Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection

Lucie Tamisier[1*], Annelies Haegeman[2], Yoika Foucart[2], Nicolas Fouillien[1], Maher Al Rwahnih[3], Nihal Buzkan[4], Thierry Candresse[5], Michela Chiumenti[6], Kris De Jonghe[2], Marie Lefebvre[5], Paolo Margaria[7], Jean Sébastien Reynard[8], Kristian Stevens[3,9], Denis Kutnjak[10], Sébastien Massart[1*]

[1] Université de Liège, Terra-Gembloux Agro-Bio Tech, Plant Pathology Laboratory, Passage des Déportés, 2, 5030 Gembloux, Belgium

[2] Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium

[3] Department of Plant Pathology, University of California, Davis, 95616

[4] Department of Plant Protection, Faculty of Agriculture, University of Sütçü Imam, Kahramanmaras 46060, Turkey

[5] Univ. Bordeaux, INRAE, UMR BFP, CS20032, 33882 Villenave d'Ornon cedex, France

[6] Institute for Sustainable Plant Protection, CNR, Via Amendola 122/D, Bari 70126, Italy

[7] Leibniz Institute - DSMZ, German Collection of Microorganisms and Cell Cultures GmbH, 38124 Braunschweig, Germany

[8] Virology, Agroscope, Nyon, Switzerland

[9] Department of Evolution and Ecology, University of California, Davis, California 95616, USA

[10] Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia

*Corresponding authors: Lucie Tamisier; E-mail: lucie.tamisier@uliege.be, Sebastien Massart; E-mail: sebastien.massart@uliege.be

Keywords

High-Throughput Sequencing, Reference data, Semi-artificial dataset, Plant virus detection, Bioinformatics pipelines, Haplotype reconstruction

Abstract

The widespread use of High-Throughput Sequencing (HTS) for detection of plant viruses and sequencing of plant virus genomes has led to the generation of large amounts of data and of bioinformatics challenges to process them. Many bioinformatics pipelines for virus detection are available, making the choice of a suitable one difficult. A robust benchmarking is needed for the unbiased comparison of the pipelines, but there is currently a lack of reference datasets that could be used for this purpose. We present 7 semi-artificial datasets composed of real RNA-seq datasets from virus-infected plants spiked with artificial virus reads. Each dataset addresses challenges that could prevent virus detection. We also present 3 real datasets

showing a challenging virus composition as well as 8 completely artificial datasets to test haplotype reconstruction software.

In the last decade, High-Throughput Sequencing (HTS) has revolutionized plant virus discovery and diagnosis (Maree *et al.*, 2018; Massart *et al.*, 2014). The main advantage of this technology is that it allows a complete characterization of the virus populations infecting a plant, without any *a priori* knowledge of the infecting viruses. Current HTS platforms can ascertain the molecular sequences of large quantities of nucleic acid fragments at a very low base pair price, allowing the simultaneous sequencing of many samples. The increased use of HTS in the diagnostic field has led to the generation of massive amounts of data and resulted in computational and bioinformatics challenges to process them (*i.e.* storage, processing speed, bioinformatics competence) (Olmos *et al.*, 2018). Many bioinformatics pipelines for plant virus detection have been developed, from easy-to-use commercial software to command line tools (Blawid *et al.*, 2017; Jones *et al.*, 2017). Most of them aim to improve virus detection and/or reduce processing time, but the high number of pipelines available complicate the choice of the most appropriate for a given goal or environment. Moreover, the sequence analysis strategy can have a significant influence on the ability to detect viruses from identical datasets, as shown by a large-scale performance testing involving 21 plant virology laboratories (Massart *et al.*, 2019). Performing a robust benchmarking is therefore essential for the unbiased comparison of the pipelines (Escalona *et al.*, 2016; Jones *et al.*, 2017). In plant disease diagnostics, validation of the bioinformatics pipelines used for the detection of viruses in HTS datasets is at its infancy and there is currently a lack of reference datasets generated for benchmarking purposes. The development of such datasets is a key step in the standardization of bioinformatics protocols, since it allows objective comparison between pipelines. These observations have led to the creation of the Plant Health Bioinformatics Network (PHBN), an Euphresco network project aiming to build a community network of bioinformaticians/computational biologists working on plant health. One of the objectives of this project is to help researchers to compare and validate their virus detection pipelines by creating open access reference datasets.

