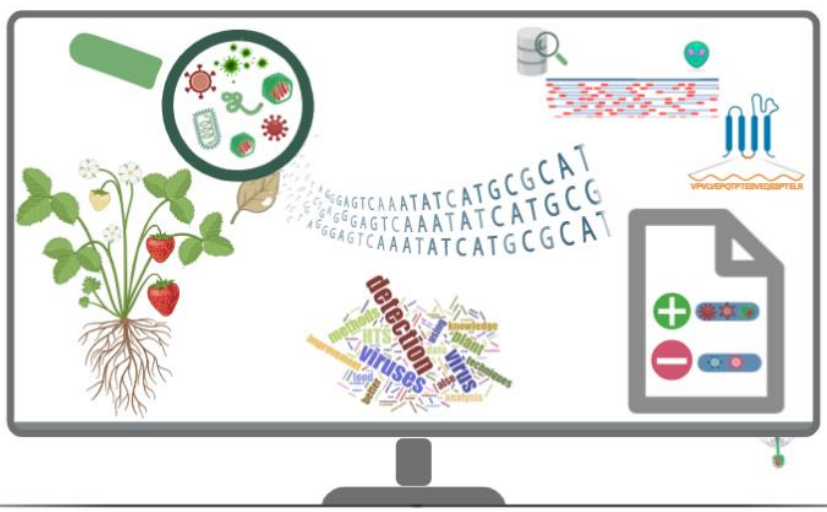
LIÈGE université
Gembloux
Agro-Bio Tech

Development and validation of innovative sequencing tools for the fast and efficient detection of plant virus

Johan Rollin



COMMUNAUTÉ FRANÇAISE DE BELGIQUE
UNIVERSITÉ DE LIÈGE – GEMBLOUX AGRO-BIO TECH

Development and validation of innovative sequencing tools for the fast and efficient detection of plant virus

Johan Rollin

Dissertation originale présentée en vue de l'obtention du grade de docteur en sciences agronomiques et ingénierie biologique

Promoteurs : Prof. Sébastien Massart
Prof. Yves Brostaux

Année civile : 2023

Abstract

Plant viruses are a major cause of crop losses and decreased agricultural productivity worldwide. Rapid and accurate detection of plant viruses is essential for the implementation of effective control measures. Traditional methods of plant virus detection, such as serological and molecular assays, often present very good performance criteria but they are targeted, and they don't detect new viruses or divergent strains of known viruses.

Overall, developing and validating innovative sequencing tools for fast and efficient detection of plant viruses gained a lot of leverage. Indeed, high throughput sequencing (HTS) tests followed by bioinformatic analyses can detect several viruses at once (including novel ones) and then characterise their genomes. This very high inclusivity allows better monitoring of agricultural pest presence than traditional methods. In addition, the sensitivity of HTS viral detection is theoretically higher than molecular and serological tests, meaning that low-level infection can be traced more efficiently.

HTS tests have several drawbacks: the price, the high technical requirements and the cross-contamination of sequences between samples nevertheless. The cost of viral detection by sequencing is higher than traditional methods, but the cost gap is reducing over time as HTS is more and more affordable. More technical skills are required for sequencing and analysis of a sample for virus detection, but the laboratory and bioinformatic protocols are becoming simpler and easier to learn and apply. Cross-contamination between samples is a recurrent phenomenon that is challenging the operational activities of laboratories aiming to detect plant pests. The high sensitivity of HTS has a drawback as it means that cross-contamination is an even more pressing issue than with traditional methods.

Cross-contamination is probably one of the main issues when using HTS for viral detection. Indeed, if an unexpected genetic material transfer happens between two samples in the laboratory, one virus can be sequenced in the other sample. Since sequencing sensitivity is high, HTS is more prone to detect this cross-contaminating virus. That may lead to a false positive virus detection (as it is really in the bioinformatic data) while it was not present in the plant. The specificities of HTS technologies (high sensitivity, high inclusivity but with the complexity of laboratory and bioinformatics steps) make their validation difficult compared to traditional tests. Therefore, this thesis describes the side-by-side comparison between traditional tests and HTS technologies for virus indexing of Musa germplasm collection. In addition, an alien control (a specific type of external control) has been used for the first time to

monitor cross-contamination in HTS. In addition, a newly described alien-based filter algorithm, called Cont-ID, has been developed and applied to find the most appropriate limit of detection that should be applied for accurate virus detection taking into account the risk of false negatives and false positives. That way, the detection prediction's confidence can be high enough to be considered for its use in plant virus diagnosis.

As written above, HTS technologies can also characterise the genome of the detected viruses. Through variant analysis, the different virus variants can be highlighted. A performance testing was conducted to better understand the difficulties and therefore improve the variants' characterisation.

This thesis has therefore addressed several drawbacks limiting potentially the use of HTS technologies for plant virus detection and genome characterisation. It has delivered several milestones to contribute to these technologies' wider and more reliable applications for plant virus detection. Overall, it has reinforced its high potential for improving the control and management of plant virus diseases.

Résumé

Les virus des plantes sont une cause majeure de pertes de récoltes et de productivité agricole dans le monde entier. La détection rapide et précise des virus des plantes est essentielle pour la mise en place de mesures de lutte efficaces. Cependant, les méthodes traditionnelles de détection des virus des plantes, telles que les tests sérologiques et moléculaires, ne détectent pas les nouveaux virus.

Le développement et la validation d'outils de séquençage innovants pour la détection rapide et efficace des virus des plantes a beaucoup avancé ces derniers temps. En effet, les méthodes de séquençage suivies d'analyses bioinformatiques peuvent détecter plusieurs virus à la fois (y compris les nouveaux) et les caractériser, permettant un meilleur suivi de la présence des virus. La sensibilité de la détection virale par HTS est supérieure à celle des tests moléculaires et sérologiques, ce qui permet d'observer les infections à bas niveau de concentration plus efficacement.

Il y a cependant des inconvénients à l'utilisation du séquençage, tels que le coût, les exigences techniques élevées et les problèmes de contamination croisées. Le coût de la détection par séquençage est supérieur à celui des méthodes traditionnelles, mais l'écart de coût diminue. Il est nécessaire de posséder des compétences techniques plus élevées pour séquencer un échantillon de plante, mais les protocoles deviennent de plus en plus connus, ce qui facilite l'apprentissage et l'application de ces techniques. Enfin, la haute sensibilité du HTS signifie que la contamination croisée est un problème encore plus pressant que pour les méthodes traditionnelles.

La contamination croisée est probablement l'un des principaux problèmes lors de l'utilisation de HTS pour la détection virale. En effet, si un transfert de matériel génétique inattendu se produit entre deux échantillons dans le laboratoire, un virus peut être séquencé dans l'autre échantillon. Comme la sensibilité du séquençage est élevée, HTS est plus enclin à détecter ce virus de contamination croisée. Cela peut entraîner une détection erronée du virus (car il est réellement présent dans les données bioinformatiques) alors qu'il n'était pas présent dans la plante. Heureusement, un contrôle alien (un type spécifique de contrôle externe) peut être utilisé pour surveiller la contamination croisée dans HTS. En effet, un filtre basé sur cet alien peut être appliqué pour trouver la limite de détection à utiliser. De cette façon, la confiance de la détection devrait-être suffisamment élevée pour être utilisée dans le diagnostic des virus de plantes.

Les inconvénients de l'utilisation de HTS pour la détection des virus des plantes sont de plus en plus limités et les avantages importants de la méthode ce qui lui permet de présenter un grand potentiel pour améliorer la gestion des maladies virales des plantes.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Sébastien Massart, for is unwavering guidance, support, and encouragement throughout the entire duration of my research. He accepted to share his experience and expertise in plant virology with me which influenced my vision of science and research and have made a significant impact on the quality of this thesis. I truly thank you.

I would also like to thank the members of my thesis committee, Yves Brostaux, Luc Willems, Marie Lefebvre and Macha Nikolski, for their valuable feedback, insightful comments, and constructive criticism, which helped me improve and refine my research. I would also like to thank Jean-Philippe Renvoise, Herve Vanderschuren and Haissam Jijakli who accepted to join the thesis' jury.

I consider myself fortunate to have been part of the INEXTVIR project and am grateful to have discussed with the participants, ESRs, supervisor and advisor. Those exchanges undoubtedly enrich myself, and this work.

I am thankful to the Gembloux plant phytopathology team member (past and present), I am going to miss our discussions, ideas and coffee breaks. Their contributions have been instrumental in making this study a success.

I would like to express my deepest appreciation to the participants who generously gave their time and energy to take part in my research.

Lastly, my thoughts also go my friends Wei Rong, Arnauld Blouin, and Benedicte Lebas which where indispensable to the advancement of the thesis and overcoming the work during COVID lockdown. And of course the Goonies, as we shared our fears and motivations in the good and bad moments. Mercé pel temps passat amassa, serà pas jamai oblidat.

Aquela tèsi seriá pas estada possibla sens la contribucion de totes e de ma familha, soi plan reconeissent per vòstre supòrt.

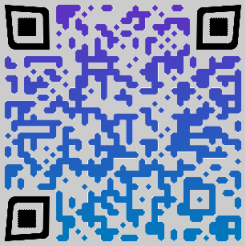
Table of contents

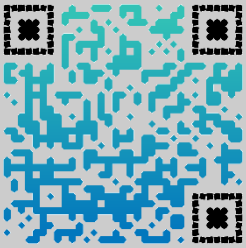
List of Abbreviations and Acronyms

Abbreviations	**Definition**
ANSES	Agence nationale de sécurité sanitaire de l'alimentation
BBTV	Banana Bunchy top virus
BLAST	Basic Local Alignment Search Tool
BSCAV	Banana streak CA virus
BSGFV	Banana streak GF virus
BSIMV	Banana streak IM virus
BSMYV	Banana streak MY virus
BSOLV	Banana streak OL virus
BSV	Banana streak virus
BWA	Burrows-Wheeler Aligner
BYDV	Barley yellow dwarf virus
CIRAD	Centre de Coopération Internationale en Recherche Agronomique
CMV	Cucumber mosaic virus
COVID	Coronavirus Disease
DNA	Deoxyribonucleic acid
DSE	Diagnostic Sensitivity
ELISA	enzyme-linked immunoassay
EPPO	European and Mediterranean Plant Protection Organization
HGT	Horizontal gene transfer
HTS	High Throuput Sequencing
ITC	International (Musa Germplasm) Transit Center
ITS	(nuclear ribosomal RNA) internal transcribed spacer
LCA	Lowest Common Ancestor
NCBI	The National Center for Biotechnology Information

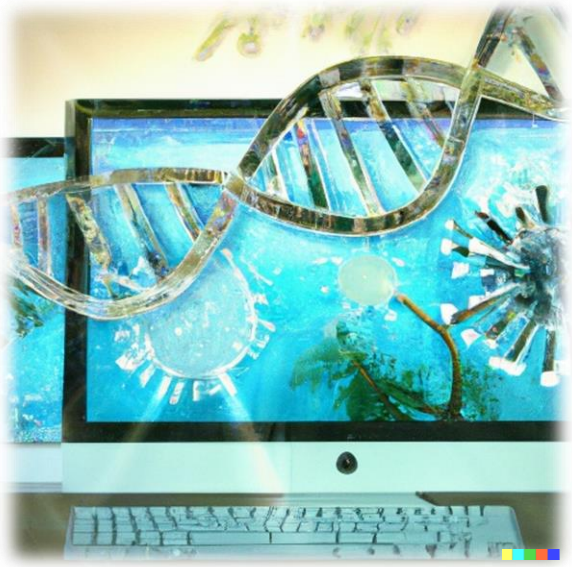
PCR	Polymerase Chain Reaction
PVY	Potato virus Y
RNA	Ribonucleic acid
RT	Reverse Transcriptase
SAM	sequence alignment and map
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
TBIA	Tissue blot immunoassay
TOL	Tree of life
VANA	Virion-Associated Nucleic Acids

List of external link

Information	Link	QR code
Chapter 3 Supplementary data	Supplementary_3	
Chapter 4 Supplementary data	Supplementary_4	
Chapter 4 scripts	Cont-ID	
Chapter 5 Supplementary data	Supplementary_5	

Chapter 5 scripts	SNP_comparison	
Presentation demultiplexing	Demultiplexing	
Presentation viral detection	Detection	
Presentation Galaxy Europe	Galaxy	
Presentation Serratus DB	Serratus	

Thesis structure and objectives



Thesis structure

This thesis's introductive section (**Chapter 1: General introduction**) provides a general view of scientific knowledge of viruses with a focus on plant viruses. This led to the explanation of how they are detected, their impact on the food system and their societal impact (food safety) to finalize their presentation on the control measures to prevent virus spread. Therefore, this chapter aims to understand plant virus detection's challenges and importance.

The second part (**Chapter 2: Sequencing of plant virus**) is a technical introduction to how to address the challenges described in Chapter 1. It focuses on sequencing and the difference between alternative techniques. First, the sequencing steps are described, followed by an overview of how the resulting data should be analysed. The confidence of plant viral detection is also discussed, highlighting the new knowledge brought by this thesis on cross-contamination handling, the improvement in detection, and the resulting viral characterisation.

How to validate virus detection by HTS taking into account the cross-contamination occurring between samples is addressed in the next chapter (**Chapter 3: Validation of virus detection**). *What are the important criteria for HTS viral detection? Can HTS be used for diagnostics tests?* Those are questions that are addressed in this chapter. The light was shed on cross-contamination and the use of alien control to monitor it. This monitoring allowed a better determination of the analytical sensitivity by setting a alien-based control metrics. This filter can help define and adapt the actual limit of detection on each independent sequencing.

Automation of virus detection (Chapter 4). With the definition in the previous chapter of filters as a limit of detection for plant virus detection in HTS, other alien control metrics can be used to further improve the detection reliability. In this chapter, other alien control-based metrics are explored, and the automatic processing of very sensitive detection for curation is proposed via a new bioinformatic tool: Cont-ID. This tool considers several metrics from the alien control to predict the viral detection as (true) infection or cross-contamination.

Genome characterization of detected virus (Chapter 5). The variant calling allows us to know the specific mutation that had occurred in the virus genome. Those mutations can change some viral traits and therefore are important to track. In this chapter, I focus on evaluating the factors influencing the variant calling analysis, allowing a better mutation prediction. Compared to the traditional exploration of variant detection, consisting in evaluating many parameters of bioinformatic tools by

a single or a few laboratories, our original methodology relied on a large-scale evaluation by end-user laboratories.

In the general discussion of this thesis **(Chapter 6: Discussion and future prospect),** the major research findings of each of the previously described aspects are discussed along the detection taxonomy level (often species level) relevance. Indeed, after formal detection, complementary analyses can be made that are also important for the evaluation of virus threats. Perspectives on plant viruses and the HTS' role in food safety are also considered.

Thesis objectives

This thesis' general aim is to strengthen the HTS usability to reliably detect plant viruses and increase food safety. Indeed, plant virus monitoring in the field via traditional methods can be enriched with sequencing to improve detection reliability. To achieve it, several axes are developed:

- I intend to show that HTS is a non-targeted, sensitive, scalable (with multiplexing) method meaning it can be a very good surveillance tool for plant pests.
- I aim to demonstrate that HTS virus detection can be more reliable by using alien control while being more sensitive that the alternative method (PCR).
- I also want to show that with alien control, automatic analysis can improve HTS detection, proving that HTS is a reliable method even with limited human interpretation.
- Virus variant tracking can be important for field surveillance. I intend to improve viral variant detection by describing important steps for a successful analysis.

1

General introduction

1. What is a virus

1.1. Viruses in nature

Viruses and tree of life

The definition of life relies on the distinction between what is considered a living organism and an inanimate object (Luisi, 1998). Through time, several definition attempts were with slight nuance. The Darwinian idea of life is what is capable of undergoing evolution through natural selection (Cleland and Chyba, 2002). A more modern definition adds the concept of a "self-sustaining chemical system" (Ruiz-Mirazo et al., 2004). The precise limit of how self-sustainable an organism's metabolism needs to be considered alive is still debated (Chodasewicz, 2014). In fact, several definitions compete depending on the field of study of their usage (Chodasewicz, 2014). The comparison of organisms' relationships is called phylogeny; it relies on studying evolution to classify organisms between them. Historically, living organisms are classified in the tree of life (TOL). Viruses are classified apart because of their non-cellular aspect(Moreira and López-García, 2009).

The phylogeny relies on nucleic acid comparison to understand evolution. The functional information of an organism is stored (and usable) in DNA or RNA (genetic information). With reproduction, this information can pass to the next generation. Some errors are made through time and generation; the information can be changed and sometimes will give advantages, leading to the change being naturally selected. The study of these modifications of information through time is a field of science called evolution (Kutschera and Niklas, 2004). Both viruses and organisms from the TOL contain genetic information. They could be compared, but because viruses are not considered living, their evolution and the evolution of TOL organisms can be considered separate. The discovery of horizontal (lateral) gene transfer (HGT) (Griffith, 1928) changes this separation since it proves that genetic information can be exchanged between organisms without heredity, even between organisms of different kingdoms. Indeed, evolution studies need to consider the genetic information acquired from other organisms (Daubin and Szöllősi, 2016). Viruses can perform HGT and exchange with living organisms (Irwin et al., 2021), meaning that the separation of the virus tree and the TOL can not be explained by the definition of life anymore (Harris and Hill, 2020). Other arguments are advanced to maintain the separation, like the likely fact that viruses are polyphyletic (Moreira and López-García, 2009). Another critical difference in the taxonomy (classification of organisms) between organisms in TOL and viruses is the presence of markers.

Each domain of TOL (Bacteria, Archea and Eukaryota) contains markers (conserved genetic information throughout the studied group) that can help the comparison of organisms within domain (Tekle et al., 2010; Wu et al., 2013). No marker has been found on viruses, even if some attempt, like (Rdrp) gene palm, can help with some virus subgroups (Koonin et al., 2020). Indeed, the virus biology diversity makes it difficult to study and edict global rules that can apply to all.

Characteristics of viruses

Viruses share common characteristics with living organisms (like using genetic material to store and produce function) and completely original ones (like the fact they do not have a cell organisation). Viruses can be defined as infectious units composed of nucleic acid and proteins that are intracellular parasites that replicate in living cells(Modrow et al., 2013). Viruses evolved over millions of years and have adapted to their hosts, leading to a huge biological diversity between them (Choua et al., 2020). Viruses present a wide range of genetic material, RNA/DNA, single/double. They can infect virtually all organisms in TOL as they have been described in Eukaryotes, Bacteria and Archea (Hull, 2014).

Viruses were originally classified based on their shape, size, and type of disease they cause (Dimmock et al., 2016). The replication of viruses occurs in three main steps. First, the enter the host cell. Once inside the host cell, the virus begins the process of gene expression and replication. This involves synthesising proteins encoded by the viral genome and replicating the viral genome itself (Dimmock et al., 2016). The viral genome can be replicated through different mechanisms, depending on the virus type (in the nucleus for DNA, in the cytoplasm for RNA). Finally, the virus assembles and releases infectious viral particles (or naked acid nucleic), which can spread the infection to other cells (Dimmock et al., 2016; Hull, 2014).

Plant viruses are composed of genetic material encased in a protein shell called a capsid. Plant viruses are comprised of a majority of single-stranded RNA of positive-sense polarity (Hull, 2014). These genomes can contain diversity in a number of genes from (replication-associated) proteins, coat proteins (particle assembly), movement proteins (cell-to-cell movement within the plant) and additional functions (like countering host defence system). Another infectious agent that can infect plants: viroid. Viroids are naked (no capsid) RNA molecules with a size from 250 to 450 nt approximately that do not code for any proteins; they use plant proteins to be replicated or move in a different cell. Throughout the studies of plants of interest, we can link diseases observed and infectious agents. For example, on grapevines, more than 70 viruses are known with 31 inducing symptoms, and seven viroids are

described with at least two linked to symptoms (Martelli, 2017). The study of plants of interest (mainly of agricultural importance) and their virus propagation and impact is important to provide good practice recommendations on food production systems.

1.2. Detection of viruses in plant

Several methods can detect viruses in plants, and the choice of method will depend on the specific virus being sought. Some common methods for detecting viruses in plants include:

Serological tests: These tests involve using antibodies to detect the presence of viral proteins in plant samples. ELISAs (enzyme-linked immunosorbent assays) (Clark and Adams, 1977) are the most popular serological test in which a specific viral antibody is immobilised on a solid surface of a wheel in a plate and exposed to a sample containing antigen. The antibodies will bind to the antigen if the sample contains the corresponding virus.

Molecular tests: These tests involve several techniques, among which the most popular are polymerase chain reaction (PCR) (VUNSH et al., 1990) or reverse transcription PCR (RT-PCR) (Roberts et al., 2000) to amplify and detect viral nucleic acid (DNA or RNA) in plant samples. PCR is a highly sensitive and specific technique that can be used to amplify small amounts of DNA or RNA from a variety of sources, including plant tissue or other biological materials. There is a lot of possible variation in the PCR processing virus sample, allowing a good adaption for better detection in specific virus cases.

Bioassays: These tests involve infecting healthy plants with a plant sample and then observing the plants for symptoms of viral infection. Bioassays can be used to confirm the presence of a virus in a plant sample, but they can be time-consuming and may not be suitable for all viruses (Legrand, 2015).

Tissue blot immunoassay (TBIA): This test involves transferring proteins from plant tissue onto a membrane, which is then probed with labelled antibodies to detect viral proteins (Whitfield et al., 2003). TBIA can detect viral proteins in tissue samples rather than in the extracted form, which can be more representative of planta situations.

Electron microscopy: This method involves using an electron microscope to observe plant samples at high magnification to look for characteristic structures of viruses, such as capsids or envelopes (Williams, 1954). Electron microscopy can be used to identify viruses based on their size, shape, and other morphological characteristics and can be a useful confirmatory test when other methods have indicated the presence

of a virus in a plant sample. It is nevertheless no more widely used in diagnostic, except for the viral particle characterization of newly identified viruses.

Sequencing: This is a regroupment of different methods, from Sanger sequencing (Sanger and Coulson, 1975), or High Throughput Sequencing (HTS) via Pyrosequencing (Roossinck et al., 2010), very small reads (sRNA) (Kreuze et al., 2009), short reads (Illumina) (Adams et al., 2009), long read (nanopore, pacbio) (Liefting et al., 2021). The common characteristics of those sequencing methods are to informatically reconstruct plant viruses' nucleic acid material (DNA or RNA) and therefore allow its detection and characterisation.

HTS is a method for rapidly sequencing large amounts of nucleic acid. It involves using specialised sequencers that simultaneously generate large numbers of sequence reads in a single run. HTS has several applications in plant virology, including identifying and characterising viruses, analysing viral gene expression, and detecting viral variants.

All the listed methods here are not equivalent as some are targeted or not, and the achievable specificity/sensitivity are different. Therefore, the choice of the method to use should be considered according to the type of virus expected and the purpose of the test. For plant virus diagnosis, the serological and molecular test seems the more used, but the more recent HTS technologies raise a growing interest as they add new possibilities it can lead to the analysis of a large amount of data/plants and increase the potential for new virus characterisation. (Jeong et al., 2014; Rubio et al., 2020).

2. Impact of plant viruses

2.1. Food production system

The food production system refers to the process of producing, processing, distributing, and consuming food. This includes all the steps involved in getting food from the farm or fishery to the consumer, including growing and harvesting crops, raising livestock, processing and packaging food products, and distributing them to markets and stores. Food production systems can vary in size and complexity, from small-scale, local operations to large, industrial systems that produce food on a worldwide scale. They can also involve different food production methods, such as conventional, organic, or aquaponics. The goal of a food production system is to produce safe, nutritious, and affordable food for a population while also considering environmental and social impacts.

Virus in food production systems

There are several ways in which viruses can impact food production systems. Some viruses, such as those that cause diseases in plants or animals, can directly affect the productivity of a food production operation. For example, plant viruses can reduce crop yields and quality, leading to economic losses for farmers and disruptions in the food supply chain. However, for plant viruses, the public health risks of the product consumed as food are low (Aguado-García et al., 2020; de Jonghe, 2020). It is important to note that the impact of viruses on food production systems can be mitigated through various measures, such as implementing effective biosecurity practices, developing resistant or tolerant cultivars, vaccines (such as CH2 PepMV) (Agüero et al., 2018) and other preventive measures, and adapting to changing market conditions.

Plant disease is an important component of global economic and human health. Indeed, One Health (Mackenzie and Jeggo, 2019) is a concept that recognizes the interconnection between the health of humans, animals, and the environment. It is based on the idea that the health of these three systems is interconnected and that addressing one system's health can impact the others. One Health approaches are often used to address complex health issues that involve multiple sectors and disciplines, such as emerging diseases, food safety, and environmental health. They involve the collaboration of professionals from various fields, including human medicine, veterinary medicine, public health, environmental science, and other disciplines.

The relationship between plant disease and food safety in the One Health framework has been studied (Rizzo et al., 2021). Even if the study did not include plant viruses, it is a legitimate claim to say that virus-induced plant disease affects food safety and human health similarly to non-virus plant diseases.

2.2. Plant virus detection and food safety

There are several ways in which plant virus detection can be used to enhance food safety. One way is through the use of rapid and sensitive detection methods, such as molecular techniques like PCR, which can detect low levels of viral nucleic acid in plant samples. This can allow for the early identification of viruses in crops or seeds, which can help prevent the spread of infection and protect the quality and safety of the crops (Food et al., 2019; Savary et al., 2019).

In addition, plant virus detection can be used to verify the effectiveness of measures taken to control viral infections in food crops. This can include the use of virus-resistant crops or the implementation of good agricultural practices. By monitoring

the presence of viruses in crops, it is possible to determine whether these measures are effective at controlling viral infections and ensuring the safety of food crops (Giovani et al., 2020; Ristaino et al., 2021).

Overall, plant virus detection plays an important role in ensuring the quality and safety of food crops and is an important tool for protecting public health and food security. The food security strategy, therefore, relies on the grower (responsible for the agricultural practice), the testing bodies (responsible for accurate virus detection) and the regulator's authorities that need the balance the burden of the testing/good practice implementation with the reality of the situation in the fields (Ward, 2016).

Action to prevent pest spread

Several studies try to evaluate the impact of plant viruses from perception (Hilaire et al., 2022) to economics (Rybicki, 2015). Direct virus impact varies from yield reduction on crops of interest or food quality deterioration. The mechanism allowing the plant virus infection is described as the most impactful one (Scholthof et al., 2011) and methods to limit the infection and/or the virus impact (while in the field) are proposed. The way to respond to plant virus varies depending on the plant(s) of interest and the method of production, the virus biology (including its transmission pathways and host range), the criticality of the potential yield loss…

Most actions to mitigate pest spread depends on factors that rely on the interaction between several entities: scientific bodies, growers and regulatory authorities.

- On the scientific aspect, several outputs are expected: the ability to confidently detect the pest with minimal time and effort, the deliverance of information such as the pest danger and spread mechanisms assessment or the possible action needed to limit the pest spread or impact.
- The regulatory authorities, along with growers, have to balance the money, time and effort to provided by the scientific community with the potential output. In addition, they have to decide on the action to perform to limit pest impact based on the danger assessment of the scientific community and the burden (cost and effort) to implement it.

The action against pests has been studied by modelling new pest spreading (Fennell et al., 2012) linked to pest management options (Cunniffe et al., 2014; Ward, 2016). We tend to observe a lag between the real status of a pest, the detection ability, and the answer to it (Figure 1-1)

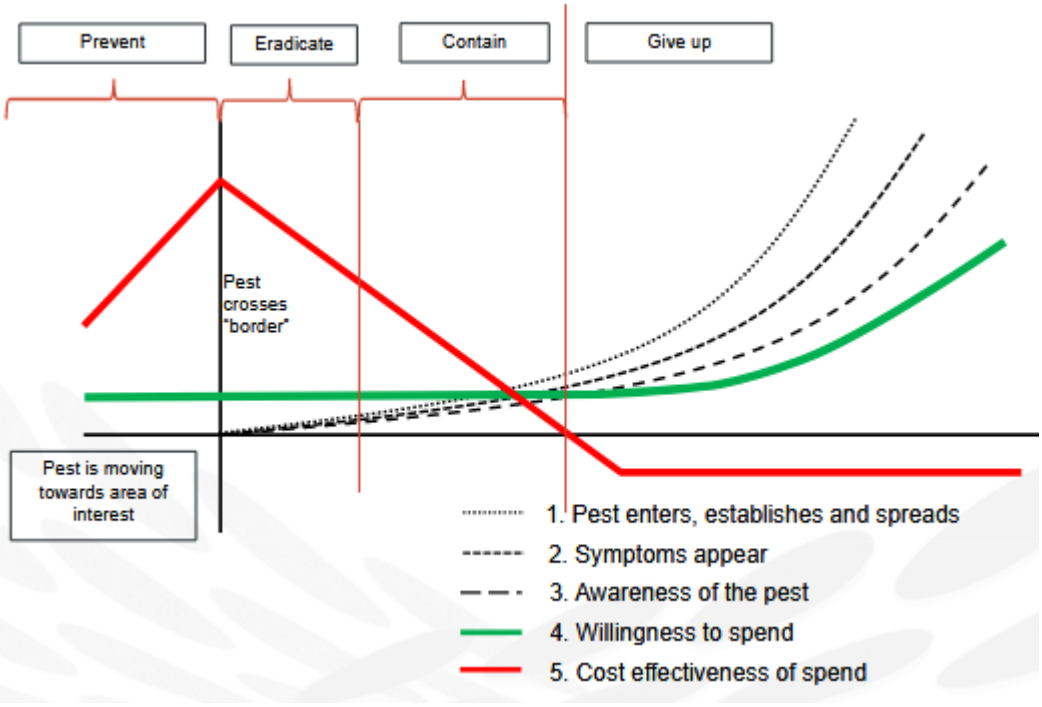


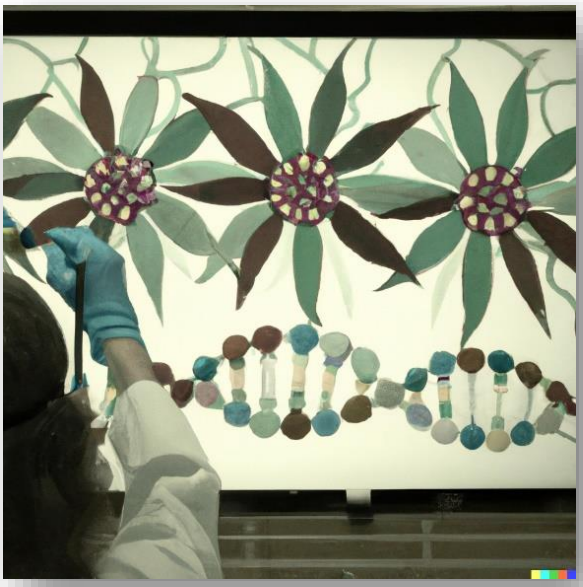
Figure 1-1 Typical progression in a standard emerging outbreak of a plant pest, illustrating the time lags between pest population, symptoms, awareness and willingness to spend, alongside the cost-effectiveness of spending. This figure is originally from (Ward, 2016) and has been adapted (colour and response phase added) by Glynn Jones (unpublished).

Figure 1-1 shows all the delays that add up during a pest entrance, first, between the pest presence and the appearance of symptoms. Secondly, with symptoms, the exploration starts with identifying the pests, followed by the danger, and spreading mechanism evaluation while raising awareness. With the awareness of the pest danger, the willingness to spend rises, but spending is less effective when made late as the pest is more present. Figure 1-1 also describes the different phases of controlling pests. When a pest exists but is not present in our area of interest, action can be taken to prevent pest entry (it relies on the ability to detect new pests). At an early stage, if the pest has just arrived in the area, action can allow the eradication of the pest from the area (it relies on a fast detection system). If the pest is already present in numbers too big to be removed from the area, a containment strategy can be implemented (it relies on an accurate detection system). Finally, if all the previous actions failed, instead of focusing on the pest presence, the action will try to limit its effect (like switching to a less impacted plant variety).

These actions rely on the scientific community to provide accurate information, so studying pests (plant viruses being one of them) is essential. The detection of pests/viruses is the cornerstone of the possible strategies. It required some adjustments depending on the action (rapidity or accuracy) envisioned and the propagation's danger (specificity or sensitivity). The decision made by the pest spread prevention decision will lead to the scientific choice of detection strategy. Technical advancement allows the choice between several methods (serological, molecular test, HTS …); inside those methods, adjustments can be made to implement the best-adapted strategy (Jones, 2014; Kalimuthu et al., 2022).

2

Sequencing of plant virus



1. Sequencing methods

1.1. Plant virus detection by sequencing

Characteristic of viral detection test

Several characteristics are important for a reliable plant virus detection test (Rubio et al., 2020):

- Analytical sensitivity: it is the metric that measures the ability of the detection to perform with small amounts of virus. This is important because early detection of a viral infection can help prevent the virus's spread to other plants. When high, it also allows testing a pool (mix) of several plants in one test.
- Analytical specificity: A plant virus detection test should be specific, meaning it should accurately detect the specific virus it is designed to detect without detecting other types of viruses or non-viral sequences.
- Accuracy: it is a metric considering both the (diagnostic) sensitivity and specificity abilities of the detection that represent the capacity to obtain accurate results that can be relied upon. This is important for deciding how to manage a viral infection and for research purposes.
- Simplicity: A plant virus detection test should be easy to perform, with clear instructions and a simple protocol to ease reproducibility.
- Reproducibility is the ability of a test to provide consistent results when applied to aliquots of the same sample tested under different conditions (e.g. time, persons, equipment, location) (EPPO standard, PM 7/76, 2018).
- Speed: the time used for detection is important. A detection test should be able to produce results in a reasonable amount of time, as early detection of a viral infection is important for preventing the spread of the virus.
- Cost-effectiveness: A plant virus detection test should be cost-effective, considering the time and resources required to perform the test and the importance of the results.

Traditionally, plant pest laboratories used molecular and/or serological tests to detect plant viruses (mainly PCR and ELISA), as they fit most of the characteristics (EPPO standard, PM 7/98, 2021). But recently, interest in using HTS for plant virus detection tests was raised, and guidelines were made available (Lebas et al., 2022; Massart et al., 2022a), as HTS fit even better the needed characteristics and provides several other advantages. HTS is an untargeted method (PCR and ELISA are not),

meaning that, in theory, HTS can detect any virus in a plant, including unexpected ones. Several new or previously uncharacterised viruses have been detected thanks to sequencing(Barba et al., 2014; Maliogka et al., 2018). Another advantage is the ability of HTS to detect mixed virus infections at once, which would require several PCR or ELISA.

The historical reason why HTS is not used more is the (lack of) simplicity and the cost that is still higher compared to PCR/ELISA. However, the cost of sequencing is going down, and the sequencing protocol is more and more optimised and shared, making it easier for the community to use and reproduce HTS testing.

1.2. Sequencing preparation

The sequencing starts with the sampling, followed by laboratory steps leading to sequencing and ending with sequencing data analysis, as described in Figure 2-1 (Lebas et al., 2022). The sampling is the same as with PCR or ELISA; it corresponds to the selection of material (most of the time plant) on which virus presence is to be tested. This material will then be used as a matrix for nucleic acid extraction. The extraction, library preparation and sequencing steps are done in the laboratory, while the later steps (reads analysis, identification, analysis, interpretation) are done behind a computer and are considered bioinformatic steps.

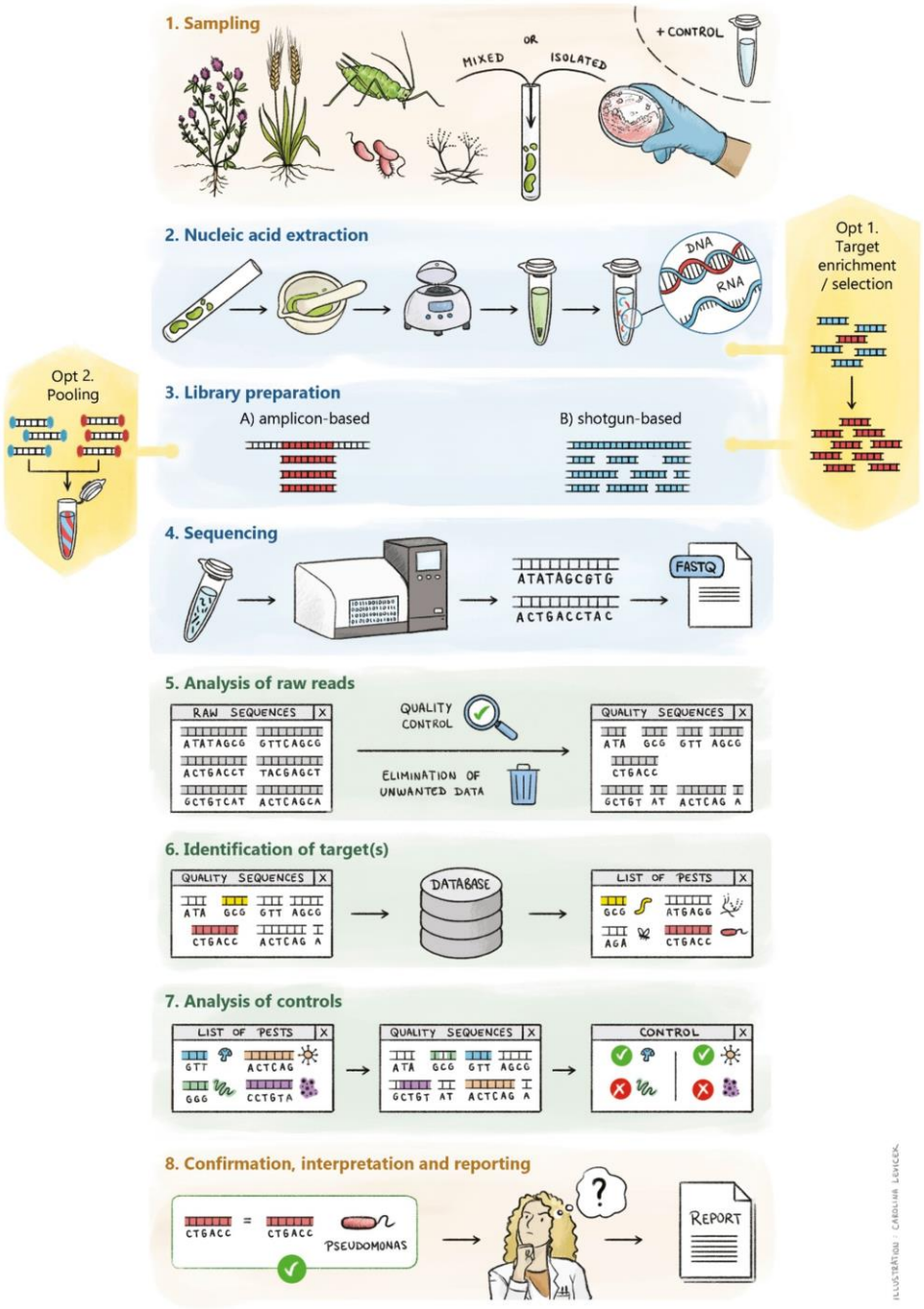


Figure 2-1: Schematic representation of the main steps of the high-throughput sequencing (HTS) process used in plant health diagnostics (Lebas et al., 2022)

Nucleic acid extraction & library preparation

From the biological material previously sampled (like plant leave), the genetic material (target of sequencing) needs to be extracted. The expected type of nucleic acid is as diverse as the virus's genetic support (DNA/RNA, single/double). In addition, to reduce cost, the sample to extract can be a mix of several plants (pool) instead of only one(Lebas et al., 2022).

Following the nucleic acid extraction, the library preparation will directly act on the genomic material. Indeed, the nucleic acid should be isolated and fragmented (if needed) and then very short (known) sequences called adapters will be added at both extremities. The adapters will allow the sequencing machine to attach the nucleic acid fragment in order to start the sequencing. To further reduce the cost, several samples can be combined (multiplexing) to sequence as one. In that case, known indexes (short nucleic acid fragments) for each sample should be previously added to the adapters to be able to assign the sequences to its original sample afterwards (Lebas et al., 2022).

In case of an RNA virus is expected, an RT-PCR should be performed to provide cDNA (with adapter) to the sequencer (although sometimes RNA sequencing is possible). In plant virus detection, it is also advised to remove non-target sequences (for example, ribodepletion) or to perform an enrichment (for example non-specific PCR based amplification) (Adams and Fox, 2016). The library preparation can be modified to be targeted to a specific genomic-part/organism (amplicon sequencing), or it can be untargeted (shotgun) (Lebas et al., 2022).

Several adaptations in extraction and library preparation should be made, as a small difference in plant/virus composition has a huge impact (Sabatier et al., 2020). For plant viruses, several protocols are described for alternative extraction/library preparation, such as Total-RNA (Oñate-Sánchez and Vicente-Carbajosa, 2008), Virion-Associated Nucleic Acids (VANA) (Filloux et al., 2015), smallRNA (Kreuze, 2014) or double-stranded (ds)RNA (Marais et al., 2018a). Methods can differ between DNA and RNA viral extraction (specially for full length genome sequencing). Careful consideration of what protocol to use, with which sample type and to what end should be done before starting the extraction.

1.3. Types of datasets obtained

The isolated nucleic acid fragments with adapters are used by a sequencing machine that will amplify (if needed) and then copy the sequences anchored to the machine. During the copy, a signal corresponding to the composition of the sequences is

obtained and informatically processed (base calling) with a quality score associated with each base (nucleotides). Finally, all the sequence fragments (reads) obtained are concatenated in a computer file (read file). All further analyses are going to be founded upon this document.

The different preparation may give different base calling since it changes the presence of certain nucleic acids (like ribodepletion). In addition, different sequencers are available with different characteristics that can also change the output (Maljkovic Berry et al., 2020). In bioinformatics, there is a distinction to make with the possible different sequencing outputs. Indeed, if the difference resides inside the data without changing its representation, the data interpretation will change, but the same bioinformatics analysis can be done. For example, two datasets can be obtained with different extraction protocols and different sequencers for the same sample. The two files obtained can have similar read characteristics (read number, read size, read quality), meaning the same bioinformatics analysis can be performed on both files (but different results in sample virus composition can be obtained). However, if the differences are also on the reads file characteristics (read number, read size, read quality), then different bioinformatics analyses will have to be considered.

The different characteristics of reads files that impacted the later bioinformatic analysis are:

- Paired or single ends (depending on whether the genomic fragments were sequenced from both ends or not). Paired ends provide more information but are more expensive
- Read numbers (sequencing depth); if the number is too low, we may lack information; the higher it is, the more computational power will be required for the analysis and the higher the price is.
- Reads quality represents the possible error during the base calling; the higher, the better (fewer errors), but more effort (and money) is required to improve the quality.
- Reads length greatly impacts the analysis because it can be very variable. From very small reads (<50 nt) (Kreuze, 2014), short reads (50>1000 nt) and long reads (>1000 nt) (Tedersoo et al., 2021).

The biologist decides these criteria as he/she can decide which read size, number, and quality to aim for. Several sequencing protocols are available as many combinations between those criteria are possible from different providers (Ion-Torren,

Illumina, Nanopore, PacBio…). The decision should be motivated by the price, time required for the sequencing and the expected impact on the bioinformatic analysis that will follow.

2. Analysis of sequencing data

2.1. Analysis steps

First, the prior events have to be considered as data reflect the type of extraction and sequencing they originated from. In general, there are two distinct phases in the bioinformatic analysis, called "analysis of raw reads" in Figure 2-1 (also refer as "read cleaning" or "pre-processing") and the analysis itself allowing the identification of targets thanks to the control and possible exploration (variant discovery, phylogeny …) corresponding to the step 6 and 7 in Figure 2-1.(Maljkovic Berry et al., 2020; Rivarez et al., 2021). In this thesis (chapter 3 and 4), I will demonstrate that the "analysis of control" (step 7) should be considered part of the "identification of targets" (step 6) as the use of control allows improvement in identification reliability.

Pre-processing

The pre-processing of sequencing data corresponds to cleaning raw data, which can include everything corresponding to the reads preparation for their accurate analysis. The reads obtained from sequencing need to be controlled to remove the low-quality ones, the adapters, and the artefacts that can be present. This step can also include demultiplexing (if needed) to separate the multiplex samples thanks to the index added to the adapters. In addition, if possible, the reads can be paired (association of 5' end and corresponding 3' end) and merged together. Finally, a procedure called host removing can be considered; it consists of identifying and removing the host-linked reads letting in the dataset only very informative reads about potential pests.

Most of these steps allow the improvement of accuracy in later analysis (including detection) (He et al., 2020). The correct level of cleaning is to be tested (depending on the data and protocol used) as cleaning too relaxed may let erroneous reads that can bias the later analysis. On the contrary, a cleaning too strict can remove too many reads, not letting enough information in the dataset to reach a conclusion. Some procedures like host removal should be used with caution as it can have an undesired effect; as in the case of an integrated virus (a virus for which their genome can be integrated into the host genome), the viral reads will be removed, as those reads can be considered as host by the tools.

Identification of target

In metagenomic analysis, the first aim is to identify the organism that constitutes the datasets. Some alternative strategies and tools provide advantages and drawbacks. Those strategies can be classified into few categories (Brinkmann et al., 2019)

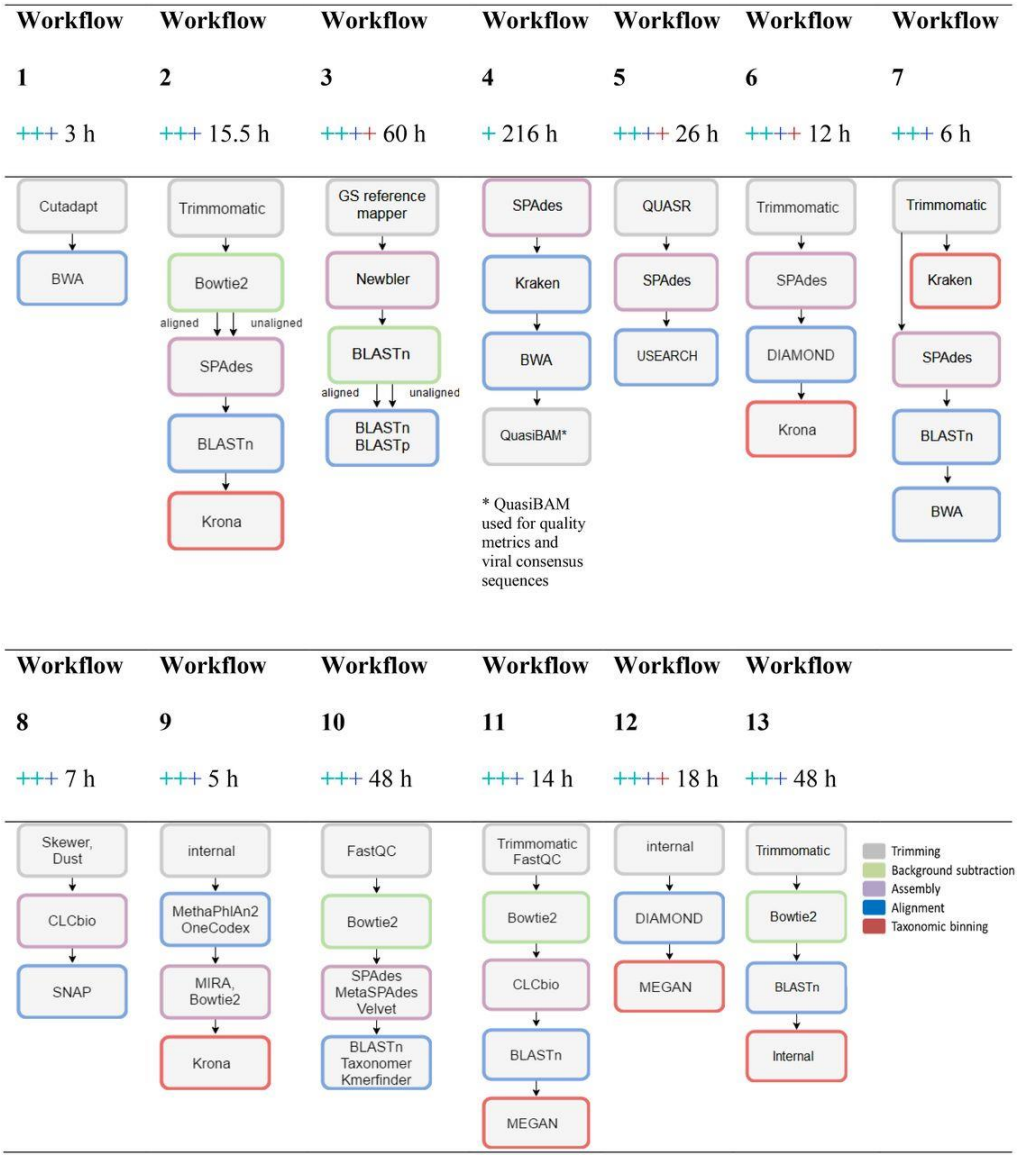


Figure 2-2: Simplified comparison of different bioinformatics workflows for virus identification used in the COMPARE virus proficiency test. The plus sign is a representation of the machine usage (Brinkmann et al., 2019)

In Figure 2-2, pre-processing step's elements trimming (quality reads removing) and background subtraction (host removing) are displayed. In addition, there are three other categories, taxonomic binning, assembly and alignment.

