequencing on plant health diagnostics. Eur. J. Plant Pathol. 152:909-919.

Adams, I. P., Skelton, A., Macarthur, R., Hodges, T., Hinds, H., Flint, L., Nath, P. D., Boonham, N., and Fox, A. 2014. Carrot yellow leaf virus is associated with carrot internal necrosis. PLoS One 9:e109125.

Al Rwahnih, M., Daubert, S., Golino, D., Islas, C., and Rowhani, A. 2015. Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. Phytopathology 105:758-763.

Bester, R., Cook, G., Breytenbach, J. H. J., Steyn, C., De Bruyn, R., and Maree, H. J. 2021. Towards the validation of high-throughput sequencing (HTS) for routine plant virus diagnostics: Measurement of variation linked to HTS detection of citrus viruses and viroids. Virol. J. 18:1-19.

Bloom, J. S., Sathe, L., Munugala, C., Jones, E. M., Gasperini, M., Lubock, N. B., Yarza, F., Thompson, E. M., Kovary, K. M., Park, J., Marquette, D., Kay, S., Lucas, M., Love, T., Booeshaghi, A. S., Brandenberg, O. F., Guo, L., Boocock, J., Hochman, M., Simpkins, S. W., Lin, I., LaPierre, N., Hong, D., Zhang, Y., Oland, G., Choe, B. J., Chandrasekaran, S., Hilt, E. E., Butte, M. J., Damoiseaux, R., Kravit, C., Cooper, A. R., Yin, Y., Pachter, L., Garner, O. B., Flint, J., Eskin, E., Luo, C., Kosuri, S., Kruglyak, L., and Arboleda, V. A. 2021. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. MedRxiv 2020.08.04.20167874.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., and Mumford, R. 2014. Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res. 186:20-31.

Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A. D. 2019. rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8:1-13.

Bushnell, B., Rood, J., and Singer, E. 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One 12:e0185056.

Calil, I. P., and Fontes, E. P. B. 2017. Plant immunity against viruses: Antiviral immune receptors in focus. Ann. Bot. 119:711-723.

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., and Muller, E. 2021. Badnaviruses and banana genomes: A long association sheds light on *Musa* phylogeny and origin. Mol. Plant Pathol. 22:216-230.

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., and Gabriel, S. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genom. 19:332.

De Clerck, C., Crew, K., Van den houwe, I., McMichael, L., Berhal, C., Lassois, L., Haissam Jijakli, M., Roux, N., Thomas, J., and Massart, S. 2017. Lessons learned from the virus indexing of *Musa* germplasm: Insights from a multiyear collaboration. Ann. Appl. Biol. 171:15-27.

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., and Marroni, F. 2022. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J. Virol. Methods 300:114383.

Fox, A., Fowkes, A. R., Skelton, A., Harju, V., Buxton-Kirk, A., Kelly, M., Forde, S. M. D., Pufal, H., Conyers, C., Ward, R., Weekes, R., Boonham, N., and Adams, I. P. 2018. Using high-throughput sequencing in support of a plant health outbreak reveals novel viruses in *Ullucus tuberosus* (Basellaceae). Plant Pathol. 68:12962.

Gaafar, Y. Z. A., Westenberg, M., Botermans, M., László, K., De Jonghe, K., Foucart, Y., Ferretti, L., Kutnjak, D., Pecman, A., Mehle, N., Kreuze, J., Muller, G., Vakirlis, N., Beris, D., Varveri, C., and Ziebell, H. 2021. Interlaboratory comparison study on ribodepleted total RNA high-throughput sequencing for plant virus diagnostics and bioinformatic competence. Pathogens 10:1174.

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., and Cosson, J. F. 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. MSystems 1:1-22.

Gauthier, M.-E. A., Lelwala, R. V., Elliott, C. E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., and Barrero, R. A. 2022. Side-by-side comparison of post-entry quarantine and high throughput sequencing methods for virus and viroid diagnosis. Biology 11:263.

Geering, A. D. W., McMichael, L. A., Dietzgen, R. G., and Thomas, J. E. 2000. Genetic diversity among banana streak virus isolates from Australia. Phytopathology 90:921-927.

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., and Massart, S. 2022. Detection of banana mild mosaic virus in *Musa* in vitro plants: High-throughput sequencing presents higher diagnostic sensitivity than (IC)-RT-PCR and identifies a new *Betaflexiviridae* species. Plants 11:226.

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., and Massart, S. 2020. Identification of divergent isolates of banana mild mosaic virus and development of a new diagnostic primer to improve detection. Pathogens 9:1045.

Kechin, A., Boyarskikh, U., Kel, A., and Filipenko, M. 2017. CutPrimers: A new tool for accurate cutting of primers from reads of targeted next generation sequencing. J. Comput. Biol. 24:1138-1143.

Kumar, P. L., Cuervo, M., Kreuze, J. F., Muller, G., Kulkarni, G., Kumari, S. G., Massart, S., Mezzalama, M., Alakonya, A., Muchugi, A., Graziosi, I., Ndjiondjop, M.-N., Sharma, R., and Negawo, A. T. 2021. Phytosanitary interventions for safe global germplasm exchange and the prevention of transboundary pest spread: The role of CGIAR germplasm health units. Plants 10:328.