Creation of the datasets

Two main kinds of reference datasets can be used: real and artificial ones. Working with real datasets offers the benefit of providing real life scenarios which are close to those encountered by plant pathologists and diagnosticians. However, the use of such purely empirical data has limitations since it is impossible to know with an absolute certainty the “true” value that should be used to benchmark the performance of the pipelines (Escalona *et al.*, 2016). Artificial datasets do not have this drawback since their composition is totally controlled and known.

However, completely artificial datasets are often unrealistic and too simple, and may thus fail to represent accurately the complexity of real HTS datasets. In order to overcome the drawbacks of these two approaches, we have chosen to create semi-artificial datasets composed each of a real HTS dataset from virus-infected plants spiked with additional *in-silico* generated viral reads. The artificial component of these semi-artificial datasets is totally known, but the datasets are still complex and close to real-life situations. We also developed and propose some real and some completely artificial datasets, which can be used for specific purposes as explained bellow.

A total of 8 real RNA-seq datasets from virus/viroid-infected plants obtained using Illumina technology have been chosen in order to cover as much as possible host plant diversity (fruit trees, vegetables and biological indicator plants), pathogen diversity (RNA and DNA viruses, viroids) and sequencing options (reads length from 50 to 301 bp, number of reads per dataset from 65,177 to 49,052,832 reads, and single-end or paired-end reads). For each real dataset, the presence of the viruses/viroids identified has been confirmed by PCR and/or ELISA. Five of these real datasets have been used to create 7 semi-artificial datasets (Datasets 1, 2, 3, 4, 5, 6 and 10) (Table 1), either by adding artificial reads of a virus/viroid (already present or not in the dataset) or by removing part of the real viral reads. The artificial viral reads were synthesized using the ART software (Huang *et al.*, 2012) which allows the generation of artificial next-generation sequencing reads showing the same quality score as the reads from a real datasets. For each semi-artificial dataset, similar headers have been assigned to the artificial and real reads, and both types of reads have been mixed in each FASTA file. The three other real datasets (Datasets 7, 8 and 9) were already showing a challenging viral composition (presence of a defective variant, presence of a cryptic virus and presence of several genomic segments showing different concentrations) and have not been modified.

Each dataset has been developed or selected to address one or several challenges that could prevent virus detection or a correct virus identification from HTS data (*i.e.* low viral concentration, new viral species, non-complete genome, etc). In addition, eight fully artificial datasets (Datasets 11-18), composed only of viral reads have also been created. These datasets can be used to test haplotype reconstruction software, the goal being to evaluate the ability to reconstruct all the strains present in a dataset. Each artificial dataset consists of a mix of several stains from the same viral species showing different frequencies. The virus species have been selected to be as divergent as possible. Therefore, the selected viruses have (i) a DNA or RNA genome, (ii) a single or double-stranded genome, (iii) a linear, circular and/or segmented genome, and (iv) show a genome length ranging from 2.8 to 17.1 kb. For each strain, artificial viral reads of 150 bp have been synthesized using the ART software (Huang *et* *al.*, 2012) from NCBI reference genomes and no single nucleotide polymorphisms (SNPs) have been added.

Availability and description of the datasets

A GitLab repository (https://gitlab.com/ilvo/VIROMOCKchallenge) is available and provides a complete description of the composition of each dataset, the methods used to create them, a link to download them and their goals. The datasets themselves are stored in Dryad (datadryad.org). We provide here a quick summary of the composition of the datasets and the challenges they address (Table 1).

- Dataset 1: The challenge addressed is the detection of several virus strains showing different concentrations, some being very low. In this case, one or more strains can be missed, especially if the sample has not been enriched in viral sequences (Barzon *et al.*, 2013; Knierim *et al.*, 2019). The real dataset is composed of mixed infections of citrus tristeza virus (CTV), citrus vein enation virus (CVEV), citrus exocortis viroid (CEVd), citrus viroid III (CVd-III) and hop stunt viroid (HSVd) on citrus. Artificial reads for three CTV strains (JQ911663 – strain T68, KU883267 – strain S1 and MH323442 – strain T36) have been added to the dataset at different read depth.