Taxonomic binning corresponds to the direct taxonomic assignation of reads via matching reads on (previously) prepared taxonomic databases. Several tools can perform it with slight differences in the matching algorithm and the (taxonomic) assignation that ensue. Indeed, in this category, there is a specificity problem because read size (often between 50-250 nt with Illumina type sequencing) sometimes doesn't allow species' specific part of the genome to be on each related read. Some of the binning tools like Kraken (Wood and Salzberg, 2014) use a lowest common ancestor strategy (LCA) which allows the assignation to the upper taxonomic level (Genus, Family …). This method should not be used with very small reads (<50 nt), as it may lead to wrong results. Nevertheless, taxonomic binning is an interesting detection method because, despite possible specificity issues and the high computational power required, it provides fast detection with high sensitivity (Wood et al., 2019).

Assembly is not a detection method per se, as it is more a preparation to further analysis. It consists of elongating reads into longer sequences (called contig) that should ideally correspond to sequenced genomes. Longer sequences are interesting as they allow more analysis possibilities because there is more chance to have organisms' specific parts on them. Among the assembler (tools doing the assembly), SPAdes (Bankevich et al., 2012) is broadly used. The assembly requires high computational power and is not very fast, but it allows a more specific detection afterwards.

Alignment is the traditional detection method; it is the alignment of the longest sequence possible (here, contig) on a database of known sequences. It can be performed in nucleotide form or amino acid to improve the specificity (due to the degenerate aspect of the genetic code). The historical tool for performing an alignment is BLAST (Camacho et al., 2009), but a huge diversity of tools can also perform sequence alignment. The range of time, computational power, specificity and sensitivity for alignment depend on the chosen tools (Steinegger and Söding, 2017). Alignment almost always relies on performing assembly before.

Direct taxonomy binning and assembly and alignment are the two main strategies to detect organisms in metagenomic datasets. Both provide different advantages and drawbacks, which make them partially complementary.

Exploration

After the detection, exploratory analysis can be done to improve the knowledge of the dataset. The range of possible analysis is very diverse: proteomic prediction leading to functional information can be done, phylogeny study can be performed to try to unveil pest origin and relationship with already sequenced isolates, epidemiologic work can help understand virus propagation, variant analysis can provide information on relative variant danger… Most of these complementary analyses will use the contigs resulting from the assembly as their starting point. In the case of a new virus or new variant detection, additional investigations are needed to establish a plan of action to eradicate or prevent the virus from spreading(Massart et al., 2017).

2.2. Analysis reliability

The reliability of the virus detection analysis refers to the level of certainty that a particular virus detected in the generated sequences is present in the sample tested (risk of false positive) and that a virus present in the sample tested is detected in the generated sequences (risk of false negative). Several factors, such as sensitivity, specificity, or sample composition, can affect confidence in virus detection analysis. A high sensitivity and specificity test provides higher confidence in the results. The sample composition impacts confidence, as the more complex the dataset (more plants pooled together, a lot of different viruses present or a high difference in relative virus concentration), the less confident the results will be. For example, if the sample is cross-contaminated (organisms not linked to the host are unwillingly added) or partially degraded, the test may not be able to accurately detect the virus. The detection test can be confirmed by testing again with another method like sequencing, PCR or ELISA.

Control

To improve confidence in detection, it is good practice to add control to the test (Massart et al., 2022a). Controls are monitoring datasets that are added to the main analysis to monitor certain criteria. There are a lot of different controls, each type monitoring a certain aspect. The most known and used for a long time are the external positive control (checking the correct detection of an already known target) and the external negative control (checking that no detection is made in the absence of targets). Internal control can also be used as a positive control for checking steps (like extraction). The use of control is very efficient in adding confidence to the detection.

Cross-contamination

Cross-contamination is a major problem for detection confidence. Indeed, an organism unrelated to the host , or its DNA/RNA, can be unwillingly added at any step of the detection process, from initial sampling to the result generation. Meaning that the test can legitimately detect an organism not originally present in the sample. Limiting or at least monitoring cross-contamination is very important for confidently detecting pests. In theory, a new type of control introduced for HTS tests and called alien can help monitor cross-contamination. "An alien control corresponds to a matrix containing one or several targets (called alien targets) which belong to the same group as the target organism(s) but cannot be present in the samples to be tested" (Massart et al., 2022a) This means that cross-contamination can be monitored by detecting alien reads in the samples of interest or unexpected reads in the alien sample. I will discuss later (in Chapter 3) the use of alien control and its improvement to plant virus detection.

Interpretation of detection

A difference between molecular/serological tests and sequencing is the interpretation range. In most of the PCR/ELISA, the interpretation is more straightforward, for sequencing leading to absence or presence prediction. It is due to the fact that those techniques are targeted and have been more studied through time allowing to improve the interpretation (Minic et al., 2020). In sequencing results, a certain number of reads (or contigs) could be considered as virus detection. The interpretation can change depending on the different metrics (or contigs). The threshold that switches the interpretation from presence to absence (considering therefore the detection in those samples as originating from contamination) is not yet addressed but it can depend on several metrics (sequence number, match quality, genome coverage, specificity/sensitivity expected) and the context of the analysis. As discussed in Chapter 1, the purpose of the test is important as it will impact the choice of a relaxed or strict threshold for the prediction. Predicting three situations might be more interesting in some aspects: absence, confirmation needed and presence. This will be discussed in more detail later (in Chapter 4).

Viral characterisation

Viral characterization consists of determining the specific genome of viruses present in a plant to further explore its biology. Indeed, HTS technologies can first detect the presence of viruses. The detection is able to identify the virus (often to the species level), characterising the virus to go a step further by determining its precise

characteristics. To that end, a consensus genome of the viruses can (in most cases) be reconstructed by keeping the major nucleotide in each position. Additional work can be done to try to obtain variants (especially below 50% frequencies) that could be hidden by the consensus sequence. This can lead to haplotype characterization (reconstruction of genomes present in the plant). The study of the mutation profile of the virus can be done by using variant calling analysis. As shown in Chapter 4, the CMV case mutation's profiles can help track cross-contamination (also discussed in Chapter 5).In addition, variant analysis is important since it has been shown that variants can change the threats (virulence, replication rate …) of viruses, like in the case of Potato Virus Y (Kogovšek et al., 2010). The important analysis points for accurate variant calling were studied in a performance testing and are discussed in Chapter 5.

Bioinformatic viral detection improvement

The main problem of virus detection from generated sequencing datasets does not seem to be the lack of methods. Indeed, there are already many tools using several different methods to detect viruses in plants (see Figure 2-1). To better understand their specificities, those tools have been benchmarked extensively (Brinkmann et al., 2019; de Vries et al., 2021; Jones et al., 2017). Workflow can be made by combining tools to efficiently use their individual strengths while limiting their weaknesses, so more accurate predictions can be obtained. The community has already done a lot of work (boosted by the COVID-19 pandemic) (Baizan-Edge et al., 2019; Ferguson et al., 2022; Rosenthal et al., 2022; Waite et al., 2022). Therefore, improving known virus detection by adding another tool or workflow to the list seems difficult. Instead, this thesis focuses on improving confidence in virus detection by clarifying the choice of the metrics leading to the interpretation and monitoring the cross-contamination. By improving the reliability of plant virus detection by sequencing, this thesis aims to increase food safety in the agricultural system.

3

Validation of virus detection

Article 1

(Published in phytofrontiers)

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]

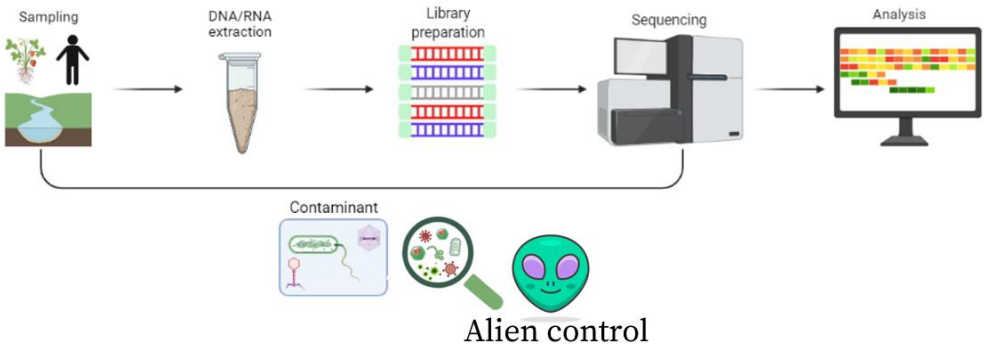
[1]*Laboratory of Plant Pathology – TERRA - Gembloux Agro-Bio Tech – University of Liège, 5030 Gembloux, Belgium*

[2]*DNAVision, 6041 Gosselies, Belgium*

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

*These authors contributed equally to this work

Correspondence: Sébastien Massart (sebastien.massart@uliege.be)



Abstract

High-throughput sequencing (HTS) technologies have brought tremendous improvements in the ability to detect plant viruses and have a great potential for application in virus routine diagnostics. The performance criteria of an HTS test need therefore to be estimated and compared to traditional virus indexing tests before it can be used in routine diagnostic. 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 analyzed for virus detection. The external alien control, a wheat sample infected by BYDV, monitored the contamination burden and determined an adaptative detection threshold. Overall, HTS test displayed a better analytical sensitivity compared to 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. Repeatability and reproducibility of virus detection were both of 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 of expert judgement in the interpretation of the results. In conclusion, the HTS test with an alien control and completed by expert evaluation has fulfilled the criteria of virus indexing protocol for *Musa* germplasm.

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

Introduction

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; S. Massart et al., 2014a; Soltani et al., 2021a). From the first discovery of a plant virus in 1898, the ability to detect plant viruses has followed the technological evolution throughout time (Maclot et al., 2020). In the first part of the 20th century, the detection of plant viruses relied on symptomatology on natural hosts 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 indicators results by bud grafting failure, relatively long testing period before release (Al Rwahnih et al., 2015; Soltani et al., 2021b). Further on, 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 (Maree et al., 2018; Thomas et al., 2015).

The development of high throughput sequencing (HTS) technologies represents a very promising method for universal virus detection (S. Massart et al., 2014b). The sequencing machine will produce millions to billions of sequences that 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, both at 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., 2018a). 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 in complement to conventional methods to inform diagnostic workflows in the identification 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 has been drafted by an international consortium of plant pest diagnostics specialists (Massart et al., Accepted EPPO bulletin). 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 tested positive and the total number of infected samples and the diagnostic specificity, corresponding to the ratio between the number of healthy samples tested negative and the total number of healthy sample, are both a key performance criterion 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 (Bester et al., 2021a; Di Gaspero et al., 2022a; Gauthier et al., 2022a; Rott et al., 2017a; Soltani et al., 2021b; Thomas et al., 2015).

An HTS test can be divided in 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., n.d.). 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 underlined that each step of an HTS test can impact its performance: sampled tissue (Di Gaspero et al., 2022b; Malapi-Wight et al., 2021), RNA extraction (Bester et al., 2021a), library protocol (Bester et al., 2021a; Di Gaspero et al., 2022b; Maachi et al., 2021; Pecman et al., 2017a), sequencing technology (Bester et al., 2021b), sequencing service provider (Gauthier et al., 2022b), number of generated sequences per sample (Gauthier et al., 2022b; Pecman et al., 2017b; Visser et al., 2016a), bioinformatics analysis (Bester et al., 2021b; Gaafar et al., 2021; Galan et al.,

2016; Gauthier et al., 2022b; S. Massart et al., 2019; Tamisier et al., 2021a) 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 or better than the other compared tests (molecular, immunological or bioassays) (Hanafi et al., 2020a; Thomas et al., 2015; Velasco and Padilla, 2021). On the other hand, the exclusivity of HTS test is intrinsically high but it depends on the number of reads generated for the detected virus (and the genome coverage of the virus), and 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 the reproducibility of HTS tests have been evaluated in the literature and were generally very high with up to 100% of repeatability and reproducibility for virus detection (Bester et al., 2021b; Di Gaspero et al., 2022b; Soltani et al., 2021b).

A key performance criterion is the analytical sensitivity and the corresponding limit of detection of the HTS test. When compared to reverse transcription PCR (RT-PCR), the limit of detection of the HTS test for virus detection was often improved, for example by 10X for PVY (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 S et al., n.d.). Although cross-contamination is long known to occur during the detection of plant viruses by HTS, very few reports 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., 2018a) 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 take into account the cross-contamination level between samples to minimize false positive rate while maintaining an analytical sensitivity (limit of detection) and a diagnostic sensitivity (limiting number of false negative due to very low level of

infection) fitting 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., 2022b).

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 diagnostic: 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 as a cross-contamination from the alien control (Massart et al., Accepted EPPO bulletin). 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 as cross-contamination (controlling the reads exchange from any sample to a single sample). In addition, the alien control also 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 high throughput sequencing 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 comprised solely of 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 they can trigger an infection with viral particle of BSV.

The performance criteria, e.g. 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, as well as 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 detailed in supplementary table 3S1. A total of 78 distinct Musa accessions were used and they were provided by Bioversity International Musa Germplasm Transit Center (ITC, Belgium), ANSES (from La Réunion, France), CIRAD (from 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 9 virus species were detected in at least one sample: banana mild mosaic virus (BanMMV), banana bract mosaic virus (BBrMV), banana bunchy top virus (BBTV), and cucumber mosaic virus (CMV), banana streak OL virus (BSOLV), banana streak CA virus (BSCAV), banana streak GF virus (BSGFV), banana streak IM virus (BSIMV), banana streak MY virus (BSMYV). These plants were either sequenced individually or in pool (corresponding to a single sequencing library) in five different batches.

For each batch of sequencing, at least one pool of five positive controls was sequenced together with the samples. These positive controls were plants maintained from more than a decade in the greenhouse of Gembloux Agro-Bio Tech (Liège University, Belgium) and used as positive control during the routine virus detection tests based on molecular tests and electron microscopy (De Clerck et al., 2017). The positive control plants from batch 1 was 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 tend to provide less read, and CMV a lot). The analytical sensitivity was also evaluated using serial

dilutions of 4 pools of 5 plants (see hereunder), always using the same proportion per sample (20 mg).

For each batch of sequencing, at least one alien control (wheat plant infected with 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., Accepted EPPO bulletin). Leaves of wheat plants infected with barley yellow dwarf virus PAS and PAV (BYDV-PAS and BYDV-PAV) and maintained in the greenhouse of Gembloux Agro-Bio Tech were sampled by the same process as for banana plants. The concentration of virus reads in the sequence dataset from such sample was observed previously as very high in the growing conditions of the greenhouse ((Tamisier et al., n.d.). 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 done with the nine virus species mentioned above.

Sampling and nucleic acids extraction

For the accessions grown in 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 leaves part) were sampled from fresh or lyophilized leaves respectively (weight used were adapted to maintain homogeneity between fresh and dry sample). The same process was carried out for the alien control.

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

PCR-based protocols

The five above mentioned banana viruses were detected by PCR and RT-PCR using cDNA generated from extracted total RNA, respectively. The sequences of these primers used are listed in Table 3S3. For targeted RT-PCR, cDNA was synthesized from 1 µg of total RNA with SuperScript III Reverse Transcriptase (Invitrogen™, CA, United States) according to the manufacturer's instructions and random hexamers (Invitrogen™, CA, United States) was used. PCRs were performed in twenty-five µl reactions. The mix used for BanMMV, BBrMV, BBTV, and CMV contained 2 µl cDNA, 1 µl forward primer (25 µmol 1^{-1}), 1 µl reverse primer (25 µmol 1^{-1}), 5 µl 5X PCR buffer, 1 µl 50mM $MgCl_2$, 0.5 µl 10 mM dNTPs mixture, 0.5 µl Mango *Taq*™ DNA Polymerase (Bioline, London, United Kingdom), and 14 µl sterilized and distilled water. The temperature cycles used for each virus are as described hereunder:

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

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

BSV detection (from cDNA): 2 µl cDNA, 2.5 µl primer mixtures (see primer list in Table 3S3, the final concentration for each primer was 1 µmol 1^{-1} and 2 µmol 1^{-1} for BSMYV detection primers), 5 µl 5X PCR buffer , 0.75 µl 50mM $MgCl_2$, 0.5 µl 10 mM dNTPs mixture, 0.5 µl Mango *Taq*™ DNA Polymerase (Bioline, London, United Kingdom), and 13.75 µl sterilized and distilled water. The PCR and RT-PCR programs used for BSV were as follow. The PCR cycle consisted of a pre-incubation step at 94 ℃ for 30 s; followed by 35 cycles of 15 s at 94 ℃, 30 s at 60 ℃, and 1 min at 72 ℃; and a final elongation step at 72 ℃ for 10 min. For confirming the HTS detection of BSMYV and BSIMV previously undetected by virus indexing, the same mastermix and temperature cycles were used with the only change in reverse primer sequence (see supplementary Table 3S3).

For evaluating 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 1x stained with GelRed® Nucleic Acid Gel Stain (Biotium, Fremont, CA, United States).

High throughput sequencing of total RNA extracts

The library preparation and the sequencing were carried out at the Interdisciplinary Center of Biomedical Research of Liège University (GIGA, Liège, Belgium). Before sequencing, RNA concentration and RNA integrity number (RIN) of each sample were analyzed (qPCR, using KAPA kit). Stranded Total RNA library Prep Human/Mouse/Rat (Illumina, CA, United States, hereafter referred as "old kit") or TruSeq® Stranded Total RNA Library Prep Plant (Illumina, CA, United States, hereafter referred as "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, CA, United States) was used before applying the old kit. The prepared libraries were quantified, pooled and paired-end sequenced on a NextSeq 500 sequencer (2 X 150 bp read length for batches 1, 2, 3, 4) and NovaSeq 6000 for batch 5 from Illumina and operated by GIGA facilities of Liège university. Sequencing batches were processed at 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 GIGA facilities using the SMARTer smRNA-Seq Kit, (Clontech – Takara Bio). The sequencing of 1x50 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 35bp 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, see Supplementary Table 3S4, Table 3S4). Except BanMMV, all the other tested viruses had more than one reference genomes 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 high sensitivity mapping method. We used the default parameters associated with "Low sensitivity / Fastest" profile allowing two iterations and manually setting 20% mismatch and maximum 3 nucleotides gaps allowed. Those parameters were used for all samples except for small RNA on which adapted parameters (2 iterations, 10% mismatch and maximum 3 nucleotides 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 new or known unexpected virus (not targeted by PCR test). For new virus detected, we combined tBlastX with NCBI non-redundant 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 (see Supplementary Table 3S2). 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 5x dilution for each virus. The pools were further diluted to reach 100-, 1,000-, and 10,000-time 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 sampling of plant tissue from the same plant at the same time. The number of replicates analyzed in parallel ranged between 2 to 5 and are indicated by "r" in sample code column from supplementary table 3S1.

The reproducibility of the HTS test over time was also evaluated on the positive control mix at each sequencing batch (five independent sequencing starting from plant material, although in different proportions in batches 3 to 5 compared to batch 2). In addition, 3 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 sampling of plant tissue. The impact of 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, SNPs profiling was done, and comparative de-duplication was performed between supposed contaminating and contaminated samples. When the reads abundance allowed it, SNP were detected using Geneious (Prime 2020.0.5, Biomatters) "find variant" functionality (only variants with more than 25% frequency were kept). Then the distribution of the SNP 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 de-duplication strategy. The mapped unique reads from sample 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 has been calculated according to recent recommendations for the statistical analysis of validation datasets (Massart et al., Accepted EPPO bulletin). The known formula has been used (True Positive/(True Positive + False Negative)) but without taking into account diluted samples nor alien controls. But 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 (**Table 3S1**). 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 3-1 and publicly available (https://dataview.ncbi.nlm.nih.gov/object/PRJNA777477?reviewer=i5afdqucfeme0tps544b5vudlo).

	Sequencing Batch 1 totalRNA	Sequencing Batch 2 totalRNA	Sequencing Batch 3 totalRNA	Sequencing Batch 4 totalRNA	Sequencing Batch 5 totalRNA	Sequencing Batch6 smallRNA
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

Table 3-1: General information on the HTS data generated for five independent sequencing batches

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 12M, excepting batch 4 with 26 M. In addition, the average numbers of contigs generated after *de novo* assembly for each library ranged from 52949 to 102211. 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 Tables 3S1a and 3S1b) show that small RNA sequencing and total RNA sequencing protocols generated 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 to 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 detections of viruses were achieved with 10

or less reads and one false negative result 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 3-2.

Batch #	1	2	3	4	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	26
Average number of cross-contaminating reads per sample	6	20	31	11	69

Table 3-2: Summary of the BYDV reads detected in the sequencing datasets for each batch of sequencing

Table 3-2 underlined several trends in the cross-contamination level by analysing 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 190x and 7,269x in the different batches. In addition, the number of Musa samples with cross-contaminating (BYDV) reads is also very variable, from less than 10% (n=2 – 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 while, in batch 2, most samples had less than 20 cross-contaminating reads. In addition, within a batch, the number of cross-contaminating reads detected in a sample could vary significantly as it ranged from 0 (five samples) to 288 within batch 3. The very 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 668x. So, the cross-contamination detected could be quite constant between samples (batch 2) or concentrated in few samples (batch 3).

Determining 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 for the viruses infecting the *Musa* samples. As reference samples were used in this study, it was possible to evaluate the 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 as cross-contamination due to the low number of viral reads) rates. We tested two simple contamination thresholds to evaluate their impact on the accuracy, the false positive rate and the false negative rate compared to the absence of any threshold. The first threshold was 10 reads which have been proposed empirically in the literature (Bloom et al., 2021a; Soltani et al., 2021b; Strong et al., 2014a). Nevertheless, the fixation of a unique threshold did not consider the variability of 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, …) 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 took into account the inter-batch variability observed and was considered as "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 (see Table 3-2).

It is important to mention that an unusual very 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 and false negative rate was presented independently for BSV species.

	Filter	non BSV virus predication			BSV virus predication		
		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%

Table 3-3: Evaluation of the impact of three levels of contamination thresholds on the accuracy, the false positive and false negative rates per batch. For each batch, the first line represents the HTS read base detection with no minimum read required (non threshold), the second with 10 read required (fixed threshold) and the third a number depending on the adaptative alien threshold (highest number of contaminating alien virus (BYDV) identified in a banana sample for the batch).

Table 3-3 shows that, whatever the filter, the false positive rate for BSV species was particularly high, confirming the necessity to address separately this viral complex. The absence of threshold allows the detection of all the viruses present in the samples at the expense of a very 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., Accepted EPPO bulletin), the serially diluted samples (up to 10,000x 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 very 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 positive for this batch corresponded to two BBrMV that are discussed later. The application of "10 reads" threshold improved the accuracy considerably (ranging between 84 and 93% for the viruses other than BSV species) with up to 10 folds 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

while it reached 2% for batch 1. The threshold based on the alien control provided overall the best accuracy (always equal or higher than the "10 reads" threshold) with, again a marked decrease in false negative rate. Nevertheless, it caused more false negative results in batches 2 (3%) and 3 (4%) while no false negative was observed for batches 1 and 5. For further analysis, the alien-based threshold was kept as 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), 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 3-3 suggested a very high rate of false positive results for BSV species. AS stated earlier, BSV sequences integrated in the genome can be activated and they can trigger an infection with viral particle of BSV. So far, the complete genome 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 while partial sequences of BSV are integrated (Chabannes et al., 2021a). 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 analyse more in depth the false negative rate, the detected BSV reads were summarized depending on the Musa genome and the presence of viral particles (see Table 3-4).

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 3-4: Summary of detection of BSV reads depending on the presence of B genome and the detection of BSV particles by IC-PCR.

For the plants without BSV particles, a marked difference was observed between accessions containing a B genome (whatever their ploidy and genome combination) and accession with only A genome. Indeed, BSV reads were detected in more than 90% of the samples (31 on 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 B genome with an average number of reads of three. Most of the samples with A genome had only one or two reads of BSV (contamination background) while 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 B genome but without BSV particle detected, 1 *Musa* sample with only A genome, one *Musa* sample with B genome and infected with BSOLV particles were further investigated. All those samples went through both PCR and inmmunocapture PCR-(IC-PCR) detection. The comparison on 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 genomic DNA of selected samples with B genome but were tested negative by IC-PCR, while no band was found by PCR or IC-PCR from the sample (ITC1833) with only A genome and a band was observed by PCR and IC-PCR for the sample infected by BSV particles.

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

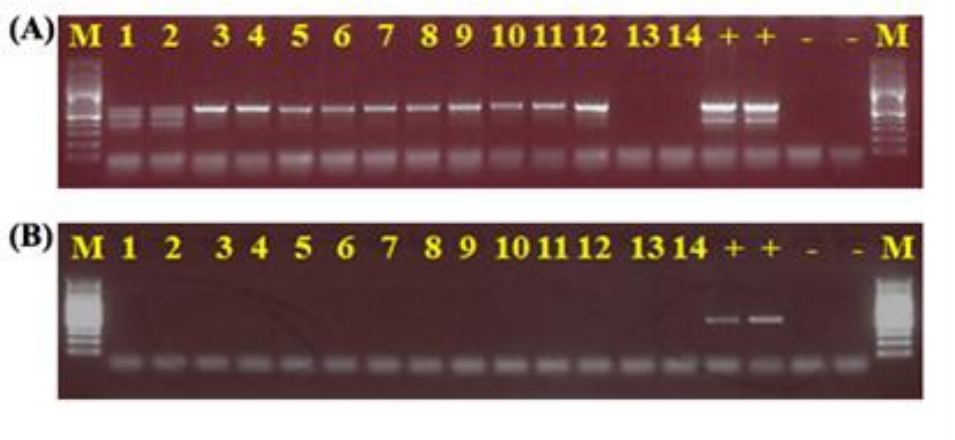


Figure 1 PCR (A) and IC-PCR (B) results of BSV species detection (BSOLV, BSMYV, BSIMV) on selected samples. 1,2: ITC1607 'Birbutia'(ABB), BSV un-infected; 3,4: ITC0148 'Isansi' (AAB), BSV un-infected; 5,6: ITC1498 'Libanga Dark Green' (AAB), BSV un-infected; 7,8: ITC1565 'Halahala' (AA), BSV un-infected; 9,10: ITC0146 'Mushaba'(AAB), BSV un-infected; 11,12: ITC1852 'Amagaba'(AAB), BSV un-infected; 13,14: ITC1833'ShweNi'(AAA), BSV un-infected; +: ITC1867 'Atili' (AAB), BSOLV infected in duplicate; -: ddH_2O as templates; M: GeneRuler 100 bp DNA ladder" from Thermoscientific. When 3 bands are observed, they corresponded to the detection of BSOLV (700 bp), BSMYV (589 bp), BSIMV (400 bp)

Higher inclusivity at isolate and species level for HTS test

The HTS test detected a BSIMV infection in accession ITC1843 and a BSMYV infection in accession ITC1599 while 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 available sequences on 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 (see Table 3S3). 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 (Figure S1) confirming the higher inclusivity of HTS testing.

The BLAST annotation of de novo contigs allowed the detection of other virus species compared to 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 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 amino acid level was 85% for one of the contig 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., in preparation). The presence of these contigs related to crinivirus or ampelovirus species was confirmed by independent RT-PCR (Rong et al., in preparation)

Diagnostic sensitivity (DSE) of HTS test

The alien control datasets and the datasets from diluted samples were not included in the in-depth analysis of diagnostic sensitivity, as recommended in the recent guidelines for statistical analysis of validation datasets (Massart et al., Accepted EPPO bulletin). So, 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 while 53 were coming from plant infected by at least one virus species (and up to five). According to PCR detection results, a total of 120 virus detection should occur within these 53 datasets.

Using the alien threshold, 115 true positives were detected, representing a DSE of 96 %. Considering only the non-integrated 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 rate of false negative 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. Noteworthy, the positive samples with BBTV corresponded to a 5x 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 10x and 72x higher than the maximum number of BBTV reads. So, 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 HTS test

Considering the problem of low BBTV reads number, four samples with a number of reads under the alien threshold were considered as 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 batch 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, hereafter referred as “old kit”) was no longer produced by the provider. On batch 3, another kit (TruSeq® Stranded Total RNA Library Prep Plant, hereafter referred as “new kit”) was used. A small-scale comparison between old and new kits (see Supplementary Table 3S1) was performed**.** The same viruses were detected between both kits (100% reproducibility) with minor differences on the number of reads mapped to each virus.

Although 100% repeatability was obtained for virus detection, a marked variation in reads number was observed between biological replicates (e.g. from different samplings from the sample tissue at the same time). For example, with 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. As the sequencing

of the same sample (sample J) generated 1,519x more BBrMV reads with batch 3 compared to batch 2 (see Supplementary table 3S1a).

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 5x dilution, while CMV could still be detected even after 1,000x 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. Also, BBrMV was detected up to 5X dilution for pool C, 100X dilution for pools B and D and 1,000x dilution for pool A. The total RNA samples with all the dilution levels from Pool 1 to 4 were tested by HTS. All the results are summarized in Table 3-5.

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

Table 3-5: Comparison of analytical sensitivity between RT-PCR and 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.

C & 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.

First, the impact of any detection threshold on the analytical sensitivity was very important. Indeed, without threshold, all the viruses are detected, even at 10,000x dilution, in some cases with a few reads. As previously shown, this very high analytical sensitivity was nevertheless counterbalanced by a very high number of false positive detection due to 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). So, 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 10X and 100X better for BanMMV, equal or 10X better for BBTV, equal to 100X better for BBrMV, and equal or 10X better for CMV. For BSOLV, the limit of detection of HTS test was equal or even lower. Overall, the HTS test presented an analytical sensitivity equal or improved compared to 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 while it was indexed positive. For further verification, the preserved lyophilized leaf and the extracted total RNA used for 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 were 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 this detection was not coming from cross-contamination between these samples. As the RNA extract was not available, an 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 very high quantity of CMV reads: on a total of 12M (10M) reads, there were 8M (5M) reads, corresponding to 708,736 (1M) unique reads, mapped on 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 n°1 (n°2) suggesting two specific cross-contamination events. The SNP list observed after mapping the 4 datasets on RNA3 reference sequence is detailed in Supplementary Table 3S6. The rate of contamination from the original sample was very low as it corresponded to 0.01% and 0.02% for ITC 1565 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, a SNP profiling analysis could not be done, instead, a de-duplication analysis was performed by comparing these 199 and 121 reads with the 38,666 reads generated from the BBrMV

most abundant sample. First, the 320 reads were deduplicated, leading to 249 unique reads (79%). Further on, 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 high throughput sequencing was carried out from new samples taken from lyophilized tissue of these two accessions but no BBrMV reads were detected in any of the sample (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

High throughput sequencing 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, contrasting results have been observed as the overall performance of small RNA was better in a study on various commodities (Gauthier et al., 2022b) while total RNA protocol was better with other commodities like tomato, lemon tree or grapevine in several studies (Di Gaspero et al., 2022b; Pecman et al., 2017b; Visser et al., 2016b). 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 underlined 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 very variable with the number of alien virus reads ranging from 0 to 288 reads per sample in batch 3 while 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 one to 10 viral reads were observed (Soltani et al., 2021b). Putative cross-contaminating reads were also observed in another study, causing a false discovery rate of up to 11% (Gauthier et al., 2022b). So, the determination of a threshold for detection, taking possible cross-contamination into account, 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., 2021b; Soltani et al., 2021b; Strong et al., 2014b) but it does not consider the possible variability in cross-contamination burden. With a threshold of 10 reads, a very high false discovery rate (21%) was observed on grapevine (Soltani et al., 2021b) but an in-depth analysis of the results showed that 40% of the false positive corresponded to samples with less than 30 reads for the detected virus. So, 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 like BBTV. We tested (unpresented results) 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 raised significantly the number of false positive and reduced the accuracy. Anyway, these results underlined 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., 2022b) that could be adapted for alien virus. As example, the maximum number of reads in a sample was 2,397 for BBTV, 288,675 for BBrMV and 86,915 for CMV. So, detecting a few reads of BBTV in a sample is less likely a cross-contamination event than for BBrMV or CMV when 36x or 120x 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 taking into account 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 control at different steps of the process (as sample tissue, RNA extract, and/or prepared library) would allow an improved cross-contamination analysis at the expense of higher cost as more samples 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 which typically present a low abundance of reads like 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 follow a trade-off between the ability to detect low level infection (raising true positive ratio) and baseline cross-contamination (raising 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 ensures 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, taking into account 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 as they might be also the origin of cross-contamination. The analysis of 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 2 samples are 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 adaptative threshold based on an alien virus control read detection in the samples of interest. With an expert review of the data taking into account the relative abundance of viral reads in samples for each viral species. The determination of such threshold should be evaluated during validation and will anyway correspond to a trade-off between the diagnostic sensitivity (improving while the threshold is diminishing and calculated at 100% without threshold) and the diagnostic specificity/ false discovery rate (worsening while the threshold is diminishing). 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 also 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 already the ability of HTS test to detect distant isolates of BanMMV that were not amplified by existing primers (Hanafi et al., 2020b). In this publication, HTS test detected isolates of two BSV species and triggered the development of new primers able to detect them. This underlined again 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 HTS test as a new virus infecting *Musa* plant was detected only by HTS test (Hanafi et al., 2022a). In this report, at least two new viral species were detected by HTS test and are currently characterized. Similar cases of new viral species discovery when evaluating HTS test on a broad range of samples have been reported previously (Pecman et al., 2017b; Rott et al., 2017b). The application of HTS test on banana *in vitro* plants (Hanafi et al., 2022b) also revealed a better diagnostic sensitivity of HTS test compared to 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 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., 2021a). So, it will be important to specifically identify the BSV species detected in A genome to evaluate the risk of presence of viral particles. More globally, other virus species, for example, badnavirus or geminivirus, are known to be integrated in 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 a 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 very good results observed during previous evaluations on grapevine or *Citrus* sp. (Bester et al., 2021b; Di Gaspero et al., 2022b; Soltani et al., 2021b). Behind this 100% value, a huge variability in reads number wasobserved 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 a complete virus indexing of an accession requires testing a pool of at least four plants with midribs and limbs from the 3 last leaves in order 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 as appropriate for some viruses although this number was not applicable to all viruses (Gauthier et al., 2022b; Visser et al., 2016b). 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., 2022b) 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, a subsampling will have no effect as 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 fitted 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 B genome to confirm the presence of viral particles.

Acknowledgment:

We would like to 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), 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

Funding

The work has been supported by (1) the NGS cross-center project from the CGIAR Fund, and in particular by the Germplasm Health Unit (GHU) of the CGIAR Genebank Platform, and (2) European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 813542T (INEXTVIR)

Supplementary data

All supplementary material are available here :

Supplementary_3

Supplementary Table 3S1a:

Samples sequenced by total RNA sequencing (5 batches) during the validation of the HTS test with the total number of reads sequenced by library and the number of reads mapping on reference genomes for 9 viral species and the alien virus. Impact of the detection threshold on the accuracy of the test. Colored cells correspond to viral

reads detection. In blue: true negative, in orange: false positive, in green: true positive and in red: false negative

Supplementary Table 3S1b:

Samples sequenced by small RNA sequencing with the total number of reads sequenced by library and the number of reads mapping on reference genomes for 9 viral species and the alien virus. Impact of the detection threshold on the accuracy of the test. Colored cells correspond to viral reads detection. In blue: true negative, in orange: false positive, in green: true positive and in red: false negative

Supplementary Table 3S2:

The final dilution of each virus (BanMMV, BSOLV, BBTV, BBrMV, and CMV) in the 8 mixed RNA pools and the origin information of samples used in this study for detection sensitivity comparison between RT-PCR and HTS

Supplementary Table 3S3:

The sequences of primers used for BanMMV, BSV, BBTV, BBrMV, and CMV detection in PCR experiments

Supplementary Table 3S4:

Genebank accession number of the reference genomes used in this study for mapping the reads generated from Musa sample

Supplementary Table 3S5a:

Mutations (frequency>25%) observed on the read datasets of ITC1565 when mapping the reads on AB369272 RNA3 CMV and comparison of their presence/frequency with original external sample

Supplementary Table 3S5b:

Mutations (frequency>25%) observed on the read datasets of ITC1498 when mapping the reads on AB369272 RNA3 CMV and comparison of their presence/frequency with original external sample

References

Adams, I. P., Fox, A., Boonham, N., Massart, S., de Jonghe, K., Tahzima, R., Foucart, Y., Peusens, G., Beliën, T., Massart, S., & de Jonghe, K. (2018). The impact of high throughput sequencing on plant health diagnostics. *European Journal of Plant Pathology*, *152*(4), 909–919. https://doi.org/10.1007/s10658-018-1570-0

Adams, I. P., Skelton, A., Macarthur, R., Hodges, T., Hinds, H., Flint, L., Nath, P. D., Boonham, N., & Fox, A. (2014). *Carrot yellow leaf virus is associated with carrot internal necrosis*. *9*(11). https://doi.org/10.1371/journal.pone.0109125

Al Rwahnih, M., Daubert, S., Golino, D., Islas, C., & Rowhani, A. (2015). Comparison of Next-Generation Sequencing Versus Biological Indexing for the Optimal Detection of Viral Pathogens in Grapevine. *Phytopathology®*, *105*(6), 758–763. https://doi.org/10.1094/PHYTO-06-14-0165-R

Bester, R., Cook, G., Breytenbach, J. H. J., Steyn, C., De Bruyn, R., & 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. *Virology Journal 2021 18:1*, *18*(1), 1–19. https://doi.org/10.1186/S12985-021-01523-1

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., … Arboleda, V. A. (2021). Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. *MedRxiv : The Preprint Server for Health Sciences*, 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., & Mumford, R. (2014). Methods in virus diagnostics: From ELISA to next generation sequencing. *Virus Res*, *186*, 20–31. https://doi.org/10.1016/j.virusres.2013.12.007

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, *12*(10). https://doi.org/10.1371/JOURNAL.PONE.0185056

Calil, I. P., & Fontes, E. P. B. (2017). Plant immunity against viruses: antiviral immune receptors in focus. *Annals of Botany*, *119*(5), 711–723. https://doi.org/10.1093/AOB/MCW200

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., & Muller, E. (2021). Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. *Molecular Plant Pathology*, *22*(2), 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

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., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant

dual indexing on massively parallel sequencing platforms. *BMC Genomics*, *19*(1), 332. https://doi.org/10.1186/s12864-018-4703-0

De Clerck, C., Crew, K., Van den houwe, I., McMichael, L., Berhal, C., Lassois, L., Haissam Jijakli, M., Roux, N., Thomas, J., & Massart, S. (2017). Lessons learned from the virus indexing of Musa germplasm: insights from a multiyear collaboration. *Annals of Applied Biology*, *171*(1), 15–27. https://doi.org/10.1111/AAB.12353

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., & Marroni, F. (2022). Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. *Journal of Virological Methods*, *300*, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

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., & Adams, I. P. (2018). Using high-throughput sequencing in support of a plant health outbreak reveals novel viruses in Ullucus tuberosus (Basellaceae). *Plant Pathology*, *68*(3), ppa.12962. https://doi.org/10.1111/ppa.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., & Ziebell, H. (2021). Interlaboratory Comparison Study on Ribodepleted Total RNA High-Throughput Sequencing for Plant Virus Diagnostics and Bioinformatic Competence. *Pathogens (Basel, Switzerland)*, *10*(9), 1174. https://doi.org/10.3390/pathogens10091174

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., & Cosson, J. F. (2016). 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. *MSystems*, *1*(4), 1–22.

Gauthier, M.-E. A., Lelwala, R. V, Elliott, C. E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., & Barrero, R. A. (2022). Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. In *Biology* (Vol. 11, Issue 2). https://doi.org/10.3390/biology11020263

Geering, A. D. W., McMichael, L. A., Dietzgen, R. G., & Thomas, J. E. (2000). Genetic Diversity Among Banana streak virus Isolates from Australia. *Phytopathology®*, *90*(8), 921–927. https://doi.org/10.1094/PHYTO.2000.90.8.921

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., & 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. In *Plants* (Vol. 11, Issue 2). https://doi.org/10.3390/plants11020226

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., & Massart, S. (2020). Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. In *Pathogens* (Vol. 9, Issue 12). https://doi.org/10.3390/pathogens9121045

Kechin, A., Boyarskikh, U., Kel, A., & Filipenko, M. (2017). CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *Journal of*

Computational Biology, *24*(11), 1138–1143. https://doi.org/10.1089/CMB.2017.0096/ASSET/IMAGES/LARGE/FIGURE3.JPEG

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., & 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. In *Plants* (Vol. 10, Issue 2). https://doi.org/10.3390/plants10020328

Lebas, B., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G., Blouin, A., Boonham N., Candresse, T., Chandelier, A., De Jonghe, K., Fox, A., Gaafar, Y., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., … Massart, S. (n.d.). Guidelines for preparing a laboratory to use high-throughput sequencing as routine diagnostic test: general and technical recommendations. *EPPO Bulletin.*

Maachi, A., Torre, C., Sempere, R. N., Hernando, Y., Aranda, M. A., & Donaire, L. (2021). Use of High-Throughput Sequencing and Two RNA Input Methods to Identify Viruses Infecting Tomato Crops. In *Microorganisms* (Vol. 9, Issue 5). https://doi.org/10.3390/microorganisms9051043

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C. M., 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*. https://doi.org/10.3389/fmicb.2020.578064

Malapi-Wight, M., Adhikari, B., Zhou, J., Hendrickson, L., Maroon-Lango, C. J., McFarland, C., Foster, J. A., & Hurtado-Gonzales, O. P. (2021). HTS-Based Diagnostics of Sugarcane Viruses: Seasonal Variation and Its Implications for Accurate Detection. In *Viruses* (Vol. 13, Issue 8). https://doi.org/10.3390/v13081627

Maree, H. J., Fox, A., Al Rwahnih, M., Boonham, N., & Candresse, T. (2018). Application of hts for routine plant virus diagnostics: state of the art and challenges. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01082

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., … Candresse, T. (2019). Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. *Phytopathology*, *109*(3). https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, S., Olmos, A., Jijakli, H., & Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research*, *188*. https://doi.org/10.1016/j.virusres.2014.03.029

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, Jonkers W., Kreuze J., Kutnjak D., & Landa B. (n.d.). *Guidelines for the reliable use of*

high throughput sequencing technologies to detect plant pathogens and pests | Zenodo. Retrieved July 11, 2022, from https://zenodo.org/record/6637519

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., … Massart, S. (2018). High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. *EPPO Bulletin*, *48*(2), 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., & Ravnikar, M. (2017). Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . In *Frontiers in Microbiology* (Vol. 8).

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*(8), 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Santala, J., & Valkonen, J. P. T. (2018). Sensitivity of Small RNA-Based Detection of Plant Viruses . In *Frontiers in Microbiology* (Vol. 9).

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., & Weissman, I. L. (2017). Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *BioRxiv*, 125724. https://doi.org/10.1101/125724

Soltani, N., Stevens, K. A., Klaassen, V., Hwang, M.-S., Golino, D. A., & Al Rwahnih, M. (2021). Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. In *Viruses* (Vol. 13, Issue 6). https://doi.org/10.3390/v13061130

Strong, M. J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., & Lin, Z. (2014). Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. *Journal of Virology*, *88*(18), 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Tamisier, L, Colson, C., Maclot, F., Zhang, P., Wang, X., Francis, F., Baurain, D., & Massart, S. (n.d.). *Deciphering the intra-host genetic diversity of viruses: insights from the experimental evolution of a plant virus*.

Tamisier, Lucie, 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., & Massart, S. (2021). Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. *Peer Community Journal*, *1*. https://doi.org/10.24072/pcjournal.62

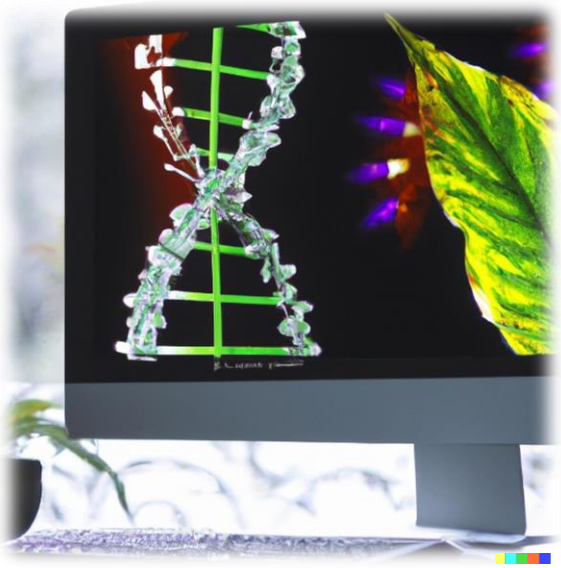
Thomas, J., Sharman, M., Lassois, L., Massart, S., De Clerck, C., Caruana, M.-L., Chabannes, M., Teycheney, P.-Y., Kumar, P. L., Van den Houwe, I., & Roux, N. (2015). *Technical guidelines for the safe movement of Musa germplasm* (J. Thomas (ed.); 3rd Editio). Bioversity International.

Velasco, L., & Padilla, C. V. (2021). High-Throughput Sequencing of Small RNAs for the Sanitary Certification of Viruses in Grapevine. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/FPLS.2021.682879/FULL

Visser, M., Bester, R., Burger, J. T., & Maree, H. J. (2016). Next-generation sequencing for virus detection: covering all the bases. *Virology Journal*, *13*, 85. https://doi.org/10.1186/s12985-016-0539-x

4

Automation of virus detection



Article 2

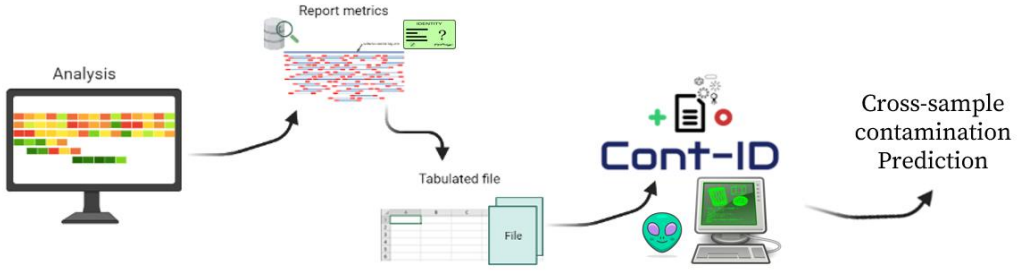
(in BMC Biology)

Cont-ID: Detection of samples cross-contamination in viral metagenomic data

Johan Rollin*[1,2], Wei Rong*[1] and Sébastien Massart[1]

*These authors contributed equally to this work

1. *University of Liège, Gembloux Agro-Bio Tech, Plant Pathology Laboratory, 5030, Gembloux, Belgium*
2. *DNAVision, 6041, Gosselies, Belgium*



Background: High Throughput sequencing (HTS) technologies completed by the bioinformatic analysis of the generated data are becoming an important detection technique for virus diagnostics. They have the potential to replace or complement the current PCR-based methods thanks to their improved inclusivity and analytical sensitivity, as well as their overall good repeatability and reproducibility. Cross-contamination is a well-known phenomenon in molecular diagnostics and corresponds to the exchange of genetic material between samples. Cross-contamination management was a key drawback during the development of PCR-based detection and is now adequately monitored in routine diagnostics. HTS technologies are facing similar difficulties due to their very high analytical sensitivity. As a single viral read could be detected in millions of sequencing reads, it is mandatory to fix a detection threshold that will be informed by estimated cross-contamination. Cross-contamination monitoring should therefore be a priority when detecting viruses by HTS technologies.

Results: We present Cont-ID, a bioinformatic tool designed to check for cross-contamination by analysing the relative abundance of virus sequencing reads identified in sequence metagenomic datasets and their duplication between samples. It can be applied when the samples in a sequencing batch have been processed in parallel in the laboratory and with at least one specific external control called Alien control. Using 273 real datasets, including 68 virus species from different hosts (fruit tree, plant, human) and several library preparation protocols (Ribodepleted total RNA, small RNA and double stranded RNA), we demonstrated that Cont-ID classifies with high accuracy (91%) viral species detection into (true) infection or (cross) contamination. This classification raises confidence in the detection and facilitates the downstream interpretation and confirmation of the results by prioritising the virus detections that should be confirmed.

Conclusions: Cross-contamination between samples when detecting viruses using HTS (Illumina technology) can be monitored and highlighted by Cont-ID (provided an alien control is present). Cont-ID is based on a flexible methodology relying on the output of bioinformatics analyses of the sequencing reads and considering the contamination pattern specific to each batch of samples. The Cont-ID method is adaptable so that each laboratory can optimise it before its validation and routine use.