Lebas, B., Adams, I., al Rwahnih, M., Baeyen, S., Bilodeau, G. J., Blouin, A. G., Boonham, N., Candresse, T., Chandelier, A., de Jonghe, K., Fox, A., Gaafar, Y. Z. A., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., Landa, B., Liu, M., Maclot, F., Malapi-Wight, M., Maree, H. J., Martoni, F., Mehle, N., Minafra, A., Mollov, D., Moreira, A., Nakhla, M., Petter, F., Piper, A. M., Ponchart, J., Rae, R., Remenant, B., Rivera, Y., Rodoni, B., Roenhorst, J. W., Rollin, J., Saldarelli, P., Santala, J., Souza-Richards, R., Spadaro, D., Studholme, D. J., Sultmanis, S., van der Vlugt, R., Tamisier, L., Trontin, C., Vazquez-Iglesias, I., Vicente, C. S. L., Vossenberg, B. T. L. H., Wetzel, T., Ziebell, H., and Massart, S. 2022. Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: A step-by-step description. EPPO Bull. 52:394-418.

Maachi, A., Torre, C., Sempere, R. N., Hernando, Y., Aranda, M. A., and Donaire, L. 2021. Use of high-throughput sequencing and two RNA input methods to identify viruses infecting tomato crops. Microorganisms 9:1043.

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C. M., Roumagnac, P., van der Vlugt, R., and Massart, S. 2020. Illuminating an ecological blackbox: Using high throughput sequencing to characterize the plant virome across scales. Front. Microbiol. 11:578064.

Malapi-Wight, M., Adhikari, B., Zhou, J., Hendrickson, L., Maroon-Lango, C. J., McFarland, C., Foster, J. A., and Hurtado-Gonzales, O. P. 2021. HTS-based diagnostics of sugarcane viruses: Seasonal variation and its implications for accurate detection. Viruses 13:1627.

Maree, H. J., Fox, A., Al Rwahnih, M., Boonham, N., and Candresse, T. 2018. Application of HTS for routine plant virus diagnostics: State of the art and challenges. Front. Plant Sci. 9:1082.

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, H. J., Olivier, T., Olmos, A., Pooggin, M. M., Reynard, J.-S., Ruiz-García, A. B., Safarova, D., Schneeberger, P. H. H., Sela, N., Turco, S., Vainio, E. J., Varallyay, E., Verdin, E., Westenberg, M., Brostaux, B., and Candresse, T. 2019. Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. Phytopathology 109:488-497.

Massart, S., Olmos, A., Jijakli, H., and Candresse, T. 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res. 188:90-96.

Massart, S., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G. J., Blouin, A., Boonham, N., Candresse, T., Chandelier, A., De Jonghe, K., Fox, A., Gaafar, Y. Z. A., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O. Jonkers, W., Kreuze, J., Kutnjak, D., and Landa, B. 2022. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. Peer Community J. 2:37.

Olmos, A., Boonham, N., Candresse, T., Gentit, P., Giovani, B., Kutnjak, D., Liefting, L., Maree, H. J., Minafra, A., Moreira, A., Nakhla, M. K., Petter, F., Ravnikar, M., Rodoni, B., Roenhorst, J. W., Rott, M., Ruiz-García, A. B., Santala, J., Stancanelli, G., van der Vlugt, R., Varveri, C., Westenberg, M., Wetzel, T., Ziebell, H., and Massart, S. 2018. High-throughput sequencing technologies for plant pest diagnosis: Challenges and opportunities. EPPO Bull. 48:219-224.

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., and Ravnikar, M. 2017. Next generation sequencing for detection and discovery of plant viruses and viroids: Comparison of two approaches. Front. Microbiol. 8:1998.

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

Santala, J., and Valkonen, J. P. T. 2018. Sensitivity of small RNA-based detection of plant viruses. Front. Microbiol. 9:939.

Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., Chan, C. K. F., Nabhan, A. N., Su, T., Morganti, R. M., Conley, S. D., Chaib, H., Red-Horse, K., Longaker, M. T., Snyder, M. P., Krasnow, M. A., and Weissman, I. L. 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724.

Soltani, N., Stevens, K. A., Klaassen, V., Hwang, M.-S., Golino, D. A., and Al Rwahnih, M. 2021. Quality assessment and validation of high-throughput sequencing for grapevine virus diagnostics. Viruses 13:1130.

Strong, M. J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., and Lin, Z. 2014. Comprehensive high-throughput RNA sequencing analysis reveals contamination of multiple nasopharyngeal carcinoma cell lines with HeLa cell genomes. J. Virol. 88:10696-10704.

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, J. S., Stevens, K., Kutnjak, D., and Massart, S. 2021. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community J. 1:e53.

Thomas, J., Sharman, M., Lassois, L., Massart, S., De Clerck, C., Caruana, M.-L., Chabannes, M., Teycheney, P.-Y., Kumar, P. L., Van den Houwe, I., and Roux, N. 2015. Technical guidelines for the safe movement of Musa germplasm, 3rd Ed. J. Thomas, ed. Bioversity International, Rome, Italy.

Velasco, L., and Padilla, C. V. 2021. High-throughput sequencing of small RNAs for the sanitary certification of viruses in grapevine. Front. Plant Sci. 12:682879.

Visser, M., Bester, R., Burger, J. T., and Maree, H. J. 2016. Next-generation sequencing for virus detection: Covering all the bases. Virol. J. 13:85.