- Dataset 2: The challenge addressed is the identification of different types of mutations at different frequencies. The viral populations infecting a plant are usually composed of closely related virus genotypes, differing by a few SNPs (substitution) or indels (insertion or deletion) and at differing relative concentrations. Some variants can be missed depending on their frequencies, the bioinformatics strategy or the presence of sequencing errors (Lefterova *et al.*, 2015). The same real data set from a naturally infected citrus as in dataset 1 has been used with the addition of artificial reads for the CTV MH323442 isolate, using 5 nearly identical sequences of this isolate, each differing by 1 substitution, 1 base deletion and 1 base insertion. Artificial reads for the unmutated MH323442 isolate have also been added to the dataset 2. The reads for the various MH323442 variants have been added at different frequencies.

- Dataset 3: The challenge addressed is the detection of several viral/viroid species showing different frequencies and incomplete genome coverage. The assembly process can result in incomplete genome sequences, making virus identification challenging (Boonham *et al.*, 2014), in particular when the whole genome is not completely covered, or when a genomic segment is absent or is covered by a low number of reads in the case of a multipartite virus. The real dataset corresponds to a mixed infection of grapevine rupestris vein feathering virus (GRVFV), grapevine rupestris stem pitting-associated virus (GRSPaV), grapevine leafroll-associated virus 2 (GLRaV2), hop stunt viroid (HSVd) and grapevine yellow speckle viroid 1 (GYSVd1) on grapevine. Reads assigned to GRSPaV, GRVFV and GLRaV2 have been randomly removed in order to obtain incomplete genome coverage for these 3 viruses.

- Dataset 4: The challenge addressed is the detection of closely related viroids. Closely related virus/viroid species within a genus can share high nucleotide identities, leading to taxonomic assignation problems and complicating the identification of the virus/viroid (Thekke-Veetil *et al.*, 2018). The real dataset is composed of mixed infections of grapevine red blotch virus (GRBV), grapevine rupestris stem pitting-associated virus (GRSPaV), hop stunt viroid (HSVd) and grapevine yellow speckle viroid 1 (GSYVd1) on grapevine (Reynard *et al.*, 2018). Artificial reads of grapevine yellow speckle viroid 2 (GYSVd2) isolate DQ377131 have been added to the dataset. This reference shows a pairwise nucleotide identity of 73.9% with the consensus sequence of the naturally present GYSVd1, a portion of the two genomes being very similar while the other part show more variability.

- Dataset 5: The challenge addressed is the detection of a recombinant strain and one of its parents in mixed infection. HTS samples can be infected by genetically close parental and recombinant strains. During the assembly process, it can sometimes be challenging to assemble and detect recombinant genomes while avoiding to create artefactual ones, in particular when using short-sequence reads (Martin *et al.*, 2011). The real dataset contains reads of two potato virus Y (PVY) isolates belonging to different strains (an isolate belonging to the NTN recombinant strain and the N605 isolate belonging to the N strain). Artificial reads to a further two isolates have been added, the parental isolate AY884983 (N strain), and isolate EF026076, a recombinant between isolates belonging to the N and O strains (Hu *et al.*, 2009). Both isolates show an overall pairwise nucleotide identity of 88.2% but the 5' part of their genomes (first ~2,000 nucleotides) are almost identical.

- Dataset 6: The challenge addressed is the detection of a new PVY strain that does not exist in the database, within a dataset already involving other PVY strains. Novel viruses can be detected by homology searches with databases. Nevertheless, viral sequences that are too divergent from known viruses might not be detected by this such searches. Other approaches like homology-independent algorithms may be needed to fully characterize such new viruses (Wu *et al.*, 2015). The real dataset is the same as dataset 5. It has been spiked with artificial reads from the FJ214726 PVY isolate, which was selected because it is among the most divergent PVY isolates available in GenBank (maximum 84% nucleotide identity with any other available PVY isolate). The amino acid sequence of the polyprotein of FJ214726 was obtained and then reverse translated into a nucleotide sequence using the online EMBOSS Backtranseq tool (Madeira *et al.*, 2019). Thanks to the degeneracy of the genetic code, the nucleotide sequence thus obtained was different from the original FJ214726 sequence. Non-synonymous substitutions were further manually added to the new artificial sequence, increasing divergence from any known PVY isolate. The final artificial sequence shows only 71.8% nucleotide identity and 98.9% amino acid identity with FJ214726 and was used to generate the artificial reads finally added to the dataset. The artificial genomic sequence is available in the GitLab repository for comparison purposes.