Keywords:(Lebas et al., 2022)

Bioinformatic, virus, detection, sequencing, contamination, metagenomic

Background

The advent of high-throughput sequencing (HTS) technologies coupled with the development of powerful bioinformatics approaches has improved our ability to detect viruses in a non-targeted way from any sample collected from diverse sources. Noteworthy, detecting viruses by HTS technologies relies on many steps in the laboratory: sampling, transport and storage, nucleic acid extraction, library preparation, and sequencing (Lebas et al., 2022). Compared to other molecular tests like (RT)-PCR, these steps are much more numerous and complex (Sebastien Massart et al., 2014).

The analytical sensitivity, e.g. the ability to detect viral species at very low concentration in a sample, has been demonstrated to be similar to or even better than RT-PCR for animals (Charlebois et al., 2020) or plant viruses(Rong et al., 2022; Soltani et al., 2021b)). In addition, the inclusivity of HTS technologies, e.g. the ability to detect all isolates from a species and all species whose nucleic acids are present in enough quantity in a nucleic acid extract, is particularly high compared to any other detection test (Maree et al., 2018; Sebastien Massart et al., 2014). Consequently, the use of HTS technologies is currently expanding at a rapid pace in research and is also progressively used for the diagnostic of viruses threatening humans (Ng et al., 2018), including SARS-Cov-2 (Kumar et al., 2020), livestock (Vereecke et al., 2021) or plant health (Olmos et al., 2018b).

The broader application of HTS technologies for virus detection, with the simultaneous analysis of tens to hundreds of samples, is raising a significant challenge that needs to be addressed: the management of cross-contamination between samples. Scientists and diagnosticians already faced such challenges decades ago during the development of PCR-based techniques for detecting plants (Grosdidier et al., 2017; Lau and Botella, 2017) or animal viruses (Moonen et al., 2003; Watzinger et al., 2006) and this phenomenon might worsen with the use of HTS for virus detection (Sebastien Massart et al., 2014). The higher complexity of laboratory operations, the intrinsically very high inclusivity, and the very low limit of detection (few viral reads are enough to detect the virus) of HTS make cross-contamination a more pressing issue. This is a frequently observed but, until recently, rarely reported observation in many, if not all, laboratories that have tested these technologies for virus detection. In many cases, these problems are frequently limited to a low number of reads and are of little consequence. Still, the specifics of the diagnostics field, with the need to detect viruses that can be at very low titre in the sample, clearly give more impact to such potential contamination problems (Sebastien Massart et al., 2014). The occurrence of contamination is, therefore, a key element to consider when interpreting the viruses detected in HTS datasets.

The consequences of erroneous detection due to cross-contamination between samples can be catastrophic, as described for tuberculosis prior to HTS (Martínez et al., 2006) but also using HTS for human and plant viruses (Bukowska-Ośko et al., 2017; Gauthier et al., 2022c).

So, even if the cross-contamination issue of HTS, and particularly with the widely adopted Illumina technology, is long known and discussed in the scientific community, proper methodologies and dedicated algorithms to address it are still missing. Until now, the burden of detection confirmation relied on the virologist's expertise and the use of laboratory tests to independently confirm the presence of the virus in the sample, which is a fastidious, costly, and time-consuming task. To minimise the confirmation burden, arbitrary thresholds (like 5 or 10 reads) (Bloom et al., 2021b; Soltani et al., 2021b) have been proposed to consider a detection valid. Still, these thresholds are subjectively fixed based on the sequencing/detection tools or the scientist's experience. In addition, it has been shown recently that the cross-sample contamination burden can be very variable between sequencing batches and that an adaptative threshold is required(Rong et al., 2022). Therefore, the need for formal bioinformatic pipelines for HTS-based data that consider the possible cross-contamination is growing (Ballenghien et al., 2017).

To handle cross-contamination, several laboratory protocol improvements have been implemented over time: laboratory or reagent decontamination, alternate dual indexes, inter-run washing (Champlot et al., 2010; Costello et al., 2018b) or, more recently, the use of a specific external control called alien control. An alien control is defined as "a matrix infected by a target (called alien target) which belongs to the same group as the target organism to be tested in the samples, but that cannot be present in the samples of interest." (Massart et al., 2022b). It is processed as external control alongside the sample to be analysed. It is preferably the same type of matrix as the analysed samples: plant tissue, water … Ideally, the alien target, in our case a virus, should be at a high concentration in the alien sample as it allows a better analysis of cross-contamination between samples. Indeed, the probability of detecting any virus at a low level due to cross-contamination rises if this virus is very abundant in at least one of the processed samples. A high abundance of the alien virus will therefore allow better monitoring of contaminations, including for other viruses highly abundant in at least one tested sample. The presence of sequencing reads from the alien virus in any tested sample can be considered the consequence of contamination from the alien control to this sample. The concentration of the alien virus should nevertheless remain close to the highest expected concentration of the viruses to be detected to avoid overestimation of cross-contamination. Such information can be used to monitor the cross-contamination level between samples within the sequencing batch.

Many generalist bioinformatic tools, such as Kraken (Wood et al., 2019) or BLAST (Camacho et al., 2009) can detect the presence of viruses in HTS datasets with very high analytical sensitivity, as the detection is possible from a single viral read or contig. Some of them, like VirHunter (Sukhorukov et al., 2022), VirAnnot (Lefebvre et al., 2019) or VirusDetect (Zheng et al., 2017), have been specifically developed for that purpose. Nevertheless, they have not been designed to detect cross-contamination in the input datasets. Instead, they will detect a virus, whatever its origin: virus

infection in the biological sample or contamination from another sample. The risk of contamination is particularly acute for viruses in very high abundance in one of the samples sequenced as a few contaminating reads can be detected by the bioinformatic tools in other samples prepared in parallel. The situation's impact is growing, especially in the diagnostic field (Sebastien Massart et al., 2014) as false positive results due to contamination can lead to inaccurate data interpretation, which can cause tremendous health and trade issues.

According to EDAM ontology (Ison et al., 2013), tools that address cross-contamination issues should be labelled as "Sequence contamination filtering". We used these EDAM terms along with usual ones (virus reads contamination, cross-contamination, …) to identify existing tools potentially useful to detect cross-contamination in HTS viral diagnostic assays. Some tools address similar issues like contamination on bacterial isolates (ConFindr- (Low et al., 2019) or bacterial metagenome (GUNC - (Orakov et al., 2021). They both use methods relying on operons organisation of genes that are not applicable for viruses. Croco (Simion et al., 2018) uses an approach mainly based on bacterial quantitative data. Finally, DecontaMiner (Sangiovanni et al., 2019) can be applied to metagenome data, including viruses but is based on a combination of detection methods (mainly mapping and Blast) that try to assign the dark matter (reads from unknown origin) more than formally detecting the cross-contamination material. To our knowledge, there is no tool specifically addressing cross-contamination during virus detection in metagenome datasets. It means that some risks of false positive results remain unmonitored for virologists, and the burden of confirmation of detection in case of false positive is still not addressed.

To solve this issue, we present Cont-ID, a method designed to check sample cross-contamination for viruses previously identified in metagenomic datasets. It relies on a simple requirement: every sample in a sequencing batch should have been processed at the same time and followed the same steps in the laboratory with at least one alien control as external control. Cont-ID uses a voting system to classify every species prediction on each sample of the sequencing batch into (true) infection (trully presence of the virus in the sample) or (cross) contamination. This tool will help the virologist to distinguish virus presence and virus cross-contamination in HTS data generated by Illumina technology, improving the reliability of viral detection and the efficiency of downstream confirmation and characterisation analyses. It can also help to improve feedback on upstream steps that might be linked to cross-contamination events. Cont-ID is an open-source python (v3) based script method freely available here: https://github.com/johrollin/viral_contamination.

Methods:

Implementation

In viral metagenomics, detecting multiple viral species in the same sample is frequent, and a virus species can be seen with different confidence levels in several samples of the same sequencing batch. Therefore, Cont-ID aims to determine whether a given detected virus in a sample is likely to be a contaminant or not by comparing it to the results from the other samples of the same sequencing batch, e.g. samples processed in parallel and following the same laboratory steps.

Cont-ID does not require any development or maintenance of database as it only relies on data generated by usual bioinformatics tools for Illumina dataset analyses and, most specifically, two elements: (i) the normalised abundance estimation (number of reads assigned per sample to each detected virus species, subtype - or any relevant taxonomical level) and (ii) the number of identical reads among pairs of samples (deduplication ratio). These input metrics are easy to obtain as the abundance estimation can be calculated by using any mapping tool like BWA (Li and Durbin, 2009a) or a read classifier like Kraken/Bracken (Lu et al., 2017; Wood et al., 2019), and the number of identical reads from a virus between two samples can be obtained by running any deduplication tool like BBduck (Kechin et al., 2017). A tabulated file containing these numbers associated with the detected virus name and the unique ID for each batch sample is used as input for Cont-ID, as shown in **Figure 4-2**. Each virus predicted on each batch sample is considered a distinct element and corresponds to a line in the generated table. A separate table is generated for the alien virus.

Computing the two elements mentioned above into three different metrics for the alien virus and each detected virus, Cont-ID can predict through three rules if a given viral species detection is likely a cross-contaminant or not in the sequencing batch, as described in **Figure 4-1**.

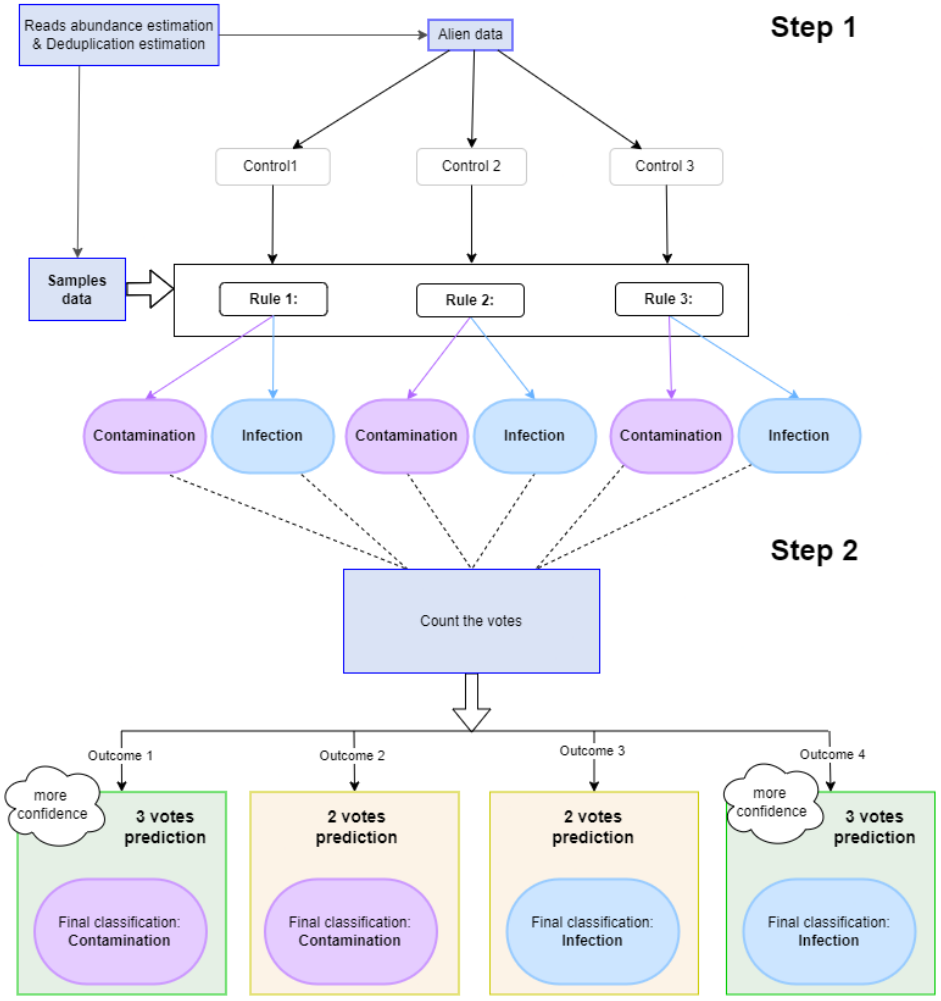


Figure 4-1: Cross-contamination prediction with Cont-ID. Controls 1, 2 and 3 and rules 1, 2 and 3 are described in detail in figure 4-2.

There are two input files, one for the alien data and one for the samples data. Alien file is used to calculate control thresholds which are then used along with the sample data to apply rules to a voting system (step 1). The votes are then counted to decide for each virus on each sample (element) either if it is a (cross)contamination or an infection (step 2).

The three rules classify as contamination or infection each element according to the pattern of reads number observed among the samples and the alien control for the alien virus and the considered viral species. Rules one and two both use the (normalised) reads abundance estimation, while rule three uses the assessment of

unique (identical) reads. Rules are calculated after normalising the number of reads per sample and are described more precisely in **Figure 4-2**.

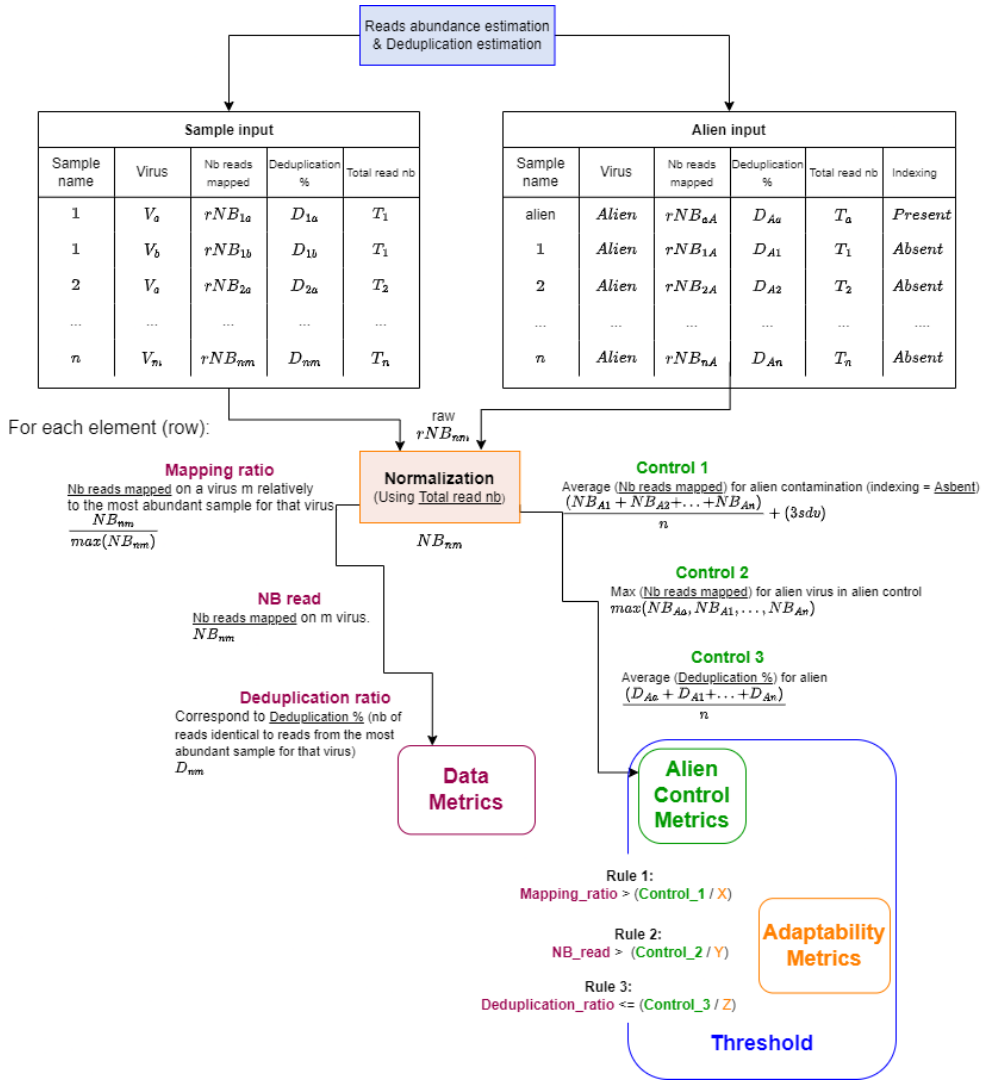


Figure 4-2: Cont-ID rules explanation.

There are two input files, one for the alien data and one for the samples data. Alien file is used to calculate each alien control metric after normalisation (except control 3). The sample file is used to calculate each data metric after normalisation (except deduplication ratio) (the small "r" in table stand for "raw"). Each alien control metric is associated with a user (manually) designed adaptability metric (X, Y or Z) to compose each rule's threshold. Finally, each Data Metric is compared to the corresponding threshold in order to obtain the three rules used in Cont-ID.

The first rule uses the mapping ratio of each virus in each sample (corresponding to an element): the number of reads of each element is divided by the maximum number of reads of the corresponding virus in one of the samples. This first rule compares this mapping ratio for the element with the Control 1 metrics calculated for the alien virus and corresponds to the average number of reads mapped on the alien virus in the samples for which the alien is a cross-contaminant, plus three times the standard deviation of this average. The second rule relies on the number of reads of the element in the sample. The rule compares it with Control 2, corresponding to the number of alien reads identified in the alien control. The third rule is based on each element's deduplication ratio, which is compared with Control 3. The average deduplication ratio of the alien virus reads between each tested sample and the alien control.

We aimed to find the most reliable formula for threshold calculation on each rule while allowing a part of adaptability according to the biological system used. As the system variability can come from the laboratory using HTS, the host and type of sample (fruit tree, herbaceous plant, human, animal …), the type of virus (integrated or non) or the extraction protocol used (dsRNA, total RNA, small RNA...), each rule includes a third number (represented by X, Y or Z) that is called adaptability metrics (see **Figure 4-2**). The X will impact the first rule that considers the relative proportion of reads of a virus in this sample compared to the sample with the maximum read of this virus. This threshold is a refinement of the "alien threshold" described earlier(Rong et al., 2022). The default value proposed is 2. The Y divides the number of reads from the alien virus in the alien control for comparing it to the number of reads of each virus in each sample. In this publication, a default value of 1,000 has been fixed for Y, and it was in the range of the expected (cross) contamination ratio (number of reads in the truly infected sample versus the number of reads in contamination one). The Z metric impacts Control 3 and the evaluation of the proportion of identical reads between different samples. The proportion of identical reads can be influenced by different factors (mutation rate, respective genome length …). The role of Z is to consider those different factors. A default value of 1.5 is proposed.

Default values of the three adaptability metrics have been provided in this publication after their optimisation on the banana datasets and their evaluation of other datasets. Nevertheless, users can independently modify them during the evaluation or validation of Cont-ID applied to their datasets. A careful evaluation of the adaptability metrics by the user is recommended to evaluate their impact on the diagnostic performance of the test. In addition, several sets of adaptability metrics can be run in parallel for further improvements in diagnostics performance. The value given to the adaptability metrics and controls resulting is always recorded in an additional log file (see Supplementary File 1 [log_file]). This log file help to ensure traceability allowing the user to check the pertinence of the chosen numbers and to adapt them when needed. As each of the three rules has two possible decisions (contamination or infection), a majority vote will be obtained with two or three votes. The decision of each vote is available in the generated result to support the result interpretation and

let the user decides on the confidence to give to each individual rule according to the biological system tested.

In addition, the proper quantitative comparison of sequencing reads datasets relies on normalising the number of reads per sample, for example, as always done for transcriptomic or microbiome studies. This assertion is also true for Cont-ID and corresponds to an adaptative parameter. To limit some bias due to the difference in sequencing depth between samples in the same batch, we also normalise by default to 5 000 000 reads in this publication. Still, it is manually changeable by the user.

Finally, Cont-ID also has another level of flexibility: the script is desing for easy change of rules in the code that can complete or replace the existing ones.

Conditions of application for Cont-ID

Three main conditions are essential to run Cont-ID. First, an alien control should be used, alien control should contain a high but realistic (close to highest expected concentration of viruses in the analyzed samples) concentration of the alien virus, so reads from that viral species are more prone to be detected in other samples when cross-contamination occurs from the alien control to the other samples. Similarly, if another virus is found in the alien control sample, that is also an indication of potential contamination (although not used so far by Cont-ID). The alien control is bioinformatically processed exactly as the samples of interest to generate the alien metrics for each sample (in a separate tabulated file). In the absence of external alien control, it is still possible to analyse sequencing batches if they include samples from different host species and some detected viruses, preferably at high abundance, are known to infect only some of the host species. In such a case, the alien file should be filled with the selected virus as if it was an alien (with the status of alien present/absent in the file). Nevertheless, the threshold set-up and the results will be less accurate and include fewer samples (the samples corresponding to species that can be hosts for the virus could not be considered). In addition, a high degree of confidence is needed regarding the actual infection of the sample selected as alien control by the virus selected as an alien virus. Cont-ID always requires at least one (cross) contamination in the alien file to be reported; otherwise, the threshold calculation will fail; in that case, the tool will state it.

The second application condition is related to the processing of the samples and the alien control. The alien control and all the other samples in a given batch should have been processed together in parallel for all the laboratory steps (RNA/DNA extraction, library preparation, sequencing) and bioinformatics (Reads cleaning, host removing …). This is a good diagnostic practice, but it is even more important here as the goal is to observe cross-contamination levels. The assumption is that the level seen with the alien represents what could have happened in samples of interest. Therefore, this assumption depends on processing all samples and control in parallel.

The third condition is that, once the user has fixed the adaptability metrics, the analysis should be carried out batch per batch. The calculation of sample and alien metrics is dynamically done for each batch as cross-contamination patterns can strongly vary between batches, as recently shown for banana samples (Rong et al., 2022).

Sequencing reads datasets

The first datasets (batches A to D) were generated in our laboratory by total RNA sequencing protocol with ribodepletion applied to RNA extracted from banana plants (belonging to the *Musa* genus) (Rong et al., 2022). These data were generated to compare the test performance criteria of high throughput sequencing with classical virus testing protocols that include ImmunoCapture (IC)-(RT)-PCR and electron microscopy (De Clerck et al., 2017). The alien control corresponded to wheat plants infected by two species of barley yellow dwarf virus (BYDV-PAS and BYDV-PAV)(Rong et al., 2022). In total, four sequencing datasets (called A, B, C, and D) composed respectively of 27, 20, 27 and 25 samples were generated independently. A fifth batch generated during this validation experiment using diluted samples for evaluating the limit of detection (analytical sensitivity) was not included in our analysis according to the recent guidelines proposed for statistical analysis of validation datasets for plant pest detection (Massart et al., 2022b). All samples used here have been tested with (RT)-PCR for all the possible virus known in *Musa*, making the viral status known for every sample (Rong et al., 2022) A total of 10 different viral species were infecting these samples, including banana mild mosaic virus (BanMMV), banana bract mosaic virus (BBrMV), banana bunchy top virus (BBTV), cucumber mosaic virus (CMV), and five species belonging to the banana streak virus (BSV) species complex. In addition, two other sequencing protocols were applied to some banana plants, small RNA sequencing (Rong et al., 2022) starting from the same RNA extract as total RNA sequencing (for 21 samples in a single batch) and double-stranded RNA (dsRNA) enrichment and sequencing protocol (Marais et al., 2018b) applied from plant tissue of 13 samples in a single sequencing batch.

Batch_ID	NB of samples	Host type	Extraction Method	Extraction kit	Library kit	Sequencing	Data link	Publication
A (1)	27	Plant (Musa)	Total RNA extraction	RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands)	Stranded Total RNA library Prep Human/Mouse/Rat illumine CA, United States) & Ribo-Zero™ Plant Leaf Kit (illumine CA, United States)	Illumina NextSeq 500 2X150	BioProject: PRJNA777477 samples starting with (1-XXX)	Wei et al., 2022 https://doi.org/10.1094/PHYTOFR-03-22-0030-FI
B (2)	20	Plant (Musa)	Total RNA extraction	RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands)	Stranded Total RNA library Prep Human/Mouse/Rat illumine CA, United States) & Ribo-Zero™ Plant Leaf Kit (illumine CA, United States)	Illumina NextSeq 500 2X150	BioProject: PRJNA777477 samples starting with (2-XXX)	Wei et al., 2022 https://doi.org/10.1094/PHYTOFR-03-22-0030-FI
C (3)	27	Plant (Musa)	Total RNA extraction	RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands)	Stranded Total RNA library Prep Human/Mouse/Rat illumine CA, United States) & Ribo-Zero™ Plant Leaf Kit (illumine CA, United States) & TruSeq® Stranded Total RNA Library Prep Plant (illumine CA, United States)	Illumina NextSeq 500 2X150	BioProject: PRJNA777477 samples starting with (3-XXX)	Wei et al., 2022 https://doi.org/10.1094/PHYTOFR-03-22-0030-FI
D (5)	25	Plant (Musa)	Total RNA extraction	RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands)	TruSeq® Stranded Total RNA Library Prep Plant (illumine CA, United States)	Illumina NovaSeq 6000 2X150	BioProject: PRJNA777477 samples starting with (5-XXX)	Wei et al., 2022 https://doi.org/10.1094/PHYTOFR-03-22-0030-FI
E (6 sRNA)	31	Plant (Musa)	Total RNA extraction	RNeasy Plus Mini Kit (Qiagen, Venlo, Netherlands)	SMARTer smRNA-Seq Kit (Clonetech)	Illumina NovaSeq 6000 2X150	BioProject: PRJNA777477 samples starting with (1sR-XXXX)	Wei et al., 2022 https://doi.org/10.1094/PHYTOFR-03-22-0030-FI
F (5 dsRNA)	9	Plant (Musa)	Double stranded RNA	see article (Armelle Marais)	NEBNext Ultra II DNA library prep kit (New England BioLabs, US)	Illumina NovaSeq 6000 2X150	BioProject: PRJNA777477 samples starting with (5ds-XXX)	Method: Marais et al., 2018 https://doi.org/10.1007/978-1-4939-7683-6_4
G (Queensland university of technology)	5	Plant (mix)	Total nucleic acid	Maxwell® Rapid Sample Concentrator instrument using SimplyRNA Tissue kit (AS1340, Promega)	TruSeq Stranded Total RNA	Illumina NovaSeq 6000 2X150	BioProject: PRJNA752836	Gauthier, M.-E. A.,et all https://doi.org/10.3390/BIOLOGY11020263
H	49	Human	Total nucleic acid + amplification	NucliSENS EasyMAG platform (bioMérieux, Marcy l'Etoile, France)	Nextera XT (Illumina, San Diego, CA, USA)	Illumina NextSeq 500 2X150	bioproject: PRJNA494633	Bal et al., 2018 https://doi.org/10.1186/S12879-018-3446-5
I	25	Human	Total nucleic acid	MagNAPure 96 DNA and Viral NA Small Volume Kit (Roche Diagnostics, Almere, the Netherlands)	EBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA)	Illumina HiSeq 4000 and NextSeq 500 depth: 10 million 2X150	bioproject: PRJNA560243	Boheemen et al., 2020 https://doi.org/10.1016/J.JMOLDX.2019.10.007
J	55	Human	Total nucleic acid	TRIzol LS reagent (Invitrogen, USA)	SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Takara Bio, USA) and the Trio RNA-Seq kit (NuGEN Technologies, USA)	Illumina HiSeq 2X150	bioproject: PRJNA540900	Li, et al., 2020 https://doi.org/10.1038/s41598-020-60992-6

Table 4-1: list of datasets used on Cont-ID

BSV is a species complex (genus: *Badnavirus*, family: *Caulimoviridae*) among which five species were included in our samples: banana streak CA virus (BSCAV), banana Goldfinger virus (BSGFV), banana streak IM virus (BSIMV), banana streak Mysore virus (BSMYV), and banana streak OL virus (BSOLV). Notably, some species of this complex have their genome fully or partially integrated into the plant genome as endogenous viral elements (EVE), most specifically in B genomes originating from *M. balbisiana*. These EVE can be transcribed in the plant, and for some BSV species, they can even trigger an infection with viral particles of BSV in the plant (Chabannes et al., 2021b). It is well documented that BSGFV, BSIMV and BSOLV are constitutive of *Musa balbisiana* (B genome) but can be activated in some conditions (Ricciuti et al., 2021). In addition, BSMyV is also integrated into the Musa B genome, although the ability to produce infectious viral particles is not yet demonstrated. This brings additional complexity as EVE can be transcribed without the presence of a viral particle. It has been recently demonstrated that the detection of BSV transcripts by a HTS test must be confirmed by an independent test such as

immunocapture (IC)-PCR for confirming the presence of viral particles (Rong et al., 2022).

The other datasets used in this work came from publicly available datasets listed in **Table 4-1** and were already included in peer-review publications. They were selected because they fit two criteria: (i) having all virus presence checked in all the samples and (ii) having a virus species that could act like an alien control for the input file. First, another set generated to detect viruses from diverse plant samples by high throughput sequencing of total RNA extraction was kindly provided by Queensland University of Technology (Gauthier et al., 2022c), corresponding to a total of 19 plant viruses and viroid in 5 samples. In addition, the datasets generated from human samples came from published data from 3 different sources, with a total of 129 samples containing 39 viral species (Bal et al., 2018; Boheemen et al., 2020; Li et al., 2020). These three human datasets allowed us to test Cont-ID with a large diversity of viruses, with different extraction and sequencing methods listed in Supplementary File 2.

In total, ten sequencing batches, including 273 samples and the presence of 68 viral species, were used to test the potential impact of a different host, extraction, and sequencing method on Cont-ID performances. All the data generated are available with the link and procedure applied to obtain them described in **Table 4-1**; the indexing status of each virus in each sample is also available in Supplementary File 2.

Bioinformatic analyses

Quality control and mapping of sequencing reads

For all datasets, read quality control (quality trimming, reads deduplication) was performed using a standard procedure described in a previous publication: "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 35bp 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, the request of 100% identity and the option of eliminating shorter reads 100% identicals." (Rong et al., 2022). The cleaned reads were then mapped to a custom-built database (DB) containing all complete genome sequences from previously detected viruses in the datasets. For banana samples, all the complete genome sequences of the viruses were downloaded from NCBI nt database on (12/12/2020) to serve as mapping DB. While the BYDV reference (KU170668 – for the alien control) was selected as it was the closest sequence from our isolate. More

information on the composition of each mapping DB is available elsewhere (Rong et al., 2022).

The reads were mapped on the custom DB using Geneious mapper (Prime 2020.0.5, Biomatters). First, the profile parameters "Low sensitivity / Fastest" were selected (with 20% mismatch and a maximum of 3 nucleotides gap allowed). To improve the results by aligning reads to each other in addition to the reference sequence, the fine-tuning for mapping was set to "Iterate 2 times". The "multiple best matches" option was set to "Randomly" (no multiple best matches between two different viruses were observed in any sample processed). In the coming result section, we will refer to these parameters as "relax". A second mapping referred to later on as "strict" was carried out using the same parameters except for the mismatch allowance that was lower than 10%. Only the second mapping was carried out for small RNA (20% mismatch is too much for small RNA). Indeed, using mismatches up to 20% should allow better inclusivity of the analysis by mapping reads from isolates that can be genetically distant from the reference sequences, especially if few reference genomes are available in the literature. Mapping with a strict parameter was done to use small RNA and confirm this hypothesis. The tolerance of mismatches of 20% is also close to many ICTV demarcation criteria to distinguish two different species (although these criteria are often considered for only one or a few genes and might vary between families). Another test with more relaxed parameters would increase the risk of adding non-specific reads (e.i. not generated from the viral genomes) and was not considered.

Deduplication of identical reads between samples

To investigate cross-contamination between samples, additional deduplication of identical reads between samples was performed using dedupe V38.37 (from BBMap) embedded in Geneious (Prime 2020.0.5, Biomatters) and with the parameters kmer seed length, maximum edit, and maximum substitutions set as "31", "0", and "0", respectively. For each virus and sample, the mapped reads from each tested sample and the sample with the highest number of mapped reads in the batch were grouped into a single pool (using "Group sequences into a list" in Geneious) and deduplicated. The deduplication percentage equalled the number of reads removed as duplicates divided by the lower number of reads between the two tested samples. The deduplication percentage was not calculated on samples if less than 5 reads from raw data were mapped to target viruses (as there might be higher random variation with few reads sequenced for a virus). For those samples, the rule (number three) automatically votes contamination. While for the samples with the highest number of reads for a given virus, the deduplication ratio was set as reference (i.e. "RF"), and the vote for rule three is infection.

Confusion matrix and performance criteria calculation

We used a confusion matrix for each batch's results to have standard metrics for comparing batches and samples. We compared the tool prediction for each element to the indexing status of the dataset assimilating infection as a positive result and contamination as a negative result, as explained in **Table 4-2**.

A

Cont-ID confusion matrix		Prediction	
		Infection (Positive)	Contamination (Negative)
Indexing status	Infection	True Positive (TP)	False Negative (FN)
	Contamination	False Positive (FP)	True Negative (TN)

B

Diagnostic sensitivity (DSE)	TP/(TP+FN)
Diagnostic specificity (DSP)	TN/(TN+FP)
False omission rate (FOR)	FN/(FN+TN)
False discovery rate (FDR)	FP/(FP+TP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

Table 4-2: (A) Confusion matrix based on Cont-ID results. (B) The formula is used to calculate accuracy.

Based on the confusion matrix, we have four possibilities after a prediction: False Positive (FP) when the tool wrongly predicted an infection, True Positive (TP) when the tool correctly predicted an infection, False Negative (FN) when the tool wrongly predicted contamination and True Negative (TN) when the tool correctly predicted contamination. In addition, we calculated and focused our analysis on the accuracy as described in table 4-2. To calculate performance criteria automatically (accuracy as well as diagnostic sensitivity/specificity, false omission/discovery rate), we used an automated script available on the same GitHub (https://github.com/johrollin/Cont_ID/tree/master/further_analysis).

Results

We used Cont-ID on ten metagenomic datasets, including a total of 273 samples, as a proof of concept (see details in method). These datasets covered a broad range of use for Cont-ID as they were generated from plant or human samples according to three library preparation protocols (8 for total RNA, 1 for small RNA and 1 for double stranded (ds) RNA). The sequencing reads were always paired 2x150 nt for totalRNA and dsRNA while, for small RNA, the reads corresponded to 1x50 nt (21-24 nt after adapter trimming (5)).

Set up adaptability metrics datasets on the banana datasets

When applying for the first time Cont-ID on banana datasets generated from reference samples with known viral status, the first objective was to determine the most appropriate values for the adaptability metrics (X, Y and Z), allowing to

minimise both FP (over-prediction of infection) and FN (over-prediction of contamination). This was particularly complex as raising the value of an adaptability metric could lead to an over-prediction of either contamination or infection by the rule, while lowering it had the opposite effect.

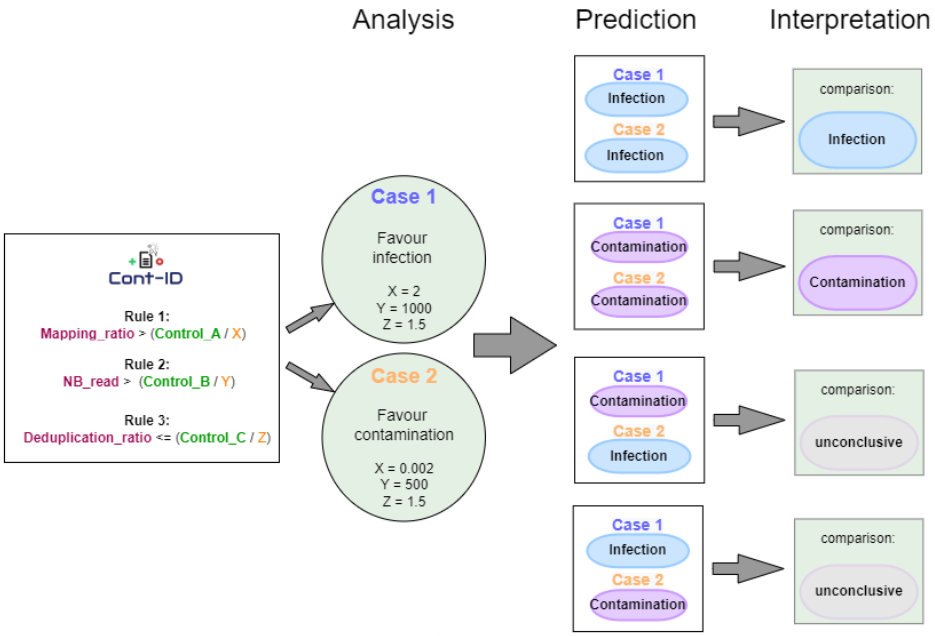


Figure 4-3: Cont-ID prediction when using the two default cases

During the set-up of the method, we looked for the most adapted set of values to balance our rule prediction on Musa datasets A, B and C. We tested several ranges of values aiming at limiting both wrong predictions (FP and FN). The optimised single set of values maintaining FP and FN low in the three datasets was not found. Indeed, variability was observed between batches, as any set limiting FP and FN in one or several batches was not optimal for the other batch(es).

Indeed, the uneven proportion and pattern of cross-contaminations observed in different sequencing batches made it very difficult to decide on a unique set of values. Instead, it seemed more efficient to apply two different sets of values (called "case 1" and "case 2" further on) that favoured the prediction of either true infection (TP - case 1) or true contamination (TN - case 2) from the same datasets. The combination of the prediction from both cases would give additional information for interpretation. We proposed values that gave the best performance criteria on our training datasets on bananas, and the purpose of the diagnostic test was to minimise the risk of false negatives (priority 1) while keeping the confirmation burden manageable (priority 2). Importantly, those sets of values can be manually adapted by the user to improve one or several performance criteria of the test, to better fit the purpose of the HTS tests

carried out and its associated risks (risk of false positive or false negative) or to limit the "grey zone" of inconclusive results (see under).

Therefore, we propose to run Cont-ID with two sets of adaptability metrics every time to compare the results so what a high level of confidence is reached for the elements with identical predictions between both cases. The combination can also highlight elements for which the prediction changed; they correspond to the "grey zone" with metrics of abundance and/or duplication close to thresholds. In such cases, the automated prediction might not be accurate. At this stage, it is mandatory to carry on additional verification, such as checking the confidence (2 or 3 votes) for each prediction or comparing the threshold numbers (also provided in the result) with the sample metrics. Cont-ID provides the list of votes for each rule in each case to facilitate this additional verification. Then according to the additional information and the test's purpose, the user can decide on the status (infection or contamination) or keep it inconclusive but decide to test the virus presence independently by another test. For the presentation of the result, the result is mentioned as "inconclusive" when both cases disagree.

Evaluation of the method accuracy on banana samples

Based on the results obtained with the two sets of adaptability metrics, the tool predictions were compared with the biological status of each reference banana sample (batches A, B, C and D Supplementary File 2), allowing us to predict the cross-contamination on the four tested batches with an average accuracy of 90%, excluding 23% of elements classified as "inconclusive" (see table 4-3A).

		A - Relax mapping				B - Strict mapping			
Batch		A	B	C	D	A	B	C	D
Sequencing method		TotalRNA				TotalRNA			
Total element tested		105	93	143	128	76	65	93	68
Expected Infection/Contamination		28/77	14/79	34/109	19/109	28/48	14/51	34/59	17/51
Case 1	3 votes accuracy	91%	87%	86%	90%	100%	92%	94%	87%
	2 votes accuracy	60%	60%	45%	80%	69%	62%	46%	69%
	overall accuracy	73%	67%	66%	85%	78%	68%	62%	75%
Case 2	3 votes accuracy	99%	95%	100%	96%	95%	92%	100%	94%
	2 votes accuracy	58%	69%	87%	67%	70%	59%	64%	44%
	overall accuracy	85%	85%	93%	91%	83%	78%	80%	82%
Cases combination	accuracy %	88%	82%	94%	95%	82%	78%	81%	90%
	inconclusive %	23%	20%	33%	16%	5%	17%	32%	28%
	correct prediction	71	61	90	103	59	42	51	44
	wrong prediction	10	13	6	5	13	12	12	5
	inconclusive (occurrence)	24	19	47	20	4	11	30	19

Table 4-3: *Percentage of the accuracy of case 1 and case 2 analysed alone or in combination on banana samples sequenced by ribodepleted totalRNA sequencing for each element (*virus detected on each sample). *Each case is presented with the proportion of correct or wrong predictions according to the number of votes obtained (2 or 3) by comparing Cont-ID prediction the the viral status (displayed in expected infection/contamination). The percentage is given by three votes confidence count only the result with three votes while the overall accuracy aggregates the 2 and 3 vote results. When combining results from both cases (described in Figure 4-3), the percentage of inconclusive results and the number*

of correct or wrong predictions are stated. Two different mapping parameters were tested, allowing respectively 20% of mismatches (part A) or 10% of mismatches (part B)

The predictions with the three votes using default mapping parameters ("relax mapping") are very trustworthy as the accuracy is higher than 86% and 95% for cases 1 and 2, respectively. These promising results are obtained on the fraction of the elements representing 25-50% and 48-84% for case 1 and case 2, respectively. The remaining elements are classified with two votes (more information is available in Supplementary File 3). The prediction accuracy with two votes is much lower, whatever the case. So, knowing the number of votes obtained by each element is crucial when the results need to be interpreted (and this number is always given in the report generated by Cont-ID). For case 1, most elements were predicted with two votes meaning that one of the (three) rules had conficting prediction, which might explain why the accuracy was lower. While for case 2, the majority of the elements were predicted with three votes. The explanation is probably in the "expected Infection/contamination" row in **Table 4-3A**: for all batches, there is more contamination than infection (from 28 infections for 77 contaminations – batch A to only 19 infections for 109 contaminations - batch D). As stated above in the text and **Figure 4-3**, Case 2 is designed to favour contamination detection at the expense of infections occurring at a low concentration that tend to be considered contamination (FN). Nevertheless, as a direct consequence, true contamination (TN) detection is high (see confusion_matrix in Supplementary File 3).

Overall, case 2 presented a higher accuracy (85-91% relax mapping) than case 1 (66-85%), while the combination of the two cases reached a similar one (82-95 % relax mapping). Those good results from combination accuracy mean that very few predictions are wrong (5 – 13) in both cases, but 16-33% of elements are not counted in the accuracy percentage because they are inconclusive. The combination's importance relies on maintaining a high accuracy while highlighting the inconclusive prediction to prioritise them for manual expertise.

The mapping parameters impacted the input files and the Cont-ID performance

In **Table 4-3**, we explored the impact on the prediction of two levels of mismatch tolerance (20% and 10%) when mapping the sequencing reads on the viral genome DB. The goal was to explore if changing a parameter from the primary bioinformatics step delivering the input files of Cont-ID could have an impact on the prediction. Strict mapping tends to lower the total number of elements tested due to a decrease in the number of samples for which we have very few reads mapped to a candidate virus (samples for which the number of reads was already very low with relaxed mapping). Cont-ID has more samples to process with a relaxed mapping, which should be better for threshold calculation. Logically, the elements lost by the strict mapping parameter should be predicted as "contamination" and present a relatively low number of reads. Indeed, those elements are most likely more distant reads (between 20 and 10 %

mismatch with the reference genome) mapped on the virus. They could correspond to non-viral reads wrongly mapped in datasets from samples tested negative by classical indexing. For example, for batch B, on the 28 differential mapping results, the number of reads mapped ranged from 1 to 10 (except for two). Of those 28 elements, 24 are classified as "contamination", 2 as infection (the two that have more than 10 reads), the two remaining are labelled inconclusive. In batch C, there are 50 differential elements between the two parameters, with 37 correct (13 from BSV), 11 inconclusive (10 from BSV) and 2 wrong (2 from BSV) classified elements. In total, 25 elements (on the 50 - batch C) are from the non-integrated virus, of which 24 are labelled contamination (one inconclusive). The separation between integrated and non-integrated viruses is explained in **Table 4-4** and another publication (Rong et al., 2022).

Using the relaxed mapping parameter seems beneficial for prediction as the accuracy is better (82-95% relax, 78-90% strict). Moreover, thanks to the combination strategy, we can focus on the proportion of inconclusive; it is uneven with an important increase, 5% (strict) to 23% (relax) for batch A, while in batch D, it decreases from 28% to 16 %. However, when we look closely at the accuracy improvement, most comes from differential elements (present only with relax) that are 'obvious' contamination with few reads. So, most of the accuracy improvement did not come from very informative elements, except in some rare occurrences where it helped classify well elements in relax parameters that were inconclusive with strict or classified inconclusive elements in relax that were wrong with strict parameters. As an example, in batch C, on the 24 elements for BanMMV, BBRMV, BBTV and CMV common in both conditions (relax and strict), elements prediction is improved (from inconclusive [strict] to correct [relax]) for three of them (sample 3B1, 3B2 and 3B14 with BanMMV).

There is, therefore, a slight improvement with relaxed mapping parameters, and we set these parameters by default to generate the input files. Indeed, with the relaxed parameters, the number of reads for each element (including alien) increases along the rise of the number of elements in the batch. This means that we change the rule's value of the threshold (see Figure 4-2), in a way that seems more representative of reality than strict mapping. In these batches, some element metrics are very close to the value sused by the rules and slightly changing those metrics or the alien metrics (the alien control metrics are obviously changed by the mapping parameters) can modify the prediction.

As we did not know the divergence of the virus genomes between different samples and the reference genomes, it seemed more logical to use relaxed mapping parameters by default. According to the virus system the user is working on and the ICTV demarcation criteria that go with it, these parameters should or could be adapted.

The virus biology can impact Cont-ID performance: the case of virus integration in the host genome

			A - Non-integrated Virus				B - Integrated Virus (BSV)			
Batch			A	B	C	D	A	B	C	D
Sequencing method			TotalRNA				TotalRNA			
Relax mapping	Total element tested		40	31	51	55	65	62	92	73
	Expected Infection/Contamination		19/21	9/22	22/29	10/45	9/56	5/57	12/80	9/64
	Case 1	3 votes accuracy	100%	91%	96%	100%	82%	83%	68%	54%
		2 votes accuracy	94%	70%	75%	100%	47%	56%	43%	78%
		overall accuracy	98%	77%	94%	100%	58%	61%	50%	74%
	Case 2	3 votes accuracy	100%	85%	100%	100%	97%	100%	100%	93%
		2 votes accuracy	50%	91%	82%	67%	62%	58%	89%	67%
		overall accuracy	88%	87%	92%	96%	83%	84%	93%	88%
	Cases combination	accuracy %	100%	88%	96%	100%	79%	79%	92%	91%
		inconclusive %	15%	16%	6%	4%	28%	23%	48%	25%
		correct prediction	34	23	46	53	37	38	44	50
		wrong prediction	0	3	2	0	10	10	4	5
		inconclusive (occurrence)	6	5	3	2	18	14	44	18

Table 4-4: *Percentage of the accuracy of case 1 and case 2 analysed alone or in combination on banana samples sequenced by ribodepleted totalRNA sequencing with relaxed mapping parameters for each element (virus detected on each sample). Each case is presented with the proportion of correct or wrong predictions according to the number of votes obtained (2 or 3) by comparing Cont-ID prediction the the viral status (displayed in expected infection/contamination). The percentage given by three votes confidence count only the result with three votes while the overall accuracy aggregates the 2 and 3 vote results. When combining results from both cases (described in Figure 4-3), the percentage of inconclusive results and the number of correct or wrong predictions are stated. Two types of viruses were tested, Non-integrated virus (part A) or integrated virus (part B).*

To highlight the potential impact of the virus biology on the results of Cont-ID, the analysis of banana batches was split between integrated and non-integrated viruses. Indeed, several species of BSV are integrated into its host genome, which complicates the reliable detection of BSV infection from sequencing datasets. Consequently, it has been recommended to confirm any detection of BSV reads by an independent PCR test combined with immunocapture of viral particles (Rong et al., 2022).

Table 4-4 shows better accuracy and a lower proportion of inconclusive results for non-integrated viruses compared to BSV. More elements with contamination status are obtained when looking for BSV than non-integrated viral species. This over-representation of contaminants might be caused by the transcription of integrated sequences of BSV even without viral particles, which will raise the number of detected reads. These are two points that reduced the efficiency of our method on BSV and, by extension, might also concern any other viral species integrated into the host genome and able to produce transcripts.

The global accuracy is lower for BSV species (79-92%) compared to the other viruses (88-100%), even if the maximum accuracy obtained with batch C (92%) was high. In addition, the proportion of inconclusive results should also be considered, and this proportion was much higher for BSV (23-45%) than for the other viruses (4-16%). So, the overall performance of Cont-ID is lower when applied on BSV and did not solve the issues of appropriate detection in sequencing data of viral infection from viruses integrated into the plant genome. Consequently, BSGFV, BSIMV, BSMYV, and BSOLV, which correspond to different but closely related species of Banana streak virus (BSV) integrated into the *Musa* genome, were excluded from the

calculation of performance criteria for the banana datasets. BSCAV was also excluded (despite not being integrated) because of its similarity with other BSV species.

Performance of Cont-ID on diverse datasets

The performance of Cont-ID using the two cases was further evaluated while diversifying the hosts (fruit trees, grasses, humans) and the sequencing protocols (total RNA, small RNA, dsRNA).

Batch		A	B	C	D	E	F	G	H	I	J	
Origin		Banana (own sequencing)						Plant Mix (Gauthier et al., 2022)	Human (Bal et al 2018)	Human (Boheemen et al 2019)	Human (Li et al 2020)	
Sequencing method		TotalRNA				SmallRNA	dsRNA	TotalRNA				Average
Total element tested		40	31	51	55	20	12	51	112	62	206	64
Expected Infection/Contamination		19/21	9/22	22/29	10/45	19/1	5/7	18/33	37/75	25/37	50/156	
Case 1	3 votes accuracy	100%	91%	96%	100%	100%	100%	100%	96%	100%	90%	97%
	2 votes accuracy	94%	70%	75%	100%	36%	67%	86%	71%	54%	63%	72%
	overall accuracy	98%	77%	94%	100%	55%	92%	92%	87%	73%	77%	84%
Case 2	3 votes accuracy	100%	85%	100%	100%	33%	100%	97%	92%	92%	95%	89%
	2 votes accuracy	50%	91%	82%	67%	63%	50%	92%	68%	84%	78%	72%
	overall accuracy	88%	87%	92%	96%	45%	92%	96%	86%	89%	87%	86%
Cases combination	accuracy	100%	88%	96%	100%	50%	100%	98%	93%	93%	92%	91%
	inconclusive %	15%	16%	6%	4%	20%	17%	8%	15%	29%	24%	15%
	correct prediction	34	23	46	53	8	10	46	88	41	144	49,3
	wrong prediction	0	3	2	0	8	0	1	7	3	12	3,6
	inconclusive (occurrence)	6	5	3	2	4	2	4	17	18	50	14,4

Table 4-5: *Percentage of the accuracy of case 1 and case 2 analysed alone or in combination from sequencing with relaxed mapping parameters (except for small RNA) for each element (virus detected on each sample). Each case is presented with the proportion of correct or wrong predictions according to the number of votes obtained (2 or 3) by comparing Cont-ID prediction the the viral status (displayed in expected infection/contamination). The percentage given by three votes confidence count only the result with three votes while the overall accuracy aggregates the 2 and 3 vote results. When combining results from both cases (described in Figure 4-3), the percentage of inconclusive results and the number of correct or wrong predictions are stated. Several virus datasets were tested, banana samples (only non-BSV viruses are considered), a mix of plants, and human datasets.*

Table 4-5 shows the method's accuracy on all datasets with relaxed mapping parameters (except for small RNA, see method). Overall, the accuracy of Cont-ID was 94%, with 15% of inconclusive results. The sRNA dataset provides a poor accuracy (50%) with 20% inconclusive; this can be explained by the (almost) absence of contamination (Expected Infection/Contamination 19/1) by the low level of reads found (see Supplementary Files 2 & 3 for more information). Apart from small RNA, the worst accuracy (88%) has been obtained from the batch B sequencing dataset of banana. Noteworthy, this protocol was independently evaluated for virus testing in banana, but its performance for virus detection was much lower than total RNA sequencing (Rong et al., 2022). The accuracy calculated from the single batch of dsRNA, with only 9 samples and 12 elements, was 100%. Even if not enough

representative dataset was used for dsRNA, the method accuracy seems not too far from what we obtained in Total RNA, indicating that, Cont-ID is independent of the extraction method. On Total RNA, for banana samples, the accuracy ranged from 88% to 100%, with 4% to 15% of inconclusive results. The accuracy of the plant mix (G) was also very high (98%), with 8% of inconclusive results. On human datasets, the accuracy remained high (92-93%), but the inconclusive results reached up to 15 - 29%. Overall, the application of Cont-ID on human datasets reached similar performance in accuracy; the slightly worse inconclusive metrics can be explained by the fact that the adaptability metrics might not be the best ones for the human dataset and underlined the importance

to adapt metrics and parameters of Cont-ID to specific dataset types.

For most of the datasets, case 1 performed worse than case 2, probably due to the design of the case metrics (see **Figure 4-2**), where case 1 values were determined to favour infection. The expected infection/contamination ratio showed that for all the datasets but E (small RNA), there was a lot more contamination than infection; therefore, case 1 overpredicted infection, lowering its accuracy. In the E dataset, case 1 (55%) performed better than case 2 (45%) as expected; it is also the case for the human dataset H (97% case 1, 96% case 2), even if the ratio (37/75) leans toward contamination.

Those results indicated that Cont-ID performed well in classifying cross-contamination in very different virus-host systems, even if some adjustments may be needed in some cases in the future. The different levels of flexibility of Cont-ID made such adjustments possible. To provide an example of analysis, all the information regarding batch C from the input file to the analysis file (including raw results) is available in Supplementary File 3.

Discussion

Despite significant efforts to limit cross-contamination (dual indexes, inter-run washing …), this still represent a concern and the appropriate distinction between low-level infection, and cross-sample contamination is crucial for the large-scale development of HTS technologies as a diagnostic test. Furthermore, it should be adequately managed because identifying and monitoring the cross-contaminations improves the detection results' reliability. In other words, it can help to find the source of contamination in the laboratory, take appropriate measures to minimise it, and raise confidence in the detected viruses.

This publication improved a preliminary work on determining an adaptative contamination threshold for the detection of plant viruses (Rong et al., 2022) which uses the maximal number of alien virus reads contaminating a sample as the threshold of detection for each sequencing batch. So, instead of using a fixed number for the contamination threshold as done in the literature, the threshold is adapted to the level of contamination monitored in the batch thanks to the alien control. The former publication used a single threshold corresponding to the maximum number of alien

virus reads in a sample. Some limitations of this previous threshold, for example, overestimating contamination when viral reads are in low number for a virus, underlined the need for improvements. This was achieved with Cont-ID through the definition of multiple formal rules, the automation of calculation and the ability to adapt the thresholds and rules by the user. The tool's prediction relies on basic and usual information generated by bioinformatics analysis of sequencing data (mapping and duplication numbers) and the use of external alien control. The criteria based on reads (relative) abundance of each virus in each sample and the (approximation of) number of identical reads for a virus between samples performed well while being relatively easy to generate. Our objective with this tool was to show that exploring data generated by standard bioinformatic procedures can facilitate the identification of cross-contamination between samples.

Cont-ID discriminated virus infection and cross-contamination between samples within a sequencing batch with a global accuracy of 91 % (median=95%) on the diverse range of datasets included in its evaluation. The diversity of situations included viral species belonging to diverse viral families with cellular hosts belonging to plant or animal kingdoms and three different library preparation protocols. Importantly, the default values of adaptability metrics determined from banana dataset predicted cross-contamination with high accuracy (96%, on banana excluding small RNA) and remain high even on human datasets (94%). Noteworthy, these additional datasets were carefully selected to check that they fit the Cont-ID requirements (alien control and samples processed in parallel). To further help the user in the analysis, we provide the detailed votes prediction in the result file (see Supplementary File 3). This is, therefore, a solid basis for the diagnostician to check the level of confidence in the generated results. Indeed, each prediction made by the method uses at least two rules to determine the classification of the element for each case. A prediction with three votes is more confident than with two votes. But all predictions with two votes do not provide the same confidence as it depends on which rules predicted what. Of our three rules, two rely more or less directly on abundance estimation, which means that when that metric is not obtainable in a reliable way, the tools' predictions will be impacted, and predictions with those rules might be less confident. On the other hand, rule three (deduplication ratio) is less effective when the read numbers are low. Depending on the scenario, users should consider the relative confidence of each rule when trying to confirm Cont-ID prediction. This underlines again the importance of proper interpretation of the obtained results based on the virus biology

The prediction quality is deeply impacted by the input data quality, meaning that the deduplication and mapping parameters are essential and should be carefully considered while evaluating their impact on the results. For example, some deduplication tools remove reads if a (small) read is contained in another (larger) read; having that option active or not will significantly impact the deduplication ratio. For example, the inclusion of PCR amplification step in the library preparation protocol, the very high abundance of viral genomes in the sample or the low

complexity of the library can impact this deduplication ratio and should be taken into account when setting up the parameters for Cont-ID application.

As shown in the results, mismatch parameters are very impactful for the mapping. Considerations like ICTV demarcation criteria or what parameters the biologist would use to reconstruct the whole viral genome are helpful in deciding the ones to use for Cont-ID input. In that regard, testing and expertise in bioinformatics analysis are heavily beneficial. Here, the 20% mismatch parameter performed well; it might be different in other datasets (viral composition) configurations or when working with databases containing many reference genomes. Indeed, independently of the mismatch parameters used, using more genome references for each expected species could also improve the ability to detect sequences from distant isolates by better covering the genetic diversity of the virus.

The biology of the virus should also be considered, as shown by the results obtained with viruses with functionally integrated genomes in the host, like BSV species. Our conclusion is that they should be considered independently from the non-integrated viruses. It was challenging to extract a reliable metric for BSV as the differentiation between reads from integrated genomes and reads from viral particles is impossible. Indeed, the biology of viruses integrated into the host genome differs from non-integrated viruses, as viral genome transcription can happen without viral particle production. We have not tested our method on species with different biological behaviour like viroids or phages. But optimisation of the adaptative metrics might likely be required in order to use Cont-ID with high accuracy. Viroid genomes are generally smaller than viruses, while the phage genomes tend to be much larger and has specific biological features. For example, a different level of identical reads and abundance (calculation based on reads number) could be obtained between the different scales of genome size.

For these reasons, Cont-ID allows the evaluation of other values for adaptability metrics (X, Y, Z) by each user to adapt the tool and optimise its diagnostic performance depending on the biological matrix, the protocol and the purpose of the test. Independently, the user can also adapt the metrics to reach the appropriate balance between FN and FP by deciding if, for the purpose of the test and the available resource for confirming detection, it is preferable to be overpredicting contamination to be confident that all the virus detection remaining are true infection or the opposite (overpredicting infection to be sure not to miss any).

In our tests, the analysis of the wrong predictions showed that none of the proposed rules (and adaptability metrics values) allowed us to reach satisfactory accuracy with a proper balance between FN and FP (see Supplementary File 1). We have observed that using two sets of adaptability metrics (one to favour contamination and the other, infection prediction) gave a higher accuracy. In a real scenario (with infection status not known for the samples), it is difficult to know if HTS virus detection (at low concentration) is in the majority due to true infection or cross-contamination. The two-case strategy allows the biologist to predict both scenarios with at least one case accurately. Indeed, if the expected ratio of infection/contamination is unknown, the

relative performance of cases 1 or 2 will be unknown, so it seems preferable to use the combination results instead of the individual.

If both cases agree, the assumption is that the prediction is correct. Nevertheless, combining the results will provide a list of interesting inconclusive results. Each inconclusive result means that the two cases delivered opposite predictions. Therefore, the scientist should address those results when analysing Cont-ID prediction by checking the number of rules for each prediction, for example, knowing if 2 or 3 rules agreed and checking the results of each rule : How close to the threshold was the read abundance/ratio and/or the duplication rate? Spotting the few errors that may occur requires excellent manual expertise as the usual manual verification methods may also indicate the wrong decision (if there are many reads from cross-contamination, the mapping results can be wrongly positive and/or the (RT)-PCR can also be wrongly positive if the contamination occurred at an early stage and the (RT)-PCR was carried out on the same nucleic acids extract). Other information about the virus-plant interaction should be considered, like virus-species-cultivar compatibility or geographical virus distribution (see investigation on unexpected viruses (Rong et al., 2022)).

Cont-ID also presents some limitations that need to be discussed. First, the number of identical reads estimation comes from the deduplication procedure, which is an approximation, and that can be a problem because it can consider the non-specific reads (reads that are not coming from cross-contamination but that are identical to another sample from a common area of the genome) as identical to the probable source of contamination by mistake. Indeed, this can be the case if, for example, two samples are infected by the same virus isolate at very different concentrations. The presence of duplicated reads might suggest contamination instead of a low-level infection. The risk of such an extreme situation is limited using two other rules, although interpreting the data will require good expertise in virus genomic variability and detailed information on the sample origins and virus prevalence and diversity. External source of contamination (for example, foreign samples in sequencing facilities) are not considered by Cont-ID and therefore can also be source of error if they involve the same virus specie as the one of interest.

Also, identification up to species subtype (like expected for Influenza A virus) was not considered during testing phase. When considered relevant, such subspecies distinction could be used with care, ensuring the unequivocal assignation of the reads to its corresponding subspecies. In other words, the genomes of the subspecies need to be different enough and the mapping parameters adapted to avoid unspecific mapping of reads between subtypes. In such case, thresholds used should be monitored carefully to unsure integrity and reliability of the analysis.In addition, the duplication metric assumes that contamination (if any) comes from the sample with the highest number of reads. This theory seems logical since the more reads in a sample, the higher the probability of detecting a few reads from it contaminating other

samples (potential of contamination). Nevertheless, it can create a bias when a virus is highly abundant in two (or more) samples and detected with a low frequency in others. In that case, it is difficult to determine the true origin of cross-contamination. Such a case could be a fundamental limit of our current method. If several samples with a very high abundance of reads are present in a batch, as developed here, Cont-ID should be applied as many times as the number of highly abundant samples. Ideally, Cont-ID should include the read duplication comparison of each sample to all other samples for a virus, but this can raise additional issues (like contamination from several origins at the same time), and, at this stage, it was not implemented.

We must also keep in mind that the relative quantity of genetic material between samples might change because the biologist normalises the quantity of DNA/RNA at two steps of the process: before starting library preparation and during the pooling of the prepared libraries, meaning that the differential in genomic material concentration (potential of contamination of a sample) is resettled. If cross-contamination happens before that step, it can lowers Cont-ID predictions accuracy. This bias in the estimation of abundance is another limitation of our method.

Using an (alien) control helps to know the expected level of contamination but is also impacted by the limit of detection inherent to the standard bioinformatic procedures. Indeed, working with very few reads for some viruses makes some analyses impossible when below their detection limit. For example, the calculation of the duplication rate below a minimal number of reads (in this study, we chose 5) of a virus did not make sense. The limit of calculation of the input metrics is another limitation of Cont-ID.

Cont-ID accuracy was high, but additional improvements can probably be explored. For example, by exploiting the ability of other metrics generated during bio-informatic analyses (like RKPM, genome coverage percentage, relative coverage depth repartition, …) to help detect contamination. In fact, some of these metrics with several thresholds were tested for Cont-ID before selecting the three rules described in **Figure 4-2** that provided the highest accuracy (in both contamination and infection determination). Importantly, values leading to a perfect scenario were not identified, and a two-cases classification system was set up (more information in Supplementary File 1).

Nevertheless, adding more metrics will also complexify the decision system. If more metrics are considered for cross-contamination prediction, other implementations (decision tree, machine learning …) might be envisioned to replace the current voting system. On the other hand, the detection in the alien control of sequencing reads of other viruses detected in the tested samples is also the consequence of contamination from one of the tested samples toward the alien control. This information is not used now but could also be considered for future improvements as it requires less complex modifications to implement. In addition, it might allow refinement of Cont-ID, potentially introducing an adaptation of threshold per virus instead of a single threshold for all samples from the sequencing batch. The idea is that two viruses present in the same batch may have different relative abundance behaviour in the

samples, so setting up a limit that can adapt for each virus should improve the tool's ability to distinguish real infection from cross-contamination. Finally, working with the combination of all/some viruses profile instead of each individually for contamination check (similarly to what is used in metabarcoding of bacteria) can also be considered. Indeed, when a sample contaminates another, it is expected that all the viruses (highly frequent) from the contaminating sample can be found in the contaminated samples. Monitoring the virus detection profile of samples can provide additional information for cross-contamination (and ease the quest for contamination origin). Even if there is still improvement to be made, Cont-ID has already delivered an excellent ability to consider the level of contamination genuinely present in a batch.

In conclusion, detection of cross-contamination is complex; in the age of sequencing, the issue of

cross-contamination across samples is increasingly important; therefore, Cont-ID will facilitate the interpretation of results by the virologist/diagnostician and reduces the confirmation burden. We demonstrated that simple metrics like relative abundance estimation and redundancies of genetic material (reads duplicates) could help monitor contamination occurring in the laboratory. The method accurately distinguished cross-contamination from infection in very diverse HTS viral datasets generated by short reads Illumina technology. Our standard parameters allowed very good accuracy (median = 95%); in addition, Cont-ID has several levels of flexibility and can be adapted by each user to take into account the specificities of the detection test (purpose of the test, type of samples, viruses to be detected, laboratory work, available resources….). We believe this is a first significant step toward increasing the monitoring and management of sample cross-contamination when using HTS technologies for virus detection.

Availability and requirements:

Project name: Cont-ID

Project home page: https://github.com/johrollin/Cont_ID

Operating system(s): Platform independent

Programming language: Python (v3.7)

Other requirements: pandas; NumPy

License: GNU GPL-3.0

Any restrictions to use by non-academics: none

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the NCBI Sequence Read Archive (SRA) repository at https://www.ncbi.nlm.nih.gov/sra (See Table 4-1). In addition, cont-ID is freely available and can be downloaded with the following command (without <>): <https://github.com/johrollin/Cont_ID>. It can be used as a command line application on a personal computer on any operating system (Linux, MacOSX or Windows) with python.

All supplementary material are available here:

Supplementary_4

All scripts are available here:

GitHub > johrollin > Cont_ID

Competing interests
The authors declare that they have no competing interests.

Funding
The work has been supported by (1) the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 813542T (INEXTVIR), and (2) the NGS cross-centre project from the CGIAR Fund and in particular by the Germplasm Health Unit (GHU) of the CGIAR Genebank Platform

Authors' contributions

WR and SM designed the data preparation and sequencing procedure. WR, JR and SM designed the bioinformatics analysis. JR implemented the program. JR ran CroCo analyses. WR and JR validated Cont-ID results with previous PCR indexing results and re-analysed outlier. JR and SM drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

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), Mathieu Chabannes and Marilyne Caruana (CIRAD, Montpellier, France) are also acknowledged for kindly providing Musa reference samples. Special thanks to Marie-Emilie Gauthier and Roberto Barrero for providing dataset G and discussing cross-contamination in viral metagenomes with us.

Reference

Adams, I., Fox, A., 2016. Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. Current Research Topics in Plant Virology 323–335. https://doi.org/10.1007/978-3-319-32919-2_14/FIGURES/3

Adams, I.P., Fox, A., Boonham, N., Massart, S., de Jonghe, K., Tahzima, R., Foucart, Y., Peusens, G., Beliën, T., Massart, S., de Jonghe, K., 2018. The impact of high throughput sequencing on plant health diagnostics. Eur J Plant Pathol 152, 909–919. https://doi.org/10.1007/s10658-018-1570-0

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10, 537. https://doi.org/10.1111/J.1364-3703.2009.00545.X

Adams, I.P., Skelton, A., Macarthur, R., Hodges, T., Hinds, H., Flint, L., Nath, P.D., Boonham, N., Fox, A., 2014. Carrot yellow leaf virus is associated with carrot internal necrosis 9. https://doi.org/10.1371/journal.pone.0109125

Aguado-García, Y., Taboada, B., Morán, P., Rivera-Gutiérrez, X., Serrano-Vázquez, A., Iša, P., Rojas-Velázquez, L., Pérez-Juárez, H., López, S., Torres, J., Ximénez, C., Arias, C.F., 2020. Tobamoviruses can be frequently present in the oropharynx and gut of infants during their first year of life. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-70684-w

Agüero, J., Gómez-Aix, C., Sempere, R.N., García-Villalba, J., García-Núñez, J., Hernando, Y., Aranda, M.A., 2018. Stable and broad spectrum cross-protection against pepino mosaic virus attained by mixed infection. Front Plant Sci 871, 1810. https://doi.org/10.3389/FPLS.2018.01810/BIBTEX

Al Rwahnih, M., Daubert, S., Golino, D., Islas, C., Rowhani, A., 2015. Comparison of Next-Generation Sequencing Versus Biological Indexing for the Optimal Detection of Viral Pathogens in Grapevine. Phytopathology 105, 758–763. https://doi.org/10.1094/PHYTO-06-14-0165-R

Anastassiadou, M., Arena, M., Auteri, D., Brancato, A., Bura, L., Carrasco Cabrera, L., Chaideftou, E., Chiusolo, A., Crivellente, F., de Lentdecker, C., Egsmose, M., Fait, G., Greco, L., Ippolito, A., Istace, F., Jarrah, S., Kardassi, D., Leuschner, R., Lostia, A., Lythgo, C., Magrans, O., Mangas, I., Miron, I., Molnar, T., Padovani, L., Parra Morte, J.M., Pedersen, R., Reich, H., Santos, M., Sharp, R., Szentes, C., Terron, A., Tiramani, M., Vagenende, B., Villamar-Bouza, L., 2021. Peer review of the pesticide risk assessment of the active substances Pepino Mosaic Virus,EU strain, mild isolate Abp1 and Pepino Mosaic Virus,CH2 strain, mild isolate Abp2. EFSA Journal 19, e06388. https://doi.org/10.2903/J.EFSA.2021.6388

Babaian, A., Edgar, R., 2022. Ribovirus classification by a polymerase barcode sequence. PeerJ 10, e14055. https://doi.org/10.7717/PEERJ.14055/SUPP-2

Bačnik, K., Kutnjak, D., Pecman, A., Mehle, N., Tušek Žnidarič, M., Gutiérrez Aguirre, I., Ravnikar, M., 2020. Viromics and infectivity analysis reveal the release of infective plant viruses from wastewater into the environment. Water Res 177, 115628. https://doi.org/10.1016/J.WATRES.2020.115628

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A. v., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Christopher Garcia, K., Grishin, N. v., Adams, P.D., Read, R.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) 373, 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Baert, L., McClure, P., Winkler, A., Karn, J., Bouwknegt, M., Klijn, A., 2021. Guidance document on the use of whole genome sequencing (WGS) for source tracking from a food industry perspective. Food Control 130, 108148. https://doi.org/10.1016/J.FOODCONT.2021.108148

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. https://doi.org/10.1099/JGV.0.001210/CITE/REFWORKS

Bal, A., Pichon, M., Picard, C., Casalegno, J.S., Valette, M., Schuffenecker, I., Billard, L., Vallet, S., Vilchez, G., Cheynet, V., Oriol, G., Trouillet-Assant, S., Gillet, Y., Lina, B., Brengel-Pesce, K., Morfin, F., Josset, L., 2018. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infectious Diseases 2018 18:1 18, 1–10. https://doi.org/10.1186/S12879-018-3446-5

Ballenghien, M., Faivre, N., Galtier, N., 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol 15. https://doi.org/10.1186/S12915-017-0366-6

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. v, Sirotkin, A. v, Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. Viruses 2014, Vol. 6, Pages 106-136 6, 106–136. https://doi.org/10.3390/V6010106

Barbitoff, Y.A., Abasov, R., Tvorogova, V.E., Glotov, A.S., Predeus, A. v., 2022. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics 23, 1–17. https://doi.org/10.1186/S12864-022-08365-3/

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., de Bruyn, R., Maree, H.J., 2021a. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., De Bruyn, R., Maree, H.J., 2021b. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

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., Arboleda, V.A., 2021a. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

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., Arboleda, V.A., 2021b. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

Boheemen, S. van, Rijn, A.L. van, Pappas, N., Carbo, E.C., Vorderman, R.H.P., Sidorov, I., 2020. Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- research that is available on the COVID-19 resource centre - including this with acknowledgement of the origin.

Bomba, L., Walter, K., Soranzo, N., 2017. The impact of rare and low-frequency genetic variants in common disease. Genome Biology 2017 18:1 18, 1–17. https://doi.org/10.1186/S13059-017-1212-4

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., Mumford, R., 2014. Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res 186, 20–31. https://doi.org/10.1016/j.virusres.2013.12.007

Bordería, A. v., Isakov, O., Moratorio, G., Henningsson, R., Agüera-González, S., Organtini, L., Gnädig, N.F., Blanc, H., Alcover, A., Hafenstein, S., Fontes, M., Shomron, N., Vignuzzi, M., 2015. Group

Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. PLoS Pathog 11, e1004838. https://doi.org/10.1371/JOURNAL.PPAT.1004838

Brinkmann, A., Andrusch, A., Belka, A., Wylezich, C., Höper, D., Pohlmann, A., Petersen, T.N., Lucas, P., Blanchard, Y., Papa, A., Melidou, A., Oude Munnink, B.B., Matthijnssens, J., Deboutte, W., Ellis, R.J., Hansmann, F., Baumgärtner, W., van der Vries, E., Osterhaus, A., Camma, C., Mangone, I., Lorusso, A., Marcacci, M., Nunes, A., Pinto, M., Borges, V., Kroneman, A., Schmitz, D., Corman, V.M., Drosten, C., Jones, T.C., Hendriksen, R.S., Aarestrup, F.M., Koopmans, M., Beer, M., Nitsche, A., 2019. Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. J Clin Microbiol 57. https://doi.org/10.1128/JCM.00466-19

Bukowska-Ośko, I., Perlejewski, K., Nakamura, S., Motooka, D., Stokowy, T., Kosińska, J., Popiel, M., Płoski, R., Horban, A., Lipowski, D., Caraballo Cortés, K., Pawełczyk, A., Demkow, U., Stępień, A., Radkowski, M., Laskus, T., 2017. Sensitivity of next-generation sequencing metagenomic analysis for detection of RNA and DNA viruses in cerebrospinal fluid: The confounding effect of background contamination. Adv Exp Med Biol 944, 53–62. https://doi.org/10.1007/5584_2016_42

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8, 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Bushnell, B., Rood, J., Singer, E., 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One 12. https://doi.org/10.1371/JOURNAL.PONE.0185056

Calil, I.P., Fontes, E.P.B., 2017. Plant immunity against viruses: antiviral immune receptors in focus. Ann Bot 119, 711–723. https://doi.org/10.1093/AOB/MCW200

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10. https://doi.org/10.1186/1471-2105-10-421

Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K.A., Giovani, B., Saunders, D.G.O., Kamoun, S., Legg, J.P., Verdier, V., Lessel, J., Neher, R.A., Day, R., Pardey, P., Gullino, M.L., Records, A.R., Bextine, B., Leach, J.E., Staiger, S., Tohme, J., 2019. A global surveillance system for crop diseases: Global preparedness minimizes the risk to food supplies. Science (1979) 364, 1237–1239. https://doi.org/10.1126/SCIENCE.AAW1572/ASSET/52AFE309-57C8-475B-88C6-17D489A74CEE/ASSETS/GRAPHIC/364_1237_F2.JPEG

Černi, S., Ruščić, J., Nolasco, G., Gatin, Ž., Krajačić, M., Škorić, D., 2008. Stem pitting and seedling yellows symptoms of Citrus tristeza virus infection may be determined by minor sequence variants. Virus Genes 36, 241–249. https://doi.org/10.1007/S11262-007-0183-Z/TABLES/1

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021a. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021b. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Champlot, S., Berthelot, C., Pruvost, M., Andrew Bennett, E., Grange, T., Geigl, E.M., 2010. An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. PLoS One 5, e13042. https://doi.org/10.1371/JOURNAL.PONE.0013042

Charlebois, R.L., Sathiamoorthy, S., Logvinoff, C., Gisonni-Lex, L., Mallet, L., Ng, S.H.S., 2020. Sensitivity and breadth of detection of high-throughput sequencing for adventitious virus detection. npj Vaccines 2020 5:1 5, 1–8. https://doi.org/10.1038/s41541-020-0207-4

Chodasewicz, K., 2014. Evolution, reproduction and definition of life. Theory in Biosciences 133, 39. https://doi.org/10.1007/S12064-013-0184-5

Choua, M., Heath, M.R., Speirs, D.C., Bonachela, J.A., 2020. The effect of viral plasticity on the persistence of host-virus systems. J Theor Biol 498, 110263. https://doi.org/10.1016/J.JTBI.2020.110263

Clark, M.F., Adams, A.N., 1977. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. J Gen Virol 34, 475–483. https://doi.org/10.1099/0022-1317-34-3-475

Cleland, C.E., Chyba, C.F., 2002. Defining "life." Orig Life Evol Biosph 32, 387–393. https://doi.org/10.1023/A:1020503324273

Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P., Jackson, S.A., 2015. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant 8, 831–846. https://doi.org/10.1016/J.MOLP.2015.02.002

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., Gabriel, S., 2018a. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 332. https://doi.org/10.1186/s12864-018-4703-0

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., Gabriel, S., 2018b. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 1–10. https://doi.org/10.1186/S12864-018-4703-0/TABLES/2

Cunniffe, N.J., Laranjeira, F.F., Neri, F.M., DeSimone, R.E., Gilligan, C.A., 2014. Cost-Effective Control of Plant Disease When Epidemiological Knowledge Is Incomplete: Modelling Bahia Bark Scaling of Citrus. PLoS Comput Biol 10, e1003753. https://doi.org/10.1371/JOURNAL.PCBI.1003753

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Daubin, V., Szöllősi, G.J., 2016. Horizontal Gene Transfer and the History of Life. Cold Spring Harb Perspect Biol 8. https://doi.org/10.1101/CSHPERSPECT.A018036

De Clerck, C., Crew, K., Van den houwe, I., McMichael, L., Berhal, C., Lassois, L., Haissam Jijakli, M., Roux, N., Thomas, J., Massart, S., 2017. Lessons learned from the virus indexing of Musa germplasm: insights from a multiyear collaboration. Annals of Applied Biology 171, 15–27. https://doi.org/10.1111/AAB.12353

de Jonghe, K., 2020. The world of plant viruses: the good, the bad and the beautiful. Floraculture International 34–36.

Deng, Z.L., Dhingra, A., Fritz, A., Götting, J., Münch, P.C., Steinbrück, L., Schulz, T.F., Ganzenmüller, T., McHardy, A.C., 2021. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. Brief Bioinform 22, 1–12. https://doi.org/10.1093/BIB/BBAA123

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022a. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022b. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Dimmock, N.J., Easton, A.J., Leppard, K.N., 2016. Introduction to Modern Virology, 7th Edition | Wiley.

Domingo, E., Perales, C., 2019. Viral quasispecies. PLoS Genet 15, e1008271. https://doi.org/10.1371/JOURNAL.PGEN.1008271

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J.F., de la Peña, M., Korobeynikov, A., Chikhi, R., Babaian, A., 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 2022 602:7895 602, 142–147. https://doi.org/10.1038/s41586-021-04332-2

Fennell, M., Murphy, J.E., Armstrong, C., Gallagher, T., Osborne, B., 2012. Plant Spread Simulator: A model for simulating large-scale directed dispersal processes across heterogeneous environments. Ecol Modell 230, 1–10. https://doi.org/10.1016/J.ECOLMODEL.2012.01.008

Ferguson, J.M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., Bowen-James, R., Deveson, I.W., 2022. InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. Bioinformatics 38, 1443. https://doi.org/10.1093/BIOINFORMATICS/BTAB846

Filloux, D., Dallot, S., Delaunay, A., Galzi, S., Jacquot, E., Roumagnac, P., 2015. Metagenomics Approaches Based on Virion-Associated Nucleic Acids (VANA): An Innovative Tool for Assessing Without A Priori Viral Diversity of Plants. Methods Mol Biol 1302, 249–257. https://doi.org/10.1007/978-1-4939-2620-6_18

Fontdevila, N., n.d. Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis.

Food, E., Authority, S., Schrader, G., Camilleri, M., Diakaki, M., Vos, S., 2019. Pest survey card on non-European isolates of citrus tristeza virus. EFSA Supporting Publications 16, 1600E. https://doi.org/10.2903/SP.EFSA.2019.EN-1600

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., 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, ppa.12962. https://doi.org/10.1111/ppa.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.,

Ziebell, H., 2021. Interlaboratory Comparison Study on Ribodepleted Total RNA High-Throughput Sequencing for Plant Virus Diagnostics and Bioinformatic Competence. Pathogens 10, 1174. https://doi.org/10.3390/pathogens10091174

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., Cosson, J.F., 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. mSystems 1, 1–22.

Garcia-Etxebarria, K., Garcia-Garcerà, M., Calafell, F., 2014. Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics 15, 1–10. https://doi.org/10.1186/1471-2105-15-90/FIGURES/4

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022a. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022b. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. v, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., Bilodeau, G.J., van de Vossenberg, B.T.L.H., 2022c. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology 2022, Vol. 11, Page 263 11, 263. https://doi.org/10.3390/BIOLOGY11020263

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G., Thomas, J.E., 2000. Genetic Diversity Among Banana streak virus Isolates from Australia. Phytopathology 90, 921–927. https://doi.org/10.1094/PHYTO.2000.90.8.921

Giovani, B., Blümel, S., Lopian, R., Teulon, D., Bloem, S., Galeano Martínez, C., Beltrán Montoya, C., Urias Morales, C.R., Dharmapuri, S., Timote, V., Horn, N., Chouibani, M., Mezui M'Ella, J.G., Herrera, V., Castinel, A., Goletsos, C., Moeller, C., Naumann, I., Stancanelli, G., Bronzwaer, S., Tramontini, S., MacDonald, P., Matheson, L., Anthoine, G., de Jonghe, K., Schenk, M., Steinmöller, S., Rodriguez, E., Cruz, M.L., Luck, J., Fraser, G., Brunel, S., Montuori, M., Fedchock, C., Steel, E., Pennington, H.G., Day, R., Rossi, J.P., Xia, J., 2020. Science diplomacy for plant health. Nature Plants 2020 6:8 6, 902–905. https://doi.org/10.1038/s41477-020-0744-x

Griffith, F., 1928. The Significance of Pneumococcal Types. J Hyg (Lond) 27, 113. https://doi.org/10.1017/S0022172400031879

Grosdidier, M., Aguayo, J., Marçais, B., Ioos, R., 2017. Detection of plant pathogens using real-time PCR: how reliable are late Ct values? Plant Pathol 66, 359–367. https://doi.org/10.1111/PPA.12591

Guirao-Rico, S., González, J., 2021. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. Mol Ecol Resour 21, 1216. https://doi.org/10.1111/1755-0998.13343

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022a. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022b. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020a. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020b. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanssen, I.M., Paeleman, A., Vandewoestijne, E., van Bergen, L., Bragard, C., Lievens, B., Vanachter, A.C.R.C., Thomma, B.P.H.J., 2009. Pepino mosaic virus isolates and differential symptomatology in tomato. Plant Pathol 58, 450–460. https://doi.org/10.1111/J.1365-3059.2008.02018.X

Harris, H.M.B., Hill, C., 2020. A Place for Viruses on the Tree of Life. Front Microbiol 11. https://doi.org/10.3389/FMICB.2020.604048

Hasiów-Jaroszewska, B., Borodynko, N., Jackowiak, P., Figlerowicz, M., Pospieszny, H., 2011. Single mutation converts mild pathotype of the Pepino mosaic virus into necrotic one. Virus Res 159, 57–61. https://doi.org/10.1016/J.VIRUSRES.2011.04.008

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., Lang, J., 2020. Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. Front Bioeng Biotechnol 8, 817. https://doi.org/10.3389/FBIOE.2020.00817/BIBTEX

Hilaire, J., Tindale, S., Jones, G., Pingarron-Cardenas, G., Bačnik, K., Ojo, M., Frewer, L.J., 2022. Risk perception associated with an emerging agri-food risk in Europe: plant viruses in agriculture. Agric Food Secur 11, 1–19. https://doi.org/10.1186/S40066-022-00366-5/TABLES/3

Hirabara, S.M., Serdan, T.D.A., Gorjao, R., Masi, L.N., Pithon-Curi, T.C., Covas, D.T., Curi, R., Durigon, E.L., 2022. SARS-COV-2 Variants: Differences and Potential of Immune Evasion. Front Cell Infect Microbiol 11, 1401. https://doi.org/10.3389/FCIMB.2021.781429/BIBTEX

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593. https://doi.org/10.1093/BIOINFORMATICS/BTR708

Hull, R., 2014. Plant Virology (Fifth Edition). Plant Virology 3–14. https://doi.org/10.1016/B978-0-12-384871-0.00001-7

Irwin, N.A.T., Pittis, A.A., Richards, T.A., Keeling, P.J., 2021. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nature Microbiology 2022 7:2 7, 327–336. https://doi.org/10.1038/s41564-021-01026-3

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332. https://doi.org/10.1093/BIOINFORMATICS/BTT113

Jeong, J.-J., Ju, H.-J., Noh, J., 2014. A Review of Detection Methods for the Plant Viruses. Research in Plant Disease 20, 173–181. https://doi.org/10.5423/RPD.2014.20.3.173

Jones, R.A.C., 2014. Trends in plant virus epidemiology: Opportunities from new or improved technologies. Virus Res 186, 3–19. https://doi.org/10.1016/J.VIRUSRES.2013.11.003

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalimuthu, K., Arivalagan, J., Mohan, M., Samuel Selvan Christyraj, J.R., Arockiaraj, J., Muthusamy, R., Ju, H.J., 2022. Point of care diagnosis of plant virus: Current trends and prospects. Mol Cell Probes 61, 101779. https://doi.org/10.1016/J.MCP.2021.101779

Kechin, A., Boyarskikh, U., Kel, A., Filipenko, M., 2017. CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. Journal of Computational Biology 24, 1138–1143. https://doi.org/10.1089/CMB.2017.0096/ASSET/IMAGES/LARGE/FIGURE3.JPEG

Koboldt, D.C., 2020a. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Koboldt, D.C., 2020b. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Kogovšek, P., Pompe-Novak, M., Baebler, S., Rotter, A., Gow, L., Gruden, K., Foster, G.D., Boonham, N., Ravnikar, M., 2010. Aggressive and mild Potato virus Y isolates trigger different specific responses in susceptible potato plants. Plant Pathol 59, 1121–1132. https://doi.org/10.1111/J.1365-3059.2010.02340.X

Koonin, E. v., Dolja, V. v., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., Kuhn, J.H., 2020. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84. https://doi.org/10.1128/MMBR.00061-19

Kreuze, J., 2014. siRNA deep sequencing and assembly: Piecing together viral infections. Detection and Diagnostics of Plant Pathogens 21–38. https://doi.org/10.1007/978-94-017-9020-8_2/FIGURES/3

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1–7. https://doi.org/10.1016/J.VIROL.2009.03.024

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res 239, 136–142. https://doi.org/10.1016/J.VIRUSRES.2017.02.002

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., 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 . https://doi.org/10.3390/plants10020328

Kumar, R., Nagpal, S., Kaushik, S., Mendiratta, S., 2020. COVID-19 diagnostic approaches: different roads to the same destination. VirusDisease 2020 31:2 31, 97–105. https://doi.org/10.1007/S13337-020-00599-7

Kutnjak, D., Rupar, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J.F., Ravnikar, M., 2015. Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. J Virol 89, 4760–4769. https://doi.org/10.1128/JVI.03685-14/SUPPL_FILE/ZJV999090298SO1.PDF

Kutnjak, D., Tamisier, L., Adams, I., Boonham, N., Candresse, T., Chiumenti, M., de Jonghe, K., Kreuze, J.F., Lefebvre, M., Silva, G., Malapi-wight, M., Margaria, P., Pleško, I.M., McGreig, S., Miozzi, L., Remenant, B., Reynard, J.S., Rollin, J., Rott, M., Schumpp, O., Massart, S., Haegeman, A., 2021. A Primer on the Analysis of High-Throughput Sequencing Data for Detection of Plant Viruses. Microorganisms 9. https://doi.org/10.3390/MICROORGANISMS9040841

Kutschera, U., Niklas, K.J., 2004. The modern theory of biological evolution: An expanded synthesis. Naturwissenschaften 91, 255–276. https://doi.org/10.1007/S00114-004-0515-Y/FIGURES/8

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 optimization of high-throughput sequencing bioinformatics pipelines for virus detection. Viruses 10. https://doi.org/10.3390/v10100528

Lau, H.Y., Botella, J.R., 2017. Advanced DNA-based point-of-care diagnostic methods for plant diseases detection. Front Plant Sci 8, 2016. https://doi.org/10.3389/FPLS.2017.02016/BIBTEX

Lebas, B., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G., Blouin, A., Boonham N., Candresse, T., Chandelier, A., De Jonghe, K., Fox, A., Gaafar, Y., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., Landa, B., Li, Massart, S., n.d. Guidelines for preparing a laboratory to use high-throughput sequencing as routine diagnostic test: general and technical recommendations. EPPO Bulletin.

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., 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 Bulletin 52, 394–418. https://doi.org/10.1111/epp.12863

Lefebvre, M., Theil, S., Ma, Y., Candresse, T., 2019. The VirAnnot Pipeline: A Resource for Automated Viral Diversity Estimation and Operational Taxonomy Units Assignation for Virome Sequencing Data. https://doi.org/10.1094/PBIOMES-07-19-0037-A 3, 256–259. https://doi.org/10.1094/PBIOMES-07-19-0037-A

Legrand, P., 2015. Biological assays for plant viruses and other graft-transmissible pathogens diagnoses: a review. EPPO Bulletin 45, 240–251. https://doi.org/10.1111/EPP.12222

Li, C.-X., Li, W., Zhou, J., Zhang, B., Feng, Y., Xu, C.-P., Lu, Y.-Y., Holmes, E.C., Shi, M., 2020. High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory infection. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-60992-6

Li, H., Durbin, R., 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, H., Durbin, R., 2009b. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Liefting, L.W., Waite, D.W., Thompson, J.R., 2021. Application of Oxford Nanopore Technology to Plant Virus Detection. Viruses 13. https://doi.org/10.3390/V13081424

Ling, K.S., Scott, J.W., 2007. Sources of Resistance to Pepino mosaic virus in Tomato Accessions. https://doi.org/10.1094/PDIS-91-6-0749 91, 749–753. https://doi.org/10.1094/PDIS-91-6-0749

Low, A.J., Koziol, A.G., Manninger, P.A., Blais, B., Carrillo, C.D., 2019. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 7, e6995. https://doi.org/10.7717/peerj.6995

Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci 2017, e104. https://doi.org/10.7717/PEERJ-CS.104/SUPP-5

Luisi, P.L., 1998. About various definitions of life. Orig Life Evol Biosph 28, 613–622. https://doi.org/10.1023/A:1006517315105

Maachi, A., Torre, C., Sempere, R.N., Hernando, Y., Aranda, M.A., Donaire, L., 2021. Use of High-Throughput Sequencing and Two RNA Input Methods to Identify Viruses Infecting Tomato Crops. Microorganisms . https://doi.org/10.3390/microorganisms9051043

Mackenzie, J.S., Jeggo, M., 2019. The One Health Approach—Why Is It So Important? Trop Med Infect Dis 4. https://doi.org/10.3390/TROPICALMED4020088

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C.M., Roumagnac, P., van der Vlugt, R., Massart, S., 2020. Illuminating an Ecological Blackbox: Using High Throughput Sequencing to Characterize the Plant Virome Across Scales. Front Microbiol 11. https://doi.org/10.3389/fmicb.2020.578064

Maksimovic Carvalho Ferreira, O., Lengar, Ž., Kogej, Z., Bačnik, K., Bajde, I., Milavec, M., Županič, A., Mehle, N., Kutnjak, D., Ravnikar, M., Gutierrez-Aguirre, I., 2022. Evaluation of Methods and Processes for Robust Monitoring of SARS-CoV-2 in Wastewater. Food Environ Virol 14, 384–400. https://doi.org/10.1007/S12560-022-09533-0/FIGURES/4

Malapi-Wight, M., Adhikari, B., Zhou, J., Hendrickson, L., Maroon-Lango, C.J., McFarland, C., Foster, J.A., Hurtado-Gonzales, O.P., 2021. HTS-Based Diagnostics of Sugarcane Viruses: Seasonal Variation and Its Implications for Accurate Detection. Viruses . https://doi.org/10.3390/v13081627

Maliogka, V.I., Minafra, A., Saldarelli, P., Ruiz-García, A.B., Glasa, M., Katis, N., Olmos, A., 2018. Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. Viruses 2018, Vol. 10, Page 436 10, 436. https://doi.org/10.3390/V10080436

Maljkovic Berry, I., Melendrez, M.C., Bishop-Lilly, K.A., Rutvisuttinunt, W., Pollett, S., Talundzic, E., Morton, L., Jarman, R.G., 2020. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations

for Development of Laboratory Capacity. J Infect Dis 221, S292–S307. https://doi.org/10.1093/INFDIS/JIZ286

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018a. Viral double-stranded RNAs (dsRNAs) from Plants: Alternative nucleic acid substrates for high-throughput sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4/COVER

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018b. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4

Maree, H.J., Fox, A., Al Rwahnih, M., Boonham, N., Candresse, T., 2018. Application of hts for routine plant virus diagnostics: state of the art and challenges. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.01082

Martelli, G.P., 2017. An overview on grapevine viruses, viroids, and the diseases they cause. Grapevine Viruses: Molecular Biology, Diagnostics and Management 31–46. https://doi.org/10.1007/978-3-319-57706-7_2/TABLES/1

Martínez, M., de Viedma, D.G., Alonso, M., Andrés, S., Bouza, E., Cabezas, T., Cabeza, I., Reyes, A., Sánchez-Yebra, W., Rodríguez, M., Sánchez, M.I., Rogado, M.C., Fernández, R., Peñafiel, T., Martínez, J., Barroso, P., Lucerna, M.A., Diez, L.F., Gutiérrez, C., 2006. Impact of Laboratory Cross-Contamination on Molecular Epidemiology Studies of Tuberculosis. J Clin Microbiol 44, 2967–2969. https://doi.org/10.1128/JCM.00754-06

Marttunen, M., Lienert, J., Belton, V., 2017. Structuring problems for Multi-Criteria Decision Analysis in practice: A literature review of method combinations. Eur J Oper Res 263, 1–17. https://doi.org/10.1016/J.EJOR.2017.04.041

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, Jonkers W., Kreuze J., Kutnjak D., Landa B., n.d. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests | Zenodo [WWW Document]. URL https://zenodo.org/record/6637519 (accessed 7.11.22).