- Dataset 7: The challenge addressed is the detection of both a defective and a normal length variant from the same sample. Related viral variants infecting a sample and showing similar genome portions can be particularly difficult to distinguish. The real dataset is composed of two variants of tomato spotted wilt virus (TSWV) from tobacco. The genome of TSWV consists of 3 negative single-stranded RNA segments named S, M and L. The variants diverge only for the L genomic segment, one being full length (8,913 bp) and the other being a shorter defective form (2,612 bp) missing the genomic region from genome position 760 to 7,060 bp. The real dataset shows already a challenging composition, and has therefore not been spiked with artificial viruses.

- Dataset 8: The challenge addressed is the detection of a low concentration persistent virus. The real dataset is composed of *P*elargonium flower break virus (PFBV) and *Chenopodium* quinoa mitovirus 1 (CqMV1), a virus from *Chenopodium* which is localized in mitochondria and presents only one ORF that encodes the RNA-dependent RNA polymerase (Nerva *et al.*, 2019). The cryptic virus CqMV1 represents a low proportion of reads (around 0.5%). The real dataset shows already a challenging composition, and has therefore not been spiked with artificial viruses.

- Dataset 9: The challenge addressed is the detection of all the genomic segments of a virus with each segment having a different concentration. The real dataset is composed of *Pistacia* emaravirus B (PiVB), a newly discovered Emaravirus from the pistachio tree (Buzkan *et al.*, 2019). The viral genome is composed of seven distinct negative-sense, single-stranded RNAs, showing different frequencies in the dataset. The real dataset shows already a challenging composition, and has therefore not been spiked with artificial viruses.

- Dataset 10: The challenge addressed is the detection of a new viral strain that does not exist in the database, thus adding a ’virus’ that is not already present in the dataset (in contrast to the challenge addressed in dataset 6). The real dataset is composed of plum bark necrosis stem pitting-associated virus (PBNSPaV) from *Prunus*. A new artificial isolate of plum pox virus (PPV) has been created as described above for the creation of the artificial PVY isolate in dataset 6. The new artificial PPV strain has finally been added to the dataset, and its sequence has been made available as well to be able to compare resulting assemblies with it.

- Datasets 11 to 18 can be used to test the ability to reconstruct haplotypes from mixed infections of virus isolates belonging to the same virus species. They are completely artificial datasets and their composition is summarized in Table 1.

The VIROMOCK challenge

The goal of all these reference datasets is to allow to perform an objective comparison of bioinformatics pipelines used to detect and analyse viruses. At first, researchers can use these datasets to check whether their current pipelines are behaving as expected, and how modifying some parameters can affect their pipeline performance depending on the challenge investigated. Second, it can be interesting for researchers to compare their results with those of other labs/pipelines. For this purpose, we propose to organize a "VIROMOCK challenge". In the frame of this challenge, researchers are encouraged to provide feedback on the results they obtained for each dataset they analyse and on the difficulties they may have encountered. This can simply be done by completing a Google spreadsheet added to each dataset page of the GitLab repository. Then, the results will be compiled for each dataset, helping to identify which pipelines perform best in approximating the real composition of the datasets and providing an idea about the robustness of the parameters used. If researchers agree, the compiled results will be open access on the GitLab repository for each dataset, allowing an easy and objective comparison of the results.

Conclusion

The two main bottlenecks slowing down the adoption of HTS in plant health diagnostics are (i) the lack of consensus on the standardization of the data analysis and (ii) the lack of expertise of some laboratories. Within the frame of PHBN project, we have generated semi-artificial, real and artificial reference datasets in order to help to overcome these bottlenecks. Firstly, the diversity of the challenges addressed by these datasets will allow to benchmark the bioinformatics pipelines used by different laboratories. Secondly, these datasets can also be viewed as open source training materials. They could be extremely valuable for laboratories with little experience, allowing them to improve their skills. Currently, there are many pipelines available, but many laboratories do not know where to start when it comes to the analysis of their HTS data in the context of virus detection. This represents a big challenge, especially in situations where HTS and data analysis are newly established or not part of the routine activities. These datasets will help them to either validate their pipelines or choose the most suitable one for their analyses.