Massart, S., Adams, I., Rwahnih, A., Baeyen, S., Bilodeau, G.J., Blouin, A.G., Boonham, N., Candresse, T., Chandellier, 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.B., Liu, M., Maclot, F., Malapi-Wight, M., Maree, H.J., Martoni, F., Mehle, N., Minafra, A., Mollov, D., Moreira, A.G., Nakhla, M., Petter, F., Piper, A.M., Ponchart, J.P., Rae, R., Remenant, B., Rivera, Y., Rodoni, B., Botermans, M., 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., van de Vossenberg, B.T.L.H., Westenberg, M., Wetzel, T., Ziebell, H., Lebas, B.S.M., al Rwahnih, M., 2022a. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. Peer Community Journal 2, 37. https://doi.org/10.24072/PCJOURNAL.181

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J.S., Rumbou, A., Saldarelli, P., Škoric, D., Vainio, E.J., Valkonen, J.P.T., 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. Front Microbiol 8, 45. https://doi.org/10.3389/FMICB.2017.00045/BIBTEX

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, Y., Candresse, T., 2019. Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. Phytopathology 109. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, Sebastien, 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, 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. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, S., Lebas, B., Chabirand, A., Chappé, A.-M., Dreo, T., Faggioli, F., Harrison, C., Macarthur, R., Mehle, N., Mezzalama, M., Petter, F., Ravnikar, M., Renvoisé, J.-P., Spadaro, D., Tomassoli, L., Tomlinson, J., Trontin, C., Vlugt, R. van der, Vučurović, A., Weekes, R., Brostaux, Y., 2022b. Guidelines for improving statistical analyses of validation datasets for plant pest diagnostic tests. EPPO Bulletin 52, 419–433. https://doi.org/10.1111/EPP.12862

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014a. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014b. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, Sebastien, Olmos, A., Jijakli, H., Candresse, T., 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188, 90–96. https://doi.org/10.1016/J.VIRUSRES.2014.03.029

Minic, R., Zivkovic, I., Minic, R., Zivkovic, I., 2020. Optimization, Validation and Standardization of ELISA. Norovirus. https://doi.org/10.5772/INTECHOPEN.94338

Modrow, S., Falke, D., Truyen, U., Schätzl, H., 2013. Viruses: Definition, Structure, Classification. Molecular Virology 17. https://doi.org/10.1007/978-3-642-20718-1_2

Moonen, P., Boonstra, J., Hakze- Van Der Honing, R., Boonstra- Leendertse, C., Jacobs, L., Dekker, A., 2003. Validation of a LightCycler-based reverse transcription polymerase chain reaction for the detection of foot-and-mouth disease virus. J Virol Methods 113, 35–41. https://doi.org/10.1016/S0166-0934(03)00220-9

Moreira, D., López-García, P., 2009. Ten reasons to exclude viruses from the tree of life. Nature Reviews Microbiology 2009 7:4 7, 306–311. https://doi.org/10.1038/nrmicro2108

Navarro, R., Ambrós, S., Butković, A., Carrasco, J.L., González, R., Martínez, F., Wu, B., Elena, S.F., 2022. Defects in plant immunity modulate the rates and patterns of RNA virus evolution. Virus Evol 8. https://doi.org/10.1093/VE/VEAC059

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D.A., Bhaya, D., Narrowe, A.B., Probst, A.J., Sczyrba, A., Kohler, A., Séguin, A., Shade, A., Campbell, B.J., Lindahl, B.D., Reese, B.K., Roque, B.M., DeRito, C., Averill, C., Cullen, D., Beck, D.A.C., Walsh, D.A., Ward, D.M., Wu, D., Eloe-Fadrosh, E., Brodie, E.L., Young, E.B., Lilleskov, E.A., Castillo, F.J., Martin, F.M., LeCleir, G.R., Attwood, G.T., Cadillo-Quiroz, H., Simon, H.M., Hewson, I., Grigoriev, I. v., Tiedje, J.M., Jansson, J.K., Lee, J., VanderGheynst, J.S., Dangl, J., Bowman, J.S., Blanchard, J.L., Bowen, J.L., Xu, J., Banfield, J.F., Deming, J.W., Kostka, J.E., Gladden, J.M., Rapp, J.Z., Sharpe, J., McMahon, K.D., Treseder, K.K., Bidle, K.D., Wrighton, K.C., Thamatrakoln, K., Nusslein, K., Meredith, L.K., Ramirez, L., Buee, M., Huntemann, M., Kalyuzhnaya, M.G., Waldrop, M.P., Sullivan, M.B., Schrenk, M.O., Hess, M., Vega, M.A., O'Malley, M.A., Medina, M., Gilbert, N.E., Delherbe, N., Mason, O.U., Dijkstra, P., Chuckran, P.F., Baldrian, P., Constant, P., Stepanauskas, R., Daly, R.A., Lamendella, R., Gruninger, R.J., McKay, R.M., Hylander, S., Lebeis, S.L., Esser, S.P., Acinas, S.G., Wilhelm, S.S., Singer, S.W., Tringe, S.S., Woyke, T., Reddy, T.B.K., Bell, T.H., Mock, T., McAllister, T., Thiel, V., Denef, V.J., Liu, W.T., Martens-Habbena, W., Allen Liu, X.J., Cooper, Z.S., Wang, Z., Krupovic, M., Dolja, V. v., Kyrpides, N.C., Koonin, E. v., Gophna, U., 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell 185, 4023-4037.e18. https://doi.org/10.1016/J.CELL.2022.08.023

Ng, S.H., Braxton, C., Eloit, M., Feng, S.F., Fragnoud, R., Mallet, L., Mee, E.T., Sathiamoorthy, S., Vandeputte, O., Khan, A.S., 2018. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: Upstream sample processing and library preparation [WWW Document]. Viruses. https://doi.org/10.3390/v10100566

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., van Helden, J., Medina-Rivera, A., Thomas-Chollier, M., 2018. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 46, W209–W214. https://doi.org/10.1093/NAR/GKY317

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., Massart, S., 2018a. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

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., Massart, S., 2018b. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

Oñate-Sánchez, L., Vicente-Carbajosa, J., 2008. DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC Res Notes 1, 1–7. https://doi.org/10.1186/1756-0500-1-93/FIGURES/3

Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology 2021 22:1 22, 1–19. https://doi.org/10.1186/S13059-021-02393-0

Pappas, N., Roux, S., Hölzer, M., Lamkiewicz, K., Mock, F., Marz, M., Dutilh, B.E., 2021. Virus Bioinformatics. Encyclopedia of Virology 124–132. https://doi.org/10.1016/B978-0-12-814515-9.00034-5

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017a. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017b. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

PM 7/76 (5) Use of EPPO Diagnostic Standards, 2018. . EPPO Bulletin 48, 373–377. https://doi.org/10.1111/EPP.12506

PM 7/98 (5) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity, 2021. . EPPO Bulletin 51, 468–498. https://doi.org/10.1111/EPP.12780

Ramesh, S., Govindarajulu, M., Parise, R.S., Neel, L., Shankar, T., Patel, S., Lowery, P., Smith, F., Dhanasekaran, M., Moore, T., 2021. Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy. Vaccines (Basel) 9. https://doi.org/10.3390/VACCINES9101195

Ricciuti, E., Laboureau, N., Noumbissié, G., Chabannes, M., Sukhikh, N., Pooggin, M.M., Iskra-Caruana, M.-L., 2021. Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. Journal of General Virology 102, 001670. https://doi.org/10.1099/JGV.0.001670

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N. v., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Martin, M.D., MacDonald, G.K., Neenan, P., Records, A., Schmale, D.G., Tateosian, L., Wei, Q., 2021. The persistent threat of emerging plant disease pandemics to global food security. Proc Natl Acad Sci U S A 118, e2022239118. https://doi.org/10.1073/PNAS.2022239118/SUPPL_FILE/PNAS.2022239118.SAPP.PDF

Rivarez, M., Faure, C., Svanella-Dumas, L., Pecman, A., Tušek-Žnidaric, M., Schönegger, D., de Jonghe, K., Blouin, A., Massart, S., Ravnikar, M., Kutnjak, D., Marais, A., Candresse, T., 2022. Diversity and pathobiology of an ilarvirus unexpectedly detected in diverse host plants and in global sequencing data. https://doi.org/10.1101/2022.12.15.520526

Rivarez, M.P.S., Vučurović, A., Mehle, N., Ravnikar, M., Kutnjak, D., 2021. Global Advances in Tomato Virome Research: Current Status and the Impact of High-Throughput Sequencing. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.671925

Rizzo, D.M., Lichtveld, M., Mazet, J.A.K., Togami, E., Miller, S.A., 2021. Plant health and its effects on food safety and security in a One Health framework: four case studies. One Health Outlook 2021 3:1 3, 1–9. https://doi.org/10.1186/S42522-021-00038-7

Roberts, C.A., Dietzgen, R.G., Heelan, L.A., MaclLean, D.J., 2000. Real-time RT-PCR fluorescent detection of tomato spotted wilt virus. J Virol Methods 88, 1–8. https://doi.org/10.1016/S0166-0934(00)00156-7

Rong, W., Rollin, J., Hanafi, M., Roux, N., Massart, S., 2022. Validation of high throughput sequencing as virus indexing test for Musa germplasm: performance criteria evaluation and contamination

monitoring using an alien control. PhytoFrontiers. https://doi.org/https://doi.org/10.1094/PHYTOFR-03-22-0030-FI

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarría, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19 Suppl 1, 81–88. https://doi.org/10.1111/J.1365-294X.2009.04470.X

Rosenthal, S.H., Gerasimova, A., Ruiz-Vega, R., Livingston, K., Kagan, R.M., Liu, Y., Anderson, B., Owen, R., Bernstein, L., Smolgovsky, A., Xu, D., Chen, R., Grupe, A., Tanpaiboon, P., Lacbawan, F., 2022. Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory. Scientific Reports 2022 12:1 12, 1–12. https://doi.org/10.1038/s41598-022-06091-0

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017a. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017b. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rubio, L., Galipienso, L., Ferriol, I., 2020. Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution. Front Plant Sci 11, 1092. https://doi.org/10.3389/FPLS.2020.01092/BIBTEX

Ruiz-Mirazo, K., Peretó, J., Moreno, A., 2004. A universal definition of life: autonomy and open-ended evolution. Orig Life Evol Biosph 34, 323–346. https://doi.org/10.1023/B:ORIG.0000016440.53346.DC

Rybicki, E.P., 2015. A Top Ten list for economically important plant viruses. Arch Virol 160, 17–20. https://doi.org/10.1007/s00705-014-2295-9

Sabatier, M., Bal, A., Destras, G., Regue, H., Quéromès, G., Cheynet, V., Lina, B., Bardel, C., Brengel-Pesce, K., Navratil, V., Josset, L., 2020. Comparison of Nucleic Acid Extraction Methods for a Viral Metagenomics Analysis of Respiratory Viruses. Microorganisms 8, 1–18. https://doi.org/10.3390/MICROORGANISMS8101539

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94. https://doi.org/10.1016/0022-2836(75)90213-2

Sangiovanni, M., Granata, I., Thind, A.S., Guarracino, M.R., 2019. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics 20. https://doi.org/10.1186/S12859-019-2684-X

Santala, J., Valkonen, J.P.T., 2018. Sensitivity of Small RNA-Based Detection of Plant Viruses . Frontiers in Microbiology .

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution 2019 3:3 3, 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schilbert, H.M., Rempel, A., Pucker, B., 2020. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants 2020, Vol. 9, Page 439 9, 439. https://doi.org/10.3390/PLANTS9040439

Silva, J.M.F., Melo, F.L., Elena, S.F., Candresse, T., Sabanadzovic, S., Tzanetakis, I.E., Blouin, A.G., Villamor, D.E.V., Mollov, D., Constable, F., Cao, M., Saldarelli, P., Cho, W.K., Nagata, T., 2022. Virus classification based on in-depth sequence analyses and development of demarcation criteria using the Betaflexiviridae as a case study. J Gen Virol 103, 001806. https://doi.org/10.1099/JGV.0.001806/CITE/REFWORKS

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J.C., Manuel, M., Philippe, H., Telford, M.J., 2018. A software tool "CroCo" detects pervasive cross-species contamination in next generation sequencing data. BMC Biol 16, 1–9. https://doi.org/10.1186/s12915-018-0486-7

Simmonds, P., Adams, M.J., Benk, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E. v., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology 2017 15:3 15, 161–168. https://doi.org/10.1038/nrmicro.2016.177

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., Weissman, I.L., 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724. https://doi.org/10.1101/125724

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., Al Rwahnih, M., 2021a. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., al Rwahnih, M., 2021b. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. https://doi.org/10.1038/nbt.3988

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014a. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014b. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., Nikolski, M., 2022. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. Frontiers in Bioinformatics 2. https://doi.org/10.3389/FBINF.2022.867111/FULL

Tamisier, L., Colson, C., Maclot, F., Zhang, P., Wang, X., Francis, F., Baurain, D., Massart, S., n.d. Deciphering the intra-host genetic diversity of viruses: insights from the experimental evolution of a plant virus.

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., Massart, S., 2021a. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/pcjournal.62

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., Massart, S., 2021b. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/PCJOURNAL.62

Tedersoo, L., Albertsen, M., Anslan, S., Callahan, B., 2021. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl Environ Microbiol 87, 1–19. https://doi.org/10.1128/AEM.00626-21/SUPPL_FILE/AEM.00626-21-S0001.PDF

Tejano, L.A., Peralta, J.P., Yap, E.E.S., Panjaitan, F.C.A., Chang, Y.W., 2019. Prediction of Bioactive Peptides from Chlorella sorokiniana Proteins Using Proteomic Techniques in Combination with Bioinformatics Analyses. International Journal of Molecular Sciences 2019, Vol. 20, Page 1786 20, 1786. https://doi.org/10.3390/IJMS20071786

Tekle, Y.I., Grant, J.R., Kovner, A.M., Townsend, J.P., Katz, L.A., 2010. Identification of new molecular markers for assembling the eukaryotic tree of life. Mol Phylogenet Evol 55, 1177–1182. https://doi.org/10.1016/J.YMPEV.2010.03.010

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., Roux, N., 2015. Technical guidelines for the safe movement of Musa germplasm, in: Thomas, J. (Ed.), Technical Guidelines for the Safe Movement of Musa Germplasm. Bioversity International, > >.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/NAR/GKAB1061

Velasco, L., Padilla, C. V., 2021. High-Throughput Sequencing of Small RNAs for the Sanitary Certification of Viruses in Grapevine. Front Plant Sci 12. https://doi.org/10.3389/FPLS.2021.682879/FULL

Vereecke, N., Carnet, F., Pronost, S., Vanschandevijl, K., Theuns, S., Nauwynck, H., 2021. Genome Sequences of Equine Herpesvirus 1 Strains from a European Outbreak of Neurological Disorders Linked to a Horse Gathering in Valencia, Spain, in 2021. Microbiol Resour Announc. https://doi.org/10.1128/MRA.00333-21

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016a. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016b. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

VUNSH, R., ROSNER, A., STEIN, A., 1990. The use of the polymerase chain reaction (PCR) for the detection of bean yellow mosaic virus in gladiolus. Annals of Applied Biology 117, 561–569. https://doi.org/10.1111/J.1744-7348.1990.TB04822.X

Waite, D.W., Liefting, L., Delmiglio, C., Chernyavtseva, A., Ha, H.J., Thompson, J.R., 2022. Development and Validation of a Bioinformatic Workflow for the Rapid Detection of Viruses in Biosecurity. Viruses 14, 2163. https://doi.org/10.3390/V14102163/S1

Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P., Dempsey, D.M., Dutilh, B.E., García, M.L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., Łobocka, M., Oksanen, H.M., Orton, R.J., Robertson, D.L., Rubino, L., Sabanadzovic, S., Simmonds, P., Smith, D.B., Suzuki, N., van Doorslaer, K., Vandamme, A.M., Varsani, A., Zerbini, F.M., 2022. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch Virol 167, 2429–2440. https://doi.org/10.1007/S00705-022-05516-5/TABLES/2

Ward, M., 2016. Action against pest spread—the case for retrospective analysis with a focus on timing. Food Secur 8, 77–81. https://doi.org/10.1007/S12571-015-0532-Z/FIGURES/1

Watzinger, F., Ebner, K., Lion, T., 2006. Detection and monitoring of virus infections by real-time PCR. Mol Aspects Med 27, 254–298. https://doi.org/10.1016/J.MAM.2005.12.001

Whitfield, A.E., Campbell, L.R., Sherwood, J.L., Ullman, D.E., 2003. Tissue Blot Immunoassay for Detection of Tomato spotted wilt virus in Ranunculus asiaticus and Other Ornamentals. Plant Dis 87, 618–622. https://doi.org/10.1094/PDIS.2003.87.6.618

Williams, R.C., 1954. Electron Microscopy of Viruses. Adv Virus Res 2, 183–239. https://doi.org/10.1016/S0065-3527(08)60533-3

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 762302. https://doi.org/10.1186/s13059-019-1891-0

Wood, D.E., Salzberg, S.L., 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, 1–12. https://doi.org/10.1186/GB-2014-15-3-R46/FIGURES/5

Wu, D., Jospin, G., Eisen, J.A., 2013. Systematic Identification of Gene Families for Use as “Markers” for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. PLoS One 8. https://doi.org/10.1371/JOURNAL.PONE.0077033

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

5

Genome characterization of detected virus

Variant analysis

After virus detection, a more precise characterization can be performed. Comparing the genetic variations of different viral strains allows the determination of different isolates within a species. As the different isolates can have different biological properties. This can provide important information on the evolution and spread of the virus, as well as its potential impact on crop yields and the development of new control measures.

Variant calling is the process of identifying specific genetic variations within a genome, which can include single nucleotide polymorphisms (SNPs) (insertions, deletions or substitutions of genetic material). The process generally involves the following steps (Koboldt, 2020a; Schilbert et al., 2020):

- Mapping: The sequenced reads are aligned to the closest known genome (reference) using alignment software such as BWA (Li and Durbin, 2009b).
- Variant calling: Software such as SAMtools (Danecek et al., 2021) is used to identify genetic variations between the reference and the reads.
- Filtering: To remove false positives, the variants calls are filtered based on (reads) alignment quality, the coverage of the region, and other criteria.
- Interpretation: Finally, the variants are interpreted in the context of the organism's biology and the specific research question. This can involve functional interpretation, phylogenetic analysis, and other methods to gain a deeper understanding of the variants and their implications.

It is important to note that different tools may be used depending on the type of organism, variant complexity and the research goals (Brinkmann et al., 2019). It is also worth noting that plant viruses can have multiple variants that can have different effects on the host (Ling and Scott, 2007).

Variant impact on host

Different variants of a plant virus can cause different symptoms and dangers depending on the genetic changes within the viral genome. Some variants may be more virulent and cause more severe symptoms, while others may be less harmful or even symptomless. The severity of the symptoms and the danger of a plant virus variant can also depend on the host plant. Some variants may be more pathogenic on certain plant species or varieties, while others may be less so (Ling and Scott, 2007).

Pepino mosaic virus (PepMV) is a virus from the family *Alphaflexiviridae* (genus *Potexvirus*) known to infect tomato, eggplant and potato plants. PepMV can cause

severe damage to tomato crops, resulting in reduced yields and economic losses. Symptoms include mosaic patterns on leaves, leaf distortion, and reduced fruit quality. PepMV is a small, non-enveloped virus with a single-stranded RNA genome. The virus has a high mutation rate, and several distinct genetic variants have been identified, which can have different effects on the host (Hanssen et al., 2009). Indeed according to the variant present, the symptoms change drastically from asymptomatic to necrotic (Hasiów-Jaroszewska et al., 2011). Those differences associated with the tendency to have variant mixed infection (as demonstrated during the Spanish tomato crop epidemic) (Agüero et al., 2018) make the study of the associated disease even more complex. It also opens new possibilities for plant protection as a pest control plan based on intentional infection of harmless variants can limit the impact of harmful ones (Agüero et al., 2018; Anastassiadou et al., 2021).

As demonstrated for PepMV, the characterization of plant virus variants is important for understanding the evolution and spread of the virus, as well as its potential impact on crop yields. It can also help develop control measures for the specific variant, such as a virus-resistant variety or a specific treatment.

Tracking viruses with mutation profile

Tracking viruses with mutation profiles is a way of identifying specific genetic variations within the viral genome and using these variations to trace the spread of the virus over time or to understand the modification of its biological properties. By analyzing the genetic makeup of a viral strain, the identification of specific mutations and classification of the virus into different variants can be done.

Once the variants have been identified, the spread of those viral strains can be tracked through various samples by comparing the mutation profiles. In addition, the mutation profile of a virus can be used to monitor the evolution of the virus over time, which can provide important information on the emergence of new variants and their potential impact on plant health (Navarro et al., 2022; Rubio et al., 2020).

Variant profiling can also help monitor sequencing cross-contamination by identifying specific genetic variations within the viral genome and using these variations to trace the origin of the contamination. An example us such use was described in Chapter 3 for plant viruses. The usage of broader mutation profiling for the bacterial source of contamination in food is also well described (Baert et al., 2021).

For all these reasons, an accurate SNPs prediction is important. Parameters leading to such prediction success or failure were studied and explained in this chapter.

Article 3

(in PeerJ)

Virus SNPs Detection by High-Throughput Sequencing: Large-Scale Performance Testing of Sequence Analysis Strategies

Johan Rollin[1,2], Rachelle Bester[3a, 3b], Yves Brostaux[4], Kadriye Caglayan[5], Kris De Jonghe[6], Ales Eichmeier[7], Yoika Foucart[6], Annelies Haegeman[6], Igor Koloniuk[8], Petr Kominek[9], Hans Jacob Maree[3a, 3b], Serkan Önder[10], Susana Posada-Céspedes[11a, 11b], Vahid Roumi[12], Dana Šafářová [13], Olivier Schumpp[14], Çiğdem Ulubaş Serçe[15], Merike Sõmera[16], Lucie Tamisier[17a, 17b], Eeva Vainio[18], Rene Vandervlugt[19] and Sebastien Massart[1]

1. *Laboratory of Plant Pathology – TERRA - Gembloux Agro-Bio Tech – University of Liège, 5030 Gembloux, Belgium*. contact person: sebastien.massart@uliege.be.
2. *DNAVision, 6041 Gosselies, Belgium*
3. *a) Citrus Research International, PO Box 2201, Matieland, 7602, South*
   b) Department of Genetics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa
4. *Laboratory of statistics, computer science and modelling applied to bioengineering, TERRA, Gembloux Agro-Bio Tech, Teaching and Research Centre, University of Liège, 5030 Gembloux, Belgium*
5. Plant Protection Department, Agricultural Faculty, Hatay Mustafa Kemal University, Hatay, 31034, Turkey
6. *Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Plant Sciences Unit, Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium*
7. *Mendeleum - Institute of Genetics, Faculty of Horticulture, Mendel University in Brno, Valtická 334, 691 44 Lednice, Czech Republic*
8. *Biology Centre CAS, Branisovska 31, Ceske Budejovice 370 05, Czech Republic*
9. *Crop Research Institute, Drnovska 507, 161 06 Praha, Czech Republic*
10. *Department of Plant Protection, Faculty of Agriculture, Eskişehir Osmangazi University, 26160, Eskişehir, Turkey*
11. *a) Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland*
    b) SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland
12. *Plant Protection Department, Faculty of Agriculture, University of Maragheh, Maragheh, Iran*
13. *Department of Cell Biology and Genetics, Faculty of Science, Palacký University Olomouc, Šlechtitelů 27, 783 71 Olomouc, Czech Republic*
14. *Plant Protection Department, Agroscope, Nyon, 1260, Switzerland*
15. *Plant Production and Technologies Department, Ayhan Şahenk Faculty of Agricultural Science and Technologies, Niğde Ömer Halisdemir University, 51240 Niğde, Turkey*
16. *Department of Chemistry and Biotechnology, Tallinn University of Technology, Akadeemia tee 15, 12618 Tallinn, Estonia*
17. *a) Pathologie Végétale, INRAE, 84140 Montfavet, France*
    b) GAFL, INRAE, 84140 Montfavet, France
18. *Natural Resources Institute Finland, Latokartanonkaari 9, FI-00790 Helsinki, Finland*
19. *Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands*

Abstract

Recent developments in high-throughput sequencing (HTS) technologies and bioinformatics have drastically changed research in virology, especially for virus discovery. Indeed, proper monitoring of the viral population requires information on the different isolates circulating in the studied area. For this purpose, HTS has greatly facilitated the sequencing of new genomes of detected viruses and their comparison. However, bioinformatics analyses allowing reconstruction of genome sequences and detection of Single Nucleotide Polymorphisms (SNPs) can potentially create bias and has not been widely addressed so far.

Therefore, more knowledge is required on the limitations of predicting SNPs based on HTS-generated sequence samples. To address this issue, we compared the ability of 14 plant virology laboratories, each employing a different bioinformatics pipeline, to detect 21 variants of pepino mosaic virus (PepMV) in three samples through large-scale Performance-Testing (PT) using three artificially designed datasets. To evaluate the impact of bioinformatics analyses, they were divided into three key steps: reads pre-processing, virus-isolate identification, and variant calling. Each step was evaluated independently through an original, PT design including discussion and validation between participants at each step.

Overall, this work underlines key parameters influencing SNPs detection and proposes recommendations for reliable variant calling for plant viruses. The identification of the closest reference, mapping parameters and manual validation of the detection were recognized as the most impactful analysis steps for the success of the SNPs detections. Strategies to improve the prediction of SNPs are also discussed.

Introduction

Establishing high-throughput sequencing (HTS) technologies and developing bioinformatics methods for analyzing the generated sequencing data have greatly improved knowledge of viral diversity, including plant viruses (Pappas et al., 2021). Complete or nearly complete virus genomes can be generated by *de novo* assembly of raw reads into longer contigs. In this process, variants can be identified, differing from the original virus (considered as the reference) by mutations, like single nucleotide polymorphisms (SNPs), i.e., insertions/deletions/substitutions (Ramesh et al., 2021). Finding individual virus variants from a mixed population via variant calling is essential for understanding viral evolution and genetic diversity. As the recent COVID-19 pandemic has shown, continuous monitoring of virus genome evolution is essential to implement societal and sanitary measures for human health (Hirabara et al., 2022). This monitoring is also key for plant viruses (Rubio et al., 2020), especially since it is well known that (+)RNA viruses mutate rapidly. Indeed, genome evolution due to the high mutation and recombination rates may allow RNA viruses to increase their host range and adapt to new environments [5]-[9]. The generation of accurate genome is therefore becoming a key challenge for proper integration of genome data into virus epidemics monitoring. Any base modification of a genome can be decisive, so the proper detection of single nucleotide polymorphisms (SNPs) in contigs built from HTS data is a crucial step in bioinformatics pipelines (Kutnjak et al., 2015) and the identification of SNPs present at low frequencies in the sequences (and therefore absent in the consensus sequence) is important too and often overlooked.

Virus low frequency variants are important to detect as they can change the dynamic of an infection. Indeed, minor variants can change the fitness of Coxsackievirus B3 in human lung cells (Bordería et al., 2015). In several citrus varieties, Citrus tristeza virus (CTV) symptoms may be caused by low frequency variants (Černi et al., 2008). Therefore, monitoring the viral populations, including the low frequency ones, is important for plant treatment or disease prevention [9]. Virus population studies often rely on consensus sequences that may hide the minor variant composition (Domingo and Perales, 2019). That is why understanding our abilities (and limits) to detect low frequency variants is important for population studies applications as shown recently on Uganda cassava brown streak virus (UCBSV) for which the analysis of minor SNPs identified three distinct haplotypes with contrasted geographical spread throughout Rwanda [28].

Most of the benchmarks done on SNPs analysis focused mainly on the variant calling step (Barbitoff et al., 2022; Deng et al., 2021; Guirao-Rico and González, 2021) or dealt with specific issues like polyploidy (Clevenger et al., 2015). It was observed that mapping (alignment) seems less impactful than the variant caller (Barbitoff et al., 2022), more amplification cycles led to false positive SNPs detection (Deng et al., 2021) or to a better understanding of the frequencies impact on the variant detection (Guirao-Rico and González, 2021). For plant viruses, reference samples have been designed recently (Tamisier et al., 2021b). They can be used in benchmarking studies for virus detection and genome characterization including, to a lesser extent, SNPs detection, as each sample addresses specific challenges (closely related isolates of the same species, presence of SNPs…). In eukaryotes, variants with frequencies below 5% are considered low-frequency variants and below 1% are rare-frequency variants [16][19]; in this study, we followed these definitions.

The studies mentioned above corresponded to systematic benchmarks carried out by a single or a few research teams. Such small-scale comparisons are necessary, but they will not necessarily reflect the performance of the algorithms once the scientific community uses them. Therefore, a valuable and complementary methodology is to organize performance testing, including many laboratories that routinely use bioinformatic analyses to generate viral genomes and detect SNPs in these genomes. Performance testing of laboratory protocols by end-users has been a common approach for over a decade. It has recently been applied also for bioinformatics analyses focused on detecting plant viruses in 10 HTS samples containing 12 plant viruses that were shared between 21 participants (Sebastien Massart et al., 2019) or for both laboratory and bioinformatics analyses (Gaafar et al., 2021). Performance testing by end-users has been useful in identifying the critical steps in bioinformatics analyses. For example, the detection results obtained (Sebastien Massart et al., 2019) underlined the importance of the reads abundance in the detection as well as expertise in result interpretation. Performance testing is a key step toward the identification of the strength and weaknesses of protocols and contributes to the definition of reliable protocols that could be further applied by many laboratories.

Prior to this publication, preliminary performance testing was conducted on real samples (tomato artificially inoculated by an infectious clone of pepino mosaic virus (PepMV)) to monitor the laboratories' abilities to detect SNPs (unpresented data). This preliminary evaluation delivered key outcomes such as the difficulty in validating the presence of low/rare frequency variants detected by some participants (limiting the evaluation of the performance of the bioinformatics analyses), the very high

complexity of bioinformatics pipelines relying on multiple steps, and the need to focus on well-defined questions. Based on these outcomes, the decision was made to reduce the complexity by generating four artificial samples on only one well-known virus, pepino mosaic virus (PepMV). PepMV is a positive sense ssRNA classified in the species *Pepino mosaic virus* (genus *Potexvirus*, family *Alphaflexiviridae*) known to infect tomato, eggplant and potato plants. Because of the specificity of this study, we created our own artificial samples keeping in mind the good practices shown in the community effort on samples production for bioinformatic validation (Tamisier et al., 2021b) and the usual frequency of PepMV sequences in dataset from infected tomato. In addition, the complexity of pipelines was managed by dividing the bioinformatics pipelines for SNPs detection into 3 fundamental steps. The three steps were decided based on our hypothesis regarding their relative importance on the final results: read quality trimming, read assembly and SNPs calling. In addition, the expert validation of SNPs detected was also independently analyzed.

We describe in this study a performance testing of bioinformatics protocols by the end users, meaning that all the steps from the reception of HTS raw data until the interpretation of SNPs predicted were evaluated. End-users corresponded to 14 plant virology laboratories with different levels of expertise in bioinformatic and variant analyses. The study reported here evaluated the sensitivity and reproducibility of variant analysis applying various bioinformatic strategies currently used in the plant virology community. Each participating laboratory received the same datasets at the same time but without information on its composition in isolates of PePMV (blind test). To precisely decipher the impact of the 3 above-mentioned steps, data were sent in different formats in consecutive steps: raw sequencing reads, quality-controlled reads and aligned reads where the next format was only sent after analysis of the previous was completed. Our approach is, therefore, complementary to the traditional benchmarking methodologies already carried out, as it evaluated the overall bioinformatic analysis, and not only the variant caller algorithm, to show the key points leading to the failure or success of the variant identification and the SNPs manual validation.

Materials and Methods

Origin of the sequencing data

Simulated samples from PepMV were generated using the sequencing reads simulator ART (Huang et al., 2012). ART (V2.5.8) was used for simulating paired-end reads using the built-in quality profile for HiSeq 2500. The reads were artificially generated from the chosen genome sequences, with mutations added manually to the

different samples using multiple criteria (frequencies, length, quality). The detailed samples composition is shown in Table1, in addition, all datasets are freely available (https://doi.org/10.5281/zenodo.7431632).

Sample	Organism	Reference	Haplotype	Number reads	Relative frequency
Sample 1	Solanum lycopersicum	HG975518.1	All	1600000 (2x800000)	80%
	PepMV	DQ000985.1		400000 (2x200000)	20%
	PepMV	DQ000985.1	A339G, A900-, ins6336T	240480 (2 × 120240)	60.12%
			C2668A, G319-, ins2191C	80160 (2 × 40080)	20.04%
			G3830C, A5439-, ins5894C	40080 (2 × 20040)	10.02%
			A2170T, A3769-, ins3912T	24048 (2 × 12024)	6.01%
			A2672C, G4090-	8016 (2 × 4008)	2.00%
			G3371C, T6381-, ins2557C	4008 (2 × 2004)	1.00%
			A5726C, G4518-, ins3885C	2004 (2 × 1002)	0.50%
			A2523G, A2065-, ins2942T	802 (2 × 401)	0.20%
			C5328T, G1580-, ins1334G	402 (2 × 201)	0.10%
Sample 2	Solanum lycopersicum	HG975518.1	All	1600000 (2x800000)	66.66%
	Dark matter	-		400000 (2x200000)	16.66%
	PepMV	DQ000985.1 & AJ606359.1		400000 (2x200000)	16.66%
	PepMV	DQ000985.1	A861C	240480 (2 × 120240)	60.12%
		AJ606359.1	A4918T	80160 (2 × 40080)	20.04%
			C339G	40080 (2 × 20040)	10.02%
			C3830T	24048 (2 × 12024)	6.01%
			T2576G	8016 (2 × 4008)	2.00%
			G2916C	4008 (2 × 2004)	1.00%
			C2523G	2004 (2 × 1002)	0.50%
			T1230C	802 (2 × 401)	0.20%
			G3024T	402 (2 × 201)	0.10%
Sample 3	Solanum lycopersicum	HG975518.1	All	1496036 (2x748018)	57.71%
	Dark matter	-		400000 (2x200000)	14.28%
	ToBRFV	MK648157.1		400000 (2x200000)	14.28%
	PepMV	DQ000985.1		400000 (2x200000)	14.28%
	PepMV	DQ000985.1	no mutation	240480 (2 × 120240)	60.12%
			G1983C, A4031C, A5067G	80160 (2 × 40080)	20.04%
			T217G, G2752A	50502 (2 × 25251)	12.63%
			A5489C, C4768A, ins3021A*	24048 (2 × 12024)	6.01%
			T1258G, T1602A*, del2363A*	4008 (2 × 2004)	1.00%
			A368-, ins1787T, ins3536C	802 (2 × 401)	0.20%

Table 5-1 Composition of samples.

*Sample 1 is composed of tomato plants (Solanum lycopersicum ch06) and the targeted virus; it comprises one isolate and nine variants (26 SNPs in total), with SNPs frequencies ranging from 60% to 0.1%. Sample 2 is a combination of PepMV, tomato plant and dark matter (sequence of unidentified origin); it contains two isolates with nine variants (9 SNPs in total), with frequencies ranging from 60% to 0.1%. Finally, sample 3 is made of PepMV, tomato plant, dark matter and ToBRFV (tomato brown rugose fruit virus); it contains one isolate with six variants (14 SNPs in total), with frequencies ranging from 60% to 0.2%. The SNPs represented with * are SNPs for which the coordinates can be different due to tandem repetition.*

Sample 1 was constructed to estimate the limit of detection of low-frequency SNPs. This sample was generated by adding one bp substitution (sometimes called *single nucleotide variants (SNVs)* by informatic tools), one bp deletion and one bp insertion in the sequence of PepMV, isolate CH2 with reference DQ000985.1. Nine underlying variants were generated, each containing several mutations. For instance, the first variant was generated by substituting an A for a G at position 339, deleting an A at position 900, and inserting a T after position 6336. This sample was relatively simple,

as only one isolate is present with 9 (artificial) variants and was a simple combination of the targeted virus and (host) tomato plant (HG975518.1). It consists of artificial sequencing reads of 125 nucleotides long with a high-quality score (Q20:94%, Q30:88%) (see supplementary file 1). The expected difficulty in SNPs detection is on the low-frequency variants (between 60 and 0.01% frequencies) that might be below the capacity of detection of most participants.

Sample 2 was constructed to evaluate whether non-identified reads not too closely resembling the targeted virus may disturb the variant calling. This sample was generated by introducing mutations on the sequence of two isolates of PepMV - Ch2 (DQ00985.1) and Spanish LE2000 (AJ606359.1; belonging to the EU genotype). In addition, dark matter reads were created and added. Dark matter is defined as sequences of unidentified origin that are not assigned to any known taxonomic group, and may represent a problem in viral metagenomics (Krishnamurthy and Wang, 2017). This dark matter was artificially generated from a randomized sequence following the organism-specific model of *Clostridioides difficile* (NZ_CM000441.1) using the "random-seq" tool of the RSAT platform (November 2019) (Nguyen et al., 2018). Overall, this sample was composed of the targeted virus, tomato plant, and dark matter with reads that were 125 nucleotides long and presented a high quality (Q20:94%, Q30:89%) (see Supplementary file 1). The difficulty of variant detection for this sample was supposed to be identifying the two isolates and detecting the low-frequency variants despite the background noise of the dark matter.

Finally, sample 3 was constructed to examine whether adding sequence noise/dark matter and reads representing a related virus would impact variant detection. Only one isolate was used (DQ000985.1), for which the most frequent (60%) variant is identical to the reference. Variant three was generated at 6% frequency with the reverse of the positive sense RNA of PepMV, therefore the prediction (T925G, G1646T, ins3392T* in forward, and reverse A5489C, C4768A, ins3021A*) can be considered as correct. On variants 3 and 4, there was a mutation in homopolymers (same nucleotide repetition) area, meaning that the coordinate predicted can change a little. As an example, the deletion del2363A* is in a region on which there are four 'A' in a row (2363-2366), meaning the SNPs predictor cannot know which one of the four 'A' was deleted; therefore, any predicted 'A' deletion within those coordinates was considered as correct. In addition to the PepMV, tomato, dark matter (same as sample 2), tomato brown rugose fruit virus (ToBRFV, MK648157.1) –were added (see Table 5-1). All reads on this sample are 250 nucleotides long and present a medium quality (Q20:88%, Q30:59%) (see Supplementary file 1). The background noise (dark matter

and ToBRFV) is more important than sample 2 and might have a bigger impact on the variant calling ability (especially with ToBRFV reads closer to PepMV than the dark matter). The potential impact of the read quality difference could be compared to sample 2. The longer reads coupled with the lower quality should allow more diversity between participants' result of the cleaning of the reads.

Organization of the performance testing

In total, 14 laboratories from eight different countries participated in the performance testing reported here with the coordination of Liège University (Belgium). Each participant was free to choose the algorithms and parameters for each of the three defined steps of the bioinformatics analysis. Importantly, the participants carried out the process step by step, as described in Figure 5-1.

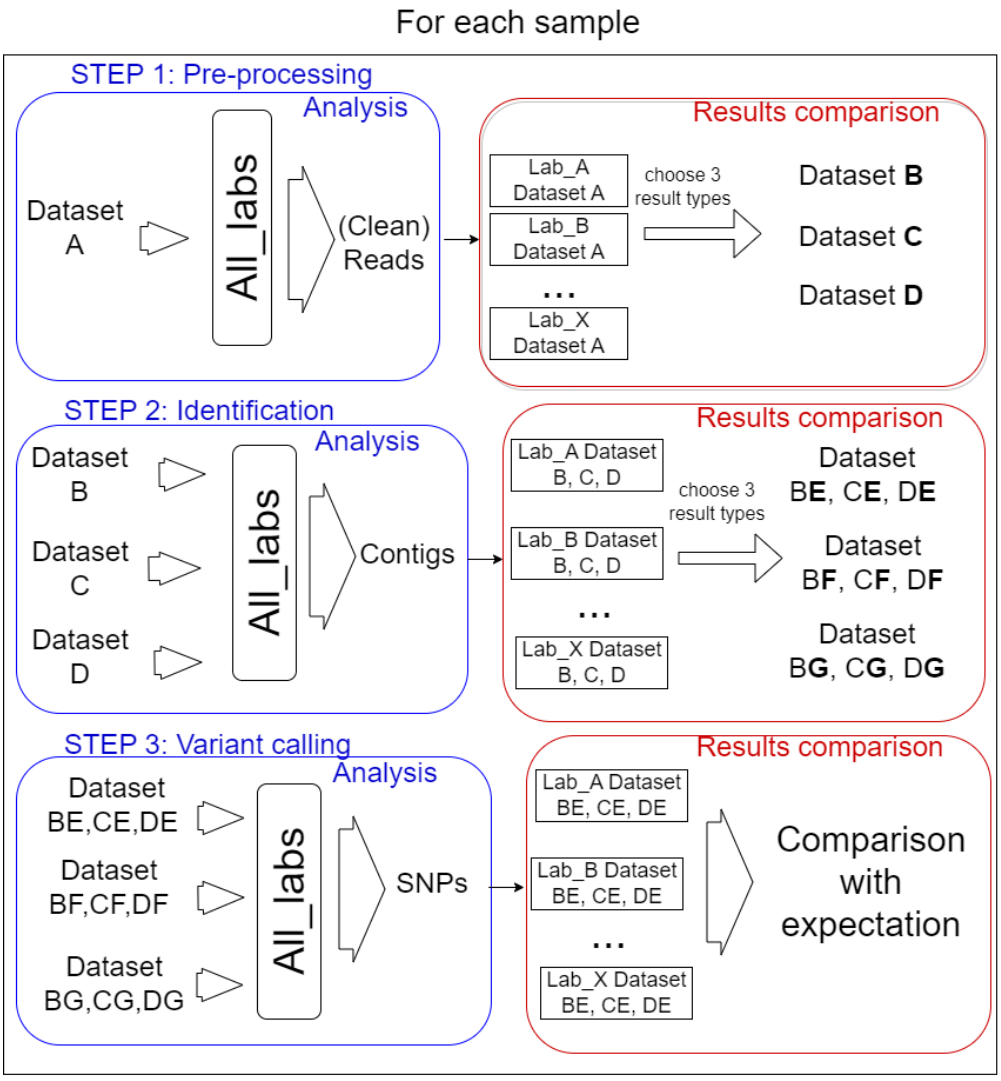


Figure 5-1: Analysis steps.

The test was organized into three steps for each sample: (I) Pre-processing of reads, three datasets were selected from the results to be performed by every lab for the second step. (II) Identification of target virus reference, three resulting mapping files were selected for the last step. (III) Variant calling, the way SNPs predictor tools were used, and manual validation of detection were compared.

In order to prevent too much diversity in the approaches used, a selection of intermediate process results was used as common datasets from which each lab had to start the next step. The overall idea was to allow the comparison of method/laboratory performance as each participant used the same input files for each step. The information on the bioinformatics analyses (anonymized by replacing laboratories name by letters) carried out for all steps by the 14 participating laboratories is available in Supplementary file 2.

I. Pre-processing step

The participants received the three designed datasets in blind. They might, in this step, perform actions to select reads to keep for further analysis, like pairing, merging, trimming, … or nothing at all. The results they were asked to send corresponded to a fastq file for each dataset containing sequences considered ready by the participant for the following steps. In addition, the bioinformatics analysis carried out had to be reported (algorithm and parameters used).

To compare the received fastq files (14 files from each sample = 42 files), the PT coordinator used several criteria (see next chapter). From each original dataset, three files (referred as B, C, and D) showing the most divergent conditions of pre-processing reads (relax cleaning, average cleaning and strict cleaning) were selected for the next step. In addition, the initial files referred to as A could have been used throughout all steps (optional for each participant) and were different for each participant (to try to evaluate the full pipeline impact) as different tools were used. The most pertinent criteria to define average or strict cleaning was decided according to the result observed.

II. Identification

The nine fastq files (3 per sample) were sent to all participants in blind (they just knew the original dataset origin). In addition, the participants carried out the analysis on their own fastq files generated during step 1. The goal for the participants was to identify the appropriate PepMV reference sequence(s) and to align the sequencing reads on it (them) (without the need to search for other viruses). Many different methods could be applied: *de novo* assembly, mapping, Kmer search and dendrogram. The expected result was the mapping of reads on the correct reference(s) in bam format.

The success of this step depended on the participant's ability to identify the correct reference(s). Further on, the coordinator analyzed the 140 bam (alignment) files received (14 laboratories x (3 samples x 3 datasets + own file)) using several criteria: number of reads mapped, the alignment quality, and coverage. From each dataset, three bam files were selected based on the

criteria differences and were called E, F, and G. In total, three bam files per dataset (9 bam files in total) were selected for step 3.

III. Variant calling

The nine bam files were sent in blind to the participants (they only knew the original dataset). In addition, each participant also processed the bam file generated from their own analysis (if any).

This step involved identifying the SNPs present in each bam file by any method of their choice and confirming the detection (manual confirmation).

Criteria for data analysis at each step

For the pre-processing step, the result comparison was based on classical reads statistics: number, quality and length. In addition, the tools and parameters used by the participants (pairing, merging, adapter removing, quality trimming, duplicate trimming) were gathered. The procedure can change the expected output significantly if duplicates are removed and reads are paired and merged, generating a lower number but sometimes longer (when merged) reads. The statistics were obtained from Geneious Prime (2020.0.5, Biomatters), and the procedure (for all steps) is available in Supplementary file 2.

In the second step, the mapping result files were compared using Geneious Prime (2020.0.5, Biomatters) to extract the following statistics: reference used, number of reads mapped, mean coverage, coverage range, identical sites) while keeping track of tools used to examine how they affected the observed results.

Finally, for SNPs calling, the comparison was more challenging as there is no standard way to compare SNPs detection. Due to the high number of files obtained from different sources (n=140), the analysis of the data files was automated by a custom script. Furthermore, the use of terminology in fields of the results files was checked and harmonized between laboratories. As an example, the “Manual validation field” could contain several different words for the same meaning (“YES”, “yes”, “ok”, ...), that needed to be replaced with only one term per meaning.

The inferred SNPs were compared to the expected SNPs using three parameters: (i) mutation type (only SNPs were considered), (ii) position of the SNPs, and (iii) reference/allele correspondence (A/T C/- . . .). The position of each predicted SNPs was compared to the expected position of the mutation in each sample. A tolerance of +/-1 was accepted for all positions as some predictors may start to count with 0 or 1. In addition, some positions were less precise, as a mutation in homopolymer regions can be reported correctly at several coordinates. Finally, for the third variant of sample 3, the mutation and reads were given from the reverse strand, and all positions for

reverse and forward strands were checked. The comparison methodology applied for the analysis of the results is available as a Python notebook (v3.7) here: https://github.com/johrollin/jupyter_variant_calling

Criteria for evaluating SNPs detection.

To be considered "correct", the detection of SNPs had to pass the three parameters described above. To obtain the "correctness" analysis, we made two confusion matrices: the first one is based on the software detection (S). In contrast, the second one uses the manual validation (MV) performed.

Software (S)	Positive	Negative
True	SNPs expected found	-
False	SNPs not expected found	SNPs missed

SNPs exist
SNPs does not exist

Manual Validation (MV)	Positive	Negative
True	SNPs expected & validated (yes)	SNPs not expected & validated (no)
False	SNPs not expected & validated (yes)	SNPs expected & validated (no)

Figure 5-2: Confusion matrices

Confusion matrices used in the frame of the performance testing for SNPs detection. The intersection of the actual and predicted results is displayed with TP=correct hit, TN=correct rejection, FP=wrong detection and FN=missed hit. Values were calculated for software detection (S) and manual validation (MV).

Confusion matrices are layouts that allow the performance monitoring of predictions. On the software part, we considered as True Positive (TPs) all the expected SNPs that were predicted (regardless of their validation status by the participant), the False Positive (FPs) as all the wrong detections of SNPs, the False Negative (FNs) as the number of expected SNPs that were not predicted. Finally, the True Negative (TNs) was not calculated as it did not make sense here (the number of non-predicted SNPs can not be known). We also calculated the True Positive and False negative rates, which correspond to the percentage of expected SNPs found (TP) or missed (FN) among the total number of expected SNPs for the sample (see Table 5-1a).

On the manual validation matrix, a True Positive (TPmv) was an expected SNPs predicted with a positive validation status. A False Positive (FPmv) corresponded to the wrong detection of SNPs that were positively validated. A False Negative (FNmv) was an expected SNPs validated as negative, and, finally, a True Negative (TNmv) represented a wrongly predicted SNPs that was negatively validated (rejected by manual validation). We also calculated the True Positive and False negative rates, which corresponded to the percentage of expected SNPs found and positively validated (TP) or negatively validated (FN) among the total number of expected SNPs predicted by the software for the sample.

Results

Result impacting decision on protocol:

Step 1: reads pre-processing

The objective of the pre-processing step was to evaluate the difference in the read-cleaning strategies of participants and to further evaluate their impact on the next steps of the analysis. We obtained several FastQ files from each participant (65 files in total); the statistics on the reads retained after this step files were studied and are available in Supplementary file 3.

The data transformations carried out by the participants, in various combinations were:

- Pairing: associate the two reads of the same pair together; the modification made on one of the reads may impact the second one.
- Merging: merge the two paired reads to obtain one longer read.
- Adapter trimming: removing small (Illumina) adapters that may be on reads because they can interfere with downstream analyses.
- Quality trimming: remove nucleotides (or read) based on their base calling quality score.
- De-duplication: remove identical reads if several are present.

Not all of these data transformations were performed by all laboratories, and the order in which they were performed was also different. These transformations and their order are important because they can impact the number of reads, their average length and the quality of the remaining reads, which were the metric for the dataset selection for the following step. The individual effect of each transformation on reads is known (Lebas et al., 2022). In our case, it is their combination that is interesting.

The results were analyzed according to two axes. First, the consistency in read number, quality, and length between the original file and the cleaned ones were checked between the three samples. In the second axis, the differences observed in applied methodology and obtained results were compared between laboratories.

An obvious trend is that a higher average base quality is obtained when more nucleotides are removed. Merging was performed by three participants and was identified as a critical step in comparing the different datasets since it significantly changed the statistics. Therefore, those datasets were analyzed separately from the non-merged ones. When merging, the quality of the remaining read is higher, but the number of remaining nucleotides is lower. That is expected since the overlapping part of a pair of reads is used for merging. On the non-merged resulting FastQ files, we saw that similar combination of methods provided similar results, with the slight difference explained by the variation of parameter/version of used tools. No unexpected results were obtained on these steps when analyzed alone (before comparing to other step).

To monitor the eventual effect of the pre-processing on the overall analysis, a selection of datasets to be used by all labs for the next step was performed. The criteria for the selection were discussed and agreed among participants. A combination of metrics corresponding to read number and length (impactful on coverage ability which is believed to be very significant for later variant calling), and quality score was used to select the datasets that are described in Table 5-2. The Illumina adapter effect on the trimming was not evaluated as they are not generated by the read simulator tool.

Condition	Sample	Reads NB	s length r	ean lengt	an confide	Q20	Q30	Q40	NB nucleotides
B	1	979286	30-230	131	38	99%	95%	71%	128377190
B	2	1031979	30-230	133	38	99%	95%	68%	137398669
B	3	1348006	68-400	400	31	94%	73%	19%	538687668
C	1	1246836	20-125	118	37	99%	95%	29%	146452744
C	2	1644810	20-125	119	37	98%	94%	21%	195589523
C	3	898660	20-250	113	30	96%	69%	0%	101325451
D	1	1969770	48-125	116	36	96%	90%	18%	228638480
D	2	2369778	45-125	118	36	96%	91%	15%	278612958
D	3	2660434	50-250	187	29	93%	67%	0%	497181717

Table 5-2 summary of quality metrics of the nine selected datasets for the following steps

The metrics number of reads (Read NB), read length range (minimum to maximum read length), mean (read) length and the number of nucleotides (NB nucleotides) were used to monitor the remaining amount of informative data. Mean confidence (based on confidence scores provided by the base calling program) and the quality score metrics were used to assess the quality/confidence of the data. Quality scores are an estimation of the probability of a base (call) being wrong, Q20 represents 1 in 100 error rates, Q30 is 1 in 1000, Q40 1 in 10000,

meaning that a higher Q score indicates a smaller probability of error, but keeping only the high quality reads means less informative data kept for further analysis.

Condition B was kept for step 2 to include one file with merged reads as its result files were more constant in terms of high quality and read number (operations used to generate each dataset are available in Supplementary file 2). It also presented a higher number of merged reads than the remaining unmerged reads (Supplementary file 3). This dataset represented high-quality reads mostly merged. Conditions C had better quality confidence than average, even if it provided fewer reads than the others, which may play a key role in the last step. Condition D is similar to the majority of other datasets by most metrics. The fact that it did not have any outlier metrics made it a good candidate for the next step as a standard read cleaning case representative.

Step 2: Reference identification and read mapping

The participants aimed to identify the correct reference(s) and to provide a mapping file for further variant calling. Information about those results was extracted and detailed in Supplementary file 4. The key result of this step was reference identification. The participants used a variation of assemblers (Geneious, CLC Genomics Workbench, Tadpole, Velvet, Spades) and/or alignment (BLAST or mapping) approaches. Alternative strategies of *de novo* assembly and mapping to all PepMV references (and no other virus species) were used by one participant, while *de novo* assembly and BLAST to a database containing all PepMV references (and no other virus species) were used by two participants. Most (8) laboratories used an approach that works without the virus prior knowledge with (*de novo*) assembly and alignment (BLAST) to a database of various (plant) viruses. No major difference between strategies or tools was found. In total, of the 14 participants, 11 found the correct (DQ000985.1) reference for sample 1, and 10 labs found the two correct references (DQ000985.1, AJ606359.1) for sample 2 (2 labs found reference DQ000985.1 only). Finally, 11 labs found the correct reference (DQ000985.1) for sample 3. We investigated the methods used to find the reference, focusing on the case(s) where it failed and identifying four different causes for not identifying the correct reference(s):

- The alignment (BLAST) database used for reference identification was too small (the correct references were not present).
- The database (BLAST) contained redundancies, meaning several identical reference IDs were found. This will not impact variant calling ability. Nevertheless, if several identical sequences are used for the mapping, the reads are likely to be divided between references depending on the mapping

options, impacting the respecting SNPs frequencies and the confidence of their detections.

- Several PepMV references are described, while only one was correct (incomplete BLAST selection criteria were applied)
- Only the best BLAST result was kept, meaning that the second reference for sample 2 was not identified.

We also evaluated the resulting mapping against the selected references using statistics extracted from the bam file (read number, identical site, mean coverage, coverage standard deviation, coverage range) (see Supplementary file 4). For samples 1 and 2, the metrics obtained by the laboratories were all close to the expectation. For sample 3, differences due to the impact of step 1 were observed. Table 5-3 displays these differences observed between conditions B, C and D.

Sample	Condition	Step 1 ID	Reference	Nb reads	Identical site	Mean coverage	Coverage range
1	E	B	DQ000985	364068	1660 (25.6%)	7182	15-7966
1	E	C	DQ000985	400001	1569 (24.2%)	7797	15-8611
1	E	D	DQ000985	400003	1571 (24.3%)	7794	15-8611
1	F	B	DQ000985	364145	1467 (22.6%)	7185	15-7967
1	F	C	DQ000985	400001	1485 (22.9%)	7798	15-8611
1	F	D	DQ000985	400007	1484 (22.9%)	7795	15-8611
1	G	B	DQ000985	366495	162 (2.1%)	7197	1-8058
1	G	C	DQ000985	400007	1474 (22.8%)	7718	1-8611
1	G	D	DQ000985	400026	1449 (22.0%)	7677	1-8611
2	E	B	AJ606359	151023	3642 (56.3%)	2957	9-3320
2	E	C	AJ606359	159530	3603 (55.7%)	3091	1-3457
2	E	D	AJ606359	159553	3599 (55.6%)	3091	9-3457
2	F	B	AJ606359	151053	3455 (51.1%)	2803	1-3320
2	F	C	AJ606359	159574	3489 (52.5%)	3006	1-3457
2	F	D	AJ606359	159634	3432 (51.6%)	3002	1-3457
2	G	B	AJ606359	152220	229 (3.0%)	2795	1-3339
2	G	C	AJ606359	162033	3384 (50.1%)	3016	2-3457
2	G	D	AJ606359	162167	3331 (48.7%)	2989	1-3457
2	E	B	DQ000985	225668	2721 (42.3%)	4450	8-4911
2	E	C	DQ000985	240481	2672 (41.5%)	4688	9-5220
2	E	D	DQ000985	240497	2670 (41.5%)	4687	9-5220
2	F	B	DQ000985	225687	2498 (38.4%)	4349	1-4911
2	F	C	DQ000985	240482	2569 (39.9%)	4625	1-5220
2	F	D	DQ000985	240498	2568 (39.5%)	4623	1-5220
2	G	B	DQ000985	227072	169 (2.3%)	4370	1-4972
2	G	C	DQ000985	240486	2581 (39.6%)	4544	1-5220
2	G	D	DQ000985	240502	2545 (39.0%)	4543	1-5220
3	E	B	DQ000985	199995	7 (0.1%)	12503	15-13588
3	E	C	DQ000985	98022	46 (0.7%)	1577	3-1814
3	E	D	DQ000985	391777	18 (0.3%)	11023	14-12033
3	F	B	DQ000985	199997	2 (0.0%)	12478	15-13593
3	F	C	DQ000985	97840	41 (0.6%)	1581	3-1819
3	F	D	DQ000985	393289	14 (0.2%)	11050	14-12079
3	G	B	DQ000985	200004	2 (0.0%)	11772	2-13593
3	G	C	DQ000985	98427	41 (0.6%)	1531	1-1821
3	G	D	DQ000985	393321	8 (0.1%)	10504	1-12079

Table 5-3 mapping metrics of the chosen dataset for the following steps

Three datasets were chosen from the participant results. The number of identical sites corresponds to identical di-nucleotide throughout all reads for a given position on the reference. Mean coverage is the average vertical coverage (also referred to as depth), corresponding to the mean number of reads covering each reference position. The coverage range is the minimum to maximum vertical coverage for each position.

All conditions of samples 1 and 2 presented very homogeneous statistics. Only condition B provided fewer reads as it represented the merged condition (comparison with non-merged should not be done). On the other hand, only sample 3 showed some differences, as condition C (High quality) provided fewer reads than others. In all conditions, the number of identical sites dropped below 1%, probably because sample 3 presented lower-quality reads than the two other samples. Because the difference between statistics was not so high, we decided to select the new conditions (E, F, G) for step 3 based on the tool/parameters used.

Condition E represented the mapping results obtained using CLC genomics (QIAGEN CLC Genomics Workbench) with standard parameters (see lab B mapping parameters in Supplementary file 2) used; the number of reads in that file was close to the expectation (see Table 5-1). Condition F is a case with Geneious standard parameters (see lab D mapping parameters in Supplementary file 2) applied and read numbers also close to expectation. Finally, Condition G used Geneious with a more relaxed parameter, using the same parameter set as dataset F but allowing more mismatches (see lab Z mapping parameters in Supplementary file 2) and presented divergence on identical site numbers.

The reasoning behind this selection was to test if the same SNPs would be predicted between the tools/parameters (meaning that the same reads were mapped at the same position). For example, the comparison between F and G should provide interesting information since the reads number was very close; the differences seen (if any) would be due to the very few read differences.

Step 3: SNPs calling

We obtained 9,624 SNPs detections, which were analyzed by comparing them to the expected SNPs using three parameters (mutation type, position of the SNPs, and reference/allele). All laboratories gave results using the forward strand coordinate, meaning that the reads were automatically reversed by the tool(s) used for the mapping.

Conditions	Combination	TP %	True positive rate	average NB of Snps predicted
BE	Merged highQC + CLC	88%	53%	11.6
BF	Merged highQC + Geneious	78%	54%	11.7
BG	Merged highQC + Geneious relax	59%	52%	29.7
CE	Intermediary QC + CLC	90%	52%	10.5
CF	Intermediary QC + Genious	77%	51%	11.2
CG	Intermediary QC + Genious relax	59%	51%	23.9
DE	Standart QC + CLC	92%	54%	9
DF	Standart QC + Geneious	73%	52%	12.8
DG	Standart QC + Geneious relax	59%	52%	36.6

Table 5-4: Impact of the conditions combination on variant calling detection for all samples.

The condition: B (merged + high-quality control), C (intermediary quality control) and D (Standard quality control) were combined with E (CLC standard mapping parameters), F (Geneious standard mapping parameters) and G (Geneious relaxed mapping parameters) to provide statistics on variant calling detection. The True Positive percentage (TP%) corresponds to the definition of TPs (expected SNPs detected Figure 5-2): percentage of expected SNPs that were predicted on all detections. The true positive rate is the percentage of expected SNPs on all SNPs expected. "average NB" corresponds to the average number of SNPs predicted.

Table 5-4 shows the expected SNPs' results compared to the predicted ones grouped by dataset conditions combined. For example, dataset BE represents the merged reads

(with high quality) mapped with CLC Genomics with standard parameters. By knowing the true positive percentages (TP%) and true positive rate, we can deduce the false positive percentages (FP%) and False negative rate (expected SNPs missed). There are no major differences between alternative pre-processing datasets (B, C, D), with 75% TP on average for all three conditions. On the alternative mapping parameter, the percentage of true positives decreases from dataset E (90% TP) to F (% 76 TP) and G (59% TP). On G (relaxed parameters for mapping), an overprediction of SNPs (last column) was observed and explains the lower percentage.

Samples	Conditions	TP % (all)	TP % (>1%)	True positive rate (all)	True positive rate (>1%)	average NB of Snps predicted (all)	average NB of Snps predicted (> 1%)
Sample1	BE	89%	90%	48%	68%	15/26	14/17
	BF	77%	85%	46%	66%	18/26	13/17
	BG	64%	68%	44%	66%	38/26	19/17
	CE	92%	93%	47%	67%	13/26	12/17
	CF	79%	88%	43%	62%	15/26	12/17
	CG	81%	88%	43%	62%	15/26	12/17
	DE	92%	93%	47%	67%	13/26	12/17
	DF	79%	88%	43%	62%	16/26	12/17
	DG	80%	87%	43%	62%	15/26	13/17
Sample2	BE	87%	95%	56%	76%	6/9	5/6
	BF	74%	76%	62%	85%	8/9	7/6
	BG	30%	30%	59%	81%	42/9	38/6
	CE	87%	91%	56%	76%	6/9	5/6
	CF	64%	63%	62%	85%	10/9	10/6
	CG	15%	14%	61%	83%	48/9	47/6
	DE	94%	97%	61%	82%	6/9	5/6
	DF	50%	49%	62%	85%	15/9	15/6
	DG	14%	13%	62%	85%	58/9	56/6
Sample3	BE	89%	89%	54%	69%	8/14	8/6
	BF	84%	84%	54%	68%	9/14	8/6
	BG	83%	83%	54%	68%	9/14	9/6
	CE	89%	89%	53%	66%	8/14	8/6
	CF	87%	87%	49%	62%	7/14	7/6
	CG	82%	81%	49%	62%	8/14	8/6
	DE	89%	89%	54%	68%	8/14	8/6
	DF	89%	89%	51%	64%	7/14	7/6
	DG	83%	83%	51%	64%	8/14	8/6
Average		**75%**	**77%**	**52%**	**71%**	17.30	14.07

Table 5-5: Impact of the samples and condition on variant calling detection with alternative frequencies filters.

The TP% and true positive rate have the same meaning as in Table 5-4. To evaluate the impact of SNPs with rare-frequency (all frequencies), the performance criteria were also calculated for the SNPs with a frequency higher than 1%. The average NB of SNPs predicted is the number of predicted SNPs / expected number of SNPs.

Contrary to what was expected, sample 3, which was designed to be the most complicated sample, showed the best results (86% TP for all conditions). Sample 1 also showed very good TP (82%), while sample 2 showed only 57% TP (average TP % (all)). If we consider the true positive rate (expected SNPs found), sample 1 performs worst with 45%, sample 3 has 52%, and sample 2 shows the best true positive rate with 60%. There is more overprediction for sample 2 compared to the number of expected SNPs, which allowed us to recover more correct SNPs than for the two other samples. There were some discrepancies between conditions depending on the sample used: BG for sample 1 and condition BG, CG, and DG for sample 2 overpredicted a lot more SNPs than other conditions. This overprediction was always linked to a worse TP/FP balance, while the TP/FN balance was not different. For sample 3, performance metrics are less variable across different conditions. The overpredictions of the relaxed mapping condition (G) are observed in the cases where the number of identical sites was lower for G than other conditions (see Table 5-3).

Most of the participating laboratories correctly identify SNPs with relative frequencies above 1%. This is largely, because as described in Supplementary file 2, most the bioinformatics pipelines are tailored to identify SNPs with frequencies above 1%. The important difference between detections for all frequencies and frequencies above 1% came from the true positive and false negative rates. Indeed, they rose when rare-frequency variants were no longer considered. First, the number of SNPs to consider is 17 (instead of 26) for sample 1, 6 (9) for sample 2 and 11 (14) for sample 3. At the same time, the true positive rate (expected SNPs predicted) rose to 65% (instead of 45%) for sample 1, 82% (60%) for sample 2 and 66% (52%) for sample 3. It indicated that a large part of the expected SNPs that were missed (false negative rate) are the ones with low frequencies. Additional analysis on SNPs frequencies showed that all laboratories obtained different frequency detections compared to the expectation for sample 2 (Supplementary file 5).

Once the SNPs detection was done, the participant could manually check the prediction to confirm or reject the SNPs detection. The criteria used varied between laboratories, but common criteria can be highlighted. Almost all participants visually examined the mapping and focused their efforts on areas more prone to error (mapping low coverage or SNPs with low frequencies). Then they evaluated if the detection might have come from mismapped reads by analyzing the quality of the alignment and quality of the read sequences. More precise information on the manual validation process is available in Supplementary file 2.

Samples	Conditions	TPmv %	FPmv %	TNmv %	FNmv %	True positive rate	average NB of Snps predicted and with validation
Sample1 (Average expected SNP within prediction: 13.04)	BE	79%	5%	5%	11%	88%	13.35
	BF	66%	16%	9%	9%	89%	16.28
	BG	51%	5%	36%	8%	80%	31.5
	CE	83%	4%	1%	12%	87%	11.42
	CF	68%	15%	7%	10%	88%	14
	CG	70%	13%	7%	10%	87%	13.5
	DE	83%	4%	1%	12%	87%	11.42
	DF	68%	15%	7%	10%	88%	13.92
	DG	69%	13%	8%	10%	88%	13.57
Sample2 (Average expected SNP within prediction: 5.97)	BE	83%	9%	5%	4%	96%	5.21
	BF	67%	8%	22%	4%	96%	7.35
	BG	25%	7%	67%	2%	96%	39.85
	CE	83%	13%	0%	4%	96%	5.21
	CF	58%	9%	29%	4%	96%	9.14
	CG	12%	15%	72%	1%	96%	44.14
	DE	93%	3%	0%	4%	96%	5.07
	DF	43%	11%	42%	4%	96%	13.71
	DG	16%	10%	73%	1%	95%	46.14
Sample3 (Average expected SNP within prediction: 8.23)	BE	88%	1%	0%	11%	89%	6.85
	BF	83%	2%	5%	10%	90%	7.5
	BG	82%	1%	6%	10%	90%	7.57
	CE	85%	0%	0%	15%	85%	6.57
	CF	82%	2%	1%	15%	84%	6.42
	CG	76%	2%	8%	15%	84%	7.14
	DE	84%	0%	0%	16%	84%	6.71
	DF	83%	0%	0%	17%	83%	6.35
	DG	75%	0%	8%	17%	83%	7
Average		**69%**	**7%**	**16%**	**9%**	**90%**	13.96

Table 5-6: Manual validation of the variant calling depending on the sample.

Table 5-6 shows the complete confusion matrix (FPmv, TPmv, TNmv, FNmv) for the manual validation of SNPs predicted by the variant caller, allowing the evaluation of the manual validation of the detection. The TPmv corresponds to the expected SNPs found and positively validated, FPmv shows the unexpected SNPs positively validated, TNmv represents unexpected SNPs negatively validated, and finally, FNmv

shows expected SNPs found and negatively validated. Here, the true positive rate represents the expected SNPs found and validated and the false negative rate the expected SNPs found but not validated. In this table, the number of SNPs to consider depends on the amount predicted by the variant calling step.

The average expected SNPs detection was 13 (instead of 26) for sample 1, 6 (9) for sample 2, and 8 (14) for sample 3. Interestingly, the TPmv follows the same pattern as in Table 5-5, where the E condition provides better results than F and G, and the overprediction of condition 1 BG and conditions 2 BG, CG and DG lower the FPmv. This is expected as most of the correct SNPs are missed by the variant calling for those conditions. As an example, in sample 2 condition BG, 25% of the SNPs were correctly found and validated (TPmv), but only 2% were correctly found and rejected by the validation (FNmv). Most of the overprediction was in TNmv (67%) as the validation rejected the wrong detections; only 7% (FPmv) of the wrong detection was validated. That means that the manual validation was very efficient in differentiating between correct and wrong detections. As evidence by the true positive and false negative rates, with true positive rates on average of 87% (sample 1), 96% (sample 2) and 86% (sample 3). The strategy of focusing on key areas of the mapping, and discarding SNPs predicted in a suspicious zone (badly aligned reads), seemed very efficient. Nevertheless, as it relies on manual (visual) examination, the quality might vary depending on the expertise of the scientist.

Discussion

The division of the overall analysis into three steps, from the reception of the reads to the SNPs detection, was very beneficial to structure the comparison between laboratories, conditions and to have checkpoints where the differences between laboratories can be traced to. It also highlighted some important points on specific steps regarding variant calling.

The relative importance of pre-processing

Our results showed that the pre-processing step was not very impactful as not much difference was observed between participants' read files (see Supplementary file 3) or sample treatment. The sample started with Q30:88% | Q20:94% (sample1), Q30:89% | Q20:94% (sample 2) and Q30:59% | Q20:88% (sample 3). The main difference was the merged reads that provided different statistics (fewer but longer reads). Neither the alternative cleaning at a different level of read quality (conditions C and D) nor the merge (condition B) impacted the SNPs detections. No real difference in the

resulting true positive rate was observed in the results. These results seemed to indicate that above a certain minimal quality limit (here Q20 > 90%), any additional quality trimming does not change the variant calling ability. No attempt on trimming with a minimum quality of 30 was made (it would have had a larger impact on sample 3). More tests with lower-quality data should be performed to confirm this hypothesis and identify the quality limit. Additionally, the effect of the presence of Illumina adaptor was not evaluated.

Importance of reference identification

The second step, including the identification of the reference genome to be used for mapping, was very important as, if it failed, the variant calling would be strongly biased. We identified the database composition as the main reason for the possible failure. Indeed, the scope of the reference database is important because if not enough representative sequences are present, the compared sequences might be too distant, and overprediction of SNPs can occur. In fact, the more reference sequences in the database, the better it is for the detection, but the analysis will take longer. In general, using the most closely related reference as possible is more beneficial, but if another (little bit less related) reference is used, this will also allow to reconstruct the genome, and investigate SNPs.

Most participants used a *de novo* assembly to obtain contigs (sequence fragments easier to identify). They then performed a first BLAST against a RefSeq database (including only one representative sequence) to have a first (quick) overview of the putative identity of the contigs. Then, as the contigs linked to PepMV were identified, they used the contigs (all of them or the pre-selected PepMV one) as input for a second BLAST analysis against a larger database containing all PepMV sequences to select the closest reference.

One of the laboratories used an alternative approach that also worked; first, the virus identification was performed using a similar approach (assembly and alignment) through the VirusDetect pipeline (Zheng et al., 2017). Once the PepMV virus was confirmed, all reference sequences for that virus were downloaded, and whole genome alignments with a phylogenetic tree construction were performed. Then one reference from each clade in the tree was selected, and all original reads were mapped to this group of diverse references. Once it was clear which clades of references were present, the procedure was repeated (with different combinations of references) to find the most closely related reference. A possible advantage of this approach is that mapping is faster than BLAST. Also, by analyzing the references in more detail

(phylogenetically), more insight was obtained in the relationships between the publicly available sequences.

Read alignment and variant calling

The alternative mapping parameters comparison (condition E, F or G) showed some differences. Indeed, a similar true positive rate was obtained with each condition but with different levels of background noise (SNPs wrongly predicted). Condition E (mapping with CLC) and condition F (mapping with Geneious) used relatively similar parameters (described in Supplementary file 2), but F gave more false positive than E. Condition G (mapping with Geneious with relax parameter) generated the most background noise with SNPs overpredictions, this was very variable between conditions. We believe that the wrong detections with relaxed parameters can be explained by erroneously adding some reads in the mapping that provided mismatches interpreted by the variant calling tool as SNPs. Even if that hypothesis seems very logical, it does not explain overpredictions in G, nor the differences seen between E and F. Table 5-3 showed that, between the three conditions, the metric on reads was very similar, which made the difference even more surprising. The differences between conditions corresponded to 100 reads, while the mean coverage was close to 10,000 reads (depth). This would indicate that if the wrong reads leading to wrong detections are among the 100 differential reads, we should have most of the false detections near a 1% frequency. Since Table 5-5 showed that most of the wrong detections are not below a 1% frequency, the explanation for the difference in mapping remains unclear. An additional analysis focusing more on mapping methods/parameter differences would be required to understand how the (minor) modifications in the read pool and placement can impact the variant calling afterwards.

Using different variant caller tools did not change the results; the only change observed was in the number of SNPs predicted when the frequency used as the limit of detection was different between laboratories. This differs from most of the variant caller benchmarks (Barbitoff et al., 2022; Deng et al., 2021; Guirao-Rico and González, 2021). But they focus mainly on the variant caller step with only sometimes considering the previous step (mapping or *de novo* assembly), which probably tends to highlight variant caller differences. One other explanation is also that we have a less complex sample (fewer SNPs to predict for each sample) compared to most of the benchmark samples (between 1 3274 SNPs (Deng et al., 2021)).

Impact of sample complexity on detection ability

The samples were designed with different levels of complexity. Sample 1 was simple, with tomato plant and PepMV reads present. The presence of two isolates (with 78% nucleotide identity) in sample 2 did not seem to impact the SNPs detection but rather the frequencies of the prediction compared to the expectation. Sample 3 was supposed to be the more complex as it provided lower sequencing quality (Q30:59%|Q20:88% instead of Q30: 88%|Q20:94%) and the additional viral background noise but with longer reads to compensate. Our method to build simulated samples was similar to the one in [16] but it differed from (Deng et al., 2021) (resequencing from known viral strain mixed). As both the read cleaning step and viral background noise were not very impactful on SNPs detection, it is not surprising that sample 3 (longer reads) provided the best result, and overall, all samples performed well.

Importance of manual validation

A manual validation step might be very important as it allows the biologist to discard SNPs that are not correct and filter the SNPs relevant for downstream analysis. In this study, most of the participants performed a visual examination of the mapping to point out the positions that could be problematic for the variant caller. This method was very efficient since, in all conditions for all samples, it allowed to discard some of the wrongly predicted SNPs. On average, the true positive rate was 90% in table 5-6 (all samples, all conditions). Meaning that the validation confirms 90% of the expected SNPs that the variant caller predicted. The manual validation discarded most of the overprediction (TNmv > 67% for condition G on sample 2). This indicates that most of the over detections were on positions where the alignment was visually doubtful.

The experience with SNPs detection was very variable among participants, as three out of 14 did not employ any manual confirmation. Of the 11 laboratories that performed it, that step was very beneficial for 10 of them (see Supplementary file 6). A too-strict manual filter caused the failure to improve the result with the manual examination (>10% frequencies) for the remaining laboratory. In real life several aspects can make the expert curation more difficult, as too many SNPs can be predicted making manual checking a slow and long process. However, expert validation of SNPs detection is a very important step of the variant calling, and even if performed by non-experienced scientists, it remains beneficial.

Conclusion

In this study, we simulated three plant samples containing PepMV with different variants at several frequencies. Then, a performance study with 14 laboratories was carried out to highlight the key points leading to the failure or success of the SNPs detection in plant viruses. We showed that steps often considered very important (Koboldt, 2020b), like pre-processing, were not that impactful when the base quality was already decent (Q30 > 88% for sample 1 and 2). The effect of the complexity of the datasets was not as conclusive because the dark matter or the viruses mix did not add the expected level of complexity. In addition, the SNPs frequency analysis only showed us that higher frequencies SNPs are easier to predict than low frequencies. The two most important factors corresponded to the strategy to identify the closest reference for mapping and the manual validation of the predicted SNPs. Finally, the mapping parameters impacted our results, with relaxed conditions performing worse. The laboratory obtained overall high TP prediction, with the same SNPs missed by most of them which show a high repeatability on the variant calling. The more experienced participants obtained improved performance thanks to their manual validation. This performance testing is useful to show a variety of strategies leading to variant calling; the community-based performed analysis shows applicable pipelines that can be used to improve end-user variant calling.

Availability:

script: https://github.com/johrollin/jupyter_variant_calling

Data: https://doi.org/10.5281/zenodo.7431632

All supplementary material are available here:

Supplementary_5

All scripts are available here:

GitHub > johrollin > Variant_calling_comparison

Acknowledgment:

We would like to thank Gladys Rufflard for administrative support and all participating author's institution. Special thanks to all participants of the previous COST-DIVAs action that helped this work to start.

Funding

This study is the follow-up of the work on COST Action FA1407 (DIVAS), supported by COST (European Cooperation in Science and Technology). It has been supported by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 813542T (INEXTVIR).

Reference:

Adams, I., Fox, A., 2016. Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. Current Research Topics in Plant Virology 323–335. https://doi.org/10.1007/978-3-319-32919-2_14/FIGURES/3

Adams, I.P., Fox, A., Boonham, N., Massart, S., de Jonghe, K., Tahzima, R., Foucart, Y., Peusens, G., Beliën, T., Massart, S., de Jonghe, K., 2018. The impact of high throughput sequencing on plant health diagnostics. Eur J Plant Pathol 152, 909–919. https://doi.org/10.1007/s10658-018-1570-0

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10, 537. https://doi.org/10.1111/J.1364-3703.2009.00545.X

Adams, I.P., Skelton, A., Macarthur, R., Hodges, T., Hinds, H., Flint, L., Nath, P.D., Boonham, N., Fox, A., 2014. Carrot yellow leaf virus is associated with carrot internal necrosis 9. https://doi.org/10.1371/journal.pone.0109125

Aguado-García, Y., Taboada, B., Morán, P., Rivera-Gutiérrez, X., Serrano-Vázquez, A., Iša, P., Rojas-Velázquez, L., Pérez-Juárez, H., López, S., Torres, J., Ximénez, C., Arias, C.F., 2020. Tobamoviruses can be frequently present in the oropharynx and gut of infants during their first year of life. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-70684-w

Agüero, J., Gómez-Aix, C., Sempere, R.N., García-Villalba, J., García-Núñez, J., Hernando, Y., Aranda, M.A., 2018. Stable and broad spectrum cross-protection against pepino mosaic virus attained by mixed infection. Front Plant Sci 871, 1810. https://doi.org/10.3389/FPLS.2018.01810/BIBTEX

Al Rwahnih, M., Daubert, S., Golino, D., Islas, C., Rowhani, A., 2015. Comparison of Next-Generation Sequencing Versus Biological Indexing for the Optimal Detection of Viral Pathogens in Grapevine. Phytopathology 105, 758–763. https://doi.org/10.1094/PHYTO-06-14-0165-R

Anastassiadou, M., Arena, M., Auteri, D., Brancato, A., Bura, L., Carrasco Cabrera, L., Chaideftou, E., Chiusolo, A., Crivellente, F., de Lentdecker, C., Egsmose, M., Fait, G., Greco, L., Ippolito, A., Istace, F., Jarrah, S., Kardassi, D., Leuschner, R., Lostia, A., Lythgo, C., Magrans, O., Mangas, I., Miron, I., Molnar, T., Padovani, L., Parra Morte, J.M., Pedersen, R., Reich, H., Santos, M., Sharp, R., Szentes, C., Terron, A., Tiramani, M., Vagenende, B., Villamar-Bouza, L., 2021. Peer review of the pesticide risk assessment of the active substances Pepino Mosaic Virus,EU strain, mild isolate Abp1 and Pepino Mosaic Virus,CH2 strain, mild isolate Abp2. EFSA Journal 19, e06388. https://doi.org/10.2903/J.EFSA.2021.6388

Babaian, A., Edgar, R., 2022. Ribovirus classification by a polymerase barcode sequence. PeerJ 10, e14055. https://doi.org/10.7717/PEERJ.14055/SUPP-2

Bačnik, K., Kutnjak, D., Pecman, A., Mehle, N., Tušek Žnidarič, M., Gutiérrez Aguirre, I., Ravnikar, M., 2020. Viromics and infectivity analysis reveal the release of infective plant viruses from wastewater into the environment. Water Res 177, 115628. https://doi.org/10.1016/J.WATRES.2020.115628

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A. v., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Christopher Garcia, K., Grishin, N. v., Adams, P.D., Read, R.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) 373, 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Baert, L., McClure, P., Winkler, A., Karn, J., Bouwknegt, M., Klijn, A., 2021. Guidance document on the use of whole genome sequencing (WGS) for source tracking from a food industry perspective. Food Control 130, 108148. https://doi.org/10.1016/J.FOODCONT.2021.108148

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. https://doi.org/10.1099/JGV.0.001210/CITE/REFWORKS

Bal, A., Pichon, M., Picard, C., Casalegno, J.S., Valette, M., Schuffenecker, I., Billard, L., Vallet, S., Vilchez, G., Cheynet, V., Oriol, G., Trouillet-Assant, S., Gillet, Y., Lina, B., Brengel-Pesce, K., Morfin, F., Josset, L., 2018. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infectious Diseases 2018 18:1 18, 1–10. https://doi.org/10.1186/S12879-018-3446-5

Ballenghien, M., Faivre, N., Galtier, N., 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol 15. https://doi.org/10.1186/S12915-017-0366-6

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. v, Sirotkin, A. v, Vyahhi, N., Tesler, G.,

Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. Viruses 2014, Vol. 6, Pages 106-136 6, 106–136. https://doi.org/10.3390/V6010106

Barbitoff, Y.A., Abasov, R., Tvorogova, V.E., Glotov, A.S., Predeus, A. v., 2022. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics 23, 1–17. https://doi.org/10.1186/S12864-022-08365-3/

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., de Bruyn, R., Maree, H.J., 2021a. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., De Bruyn, R., Maree, H.J., 2021b. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

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., Arboleda, V.A., 2021a. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

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., Arboleda, V.A., 2021b. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

Boheemen, S. van, Rijn, A.L. van, Pappas, N., Carbo, E.C., Vorderman, R.H.P., Sidorov, I., 2020. Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- research that is available on the COVID-19 resource centre - including this with acknowledgement of the origin.

Bomba, L., Walter, K., Soranzo, N., 2017. The impact of rare and low-frequency genetic variants in common disease. Genome Biology 2017 18:1 18, 1–17. https://doi.org/10.1186/S13059-017-1212-4

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., Mumford, R., 2014. Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res 186, 20–31. https://doi.org/10.1016/j.virusres.2013.12.007

Bordería, A. v., Isakov, O., Moratorio, G., Henningsson, R., Agüera-González, S., Organtini, L., Gnädig, N.F., Blanc, H., Alcover, A., Hafenstein, S., Fontes, M., Shomron, N., Vignuzzi, M., 2015. Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. PLoS Pathog 11, e1004838. https://doi.org/10.1371/JOURNAL.PPAT.1004838

Brinkmann, A., Andrusch, A., Belka, A., Wylezich, C., Höper, D., Pohlmann, A., Petersen, T.N., Lucas, P., Blanchard, Y., Papa, A., Melidou, A., Oude Munnink, B.B., Matthijnssens, J., Deboutte, W., Ellis, R.J., Hansmann, F., Baumgärtner, W., van der Vries, E., Osterhaus, A., Camma, C., Mangone, I., Lorusso, A., Marcacci, M., Nunes, A., Pinto, M., Borges, V., Kroneman, A., Schmitz, D., Corman, V.M., Drosten, C., Jones, T.C., Hendriksen, R.S., Aarestrup, F.M., Koopmans, M., Beer, M., Nitsche, A., 2019. Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. J Clin Microbiol 57. https://doi.org/10.1128/JCM.00466-19

Bukowska-Ośko, I., Perlejewski, K., Nakamura, S., Motooka, D., Stokowy, T., Kosińska, J., Popiel, M., Płoski, R., Horban, A., Lipowski, D., Caraballo Cortés, K., Pawełczyk, A., Demkow, U., Stępień, A., Radkowski, M., Laskus, T., 2017. Sensitivity of next-generation sequencing metagenomic analysis for detection of RNA and DNA viruses in cerebrospinal fluid: The confounding effect of background contamination. Adv Exp Med Biol 944, 53–62. https://doi.org/10.1007/5584_2016_42

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8, 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Bushnell, B., Rood, J., Singer, E., 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One 12. https://doi.org/10.1371/JOURNAL.PONE.0185056

Calil, I.P., Fontes, E.P.B., 2017. Plant immunity against viruses: antiviral immune receptors in focus. Ann Bot 119, 711–723. https://doi.org/10.1093/AOB/MCW200

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10. https://doi.org/10.1186/1471-2105-10-421

Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K.A., Giovani, B., Saunders, D.G.O., Kamoun, S., Legg, J.P., Verdier, V., Lessel, J., Neher, R.A., Day, R., Pardey, P., Gullino, M.L., Records, A.R., Bextine, B., Leach, J.E., Staiger, S., Tohme, J., 2019. A global surveillance system for crop diseases: Global preparedness minimizes the risk to food supplies. Science (1979) 364, 1237–1239. https://doi.org/10.1126/SCIENCE.AAW1572/ASSET/52AFE309-57C8-475B-88C6-17D489A74CEE/ASSETS/GRAPHIC/364_1237_F2.JPEG

Černi, S., Ruščić, J., Nolasco, G., Gatin, Ž., Krajačić, M., Škorić, D., 2008. Stem pitting and seedling yellows symptoms of Citrus tristeza virus infection may be determined by minor sequence variants. Virus Genes 36, 241–249. https://doi.org/10.1007/S11262-007-0183-Z/TABLES/1

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021a. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021b. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Champlot, S., Berthelot, C., Pruvost, M., Andrew Bennett, E., Grange, T., Geigl, E.M., 2010. An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. PLoS One 5, e13042. https://doi.org/10.1371/JOURNAL.PONE.0013042

Charlebois, R.L., Sathiamoorthy, S., Logvinoff, C., Gisonni-Lex, L., Mallet, L., Ng, S.H.S., 2020. Sensitivity and breadth of detection of high-throughput sequencing for adventitious virus detection. npj Vaccines 2020 5:1 5, 1–8. https://doi.org/10.1038/s41541-020-0207-4

Chodasewicz, K., 2014. Evolution, reproduction and definition of life. Theory in Biosciences 133, 39. https://doi.org/10.1007/S12064-013-0184-5

Choua, M., Heath, M.R., Speirs, D.C., Bonachela, J.A., 2020. The effect of viral plasticity on the persistence of host-virus systems. J Theor Biol 498, 110263. https://doi.org/10.1016/J.JTBI.2020.110263

Clark, M.F., Adams, A.N., 1977. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. J Gen Virol 34, 475–483. https://doi.org/10.1099/0022-1317-34-3-475

Cleland, C.E., Chyba, C.F., 2002. Defining "life." Orig Life Evol Biosph 32, 387–393. https://doi.org/10.1023/A:1020503324273

Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P., Jackson, S.A., 2015. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant 8, 831–846. https://doi.org/10.1016/J.MOLP.2015.02.002

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., Gabriel, S., 2018a. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 332. https://doi.org/10.1186/s12864-018-4703-0

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., Gabriel, S., 2018b. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 1–10. https://doi.org/10.1186/S12864-018-4703-0/TABLES/2

Cunniffe, N.J., Laranjeira, F.F., Neri, F.M., DeSimone, R.E., Gilligan, C.A., 2014. Cost-Effective Control of Plant Disease When Epidemiological Knowledge Is Incomplete: Modelling Bahia Bark Scaling of Citrus. PLoS Comput Biol 10, e1003753. https://doi.org/10.1371/JOURNAL.PCBI.1003753

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Daubin, V., Szöllősi, G.J., 2016. Horizontal Gene Transfer and the History of Life. Cold Spring Harb Perspect Biol 8. https://doi.org/10.1101/CSHPERSPECT.A018036

De Clerck, C., Crew, K., Van den houwe, I., McMichael, L., Berhal, C., Lassois, L., Haissam Jijakli, M., Roux, N., Thomas, J., Massart, S., 2017. Lessons learned from the virus indexing of Musa germplasm: insights from a multiyear collaboration. Annals of Applied Biology 171, 15–27. https://doi.org/10.1111/AAB.12353

de Jonghe, K., 2020. The world of plant viruses: the good, the bad and the beautiful. Floraculture International 34–36.

Deng, Z.L., Dhingra, A., Fritz, A., Götting, J., Münch, P.C., Steinbrück, L., Schulz, T.F., Ganzenmüller, T., McHardy, A.C., 2021. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. Brief Bioinform 22, 1–12. https://doi.org/10.1093/BIB/BBAA123

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022a. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022b. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Dimmock, N.J., Easton, A.J., Leppard, K.N., 2016. Introduction to Modern Virology, 7th Edition | Wiley.

Domingo, E., Perales, C., 2019. Viral quasispecies. PLoS Genet 15, e1008271. https://doi.org/10.1371/JOURNAL.PGEN.1008271

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J.F., de la Peña, M., Korobeynikov, A., Chikhi, R., Babaian, A., 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 2022 602:7895 602, 142–147. https://doi.org/10.1038/s41586-021-04332-2

Fennell, M., Murphy, J.E., Armstrong, C., Gallagher, T., Osborne, B., 2012. Plant Spread Simulator: A model for simulating large-scale directed dispersal processes across heterogeneous environments. Ecol Modell 230, 1–10. https://doi.org/10.1016/J.ECOLMODEL.2012.01.008

Ferguson, J.M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., Bowen-James, R., Deveson, I.W., 2022. InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. Bioinformatics 38, 1443. https://doi.org/10.1093/BIOINFORMATICS/BTAB846

Filloux, D., Dallot, S., Delaunay, A., Galzi, S., Jacquot, E., Roumagnac, P., 2015. Metagenomics Approaches Based on Virion-Associated Nucleic Acids (VANA): An Innovative Tool for Assessing Without A Priori Viral Diversity of Plants. Methods Mol Biol 1302, 249–257. https://doi.org/10.1007/978-1-4939-2620-6_18

Fontdevila, N., n.d. Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis.

Food, E., Authority, S., Schrader, G., Camilleri, M., Diakaki, M., Vos, S., 2019. Pest survey card on non-European isolates of citrus tristeza virus. EFSA Supporting Publications 16, 1600E. https://doi.org/10.2903/SP.EFSA.2019.EN-1600

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., 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, ppa.12962. https://doi.org/10.1111/ppa.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., Ziebell, H., 2021. Interlaboratory Comparison Study on Ribodepleted Total RNA High-Throughput Sequencing for Plant Virus Diagnostics and Bioinformatic Competence. Pathogens 10, 1174. https://doi.org/10.3390/pathogens10091174

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., Cosson, J.F., 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. mSystems 1, 1–22.

Garcia-Etxebarria, K., Garcia-Garcerà, M., Calafell, F., 2014. Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics 15, 1–10. https://doi.org/10.1186/1471-2105-15-90/FIGURES/4

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022a. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022b. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. v, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., Bilodeau, G.J., van de Vossenberg, B.T.L.H., 2022c. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology 2022, Vol. 11, Page 263 11, 263. https://doi.org/10.3390/BIOLOGY11020263

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G., Thomas, J.E., 2000. Genetic Diversity Among Banana streak virus Isolates from Australia. Phytopathology 90, 921–927. https://doi.org/10.1094/PHYTO.2000.90.8.921

Giovani, B., Blümel, S., Lopian, R., Teulon, D., Bloem, S., Galeano Martínez, C., Beltrán Montoya, C., Urias Morales, C.R., Dharmapuri, S., Timote, V., Horn, N., Chouibani, M., Mezui M'Ella, J.G., Herrera, V., Castinel, A., Goletsos, C., Moeller, C., Naumann, I., Stancanelli, G., Bronzwaer, S., Tramontini, S., MacDonald, P., Matheson, L., Anthoine, G., de Jonghe, K., Schenk, M., Steinmöller, S., Rodriguez, E., Cruz, M.L., Luck, J., Fraser, G., Brunel, S., Montuori, M., Fedchock, C., Steel, E., Pennington, H.G., Day, R., Rossi, J.P., Xia, J., 2020. Science diplomacy for plant health. Nature Plants 2020 6:8 6, 902–905. https://doi.org/10.1038/s41477-020-0744-x

Griffith, F., 1928. The Significance of Pneumococcal Types. J Hyg (Lond) 27, 113. https://doi.org/10.1017/S0022172400031879

Grosdidier, M., Aguayo, J., Marçais, B., Ioos, R., 2017. Detection of plant pathogens using real-time PCR: how reliable are late Ct values? Plant Pathol 66, 359–367. https://doi.org/10.1111/PPA.12591

Guirao-Rico, S., González, J., 2021. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. Mol Ecol Resour 21, 1216. https://doi.org/10.1111/1755-0998.13343

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022a. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022b. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020a. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020b. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanssen, I.M., Paeleman, A., Vandewoestijne, E., van Bergen, L., Bragard, C., Lievens, B., Vanachter, A.C.R.C., Thomma, B.P.H.J., 2009. Pepino mosaic virus isolates and differential symptomatology in tomato. Plant Pathol 58, 450–460. https://doi.org/10.1111/J.1365-3059.2008.02018.X

Harris, H.M.B., Hill, C., 2020. A Place for Viruses on the Tree of Life. Front Microbiol 11. https://doi.org/10.3389/FMICB.2020.604048

Hasiów-Jaroszewska, B., Borodynko, N., Jackowiak, P., Figlerowicz, M., Pospieszny, H., 2011. Single mutation converts mild pathotype of the Pepino mosaic virus into necrotic one. Virus Res 159, 57–61. https://doi.org/10.1016/J.VIRUSRES.2011.04.008

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., Lang, J., 2020. Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. Front Bioeng Biotechnol 8, 817. https://doi.org/10.3389/FBIOE.2020.00817/BIBTEX

Hilaire, J., Tindale, S., Jones, G., Pingarron-Cardenas, G., Bačnik, K., Ojo, M., Frewer, L.J., 2022. Risk perception associated with an emerging agri-food risk in Europe: plant viruses in agriculture. Agric Food Secur 11, 1–19. https://doi.org/10.1186/S40066-022-00366-5/TABLES/3

Hirabara, S.M., Serdan, T.D.A., Gorjao, R., Masi, L.N., Pithon-Curi, T.C., Covas, D.T., Curi, R., Durigon, E.L., 2022. SARS-COV-2 Variants: Differences and Potential of Immune Evasion. Front Cell Infect Microbiol 11, 1401. https://doi.org/10.3389/FCIMB.2021.781429/BIBTEX

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593. https://doi.org/10.1093/BIOINFORMATICS/BTR708

Hull, R., 2014. Plant Virology (Fifth Edition). Plant Virology 3–14. https://doi.org/10.1016/B978-0-12-384871-0.00001-7

Irwin, N.A.T., Pittis, A.A., Richards, T.A., Keeling, P.J., 2021. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nature Microbiology 2022 7:2 7, 327–336. https://doi.org/10.1038/s41564-021-01026-3

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332. https://doi.org/10.1093/BIOINFORMATICS/BTT113

Jeong, J.-J., Ju, H.-J., Noh, J., 2014. A Review of Detection Methods for the Plant Viruses. Research in Plant Disease 20, 173–181. https://doi.org/10.5423/RPD.2014.20.3.173

Jones, R.A.C., 2014. Trends in plant virus epidemiology: Opportunities from new or improved technologies. Virus Res 186, 3–19. https://doi.org/10.1016/J.VIRUSRES.2013.11.003

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalimuthu, K., Arivalagan, J., Mohan, M., Samuel Selvan Christyraj, J.R., Arockiaraj, J., Muthusamy, R., Ju, H.J., 2022. Point of care diagnosis of plant virus: Current trends and prospects. Mol Cell Probes 61, 101779. https://doi.org/10.1016/J.MCP.2021.101779

Kechin, A., Boyarskikh, U., Kel, A., Filipenko, M., 2017. CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. Journal of Computational Biology 24, 1138–1143. https://doi.org/10.1089/CMB.2017.0096/ASSET/IMAGES/LARGE/FIGURE3.JPEG

Koboldt, D.C., 2020a. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Koboldt, D.C., 2020b. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Kogovšek, P., Pompe-Novak, M., Baebler, S., Rotter, A., Gow, L., Gruden, K., Foster, G.D., Boonham, N., Ravnikar, M., 2010. Aggressive and mild Potato virus Y isolates trigger different specific responses in susceptible potato plants. Plant Pathol 59, 1121–1132. https://doi.org/10.1111/J.1365-3059.2010.02340.X

Koonin, E. v., Dolja, V. v., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., Kuhn, J.H., 2020. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84. https://doi.org/10.1128/MMBR.00061-19

Kreuze, J., 2014. siRNA deep sequencing and assembly: Piecing together viral infections. Detection and Diagnostics of Plant Pathogens 21–38. https://doi.org/10.1007/978-94-017-9020-8_2/FIGURES/3

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1–7. https://doi.org/10.1016/J.VIROL.2009.03.024

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res 239, 136–142. https://doi.org/10.1016/J.VIRUSRES.2017.02.002

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., 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 . https://doi.org/10.3390/plants10020328

Kumar, R., Nagpal, S., Kaushik, S., Mendiratta, S., 2020. COVID-19 diagnostic approaches: different roads to the same destination. VirusDisease 2020 31:2 31, 97–105. https://doi.org/10.1007/S13337-020-00599-7

Kutnjak, D., Rupar, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J.F., Ravnikar, M., 2015. Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. J Virol 89, 4760–4769. https://doi.org/10.1128/JVI.03685-14/SUPPL_FILE/ZJV999090298SO1.PDF

Kutnjak, D., Tamisier, L., Adams, I., Boonham, N., Candresse, T., Chiumenti, M., de Jonghe, K., Kreuze, J.F., Lefebvre, M., Silva, G., Malapi-wight, M., Margaria, P., Pleško, I.M., McGreig, S., Miozzi, L., Remenant, B., Reynard, J.S., Rollin, J., Rott, M., Schumpp, O., Massart, S., Haegeman, A., 2021. A Primer on the Analysis of High-Throughput Sequencing Data for Detection of Plant Viruses. Microorganisms 9. https://doi.org/10.3390/MICROORGANISMS9040841

Kutschera, U., Niklas, K.J., 2004. The modern theory of biological evolution: An expanded synthesis. Naturwissenschaften 91, 255–276. https://doi.org/10.1007/S00114-004-0515-Y/FIGURES/8

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 optimization of high-throughput sequencing bioinformatics pipelines for virus detection. Viruses 10. https://doi.org/10.3390/v10100528

Lau, H.Y., Botella, J.R., 2017. Advanced DNA-based point-of-care diagnostic methods for plant diseases detection. Front Plant Sci 8, 2016. https://doi.org/10.3389/FPLS.2017.02016/BIBTEX

Lebas, B., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G., Blouin, A., Boonham N., Candresse, T., Chandelier, A., De Jonghe, K., Fox, A., Gaafar, Y., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., Landa, B., Li, Massart, S., n.d. Guidelines for preparing a laboratory to use high-throughput sequencing as routine diagnostic test: general and technical recommendations. EPPO Bulletin.