Table 1: Characteristics of each dataset

Dataset	Dataset type	Plant species	Virus/Viroids already present[1]	Modification	Reads (bp)	Total number of reads[2]	Challenge	Dryad DOI
1	Semi-artificial	Citrus	CTV, CVEV, CEVd, CVd-III, HSVd	Addition of CTV	2 x 150	21,703,434 (R1) 21,703,434 (R2)	Viral concentration (at the strain level)	10.5061/dryad.crjdfn32c
2	Semi-artificial	Citrus	CTV, CVEV, CEVd, CVd-III, HSVd	Addition of CTV	2 x 150	21,756,961 (R1) 21,756,961 (R2)	Mutation	10.5061/dryad.ns1rn8pq9
3	Semi-artificial	Grapevine	GRSPaV, GLRaV2, GRVFV, HSVd, GYSVd1	Removing of real viral reads	2 x 150	24,526,416 (R1) 24,526,416 (R2)	Viral concentration (at the species level) + Non complete genome	10.5061/dryad.zs7h44j6d
4	Semi-artificial	Grapevine	GRBV, GRSPaV, HSVd, GYSVd1	Addition of GYSVd2	2 x 75	10,054,658 (R1) 10,054,658 (R2)	Viroids with very similar sequence	10.5061/dryad.jsxksn06w
5	Semi-artificial	Potato	PVY	Addition of PVY	1 x 50	31,277,475	Mix of recombinant and parental viral strains	10.5061/dryad.xgxd254dw
6	Semi-artificial	Potato	PVY	Addition of PVY	1 x 50	31,327,327	New strain	10.5061/dryad.tx95x69vw
7	Real	Tobacco	TSWV	-	2 x 301	1,904,369 (R1) 1,904,369 (R2)	Complete genome + defective form	10.5061/dryad.c2fqz615w
8	Real	Chenopodium	PFBV + mitovirus	-	2 x 301	65,177 (R1) 65,177 (R2)	Cryptic virus + low concentration	10.5061/dryad.wpzgmsbjj
9	Real	Pistachio	PiVB	-	2 x 151 (R1) 2 x 84 (R2)	5,259,903 (R1) 5,259,903 (R2)	Concentration of different genomic segments	10.5061/dryad.p5hqbzkmx
10	Semi-artificial	Prunus	PBNSPaV	Addition of PPV	1 x 75	24,573,681	New strain	10.5061/dryad.rr4xgxd6n
11	Artificial	-	PepMV	-	2 x 150	48,578 (R1) 48,578 (R2)	Haplotype reconstruction of 6 strains	10.5061/dryad.866t1g1nx
12	Artificial	-	*Cassava mosaic virus*	-	2 x 150	48,222 (R1) 48,222 (R2)	Haplotype reconstruction of 4 strains	10.5061/dryad.ns1rn8pqb
13	Artificial	-	BSV	-	2 x 150	47,240 (R1) 47,240 (R2)	Haplotype reconstruction of 6 strains	10.5061/dryad.573n5tb59
14	Artificial	-	PVY	-	2 x 150	52,333 (R1) 52,333 (R2)	Haplotype reconstruction of 5 strains	10.5061/dryad.pc866t1m5

15	Artificial	-	EMDV	-	2 x 150	48,504 (R1) 48,504 (R2)	Haplotype reconstruction of 3 strains	10.5061/dryad.p2ngf1vnq
16	Artificial	-	BPEV	-	2 x 150	49,980 (R1) 49,980 (R2)	Haplotype reconstruction of 4 strains	10.5061/dryad.xpnvx0kcn
17	Artificial	-	LChV1	-	2 x 150	49,513 (R1) 49,513 (R2)	Haplotype reconstruction of 5 strains	10.5061/dryad.9p8cz8wdh
18	Artificial	-	BYDV	-	2 x 150	46,917 (R1) 46,917 (R2)	Haplotype reconstruction of 6 strains	10.5061/dryad.zkh18937t

[1] R1: Forward read, R2: Reverse read.