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., 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 Bulletin 52, 394–418. https://doi.org/10.1111/epp.12863

Lefebvre, M., Theil, S., Ma, Y., Candresse, T., 2019. The VirAnnot Pipeline: A Resource for Automated Viral Diversity Estimation and Operational Taxonomy Units Assignation for Virome Sequencing

Data. https://doi.org/10.1094/PBIOMES-07-19-0037-A 3, 256–259. https://doi.org/10.1094/PBIOMES-07-19-0037-A

Legrand, P., 2015. Biological assays for plant viruses and other graft-transmissible pathogens diagnoses: a review. EPPO Bulletin 45, 240–251. https://doi.org/10.1111/EPP.12222

Li, C.-X., Li, W., Zhou, J., Zhang, B., Feng, Y., Xu, C.-P., Lu, Y.-Y., Holmes, E.C., Shi, M., 2020. High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory infection. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-60992-6

Li, H., Durbin, R., 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, H., Durbin, R., 2009b. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Liefting, L.W., Waite, D.W., Thompson, J.R., 2021. Application of Oxford Nanopore Technology to Plant Virus Detection. Viruses 13. https://doi.org/10.3390/V13081424

Ling, K.S., Scott, J.W., 2007. Sources of Resistance to Pepino mosaic virus in Tomato Accessions. https://doi.org/10.1094/PDIS-91-6-0749 91, 749–753. https://doi.org/10.1094/PDIS-91-6-0749

Low, A.J., Koziol, A.G., Manninger, P.A., Blais, B., Carrillo, C.D., 2019. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 7, e6995. https://doi.org/10.7717/peerj.6995

Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci 2017, e104. https://doi.org/10.7717/PEERJ-CS.104/SUPP-5

Luisi, P.L., 1998. About various definitions of life. Orig Life Evol Biosph 28, 613–622. https://doi.org/10.1023/A:1006517315105

Maachi, A., Torre, C., Sempere, R.N., Hernando, Y., Aranda, M.A., Donaire, L., 2021. Use of High-Throughput Sequencing and Two RNA Input Methods to Identify Viruses Infecting Tomato Crops. Microorganisms . https://doi.org/10.3390/microorganisms9051043

Mackenzie, J.S., Jeggo, M., 2019. The One Health Approach—Why Is It So Important? Trop Med Infect Dis 4. https://doi.org/10.3390/TROPICALMED4020088

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C.M., Roumagnac, P., van der Vlugt, R., Massart, S., 2020. Illuminating an Ecological Blackbox: Using High Throughput Sequencing to Characterize the Plant Virome Across Scales. Front Microbiol 11. https://doi.org/10.3389/fmicb.2020.578064

Maksimovic Carvalho Ferreira, O., Lengar, Ž., Kogej, Z., Bačnik, K., Bajde, I., Milavec, M., Županič, A., Mehle, N., Kutnjak, D., Ravnikar, M., Gutierrez-Aguirre, I., 2022. Evaluation of Methods and Processes for Robust Monitoring of SARS-CoV-2 in Wastewater. Food Environ Virol 14, 384–400. https://doi.org/10.1007/S12560-022-09533-0/FIGURES/4

Malapi-Wight, M., Adhikari, B., Zhou, J., Hendrickson, L., Maroon-Lango, C.J., McFarland, C., Foster, J.A., Hurtado-Gonzales, O.P., 2021. HTS-Based Diagnostics of Sugarcane Viruses: Seasonal Variation and Its Implications for Accurate Detection. Viruses . https://doi.org/10.3390/v13081627

Maliogka, V.I., Minafra, A., Saldarelli, P., Ruiz-García, A.B., Glasa, M., Katis, N., Olmos, A., 2018. Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput

Sequencing Technologies. Viruses 2018, Vol. 10, Page 436 10, 436. https://doi.org/10.3390/V10080436

Maljkovic Berry, I., Melendrez, M.C., Bishop-Lilly, K.A., Rutvisuttinunt, W., Pollett, S., Talundzic, E., Morton, L., Jarman, R.G., 2020. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis 221, S292–S307. https://doi.org/10.1093/INFDIS/JIZ286

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018a. Viral double-stranded RNAs (dsRNAs) from Plants: Alternative nucleic acid substrates for high-throughput sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4/COVER

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018b. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4

Maree, H.J., Fox, A., Al Rwahnih, M., Boonham, N., Candresse, T., 2018. Application of hts for routine plant virus diagnostics: state of the art and challenges. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.01082

Martelli, G.P., 2017. An overview on grapevine viruses, viroids, and the diseases they cause. Grapevine Viruses: Molecular Biology, Diagnostics and Management 31–46. https://doi.org/10.1007/978-3-319-57706-7_2/TABLES/1

Martínez, M., de Viedma, D.G., Alonso, M., Andrés, S., Bouza, E., Cabezas, T., Cabeza, I., Reyes, A., Sánchez-Yebra, W., Rodríguez, M., Sánchez, M.I., Rogado, M.C., Fernández, R., Peñafiel, T., Martínez, J., Barroso, P., Lucerna, M.A., Diez, L.F., Gutiérrez, C., 2006. Impact of Laboratory Cross-Contamination on Molecular Epidemiology Studies of Tuberculosis. J Clin Microbiol 44, 2967–2969. https://doi.org/10.1128/JCM.00754-06

Marttunen, M., Lienert, J., Belton, V., 2017. Structuring problems for Multi-Criteria Decision Analysis in practice: A literature review of method combinations. Eur J Oper Res 263, 1–17. https://doi.org/10.1016/J.EJOR.2017.04.041

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, Jonkers W., Kreuze J., Kutnjak D., Landa B., n.d. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests | Zenodo [WWW Document]. URL https://zenodo.org/record/6637519 (accessed 7.11.22).

Massart, S., Adams, I., Rwahnih, A., Baeyen, S., Bilodeau, G.J., Blouin, A.G., Boonham, N., Candresse, T., Chandellier, 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.B., Liu, M., Maclot, F., Malapi-Wight, M., Maree, H.J., Martoni, F., Mehle, N., Minafra, A., Mollov, D., Moreira, A.G., Nakhla, M., Petter, F., Piper, A.M., Ponchart, J.P., Rae, R., Remenant, B., Rivera, Y., Rodoni, B., Botermans, M., 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., van de Vossenberg, B.T.L.H., Westenberg, M., Wetzel, T., Ziebell, H., Lebas, B.S.M., al Rwahnih, M., 2022a. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. Peer Community Journal 2, 37. https://doi.org/10.24072/PCJOURNAL.181

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J.S., Rumbou, A., Saldarelli, P., Škoric, D., Vainio, E.J., Valkonen, J.P.T., 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. Front Microbiol 8, 45. https://doi.org/10.3389/FMICB.2017.00045/BIBTEX

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, Y., Candresse, T., 2019. Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. Phytopathology 109. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, Sebastien, 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, 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. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, S., Lebas, B., Chabirand, A., Chappé, A.-M., Dreo, T., Faggioli, F., Harrison, C., Macarthur, R., Mehle, N., Mezzalama, M., Petter, F., Ravnikar, M., Renvoisé, J.-P., Spadaro, D., Tomassoli, L., Tomlinson, J., Trontin, C., Vlugt, R. van der, Vučurović, A., Weekes, R., Brostaux, Y., 2022b. Guidelines for improving statistical analyses of validation datasets for plant pest diagnostic tests. EPPO Bulletin 52, 419–433. https://doi.org/10.1111/EPP.12862

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014a. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014b. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, Sebastien, Olmos, A., Jijakli, H., Candresse, T., 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188, 90–96. https://doi.org/10.1016/J.VIRUSRES.2014.03.029

Minic, R., Zivkovic, I., Minic, R., Zivkovic, I., 2020. Optimization, Validation and Standardization of ELISA. Norovirus. https://doi.org/10.5772/INTECHOPEN.94338

Modrow, S., Falke, D., Truyen, U., Schätzl, H., 2013. Viruses: Definition, Structure, Classification. Molecular Virology 17. https://doi.org/10.1007/978-3-642-20718-1_2

Moonen, P., Boonstra, J., Hakze- Van Der Honing, R., Boonstra- Leendertse, C., Jacobs, L., Dekker, A., 2003. Validation of a LightCycler-based reverse transcription polymerase chain reaction for the detection of foot-and-mouth disease virus. J Virol Methods 113, 35–41. https://doi.org/10.1016/S0166-0934(03)00220-9

Moreira, D., López-García, P., 2009. Ten reasons to exclude viruses from the tree of life. Nature Reviews Microbiology 2009 7:4 7, 306–311. https://doi.org/10.1038/nrmicro2108

Navarro, R., Ambrós, S., Butković, A., Carrasco, J.L., González, R., Martínez, F., Wu, B., Elena, S.F., 2022. Defects in plant immunity modulate the rates and patterns of RNA virus evolution. Virus Evol 8. https://doi.org/10.1093/VE/VEAC059

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D.A., Bhaya, D., Narrowe, A.B., Probst, A.J., Sczyrba, A., Kohler, A., Séguin, A., Shade, A., Campbell, B.J., Lindahl, B.D., Reese, B.K., Roque, B.M., DeRito, C., Averill, C., Cullen, D., Beck, D.A.C., Walsh, D.A., Ward, D.M., Wu, D., Eloe-Fadrosh, E., Brodie, E.L., Young, E.B., Lilleskov, E.A., Castillo, F.J., Martin, F.M., LeCleir, G.R., Attwood, G.T., Cadillo-Quiroz, H., Simon, H.M., Hewson, I., Grigoriev, I. v., Tiedje, J.M., Jansson, J.K., Lee, J., VanderGheynst, J.S., Dangl, J., Bowman, J.S., Blanchard, J.L., Bowen, J.L., Xu, J., Banfield, J.F., Deming, J.W., Kostka, J.E., Gladden, J.M., Rapp, J.Z., Sharpe, J., McMahon, K.D., Treseder, K.K., Bidle, K.D., Wrighton, K.C., Thamatrakoln, K., Nusslein, K., Meredith, L.K., Ramirez, L., Buee, M., Huntemann, M., Kalyuzhnaya, M.G., Waldrop, M.P., Sullivan, M.B., Schrenk, M.O., Hess, M., Vega, M.A., O'Malley, M.A., Medina, M., Gilbert, N.E., Delherbe, N., Mason, O.U., Dijkstra, P., Chuckran, P.F., Baldrian, P., Constant, P., Stepanauskas, R., Daly, R.A., Lamendella, R., Gruninger, R.J., McKay, R.M., Hylander, S., Lebeis, S.L., Esser, S.P., Acinas, S.G., Wilhelm, S.S., Singer, S.W., Tringe, S.S., Woyke, T., Reddy, T.B.K., Bell, T.H., Mock, T., McAllister, T., Thiel, V., Denef, V.J., Liu, W.T., Martens-Habbena, W., Allen Liu, X.J., Cooper, Z.S., Wang, Z., Krupovic, M., Dolja, V. v., Kyrpides, N.C., Koonin, E. v., Gophna, U., 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell 185, 4023-4037.e18. https://doi.org/10.1016/J.CELL.2022.08.023

Ng, S.H., Braxton, C., Eloit, M., Feng, S.F., Fragnoud, R., Mallet, L., Mee, E.T., Sathiamoorthy, S., Vandeputte, O., Khan, A.S., 2018. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: Upstream sample processing and library preparation [WWW Document]. Viruses. https://doi.org/10.3390/v10100566

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., van Helden, J., Medina-Rivera, A., Thomas-Chollier, M., 2018. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 46, W209–W214. https://doi.org/10.1093/NAR/GKY317

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., Massart, S., 2018a. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

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., Massart, S., 2018b. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

Oñate-Sánchez, L., Vicente-Carbajosa, J., 2008. DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC Res Notes 1, 1–7. https://doi.org/10.1186/1756-0500-1-93/FIGURES/3

Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology 2021 22:1 22, 1–19. https://doi.org/10.1186/S13059-021-02393-0

Pappas, N., Roux, S., Hölzer, M., Lamkiewicz, K., Mock, F., Marz, M., Dutilh, B.E., 2021. Virus Bioinformatics. Encyclopedia of Virology 124–132. https://doi.org/10.1016/B978-0-12-814515-9.00034-5

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017a. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017b. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

PM 7/76 (5) Use of EPPO Diagnostic Standards, 2018. . EPPO Bulletin 48, 373–377. https://doi.org/10.1111/EPP.12506

PM 7/98 (5) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity, 2021. . EPPO Bulletin 51, 468–498. https://doi.org/10.1111/EPP.12780

Ramesh, S., Govindarajulu, M., Parise, R.S., Neel, L., Shankar, T., Patel, S., Lowery, P., Smith, F., Dhanasekaran, M., Moore, T., 2021. Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy. Vaccines (Basel) 9. https://doi.org/10.3390/VACCINES9101195

Ricciuti, E., Laboureau, N., Noumbissié, G., Chabannes, M., Sukhikh, N., Pooggin, M.M., Iskra-Caruana, M.-L., 2021. Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. Journal of General Virology 102, 001670. https://doi.org/10.1099/JGV.0.001670

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N. v., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Martin, M.D., MacDonald, G.K., Neenan, P., Records, A., Schmale, D.G., Tateosian, L., Wei, Q., 2021. The persistent threat of emerging plant disease pandemics to global food security. Proc Natl Acad Sci U S A 118, e2022239118. https://doi.org/10.1073/PNAS.2022239118/SUPPL_FILE/PNAS.2022239118.SAPP.PDF

Rivarez, M., Faure, C., Svanella-Dumas, L., Pecman, A., Tušek-Žnidaric, M., Schönegger, D., de Jonghe, K., Blouin, A., Massart, S., Ravnikar, M., Kutnjak, D., Marais, A., Candresse, T., 2022. Diversity and pathobiology of an ilarvirus unexpectedly detected in diverse host plants and in global sequencing data. https://doi.org/10.1101/2022.12.15.520526

Rivarez, M.P.S., Vučurović, A., Mehle, N., Ravnikar, M., Kutnjak, D., 2021. Global Advances in Tomato Virome Research: Current Status and the Impact of High-Throughput Sequencing. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.671925

Rizzo, D.M., Lichtveld, M., Mazet, J.A.K., Togami, E., Miller, S.A., 2021. Plant health and its effects on food safety and security in a One Health framework: four case studies. One Health Outlook 2021 3:1 3, 1–9. https://doi.org/10.1186/S42522-021-00038-7

Roberts, C.A., Dietzgen, R.G., Heelan, L.A., MaclLean, D.J., 2000. Real-time RT-PCR fluorescent detection of tomato spotted wilt virus. J Virol Methods 88, 1–8. https://doi.org/10.1016/S0166-0934(00)00156-7

Rong, W., Rollin, J., Hanafi, M., Roux, N., Massart, S., 2022. Validation of high throughput sequencing as virus indexing test for Musa germplasm: performance criteria evaluation and contamination monitoring using an alien control. PhytoFrontiers. https://doi.org/https://doi.org/10.1094/PHYTOFR-03-22-0030-FI

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarría, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19 Suppl 1, 81–88. https://doi.org/10.1111/J.1365-294X.2009.04470.X

Rosenthal, S.H., Gerasimova, A., Ruiz-Vega, R., Livingston, K., Kagan, R.M., Liu, Y., Anderson, B., Owen, R., Bernstein, L., Smolgovsky, A., Xu, D., Chen, R., Grupe, A., Tanpaiboon, P., Lacbawan, F., 2022. Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory. Scientific Reports 2022 12:1 12, 1–12. https://doi.org/10.1038/s41598-022-06091-0

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017a. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017b. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rubio, L., Galipienso, L., Ferriol, I., 2020. Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution. Front Plant Sci 11, 1092. https://doi.org/10.3389/FPLS.2020.01092/BIBTEX

Ruiz-Mirazo, K., Peretó, J., Moreno, A., 2004. A universal definition of life: autonomy and open-ended evolution. Orig Life Evol Biosph 34, 323–346. https://doi.org/10.1023/B:ORIG.0000016440.53346.DC

Rybicki, E.P., 2015. A Top Ten list for economically important plant viruses. Arch Virol 160, 17–20. https://doi.org/10.1007/s00705-014-2295-9

Sabatier, M., Bal, A., Destras, G., Regue, H., Quéromès, G., Cheynet, V., Lina, B., Bardel, C., Brengel-Pesce, K., Navratil, V., Josset, L., 2020. Comparison of Nucleic Acid Extraction Methods for a Viral Metagenomics Analysis of Respiratory Viruses. Microorganisms 8, 1–18. https://doi.org/10.3390/MICROORGANISMS8101539

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94. https://doi.org/10.1016/0022-2836(75)90213-2

Sangiovanni, M., Granata, I., Thind, A.S., Guarracino, M.R., 2019. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics 20. https://doi.org/10.1186/S12859-019-2684-X

Santala, J., Valkonen, J.P.T., 2018. Sensitivity of Small RNA-Based Detection of Plant Viruses . Frontiers in Microbiology .

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution 2019 3:3 3, 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schilbert, H.M., Rempel, A., Pucker, B., 2020. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants 2020, Vol. 9, Page 439 9, 439. https://doi.org/10.3390/PLANTS9040439

Silva, J.M.F., Melo, F.L., Elena, S.F., Candresse, T., Sabanadzovic, S., Tzanetakis, I.E., Blouin, A.G., Villamor, D.E.V., Mollov, D., Constable, F., Cao, M., Saldarelli, P., Cho, W.K., Nagata, T., 2022. Virus classification based on in-depth sequence analyses and development of demarcation criteria using the Betaflexiviridae as a case study. J Gen Virol 103, 001806. https://doi.org/10.1099/JGV.0.001806/CITE/REFWORKS

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J.C., Manuel, M., Philippe, H., Telford, M.J., 2018. A software tool "CroCo" detects pervasive cross-species contamination in next generation sequencing data. BMC Biol 16, 1–9. https://doi.org/10.1186/s12915-018-0486-7

Simmonds, P., Adams, M.J., Benk, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E. v., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology 2017 15:3 15, 161–168. https://doi.org/10.1038/nrmicro.2016.177

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., Weissman, I.L., 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724. https://doi.org/10.1101/125724

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., Al Rwahnih, M., 2021a. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., al Rwahnih, M., 2021b. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. https://doi.org/10.1038/nbt.3988

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014a. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014b. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., Nikolski, M., 2022. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. Frontiers in Bioinformatics 2. https://doi.org/10.3389/FBINF.2022.867111/FULL

Tamisier, L., Colson, C., Maclot, F., Zhang, P., Wang, X., Francis, F., Baurain, D., Massart, S., n.d. Deciphering the intra-host genetic diversity of viruses: insights from the experimental evolution of a plant virus.

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., Massart, S., 2021a. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/pcjournal.62

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., Massart, S., 2021b. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/PCJOURNAL.62

Tedersoo, L., Albertsen, M., Anslan, S., Callahan, B., 2021. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl Environ Microbiol 87, 1–19. https://doi.org/10.1128/AEM.00626-21/SUPPL_FILE/AEM.00626-21-S0001.PDF

Tejano, L.A., Peralta, J.P., Yap, E.E.S., Panjaitan, F.C.A., Chang, Y.W., 2019. Prediction of Bioactive Peptides from Chlorella sorokiniana Proteins Using Proteomic Techniques in Combination with Bioinformatics Analyses. International Journal of Molecular Sciences 2019, Vol. 20, Page 1786 20, 1786. https://doi.org/10.3390/IJMS20071786

Tekle, Y.I., Grant, J.R., Kovner, A.M., Townsend, J.P., Katz, L.A., 2010. Identification of new molecular markers for assembling the eukaryotic tree of life. Mol Phylogenet Evol 55, 1177–1182. https://doi.org/10.1016/J.YMPEV.2010.03.010

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., Roux, N., 2015. Technical guidelines for the safe movement of Musa germplasm, in: Thomas, J. (Ed.), Technical Guidelines for the Safe Movement of Musa Germplasm. Bioversity International, > >.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/NAR/GKAB1061

Velasco, L., Padilla, C. V., 2021. High-Throughput Sequencing of Small RNAs for the Sanitary Certification of Viruses in Grapevine. Front Plant Sci 12. https://doi.org/10.3389/FPLS.2021.682879/FULL

Vereecke, N., Carnet, F., Pronost, S., Vanschandevijl, K., Theuns, S., Nauwynck, H., 2021. Genome Sequences of Equine Herpesvirus 1 Strains from a European Outbreak of Neurological Disorders

Linked to a Horse Gathering in Valencia, Spain, in 2021. Microbiol Resour Announc. https://doi.org/10.1128/MRA.00333-21

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016a. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016b. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

VUNSH, R., ROSNER, A., STEIN, A., 1990. The use of the polymerase chain reaction (PCR) for the detection of bean yellow mosaic virus in gladiolus. Annals of Applied Biology 117, 561–569. https://doi.org/10.1111/J.1744-7348.1990.TB04822.X

Waite, D.W., Liefting, L., Delmiglio, C., Chernyavtseva, A., Ha, H.J., Thompson, J.R., 2022. Development and Validation of a Bioinformatic Workflow for the Rapid Detection of Viruses in Biosecurity. Viruses 14, 2163. https://doi.org/10.3390/V14102163/S1

Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P., Dempsey, D.M., Dutilh, B.E., García, M.L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., Łobocka, M., Oksanen, H.M., Orton, R.J., Robertson, D.L., Rubino, L., Sabanadzovic, S., Simmonds, P., Smith, D.B., Suzuki, N., van Doorslaer, K., Vandamme, A.M., Varsani, A., Zerbini, F.M., 2022. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch Virol 167, 2429–2440. https://doi.org/10.1007/S00705-022-05516-5/TABLES/2

Ward, M., 2016. Action against pest spread—the case for retrospective analysis with a focus on timing. Food Secur 8, 77–81. https://doi.org/10.1007/S12571-015-0532-Z/FIGURES/1

Watzinger, F., Ebner, K., Lion, T., 2006. Detection and monitoring of virus infections by real-time PCR. Mol Aspects Med 27, 254–298. https://doi.org/10.1016/J.MAM.2005.12.001

Whitfield, A.E., Campbell, L.R., Sherwood, J.L., Ullman, D.E., 2003. Tissue Blot Immunoassay for Detection of Tomato spotted wilt virus in Ranunculus asiaticus and Other Ornamentals. Plant Dis 87, 618–622. https://doi.org/10.1094/PDIS.2003.87.6.618

Williams, R.C., 1954. Electron Microscopy of Viruses. Adv Virus Res 2, 183–239. https://doi.org/10.1016/S0065-3527(08)60533-3

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 762302. https://doi.org/10.1186/s13059-019-1891-0

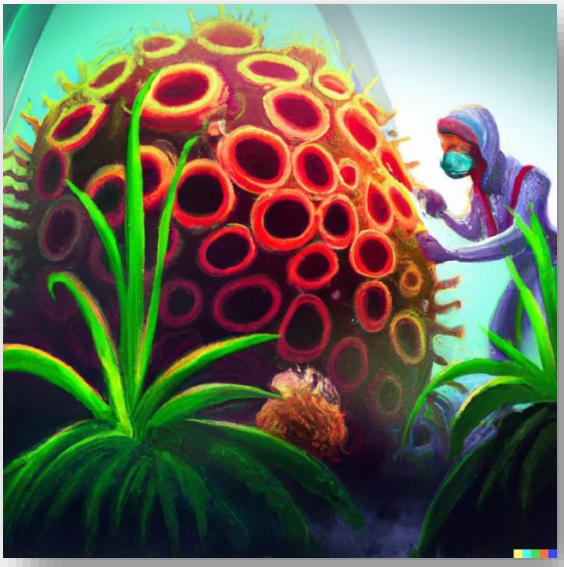
Wood, D.E., Salzberg, S.L., 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, 1–12. https://doi.org/10.1186/GB-2014-15-3-R46/FIGURES/5

Wu, D., Jospin, G., Eisen, J.A., 2013. Systematic Identification of Gene Families for Use as “Markers” for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. PLoS One 8. https://doi.org/10.1371/JOURNAL.PONE.0077033

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

6

Discussion and future prospect



Viral Detection

The detection depends on the balance between the intent of not missing a virus and not (over)predicting viruses that are not truly present. In the previous chapters, I considered HTS viral detection when the metrics for this virus in the sample are above an (alien) filter. In chapter two, I showed that several other categories (mapping, read assignation, alignment) for HTS analysis leading to detection existed. As those categories can be complementary on several points, combining them for detection seems to be a good idea (Brinkmann et al., 2019; Garcia-Etxebarria et al., 2014).

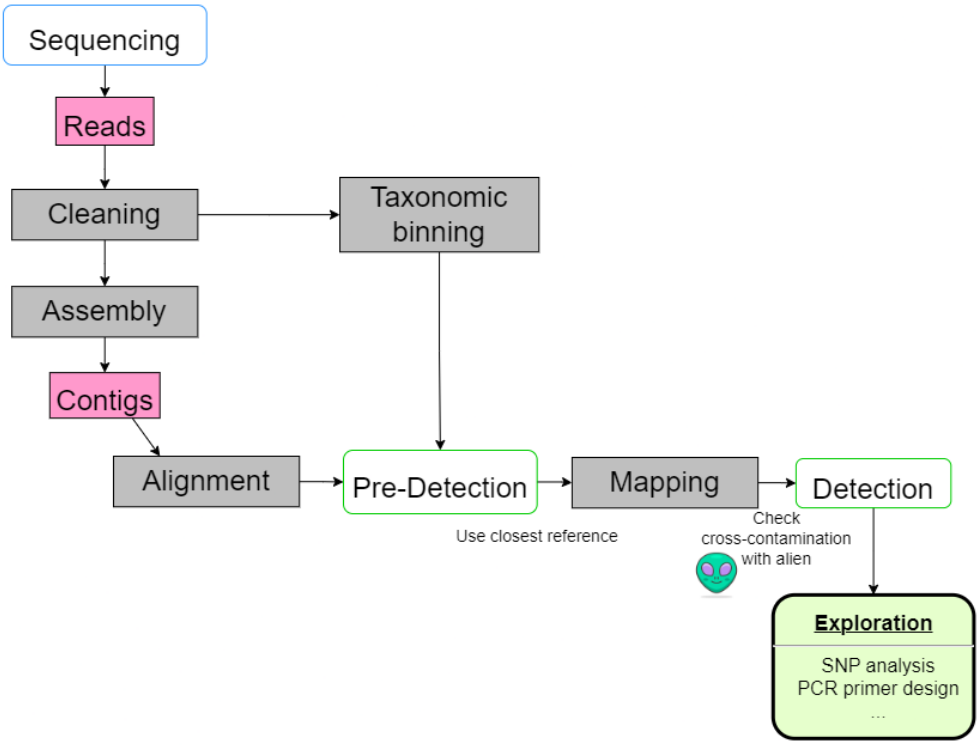


Figure 6-1: Detection workflow for known viruses.

In Figure 6-1, a workflow for plant virus detection is proposed. It combines the standard assembly and alignment with a taxonomic binning tool for a large-scale pre-detection. The idea is to combine methods that make sense and are easy to apply for non-bioinformatician users. The use of method combinations for improving analysis has been documented (Garcia-Etxebarria et al., 2014; Marttunen et al., 2017; Tejano et al., 2019). For virus detection in metagenomic datasets, such a combination of analysis methods is also advised (Lambert et al., 2018). The combination I propose

here is especially strong since taxonomic binning tends to overpredict virus presence while alignment (with strict parameters) tends to predict only the most concentrated viruses. The predictions that are considered reliable (based on reads number and alignment metrics) can be confirmed by performing a mapping on the selected reference. The mapping results can then be used to check cross-contamination before determining the list of present viruses. In addition, on the technical side, all analysis can be done using a user interface (Geneious, CLC genomic or Galaxy) or in command lines on a personal computer or on a High-Performance Computing cluster (HPC).

Exploratory analysis can be done on the virus identified to better understand them (like their mutation profiles or geographical dispersion). Other in case of suspicion of novel virus presence, other analysis should be performed.

Novel viruses

There are two cases for not yet known viruses: either we can detect it with a workflow like described in Figure 6-1, meaning that the evolutionary distance with the known viruses is not high, or we missed it. If missed, (machine learning) tools aiming at new virus detection, like VirHunter (Sukhorukov et al., 2022), can help detect them. However, no method allows us to be absolutely certain that the analysis does not miss a new virus.

If a new virus is detected, an alignment can be done to try to compare it with the already discovered viral diversity. That alignment success depends on the distance between the novel virus and the closest reference in the database, it is then logical to use the most complete database possible. Proteomic analysis should be done if the genome size is large enough and the genes used in taxonomic demarcation criteria are sequenced. It may improve the initial alignment interpretation. Indeed, due to the degenerate property of the genetic code, amino acid (protein) based alignment can hit a more distant target. In addition, from proteins, information related to function can be found by looking at their structural domain. Finally, protein architecture on the genome can help linking the new virus to a known viral family that may show similar architecture.

A virus can be considered new if the ICTV (International Committee on Taxonomy of Viruses) species demarcation criteria are not met. These criteria are decided by experts from the virus family and can vary according to where on the taxonomy the virus is supposed to be close (Simmonds et al., 2017). Most of the time the criteria are based on genes similarities such as often RdRp and CP. Those criteria and the linked

taxonomy are updated regularly to continue to fit the current knowledge on viruses (Walker et al., 2022). The generalization of sequencing and the more frequent (and numerous) new virus discoveries create a lot of debate over the rules that should be used for viral classification, as shown for the *Betaflexiviridae* family (Silva et al., 2022). In addition, new technologies allowing the screening of sequencing datasets present in SRA found 105 potential novel RNA viruses (Edgar et al., 2022; Neri et al., 2022). The idea of using polymerase barcode sequences is gaining more ground (Babaian and Edgar, 2022). With such an important number of data to process and the current limitation of the system, the debate over the taxonomy and the classification criteria is going to gain even more importance in the near future.

As a number of new sequences of viral origin are being discovered at a previously unseen rate, an update of the framework for the detection and characterization of new viruses is currently under writing to take into account the numerous new genomes produced and the new possibilities of analysis (Fontdevila et al., n.d.) ("Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis"). It corresponds to a framework proposal for prioritizing the biological characterization steps after discovering a new plant virus to evaluate its impact at different levels. The main goal is to allow efficient biological characterization to follow the increased virus discovery rate by prioritizing steps for filling knowledge gaps via diverse methods.

The improvement described in HTS for viral detection makes it a good candidate for plant diagnostic. In addition, there is still room for improvement in the HTS method. Indeed, sequencing cost is expected to continue to decrease while easiness to use is improving. On virus characterisation, the new screening methodology available with the release of Serratus (Edgar et al., 2022) helps to overcome (a little) the knowledge gap on virus taxonomy. Palm-ID (Babaian and Edgar, 2022) and the associated serratus database allow the exploration of RNA viruses in all the sequencing datasets available in SRA (10.2 petabases). This can improve the ability to characterize novel viruses and improve the host range and geographical knowledge of RNA viruses (Rivarez et al., 2022). As more and more viruses are known, more and more data are generated and our ability to explore them is improving, the potential for virus discovery is only raising. The question left is:

Are the bioinformatics analyses able to follow the amount of data generated?

Viral analysis prospect

Recent machine learning improvements in protein structures may also have an indirect benefit for viral detection and exploration. New structural predictors Alphafold (Jumper et al., 2021; Varadi et al., 2022) and RosettaFold (Baek et al., 2021) have already changed protein structural knowledge by adding a huge amount of protein structure in databases. In virology, new characterization (or even detection) based on structure similarity (instead of sequence) can now be envisioned for future analyses. It should allow raising the taxonomical range of analysis to even more different viruses than possible at the moment.

A direct improvement in sequencing technologies with the now possible combination of short reads (Illumina) and long reads (Nanopore or PacBio technologies) sequencing may change the bioinformatic analyses (Kutnjak et al., 2021). Long and short reads association change the dynamics of the genome reconstruction (*de novo* assembly). Those changes are likely to improve the accuracy of the genome obtained, meaning more precise alignment possibilities (so better detection) and better characterization as the resulting protein exploration might also be improved. In addition, the long reads methods can lower the time required before starting the analysis, which would be beneficial in plant pest monitoring (Liefting et al., 2021).

Timing and accuracy are keys to action to prevent pests from spreading. Reducing time and raising accuracy are very beneficial to efficiency for eradication and containing new pests' entry (see **Figure 1-1**). Even more important, the potential for preventing new pest entry is very high with HTS. Indeed, global monitoring of viruses can be done with more possibility of detecting unknown or emergent viruses. This monitoring can be done for border crossings to lower the risk of pest importation with collaboration between plant protection organizations (Carvajal-Yepes et al., 2019). This approach will be very dependent on the number of tests made during inspection. Water surveillance can also be useful for monitoring the spread of plant viruses in agricultural systems and identifying the source of outbreaks. Irrigation water surveillance is a method used to monitor plant viruses by analyzing the water used for irrigation. It can provide important information for developing effective prevention and control strategies as it is close to real-time monitoring of virus spread (Bačnik et al., 2020; Maksimovic Carvalho Ferreira et al., 2022).

It's worth noting that this method can be useful for early warning, but it should be used in combination with other methods. A combinatory approach, using HTS for global pest surveillance allows faster response and monitoring of not yet quarantined

viruses and can be coupled with traditional PCR/ELISA methods for precise and cost-effective checking of the well-known dangerous virus, should be considered for improving the control of pests spread.

Cross-contamination monitoring

The general aim of this research was to improve the ability of high-throughput sequencing (HTS) to detect plant viruses in order to increase food safety. Traditional methods for plant virus monitoring can be limited in their ability to detect all viruses present. However, the detection reliability can be improved by incorporating sequencing into the monitoring process. Indeed, HTS is a powerful tool for the detection of plant viruses. It allows a comprehensive analysis of the viral population in a sample, providing information on the presence of known and unknown viruses. Additionally, the characterization of viral variants can be done, which provides important information on the evolution and spread of the virus, as well as its potential impact on crop yields and the development of new control measures.

The performance characteristics of HTS-based protocols were evaluated during this thesis, showing a better analytical sensitivity and an improved analytical specificity for HTS compared to traditional test protocols based on PCR and electron microscopy (Chapter 3). The cross-contamination issue was handled for the first time with alien control that can monitor the transfer of genetic material (DNA/RNA) between samples. This alien-based monitoring allows quantifying the material exchange and determining an alien filter to distinguish low-level infection from contamination. The filter represents the limit of detection adapted to the specific cross-contamination situation for the current sequencing batch. The analytical sensitivity of HTS for plant virus detection, even with the alien filter, remained better than with the traditional detection methods.

Limitations and improvements

Other metrics allowing for tracking cross-contamination were described (Chapter 4) with an attempt at detection automatization. The pertinence of the selecting metrics allowing for monitoring cross-contamination was proved along with the failure to find a universal formula for such monitoring for all types of viral datasets. At the moment, the combination of the (automatic) metrics usage is more efficient when linked to an expert (manual) review that can adapt the virus-specific levels of presence (concentration) during sequencing. In the future, such limitation might be overcome by using a combination of aliens representing several expression levels that could act

like an “amplification ladder”. This would allow us to calculate the cross-contamination filter adapted to the level of cross-contamination observed in the sequencing batch (like now) and also adapted to the relative expression level of each virus in the sample.

Using an (alien) control provides some information regarding the expected level of contamination but with some limitations. Starting with the current limit of detection inherent to the standard bioinformatic procedures. Indeed, the difficulties of detection due to a (very) low level of expression that can be below the limit of detection threshold, or the new viruses difficult identification are also preventing cross-contamination monitoring. It is documented that when internal (spike) controls are used in too high concentration it can mask the virus of interest (Massart et al., 2022a) because of competition for replication in the tube. We can logically have the same events happening in datasets with mixed viral infections (with different concentrations). As the dynamic of the change in relative proportions in the tube is not well known, the impact it has on viral detection and cross-contamination monitoring is unclear. To track cross-contamination, we rely on the identification of the origin of contamination, with the idea that the highest the concentration of a virus is the highest the cross-contamination potential is. There are biases that can change the relative proportion of each virus in each preparation step (sampling, extraction, amplification, sequencing). In addition, we have to consider the population of different viruses and how they impact each other in relative proportion during those steps. All that is a current limitation on our ability to clearly identify to cross-contamination potential of viruses in samples and, therefore, to monitor the cross-contamination event. A more precise understanding of virus concentration change through sequencing steps and population dynamics can help to improve cross-contamination monitoring.

The current monitoring method is also missing local cross-contamination. If two samples have material exchange but not the alien, the monitoring is not going to work. To solve this, a different internal alien could be added to each sample, but the biases (especially on the “concentration mask potential” explained in the previous paragraph) and the added complexity (in both feasibility and analysis) are the reason why it was not tested. The cross-contamination between batches is also not really addressed in this thesis apart from showing (in Chapter 3) a deduplication and SNP-profile comparison to manually track down an external source of contamination. To improve that aspect, a more systematic rotation of alternative alien control used between several sequencing from laboratories work should be implemented. Ensuring

that sequencing A (with alien A), sequencing B (alien B) and sequencing C (alien C) contains different aliens. It can help to monitor external contamination if alien A is found in one or the two other sequencings. With strict registration on a date and material used in the lab, the traceback of which tools can be the cause of such contamination (and what this tool have been used for in the meantime as it could have carried contaminating viruses) can be made easier all thanks to a simple alien rotation.

Other metrics (Reads per kilobase of transcript per Million reads mapped [RPKM], coverage) could be indicative of cross-contamination and therefore be useful to monitor it. Maybe a machine learning method can be developed to identify and combine the metrics linked to cross-contamination in order to improve its monitoring furthermore.

Despite those limitations, the use of aliens is improving the reliability of HTS viral detection up to a point where it can be envisioned as an efficient diagnostic method. I think the main reason of not being widely used is that the information about aliens and how to use them in diagnostic is very recent (the guidelines that explain the alien role were published in 2022) and the diagnostic protocols require time for their validation prior to their application in routine testing. In addition, the interpretation of alien sequences (when found in unexpected samples) requires a little bit of experience, meaning that the accessibility level of the interpretation should be improved (which I tried to do during this thesis).

Once the detection is done, more exploration can be performed to better characterize viruses. Different exploratory analyses can be performed (Chapter 5). Among them, the identification of minor SNPs is among the most important as they allow the tracking of minor and potentially emerging variants for field surveillance. The steps for accurate SNPs predictions were described (Chapter 5), improving the community's understanding of predicting virus variants.

By strengthening the usability of HTS for plant virus detection, via the description of important analysis steps and the use of alien control, sequencing has been shown to be a reliable, comprehensive method for monitoring plant viruses in the tested samples thant can correspond to a single plant up to field based sequencing with bulked plants, therefore improving food safety. This can help to identify outbreaks and ensure that crops are safe, which can prevent economic losses for farmers and ultimately protect the consumers' food diversity supply.

Reference

Adams, I., Fox, A., 2016. Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. Current Research Topics in Plant Virology 323–335. https://doi.org/10.1007/978-3-319-32919-2_14/FIGURES/3

Adams, I.P., Fox, A., Boonham, N., Massart, S., de Jonghe, K., Tahzima, R., Foucart, Y., Peusens, G., Beliën, T., Massart, S., de Jonghe, K., 2018. The impact of high throughput sequencing on plant health diagnostics. Eur J Plant Pathol 152, 909–919. https://doi.org/10.1007/s10658-018-1570-0

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10, 537. https://doi.org/10.1111/J.1364-3703.2009.00545.X

Adams, I.P., Skelton, A., Macarthur, R., Hodges, T., Hinds, H., Flint, L., Nath, P.D., Boonham, N., Fox, A., 2014. Carrot yellow leaf virus is associated with carrot internal necrosis 9. https://doi.org/10.1371/journal.pone.0109125

Aguado-García, Y., Taboada, B., Morán, P., Rivera-Gutiérrez, X., Serrano-Vázquez, A., Iša, P., Rojas-Velázquez, L., Pérez-Juárez, H., López, S., Torres, J., Ximénez, C., Arias, C.F., 2020. Tobamoviruses can be frequently present in the oropharynx and gut of infants during their first year of life. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-70684-w

Agüero, J., Gómez-Aix, C., Sempere, R.N., García-Villalba, J., García-Núñez, J., Hernando, Y., Aranda, M.A., 2018. Stable and broad spectrum cross-protection against pepino mosaic virus attained by mixed infection. Front Plant Sci 871, 1810. https://doi.org/10.3389/FPLS.2018.01810/BIBTEX

Al Rwahnih, M., Daubert, S., Golino, D., Islas, C., Rowhani, A., 2015. Comparison of Next-Generation Sequencing Versus Biological Indexing for the Optimal Detection of Viral Pathogens in Grapevine. Phytopathology 105, 758–763. https://doi.org/10.1094/PHYTO-06-14-0165-R

Anastassiadou, M., Arena, M., Auteri, D., Brancato, A., Bura, L., Carrasco Cabrera, L., Chaideftou, E., Chiusolo, A., Crivellente, F., de Lentdecker, C., Egsmose, M., Fait, G., Greco, L., Ippolito, A., Istace, F., Jarrah, S., Kardassi, D., Leuschner, R., Lostia, A., Lythgo, C., Magrans, O., Mangas, I., Miron, I., Molnar, T., Padovani, L., Parra Morte, J.M., Pedersen, R., Reich, H., Santos, M., Sharp, R., Szentes, C., Terron, A., Tiramani, M., Vagenende, B., Villamar-Bouza, L., 2021. Peer review of the pesticide risk assessment of the active substances Pepino Mosaic Virus,EU strain, mild isolate Abp1 and Pepino Mosaic Virus,CH2 strain, mild isolate Abp2. EFSA Journal 19, e06388. https://doi.org/10.2903/J.EFSA.2021.6388

Babaian, A., Edgar, R., 2022. Ribovirus classification by a polymerase barcode sequence. PeerJ 10, e14055. https://doi.org/10.7717/PEERJ.14055/SUPP-2

Bačnik, K., Kutnjak, D., Pecman, A., Mehle, N., Tušek Žnidarič, M., Gutiérrez Aguirre, I., Ravnikar, M., 2020. Viromics and infectivity analysis reveal the release of infective plant viruses from wastewater into the environment. Water Res 177, 115628. https://doi.org/10.1016/J.WATRES.2020.115628

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni,

A., Pereira, J.H., Rodrigues, A. v., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Christopher Garcia, K., Grishin, N. v., Adams, P.D., Read, R.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) 373, 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Baert, L., McClure, P., Winkler, A., Karn, J., Bouwknegt, M., Klijn, A., 2021. Guidance document on the use of whole genome sequencing (WGS) for source tracking from a food industry perspective. Food Control 130, 108148. https://doi.org/10.1016/J.FOODCONT.2021.108148

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. https://doi.org/10.1099/JGV.0.001210/CITE/REFWORKS

Bal, A., Pichon, M., Picard, C., Casalegno, J.S., Valette, M., Schuffenecker, I., Billard, L., Vallet, S., Vilchez, G., Cheynet, V., Oriol, G., Trouillet-Assant, S., Gillet, Y., Lina, B., Brengel-Pesce, K., Morfin, F., Josset, L., 2018. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infectious Diseases 2018 18:1 18, 1–10. https://doi.org/10.1186/S12879-018-3446-5

Ballenghien, M., Faivre, N., Galtier, N., 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol 15. https://doi.org/10.1186/S12915-017-0366-6

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. v, Sirotkin, A. v, Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. Viruses 2014, Vol. 6, Pages 106-136 6, 106–136. https://doi.org/10.3390/V6010106

Barbitoff, Y.A., Abasov, R., Tvorogova, V.E., Glotov, A.S., Predeus, A. v., 2022. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics 23, 1–17. https://doi.org/10.1186/S12864-022-08365-3/

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., de Bruyn, R., Maree, H.J., 2021a. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

Bester, R., Cook, G., Breytenbach, J.H.J., Steyn, C., De Bruyn, R., Maree, H.J., 2021b. 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. Virology Journal 2021 18:1 18, 1–19. https://doi.org/10.1186/S12985-021-01523-1

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., Arboleda, V.A., 2021a. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

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., Arboleda, V.A., 2021b. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. medRxiv 2020.08.04.20167874. https://doi.org/10.1101/2020.08.04.20167874

Boheemen, S. van, Rijn, A.L. van, Pappas, N., Carbo, E.C., Vorderman, R.H.P., Sidorov, I., 2020. Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- research that is available on the COVID-19 resource centre - including this with acknowledgement of the origin.

Bomba, L., Walter, K., Soranzo, N., 2017. The impact of rare and low-frequency genetic variants in common disease. Genome Biology 2017 18:1 18, 1–17. https://doi.org/10.1186/S13059-017-1212-4

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., Mumford, R., 2014. Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res 186, 20–31. https://doi.org/10.1016/j.virusres.2013.12.007

Bordería, A. v., Isakov, O., Moratorio, G., Henningsson, R., Agüera-González, S., Organtini, L., Gnädig, N.F., Blanc, H., Alcover, A., Hafenstein, S., Fontes, M., Shomron, N., Vignuzzi, M., 2015. Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. PLoS Pathog 11, e1004838. https://doi.org/10.1371/JOURNAL.PPAT.1004838

Brinkmann, A., Andrusch, A., Belka, A., Wylezich, C., Höper, D., Pohlmann, A., Petersen, T.N., Lucas, P., Blanchard, Y., Papa, A., Melidou, A., Oude Munnink, B.B., Matthijnssens, J., Deboutte, W., Ellis, R.J., Hansmann, F., Baumgärtner, W., van der Vries, E., Osterhaus, A., Camma, C., Mangone, I., Lorusso, A., Marcacci, M., Nunes, A., Pinto, M., Borges, V., Kroneman, A., Schmitz, D., Corman, V.M., Drosten, C., Jones, T.C., Hendriksen, R.S., Aarestrup, F.M., Koopmans, M., Beer, M., Nitsche, A., 2019. Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. J Clin Microbiol 57. https://doi.org/10.1128/JCM.00466-19

Bukowska-Ośko, I., Perlejewski, K., Nakamura, S., Motooka, D., Stokowy, T., Kosińska, J., Popiel, M., Płoski, R., Horban, A., Lipowski, D., Caraballo Cortés, K., Pawełczyk, A., Demkow, U., Stępień, A., Radkowski, M., Laskus, T., 2017. Sensitivity of next-generation sequencing metagenomic

analysis for detection of RNA and DNA viruses in cerebrospinal fluid: The confounding effect of background contamination. Adv Exp Med Biol 944, 53–62. https://doi.org/10.1007/5584_2016_42

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8, 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Bushnell, B., Rood, J., Singer, E., 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One 12. https://doi.org/10.1371/JOURNAL.PONE.0185056

Calil, I.P., Fontes, E.P.B., 2017. Plant immunity against viruses: antiviral immune receptors in focus. Ann Bot 119, 711–723. https://doi.org/10.1093/AOB/MCW200

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10. https://doi.org/10.1186/1471-2105-10-421

Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K.A., Giovani, B., Saunders, D.G.O., Kamoun, S., Legg, J.P., Verdier, V., Lessel, J., Neher, R.A., Day, R., Pardey, P., Gullino, M.L., Records, A.R., Bextine, B., Leach, J.E., Staiger, S., Tohme, J., 2019. A global surveillance system for crop diseases: Global preparedness minimizes the risk to food supplies. Science (1979) 364, 1237–1239. https://doi.org/10.1126/SCIENCE.AAW1572/ASSET/52AFE309-57C8-475B-88C6-17D489A74CEE/ASSETS/GRAPHIC/364_1237_F2.JPEG

Černi, S., Ruščić, J., Nolasco, G., Gatin, Ž., Krajačić, M., Škorić, D., 2008. Stem pitting and seedling yellows symptoms of Citrus tristeza virus infection may be determined by minor sequence variants. Virus Genes 36, 241–249. https://doi.org/10.1007/S11262-007-0183-Z/TABLES/1

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021a. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Chabannes, M., Gabriel, M., Aksa, A., Galzi, S., Dufayard, J.-F., Iskra-Caruana, M.-L., Muller, E., 2021b. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. Mol Plant Pathol 22, 216–230. https://doi.org/https://doi.org/10.1111/mpp.13019

Champlot, S., Berthelot, C., Pruvost, M., Andrew Bennett, E., Grange, T., Geigl, E.M., 2010. An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. PLoS One 5, e13042. https://doi.org/10.1371/JOURNAL.PONE.0013042

Charlebois, R.L., Sathiamoorthy, S., Logvinoff, C., Gisonni-Lex, L., Mallet, L., Ng, S.H.S., 2020. Sensitivity and breadth of detection of high-throughput sequencing for adventitious virus detection. npj Vaccines 2020 5:1 5, 1–8. https://doi.org/10.1038/s41541-020-0207-4

Chodasewicz, K., 2014. Evolution, reproduction and definition of life. Theory in Biosciences 133, 39. https://doi.org/10.1007/S12064-013-0184-5

Choua, M., Heath, M.R., Speirs, D.C., Bonachela, J.A., 2020. The effect of viral plasticity on the persistence of host-virus systems. J Theor Biol 498, 110263. https://doi.org/10.1016/J.JTBI.2020.110263

Clark, M.F., Adams, A.N., 1977. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. J Gen Virol 34, 475–483. https://doi.org/10.1099/0022-1317-34-3-475

Cleland, C.E., Chyba, C.F., 2002. Defining "life." Orig Life Evol Biosph 32, 387–393. https://doi.org/10.1023/A:1020503324273

Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P., Jackson, S.A., 2015. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant 8, 831–846. https://doi.org/10.1016/J.MOLP.2015.02.002

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., Gabriel, S., 2018a. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 332. https://doi.org/10.1186/s12864-018-4703-0

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., Gabriel, S., 2018b. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 1–10. https://doi.org/10.1186/S12864-018-4703-0/TABLES/2

Cunniffe, N.J., Laranjeira, F.F., Neri, F.M., DeSimone, R.E., Gilligan, C.A., 2014. Cost-Effective Control of Plant Disease When Epidemiological Knowledge Is Incomplete: Modelling Bahia Bark Scaling of Citrus. PLoS Comput Biol 10, e1003753. https://doi.org/10.1371/JOURNAL.PCBI.1003753

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Daubin, V., Szöllősi, G.J., 2016. Horizontal Gene Transfer and the History of Life. Cold Spring Harb Perspect Biol 8. https://doi.org/10.1101/CSHPERSPECT.A018036

De Clerck, C., Crew, K., Van den houwe, I., McMichael, L., Berhal, C., Lassois, L., Haissam Jijakli, M., Roux, N., Thomas, J., Massart, S., 2017. Lessons learned from the virus indexing of Musa germplasm: insights from a multiyear collaboration. Annals of Applied Biology 171, 15–27. https://doi.org/10.1111/AAB.12353

de Jonghe, K., 2020. The world of plant viruses: the good, the bad and the beautiful. Floraculture International 34–36.

Deng, Z.L., Dhingra, A., Fritz, A., Götting, J., Münch, P.C., Steinbrück, L., Schulz, T.F., Ganzenmüller, T., McHardy, A.C., 2021. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. Brief Bioinform 22, 1–12. https://doi.org/10.1093/BIB/BBAA123

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022a. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Di Gaspero, G., Radovic, S., De Luca, E., Spadotto, A., Magris, G., Falginella, L., Cattonaro, F., Marroni, F., 2022b. Evaluation of sensitivity and specificity in RNA-Seq-based detection of grapevine viral pathogens. J Virol Methods 300, 114383. https://doi.org/https://doi.org/10.1016/j.jviromet.2021.114383

Dimmock, N.J., Easton, A.J., Leppard, K.N., 2016. Introduction to Modern Virology, 7th Edition | Wiley.

Domingo, E., Perales, C., 2019. Viral quasispecies. PLoS Genet 15, e1008271. https://doi.org/10.1371/JOURNAL.PGEN.1008271

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J.F., de la Peña, M., Korobeynikov, A., Chikhi, R., Babaian, A., 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 2022 602:7895 602, 142–147. https://doi.org/10.1038/s41586-021-04332-2

Fennell, M., Murphy, J.E., Armstrong, C., Gallagher, T., Osborne, B., 2012. Plant Spread Simulator: A model for simulating large-scale directed dispersal processes across heterogeneous environments. Ecol Modell 230, 1–10. https://doi.org/10.1016/J.ECOLMODEL.2012.01.008

Ferguson, J.M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., Bowen-James, R., Deveson, I.W., 2022. InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. Bioinformatics 38, 1443. https://doi.org/10.1093/BIOINFORMATICS/BTAB846

Filloux, D., Dallot, S., Delaunay, A., Galzi, S., Jacquot, E., Roumagnac, P., 2015. Metagenomics Approaches Based on Virion-Associated Nucleic Acids (VANA): An Innovative Tool for Assessing Without A Priori Viral Diversity of Plants. Methods Mol Biol 1302, 249–257. https://doi.org/10.1007/978-1-4939-2620-6_18

Fontdevila, N., n.d. Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis.

Food, E., Authority, S., Schrader, G., Camilleri, M., Diakaki, M., Vos, S., 2019. Pest survey card on non-European isolates of citrus tristeza virus. EFSA Supporting Publications 16, 1600E. https://doi.org/10.2903/SP.EFSA.2019.EN-1600

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., 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, ppa.12962. https://doi.org/10.1111/ppa.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., Ziebell, H., 2021. Interlaboratory Comparison Study on Ribodepleted Total RNA High-Throughput Sequencing for Plant Virus Diagnostics and Bioinformatic Competence. Pathogens 10, 1174. https://doi.org/10.3390/pathogens10091174

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., Cosson, J.F., 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. mSystems 1, 1–22.

Garcia-Etxebarria, K., Garcia-Garcerà, M., Calafell, F., 2014. Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics 15, 1–10. https://doi.org/10.1186/1471-2105-15-90/FIGURES/4

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022a. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. V, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., 2022b. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology . https://doi.org/10.3390/biology11020263

Gauthier, M.-E.A., Lelwala, R. v, Elliott, C.E., Windell, C., Fiorito, S., Dinsdale, A., Whattam, M., Pattemore, J., Barrero, R.A., Bilodeau, G.J., van de Vossenberg, B.T.L.H., 2022c. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. Biology 2022, Vol. 11, Page 263 11, 263. https://doi.org/10.3390/BIOLOGY11020263

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G., Thomas, J.E., 2000. Genetic Diversity Among Banana streak virus Isolates from Australia. Phytopathology 90, 921–927. https://doi.org/10.1094/PHYTO.2000.90.8.921

Giovani, B., Blümel, S., Lopian, R., Teulon, D., Bloem, S., Galeano Martínez, C., Beltrán Montoya, C., Urias Morales, C.R., Dharmapuri, S., Timote, V., Horn, N., Chouibani, M., Mezui M'Ella, J.G., Herrera, V., Castinel, A., Goletsos, C., Moeller, C., Naumann, I., Stancanelli, G., Bronzwaer, S., Tramontini, S., MacDonald, P., Matheson, L., Anthoine, G., de Jonghe, K., Schenk, M., Steinmöller, S., Rodriguez, E., Cruz, M.L., Luck, J., Fraser, G., Brunel, S., Montuori, M., Fedchock, C., Steel, E., Pennington, H.G., Day, R., Rossi, J.P., Xia, J., 2020. Science diplomacy for plant health. Nature Plants 2020 6:8 6, 902–905. https://doi.org/10.1038/s41477-020-0744-x

Griffith, F., 1928. The Significance of Pneumococcal Types. J Hyg (Lond) 27, 113. https://doi.org/10.1017/S0022172400031879

Grosdidier, M., Aguayo, J., Marçais, B., Ioos, R., 2017. Detection of plant pathogens using real-time PCR: how reliable are late Ct values? Plant Pathol 66, 359–367. https://doi.org/10.1111/PPA.12591

Guirao-Rico, S., González, J., 2021. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. Mol Ecol Resour 21, 1216. https://doi.org/10.1111/1755-0998.13343

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022a. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Rong, W., Tamisier, L., Berhal, C., Roux, N., Massart, S., 2022b. 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 . https://doi.org/10.3390/plants11020226

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020a. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanafi, M., Tahzima, R., Ben Kaab, S., Tamisier, L., Roux, N., Massart, S., 2020b. Identification of Divergent Isolates of Banana Mild Mosaic Virus and Development of a New Diagnostic Primer to Improve Detection. Pathogens . https://doi.org/10.3390/pathogens9121045

Hanssen, I.M., Paeleman, A., Vandewoestijne, E., van Bergen, L., Bragard, C., Lievens, B., Vanachter, A.C.R.C., Thomma, B.P.H.J., 2009. Pepino mosaic virus isolates and differential symptomatology in tomato. Plant Pathol 58, 450–460. https://doi.org/10.1111/J.1365-3059.2008.02018.X

Harris, H.M.B., Hill, C., 2020. A Place for Viruses on the Tree of Life. Front Microbiol 11. https://doi.org/10.3389/FMICB.2020.604048

Hasiów-Jaroszewska, B., Borodynko, N., Jackowiak, P., Figlerowicz, M., Pospieszny, H., 2011. Single mutation converts mild pathotype of the Pepino mosaic virus into necrotic one. Virus Res 159, 57–61. https://doi.org/10.1016/J.VIRUSRES.2011.04.008

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., Lang, J., 2020. Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. Front Bioeng Biotechnol 8, 817. https://doi.org/10.3389/FBIOE.2020.00817/BIBTEX

Hilaire, J., Tindale, S., Jones, G., Pingarron-Cardenas, G., Bačnik, K., Ojo, M., Frewer, L.J., 2022. Risk perception associated with an emerging agri-food risk in Europe: plant viruses in agriculture. Agric Food Secur 11, 1–19. https://doi.org/10.1186/S40066-022-00366-5/TABLES/3

Hirabara, S.M., Serdan, T.D.A., Gorjao, R., Masi, L.N., Pithon-Curi, T.C., Covas, D.T., Curi, R., Durigon, E.L., 2022. SARS-COV-2 Variants: Differences and Potential of Immune Evasion. Front Cell Infect Microbiol 11, 1401. https://doi.org/10.3389/FCIMB.2021.781429/BIBTEX

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593. https://doi.org/10.1093/BIOINFORMATICS/BTR708

Hull, R., 2014. Plant Virology (Fifth Edition). Plant Virology 3–14. https://doi.org/10.1016/B978-0-12-384871-0.00001-7

Irwin, N.A.T., Pittis, A.A., Richards, T.A., Keeling, P.J., 2021. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nature Microbiology 2022 7:2 7, 327–336. https://doi.org/10.1038/s41564-021-01026-3

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332. https://doi.org/10.1093/BIOINFORMATICS/BTT113

Jeong, J.-J., Ju, H.-J., Noh, J., 2014. A Review of Detection Methods for the Plant Viruses. Research in Plant Disease 20, 173–181. https://doi.org/10.5423/RPD.2014.20.3.173

Jones, R.A.C., 2014. Trends in plant virus epidemiology: Opportunities from new or improved technologies. Virus Res 186, 3–19. https://doi.org/10.1016/J.VIRUSRES.2013.11.003

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalimuthu, K., Arivalagan, J., Mohan, M., Samuel Selvan Christyraj, J.R., Arockiaraj, J., Muthusamy, R., Ju, H.J., 2022. Point of care diagnosis of plant virus: Current trends and prospects. Mol Cell Probes 61, 101779. https://doi.org/10.1016/J.MCP.2021.101779

Kechin, A., Boyarskikh, U., Kel, A., Filipenko, M., 2017. CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. Journal of Computational Biology 24, 1138–1143. https://doi.org/10.1089/CMB.2017.0096/ASSET/IMAGES/LARGE/FIGURE3.JPEG

Koboldt, D.C., 2020a. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Koboldt, D.C., 2020b. Best practices for variant calling in clinical sequencing. Genome Medicine 2020 12:1 12, 1–13. https://doi.org/10.1186/S13073-020-00791-W

Kogovšek, P., Pompe-Novak, M., Baebler, S., Rotter, A., Gow, L., Gruden, K., Foster, G.D., Boonham, N., Ravnikar, M., 2010. Aggressive and mild Potato virus Y isolates trigger different specific responses in susceptible potato plants. Plant Pathol 59, 1121–1132. https://doi.org/10.1111/J.1365-3059.2010.02340.X

Koonin, E. v., Dolja, V. v., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., Kuhn, J.H., 2020. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84. https://doi.org/10.1128/MMBR.00061-19

Kreuze, J., 2014. siRNA deep sequencing and assembly: Piecing together viral infections. Detection and Diagnostics of Plant Pathogens 21–38. https://doi.org/10.1007/978-94-017-9020-8_2/FIGURES/3

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1–7. https://doi.org/10.1016/J.VIROL.2009.03.024

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res 239, 136–142. https://doi.org/10.1016/J.VIRUSRES.2017.02.002

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., 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 . https://doi.org/10.3390/plants10020328

Kumar, R., Nagpal, S., Kaushik, S., Mendiratta, S., 2020. COVID-19 diagnostic approaches: different roads to the same destination. VirusDisease 2020 31:2 31, 97–105. https://doi.org/10.1007/S13337-020-00599-7

Kutnjak, D., Rupar, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J.F., Ravnikar, M., 2015. Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. J Virol 89, 4760–4769. https://doi.org/10.1128/JVI.03685-14/SUPPL_FILE/ZJV999090298SO1.PDF

Kutnjak, D., Tamisier, L., Adams, I., Boonham, N., Candresse, T., Chiumenti, M., de Jonghe, K., Kreuze, J.F., Lefebvre, M., Silva, G., Malapi-wight, M., Margaria, P., Pleško, I.M., McGreig, S., Miozzi, L., Remenant, B., Reynard, J.S., Rollin, J., Rott, M., Schumpp, O., Massart, S., Haegeman, A.,

2021. A Primer on the Analysis of High-Throughput Sequencing Data for Detection of Plant Viruses. Microorganisms 9. https://doi.org/10.3390/MICROORGANISMS9040841

Kutschera, U., Niklas, K.J., 2004. The modern theory of biological evolution: An expanded synthesis. Naturwissenschaften 91, 255–276. https://doi.org/10.1007/S00114-004-0515-Y/FIGURES/8

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 optimization of high-throughput sequencing bioinformatics pipelines for virus detection. Viruses 10. https://doi.org/10.3390/v10100528

Lau, H.Y., Botella, J.R., 2017. Advanced DNA-based point-of-care diagnostic methods for plant diseases detection. Front Plant Sci 8, 2016. https://doi.org/10.3389/FPLS.2017.02016/BIBTEX

Lebas, B., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G., Blouin, A., Boonham N., Candresse, T., Chandelier, A., De Jonghe, K., Fox, A., Gaafar, Y., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., Landa, B., Li, Massart, S., n.d. Guidelines for preparing a laboratory to use high-throughput sequencing as routine diagnostic test: general and technical recommendations. EPPO Bulletin.

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., 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 Bulletin 52, 394–418. https://doi.org/10.1111/epp.12863

Lefebvre, M., Theil, S., Ma, Y., Candresse, T., 2019. The VirAnnot Pipeline: A Resource for Automated Viral Diversity Estimation and Operational Taxonomy Units Assignation for Virome Sequencing Data. https://doi.org/10.1094/PBIOMES-07-19-0037-A 3, 256–259. https://doi.org/10.1094/PBIOMES-07-19-0037-A

Legrand, P., 2015. Biological assays for plant viruses and other graft-transmissible pathogens diagnoses: a review. EPPO Bulletin 45, 240–251. https://doi.org/10.1111/EPP.12222

Li, C.-X., Li, W., Zhou, J., Zhang, B., Feng, Y., Xu, C.-P., Lu, Y.-Y., Holmes, E.C., Shi, M., 2020. High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory infection. Scientific Reports 2020 10:1 10, 1–11. https://doi.org/10.1038/s41598-020-60992-6

Li, H., Durbin, R., 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, H., Durbin, R., 2009b. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Liefting, L.W., Waite, D.W., Thompson, J.R., 2021. Application of Oxford Nanopore Technology to Plant Virus Detection. Viruses 13. https://doi.org/10.3390/V13081424

Ling, K.S., Scott, J.W., 2007. Sources of Resistance to Pepino mosaic virus in Tomato Accessions. https://doi.org/10.1094/PDIS-91-6-0749 91, 749–753. https://doi.org/10.1094/PDIS-91-6-0749

Low, A.J., Koziol, A.G., Manninger, P.A., Blais, B., Carrillo, C.D., 2019. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 7, e6995. https://doi.org/10.7717/peerj.6995

Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci 2017, e104. https://doi.org/10.7717/PEERJ-CS.104/SUPP-5

Luisi, P.L., 1998. About various definitions of life. Orig Life Evol Biosph 28, 613–622. https://doi.org/10.1023/A:1006517315105

Maachi, A., Torre, C., Sempere, R.N., Hernando, Y., Aranda, M.A., Donaire, L., 2021. Use of High-Throughput Sequencing and Two RNA Input Methods to Identify Viruses Infecting Tomato Crops. Microorganisms . https://doi.org/10.3390/microorganisms9051043

Mackenzie, J.S., Jeggo, M., 2019. The One Health Approach—Why Is It So Important? Trop Med Infect Dis 4. https://doi.org/10.3390/TROPICALMED4020088

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C.M., Roumagnac, P., van der Vlugt, R., Massart, S., 2020. Illuminating an Ecological Blackbox: Using High Throughput Sequencing to Characterize the Plant Virome Across Scales. Front Microbiol 11. https://doi.org/10.3389/fmicb.2020.578064

Maksimovic Carvalho Ferreira, O., Lengar, Ž., Kogej, Z., Bačnik, K., Bajde, I., Milavec, M., Županič, A., Mehle, N., Kutnjak, D., Ravnikar, M., Gutierrez-Aguirre, I., 2022. Evaluation of Methods and Processes for Robust Monitoring of SARS-CoV-2 in Wastewater. Food Environ Virol 14, 384–400. https://doi.org/10.1007/S12560-022-09533-0/FIGURES/4

Malapi-Wight, M., Adhikari, B., Zhou, J., Hendrickson, L., Maroon-Lango, C.J., McFarland, C., Foster, J.A., Hurtado-Gonzales, O.P., 2021. HTS-Based Diagnostics of Sugarcane Viruses: Seasonal Variation and Its Implications for Accurate Detection. Viruses . https://doi.org/10.3390/v13081627

Maliogka, V.I., Minafra, A., Saldarelli, P., Ruiz-García, A.B., Glasa, M., Katis, N., Olmos, A., 2018. Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. Viruses 2018, Vol. 10, Page 436 10, 436. https://doi.org/10.3390/V10080436

Maljkovic Berry, I., Melendrez, M.C., Bishop-Lilly, K.A., Rutvisuttinunt, W., Pollett, S., Talundzic, E., Morton, L., Jarman, R.G., 2020. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis 221, S292–S307. https://doi.org/10.1093/INFDIS/JIZ286

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018a. Viral double-stranded RNAs (dsRNAs) from Plants: Alternative nucleic acid substrates for high-throughput sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4/COVER

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018b. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods in Molecular Biology 1746, 45–53. https://doi.org/10.1007/978-1-4939-7683-6_4

Maree, H.J., Fox, A., Al Rwahnih, M., Boonham, N., Candresse, T., 2018. Application of hts for routine plant virus diagnostics: state of the art and challenges. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.01082

Martelli, G.P., 2017. An overview on grapevine viruses, viroids, and the diseases they cause. Grapevine Viruses: Molecular Biology, Diagnostics and Management 31–46. https://doi.org/10.1007/978-3-319-57706-7_2/TABLES/1

Martínez, M., de Viedma, D.G., Alonso, M., Andrés, S., Bouza, E., Cabezas, T., Cabeza, I., Reyes, A., Sánchez-Yebra, W., Rodríguez, M., Sánchez, M.I., Rogado, M.C., Fernández, R., Peñafiel, T., Martínez, J., Barroso, P., Lucerna, M.A., Diez, L.F., Gutiérrez, C., 2006. Impact of Laboratory Cross-Contamination on Molecular Epidemiology Studies of Tuberculosis. J Clin Microbiol 44, 2967–2969. https://doi.org/10.1128/JCM.00754-06

Marttunen, M., Lienert, J., Belton, V., 2017. Structuring problems for Multi-Criteria Decision Analysis in practice: A literature review of method combinations. Eur J Oper Res 263, 1–17. https://doi.org/10.1016/J.EJOR.2017.04.041

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, Jonkers W., Kreuze J., Kutnjak D., Landa B., n.d. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests | Zenodo [WWW Document]. URL https://zenodo.org/record/6637519 (accessed 7.11.22).

Massart, S., Adams, I., Rwahnih, A., Baeyen, S., Bilodeau, G.J., Blouin, A.G., Boonham, N., Candresse, T., Chandellier, 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.B., Liu, M., Maclot, F., Malapi-Wight, M., Maree, H.J., Martoni, F., Mehle, N., Minafra, A., Mollov, D., Moreira, A.G., Nakhla, M., Petter, F., Piper, A.M., Ponchart, J.P., Rae, R., Remenant, B., Rivera, Y., Rodoni, B., Botermans, M., 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., van de Vossenberg, B.T.L.H., Westenberg, M., Wetzel, T., Ziebell, H., Lebas, B.S.M., al Rwahnih, M., 2022a. Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. Peer Community Journal 2, 37. https://doi.org/10.24072/PCJOURNAL.181

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J.S., Rumbou, A., Saldarelli, P., Škoric, D., Vainio, E.J., Valkonen, J.P.T., 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. Front Microbiol 8, 45. https://doi.org/10.3389/FMICB.2017.00045/BIBTEX

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, Y., Candresse, T., 2019. Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. Phytopathology 109. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, Sebastien, 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, 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. https://doi.org/10.1094/PHYTO-02-18-0067-R

Massart, S., Lebas, B., Chabirand, A., Chappé, A.-M., Dreo, T., Faggioli, F., Harrison, C., Macarthur, R., Mehle, N., Mezzalama, M., Petter, F., Ravnikar, M., Renvoisé, J.-P., Spadaro, D., Tomassoli, L., Tomlinson, J., Trontin, C., Vlugt, R. van der, Vučurović, A., Weekes, R., Brostaux, Y., 2022b. Guidelines for improving statistical analyses of validation datasets for plant pest diagnostic tests. EPPO Bulletin 52, 419–433. https://doi.org/10.1111/EPP.12862

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014a. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014b. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188. https://doi.org/10.1016/j.virusres.2014.03.029

Massart, Sebastien, Olmos, A., Jijakli, H., Candresse, T., 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188, 90–96. https://doi.org/10.1016/J.VIRUSRES.2014.03.029

Minic, R., Zivkovic, I., Minic, R., Zivkovic, I., 2020. Optimization, Validation and Standardization of ELISA. Norovirus. https://doi.org/10.5772/INTECHOPEN.94338

Modrow, S., Falke, D., Truyen, U., Schätzl, H., 2013. Viruses: Definition, Structure, Classification. Molecular Virology 17. https://doi.org/10.1007/978-3-642-20718-1_2

Moonen, P., Boonstra, J., Hakze- Van Der Honing, R., Boonstra- Leendertse, C., Jacobs, L., Dekker, A., 2003. Validation of a LightCycler-based reverse transcription polymerase chain reaction for the detection of foot-and-mouth disease virus. J Virol Methods 113, 35–41. https://doi.org/10.1016/S0166-0934(03)00220-9

Moreira, D., López-García, P., 2009. Ten reasons to exclude viruses from the tree of life. Nature Reviews Microbiology 2009 7:4 7, 306–311. https://doi.org/10.1038/nrmicro2108

Navarro, R., Ambrós, S., Butković, A., Carrasco, J.L., González, R., Martínez, F., Wu, B., Elena, S.F., 2022. Defects in plant immunity modulate the rates and patterns of RNA virus evolution. Virus Evol 8. https://doi.org/10.1093/VE/VEAC059

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D.A., Bhaya, D., Narrowe, A.B., Probst, A.J., Sczyrba, A., Kohler, A., Séguin, A., Shade, A., Campbell, B.J., Lindahl, B.D., Reese, B.K., Roque, B.M., DeRito, C., Averill, C., Cullen, D., Beck, D.A.C., Walsh, D.A., Ward, D.M., Wu, D., Eloe-Fadrosh, E., Brodie, E.L., Young, E.B., Lilleskov, E.A., Castillo, F.J., Martin, F.M., LeCleir, G.R., Attwood, G.T., Cadillo-Quiroz, H., Simon, H.M., Hewson, I., Grigoriev, I. v., Tiedje, J.M., Jansson, J.K., Lee, J., VanderGheynst, J.S., Dangl, J., Bowman, J.S., Blanchard, J.L., Bowen, J.L., Xu, J., Banfield, J.F., Deming, J.W., Kostka, J.E., Gladden, J.M., Rapp, J.Z., Sharpe, J., McMahon,

K.D., Treseder, K.K., Bidle, K.D., Wrighton, K.C., Thamatrakoln, K., Nusslein, K., Meredith, L.K., Ramirez, L., Buee, M., Huntemann, M., Kalyuzhnaya, M.G., Waldrop, M.P., Sullivan, M.B., Schrenk, M.O., Hess, M., Vega, M.A., O'Malley, M.A., Medina, M., Gilbert, N.E., Delherbe, N., Mason, O.U., Dijkstra, P., Chuckran, P.F., Baldrian, P., Constant, P., Stepanauskas, R., Daly, R.A., Lamendella, R., Gruninger, R.J., McKay, R.M., Hylander, S., Lebeis, S.L., Esser, S.P., Acinas, S.G., Wilhelm, S.S., Singer, S.W., Tringe, S.S., Woyke, T., Reddy, T.B.K., Bell, T.H., Mock, T., McAllister, T., Thiel, V., Denef, V.J., Liu, W.T., Martens-Habbena, W., Allen Liu, X.J., Cooper, Z.S., Wang, Z., Krupovic, M., Dolja, V. v., Kyrpides, N.C., Koonin, E. v., Gophna, U., 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell 185, 4023-4037.e18. https://doi.org/10.1016/J.CELL.2022.08.023

Ng, S.H., Braxton, C., Eloit, M., Feng, S.F., Fragnoud, R., Mallet, L., Mee, E.T., Sathiamoorthy, S., Vandeputte, O., Khan, A.S., 2018. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: Upstream sample processing and library preparation [WWW Document]. Viruses. https://doi.org/10.3390/v10100566

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., van Helden, J., Medina-Rivera, A., Thomas-Chollier, M., 2018. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 46, W209–W214. https://doi.org/10.1093/NAR/GKY317

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., Massart, S., 2018a. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

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., Massart, S., 2018b. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin 48, 219–224. https://doi.org/https://doi.org/10.1111/epp.12472

Oñate-Sánchez, L., Vicente-Carbajosa, J., 2008. DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. BMC Res Notes 1, 1–7. https://doi.org/10.1186/1756-0500-1-93/FIGURES/3

Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology 2021 22:1 22, 1–19. https://doi.org/10.1186/S13059-021-02393-0

Pappas, N., Roux, S., Hölzer, M., Lamkiewicz, K., Mock, F., Marz, M., Dutilh, B.E., 2021. Virus Bioinformatics. Encyclopedia of Virology 124–132. https://doi.org/10.1016/B978-0-12-814515-9.00034-5

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017a. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., Ravnikar, M., 2017b. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches . Frontiers in Microbiology .

PM 7/76 (5) Use of EPPO Diagnostic Standards, 2018. . EPPO Bulletin 48, 373–377. https://doi.org/10.1111/EPP.12506

PM 7/98 (5) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity, 2021. . EPPO Bulletin 51, 468–498. https://doi.org/10.1111/EPP.12780

Ramesh, S., Govindarajulu, M., Parise, R.S., Neel, L., Shankar, T., Patel, S., Lowery, P., Smith, F., Dhanasekaran, M., Moore, T., 2021. Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy. Vaccines (Basel) 9. https://doi.org/10.3390/VACCINES9101195

Ricciuti, E., Laboureau, N., Noumbissié, G., Chabannes, M., Sukhikh, N., Pooggin, M.M., Iskra-Caruana, M.-L., 2021. Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. Journal of General Virology 102, 001670. https://doi.org/10.1099/JGV.0.001670

Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N. v., Finegold, C., Garrett, K.A., Gilligan, C.A., Jones, C.M., Martin, M.D., MacDonald, G.K., Neenan, P., Records, A., Schmale, D.G., Tateosian, L., Wei, Q., 2021. The persistent threat of emerging plant disease pandemics to global food security. Proc Natl Acad Sci U S A 118, e2022239118. https://doi.org/10.1073/PNAS.2022239118/SUPPL_FILE/PNAS.2022239118.SAPP.PDF

Rivarez, M., Faure, C., Svanella-Dumas, L., Pecman, A., Tušek-Žnidaric, M., Schönegger, D., de Jonghe, K., Blouin, A., Massart, S., Ravnikar, M., Kutnjak, D., Marais, A., Candresse, T., 2022. Diversity and pathobiology of an ilarvirus unexpectedly detected in diverse host plants and in global sequencing data. https://doi.org/10.1101/2022.12.15.520526

Rivarez, M.P.S., Vučurović, A., Mehle, N., Ravnikar, M., Kutnjak, D., 2021. Global Advances in Tomato Virome Research: Current Status and the Impact of High-Throughput Sequencing. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.671925

Rizzo, D.M., Lichtveld, M., Mazet, J.A.K., Togami, E., Miller, S.A., 2021. Plant health and its effects on food safety and security in a One Health framework: four case studies. One Health Outlook 2021 3:1 3, 1–9. https://doi.org/10.1186/S42522-021-00038-7

Roberts, C.A., Dietzgen, R.G., Heelan, L.A., MaclLean, D.J., 2000. Real-time RT-PCR fluorescent detection of tomato spotted wilt virus. J Virol Methods 88, 1–8. https://doi.org/10.1016/S0166-0934(00)00156-7

Rong, W., Rollin, J., Hanafi, M., Roux, N., Massart, S., 2022. Validation of high throughput sequencing as virus indexing test for Musa germplasm: performance criteria evaluation and contamination monitoring using an alien control. PhytoFrontiers. https://doi.org/https://doi.org/10.1094/PHYTOFR-03-22-0030-FI

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarría, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19 Suppl 1, 81–88. https://doi.org/10.1111/J.1365-294X.2009.04470.X

Rosenthal, S.H., Gerasimova, A., Ruiz-Vega, R., Livingston, K., Kagan, R.M., Liu, Y., Anderson, B., Owen, R., Bernstein, L., Smolgovsky, A., Xu, D., Chen, R., Grupe, A., Tanpaiboon, P., Lacbawan, F., 2022. Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory. Scientific Reports 2022 12:1 12, 1–12. https://doi.org/10.1038/s41598-022-06091-0

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017a. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., Rast, H., 2017b. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101, 1489–1499. https://doi.org/10.1094/PDIS-03-17-0306-RE

Rubio, L., Galipienso, L., Ferriol, I., 2020. Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution. Front Plant Sci 11, 1092. https://doi.org/10.3389/FPLS.2020.01092/BIBTEX

Ruiz-Mirazo, K., Peretó, J., Moreno, A., 2004. A universal definition of life: autonomy and open-ended evolution. Orig Life Evol Biosph 34, 323–346. https://doi.org/10.1023/B:ORIG.0000016440.53346.DC

Rybicki, E.P., 2015. A Top Ten list for economically important plant viruses. Arch Virol 160, 17–20. https://doi.org/10.1007/s00705-014-2295-9

Sabatier, M., Bal, A., Destras, G., Regue, H., Quéromès, G., Cheynet, V., Lina, B., Bardel, C., Brengel-Pesce, K., Navratil, V., Josset, L., 2020. Comparison of Nucleic Acid Extraction Methods for a Viral Metagenomics Analysis of Respiratory Viruses. Microorganisms 8, 1–18. https://doi.org/10.3390/MICROORGANISMS8101539

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94. https://doi.org/10.1016/0022-2836(75)90213-2

Sangiovanni, M., Granata, I., Thind, A.S., Guarracino, M.R., 2019. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics 20. https://doi.org/10.1186/S12859-019-2684-X

Santala, J., Valkonen, J.P.T., 2018. Sensitivity of Small RNA-Based Detection of Plant Viruses . Frontiers in Microbiology .

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A., 2019. The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution 2019 3:3 3, 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schilbert, H.M., Rempel, A., Pucker, B., 2020. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants 2020, Vol. 9, Page 439 9, 439. https://doi.org/10.3390/PLANTS9040439

Silva, J.M.F., Melo, F.L., Elena, S.F., Candresse, T., Sabanadzovic, S., Tzanetakis, I.E., Blouin, A.G., Villamor, D.E.V., Mollov, D., Constable, F., Cao, M., Saldarelli, P., Cho, W.K., Nagata, T., 2022. Virus classification based on in-depth sequence analyses and development of demarcation criteria

using the Betaflexiviridae as a case study. J Gen Virol 103, 001806. https://doi.org/10.1099/JGV.0.001806/CITE/REFWORKS

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J.C., Manuel, M., Philippe, H., Telford, M.J., 2018. A software tool "CroCo" detects pervasive cross-species contamination in next generation sequencing data. BMC Biol 16, 1–9. https://doi.org/10.1186/s12915-018-0486-7

Simmonds, P., Adams, M.J., Benk, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E. v., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology 2017 15:3 15, 161–168. https://doi.org/10.1038/nrmicro.2016.177

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., Weissman, I.L., 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv 125724. https://doi.org/10.1101/125724

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., Al Rwahnih, M., 2021a. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Soltani, N., Stevens, K.A., Klaassen, V., Hwang, M.-S., Golino, D.A., al Rwahnih, M., 2021b. Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics. Viruses . https://doi.org/10.3390/v13061130

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. https://doi.org/10.1038/nbt.3988

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014a. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Strong, M.J., Baddoo, M., Nanbo, A., Xu, M., Puetter, A., Lin, Z., 2014b. Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes. J Virol 88, 10696–10704. https://doi.org/10.1128/JVI.01457-14/SUPPL_FILE/ZJV999099495SO1.PDF

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., Nikolski, M., 2022. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. Frontiers in Bioinformatics 2. https://doi.org/10.3389/FBINF.2022.867111/FULL

Tamisier, L., Colson, C., Maclot, F., Zhang, P., Wang, X., Francis, F., Baurain, D., Massart, S., n.d. Deciphering the intra-host genetic diversity of viruses: insights from the experimental evolution of a plant virus.

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., Massart, S., 2021a. Semi-artificial datasets as a resource for validation of bioinformatics

pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/pcjournal.62

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., Massart, S., 2021b. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. Peer Community Journal 1. https://doi.org/10.24072/PCJOURNAL.62

Tedersoo, L., Albertsen, M., Anslan, S., Callahan, B., 2021. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl Environ Microbiol 87, 1–19. https://doi.org/10.1128/AEM.00626-21/SUPPL_FILE/AEM.00626-21-S0001.PDF

Tejano, L.A., Peralta, J.P., Yap, E.E.S., Panjaitan, F.C.A., Chang, Y.W., 2019. Prediction of Bioactive Peptides from Chlorella sorokiniana Proteins Using Proteomic Techniques in Combination with Bioinformatics Analyses. International Journal of Molecular Sciences 2019, Vol. 20, Page 1786 20, 1786. https://doi.org/10.3390/IJMS20071786

Tekle, Y.I., Grant, J.R., Kovner, A.M., Townsend, J.P., Katz, L.A., 2010. Identification of new molecular markers for assembling the eukaryotic tree of life. Mol Phylogenet Evol 55, 1177–1182. https://doi.org/10.1016/J.YMPEV.2010.03.010

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., Roux, N., 2015. Technical guidelines for the safe movement of Musa germplasm, in: Thomas, J. (Ed.), Technical Guidelines for the Safe Movement of Musa Germplasm. Bioversity International, > >.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/NAR/GKAB1061

Velasco, L., Padilla, C. V., 2021. High-Throughput Sequencing of Small RNAs for the Sanitary Certification of Viruses in Grapevine. Front Plant Sci 12. https://doi.org/10.3389/FPLS.2021.682879/FULL

Vereecke, N., Carnet, F., Pronost, S., Vanschandevijl, K., Theuns, S., Nauwynck, H., 2021. Genome Sequences of Equine Herpesvirus 1 Strains from a European Outbreak of Neurological Disorders Linked to a Horse Gathering in Valencia, Spain, in 2021. Microbiol Resour Announc. https://doi.org/10.1128/MRA.00333-21

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016a. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

Visser, M., Bester, R., Burger, J.T., Maree, H.J., 2016b. Next-generation sequencing for virus detection: covering all the bases. Virol J 13, 85. https://doi.org/10.1186/s12985-016-0539-x

VUNSH, R., ROSNER, A., STEIN, A., 1990. The use of the polymerase chain reaction (PCR) for the detection of bean yellow mosaic virus in gladiolus. Annals of Applied Biology 117, 561–569. https://doi.org/10.1111/J.1744-7348.1990.TB04822.X

Waite, D.W., Liefting, L., Delmiglio, C., Chernyavtseva, A., Ha, H.J., Thompson, J.R., 2022. Development and Validation of a Bioinformatic Workflow for the Rapid Detection of Viruses in Biosecurity. Viruses 14, 2163. https://doi.org/10.3390/V14102163/S1

Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P., Dempsey, D.M., Dutilh, B.E., García, M.L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., Łobocka, M., Oksanen, H.M., Orton, R.J., Robertson, D.L., Rubino, L., Sabanadzovic, S., Simmonds, P., Smith, D.B., Suzuki, N., van Doorslaer, K., Vandamme, A.M., Varsani, A., Zerbini, F.M., 2022. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch Virol 167, 2429–2440. https://doi.org/10.1007/S00705-022-05516-5/TABLES/2

Ward, M., 2016. Action against pest spread—the case for retrospective analysis with a focus on timing. Food Secur 8, 77–81. https://doi.org/10.1007/S12571-015-0532-Z/FIGURES/1

Watzinger, F., Ebner, K., Lion, T., 2006. Detection and monitoring of virus infections by real-time PCR. Mol Aspects Med 27, 254–298. https://doi.org/10.1016/J.MAM.2005.12.001

Whitfield, A.E., Campbell, L.R., Sherwood, J.L., Ullman, D.E., 2003. Tissue Blot Immunoassay for Detection of Tomato spotted wilt virus in Ranunculus asiaticus and Other Ornamentals. Plant Dis 87, 618–622. https://doi.org/10.1094/PDIS.2003.87.6.618

Williams, R.C., 1954. Electron Microscopy of Viruses. Adv Virus Res 2, 183–239. https://doi.org/10.1016/S0065-3527(08)60533-3

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 762302. https://doi.org/10.1186/s13059-019-1891-0

Wood, D.E., Salzberg, S.L., 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, 1–12. https://doi.org/10.1186/GB-2014-15-3-R46/FIGURES/5

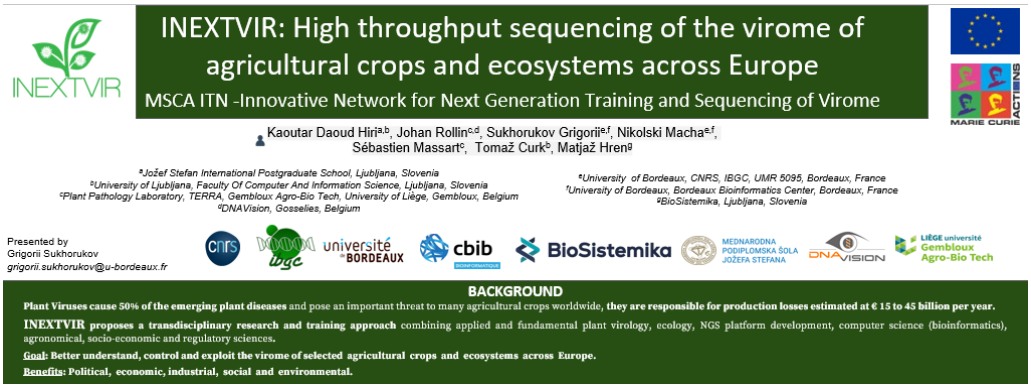
Wu, D., Jospin, G., Eisen, J.A., 2013. Systematic Identification of Gene Families for Use as “Markers” for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. PLoS One 8. https://doi.org/10.1371/JOURNAL.PONE.0077033

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

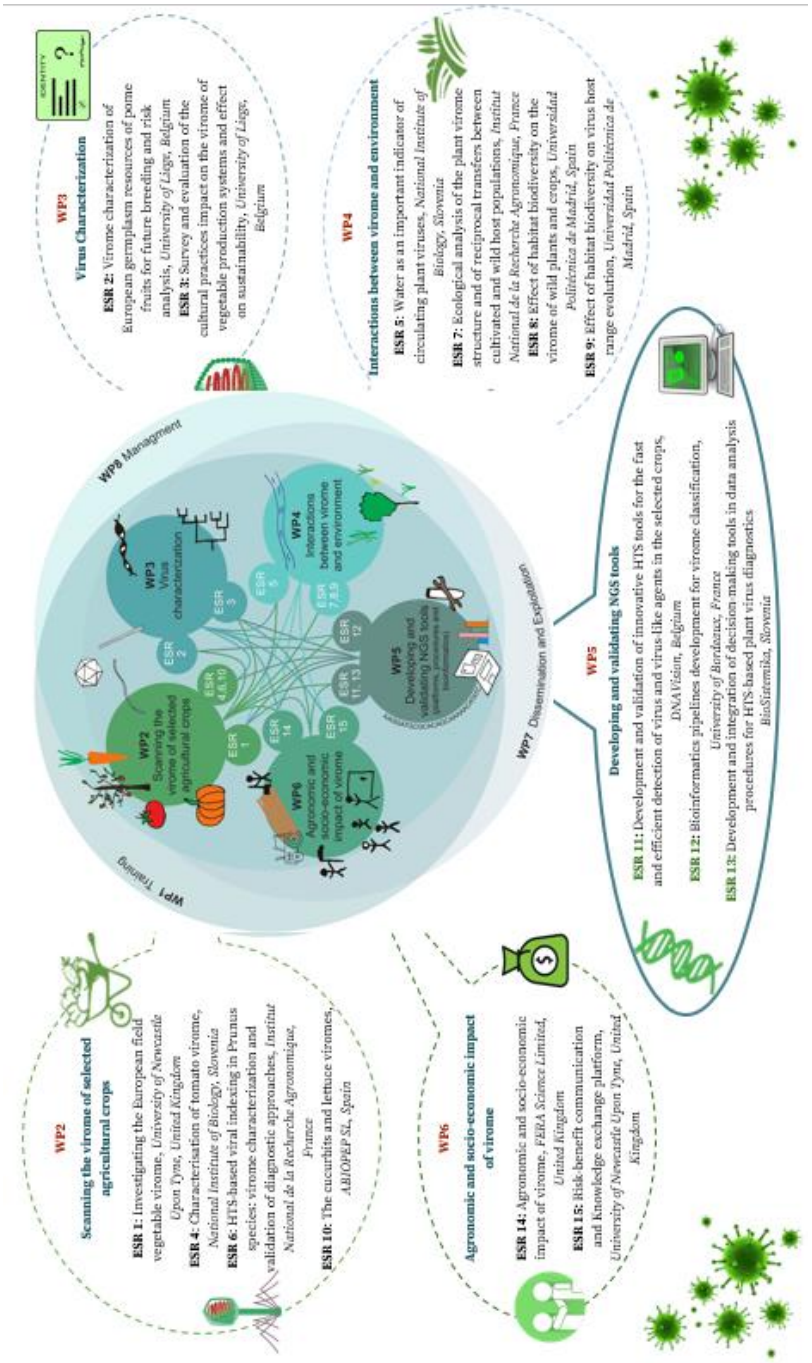
Appendix


burden
analytical
and/or
monitor
previously
electron
extracts
tested
isolates
replicates
inclusivity
distant
although
potential
diagnostic
allowed
analyzed
false
threshold
accessions
high-throughput
now
dilutions
brought
based
used
interpretation
rt-pcr
expert
reads
study
using
overall
testing
high
therefore
positive
finally
five
completed
libraries
molecular
great
huge
per
serial
conclusion
pools
detect
can
ability
viruses
ribodepletion
sequenced
technologies
criteria
fulfilled
germplasm
viral
number
evaluation
routine
repeatability
improvements

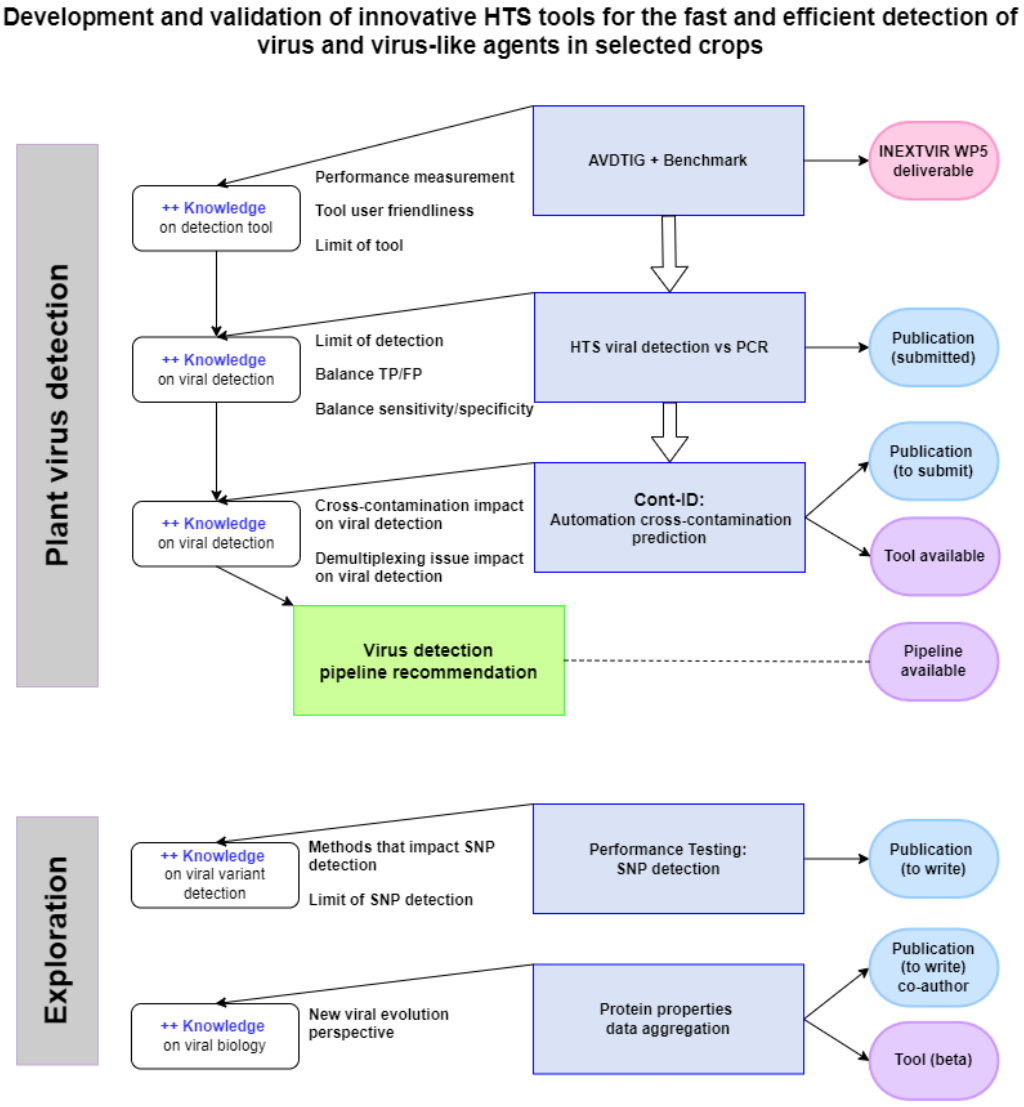
Appendix 1A. Poster European Student Council Symposium (ECSC) 2020.



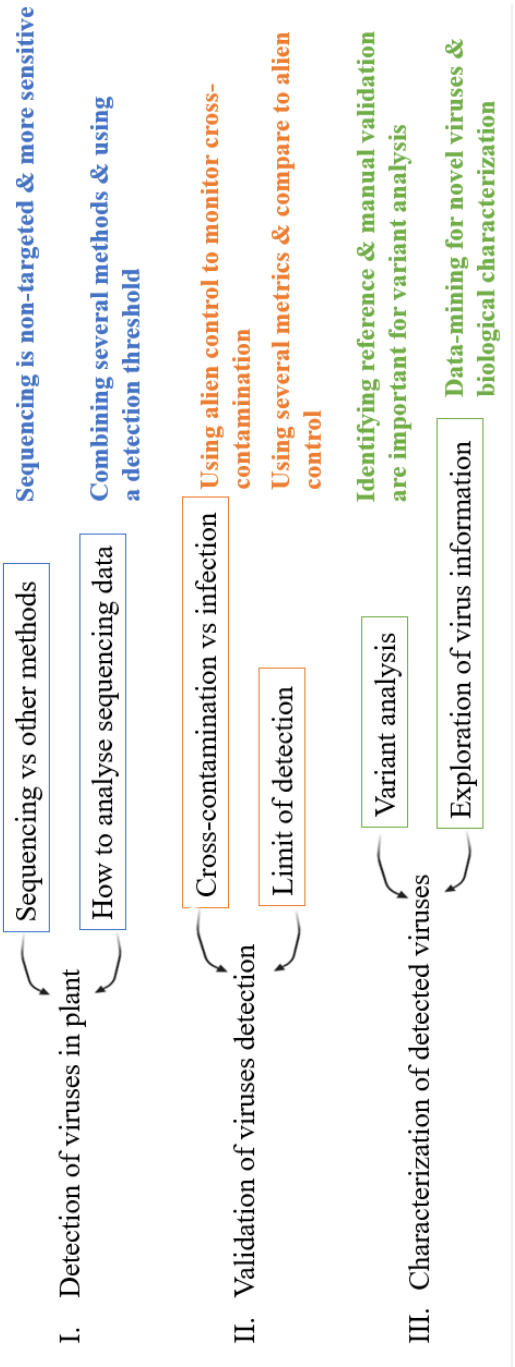
Appendix 1B. Poster European Student Council Symposium (ECSC) 2020.



Appendix 2. Overview of the thesis.



Appendix 3. Sumary achievement of the thesis



Appendix 4. Aknowledgment

This Project has received funding from the European Union´s **Horizon2020** Research and Innovation program under the Marie Sklodowska-Curie Grant Agreement no. **813542**


detection
viruses
virus
plant
techniques
analysis
knowledge
using
data
methods
HTS
improvement
food
better
also