[2] CTV: citrus tristeza virus, CVEV: citrus vein enation virus, CEVd: citrus exocortis viroid, CVd-III: citrus viroid III, HSVd: hop stunt viroid, GRSPaV: grapevine rupestris stem pitting-associated virus, GLRaV2: grapevine leafroll-associated virus 2, GRVFV: grapevine rupestris vein feathering virus, GYSVd1: grapevine yellow speckle viroid 1, GRBV: grapevine red blotch virus, PVY: potato virus Y, TSWV: tomato spotted wilt virus, PFBV: *Pelargonium* flower break virus, PiVB: *Pistacia* emaravirus B, PBNSPaV: plum bark necrosis stem pitting-associated virus, PepMV: pepino mosaic virus, BSV: banana streak virus, EMDV: eggplant mottled dwarf virus, BPE: bell pepper endornavirus, LChV1: little cherry virus 1, BYDV: barley yellow dwarf virus

Funding

This work reports the results of the Plant Health Bioinformatics Network (PHBN) Euphresco project (European Phytosanitary Research Coordination), funded by Special Research Funds (FSR) of Liège University (byPOP project), and the Belgian Federal Government (FPS Health project RI 18/A-289 PHBN).

References

Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S. and Palù, G. (2013) Next-generation sequencing technologies in diagnostic virology. *J. Clin. Virol.* **58**, 346–350.

Blawid, R., Silva, J. and Nagata, T. (2017) Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Ann. Appl. Biol.* **170**, 301–314.

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

Buzkan, N., Chiumenti, M., Massart, S., Sarpkaya, K., Karadağ, S. and Minafra, A. (2019) A new emaravirus discovered in Pistacia from Turkey. *Virus Res.* **263**, 159–163.

Escalona, M., Rocha, S. and Posada, D. (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* **17**, 459.

Hu, X., Karasev, A.V., Brown, C.J. and Lorenzen, J.H. (2009) Sequence characteristics of potato virus Y recombinants. *J. Gen. Virol.* **90**, 3033–3041.

Huang, W., Li, L., Myers, J.R. and Marth, G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594.

Jones, S., Baizan-Edge, A., MacFarlane, S. and Torrance, L. (2017) Viral diagnostics in plants using next generation sequencing: computational analysis in practice. *Front. Plant Sci.* **8**, 1770.

Knierim, D., Menzel, W. and Winter, S. (2019) Immunocapture of virions with virus-specific antibodies prior to high-throughput sequencing effectively enriches for virus-specific sequences. *PloS One* **14**, e0216713.

Lefterova, M.I., Suarez, C.J., Banaei, N. and Pinsky, B.A. (2015) Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. *J. Mol. Diagn.* **17**, 623–634.

Madeira, F., Park, Y.M., Lee, J., et al. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641.

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

Martin, D.P., Lemey, P. and Posada, D. (2011) Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.* **11**, 943–955.

Massart, S., Chiumenti, M., De Jonghe, K., et al. (2019) Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. *Phytopathology* **109**, 488–497.

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

Nerva, L., Vigani, G., Di Silvestre, D., Ciuffo, M., Forgia, M., Chitarra, W. and Turina, M. (2019) Biological and molecular characterization of Chenopodium quinoa mitovirus 1 reveals a distinct small RNA response compared to those of cytoplasmic RNA viruses. *J. Virol.* **93**.

Olmos, A., Boonham, N., Candresse, T., et al. (2018) High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. *EPPO Bull.* **48**, 219–224.

Reynard, J.-S., Brodard, J., Dubuis, N., Zufferey, V., Schumpp, O., Schaerer, S. and Gugerli, P. (2018) Grapevine red blotch virus: Absence in Swiss vineyards and analysis of potential detrimental effect on viticultural performance. *Plant Dis.* **102**, 651–655.

Thekke-Veetil, T., Ho, T., Postman, J., Martin, R. and Tzanetakis, I. (2018) A Virus in American Blackcurrant (Ribes americanum) with Distinct Genome Features Reshapes Classification in the Tymovirales. *Viruses* **10**, 406.

Wu, Q., Ding, S.-W., Zhang, Y. and Zhu, S. (2015) Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. *Annu. Rev. Phytopathol.* **53**, 425–444.