Final Report

Project title (Acronym)
Plant Health Bioinformatics Network (PHBN)

Project duration:

Start date:	19-03-01
End date:	22-02-28

Contents

1. Research consortium partners

Partner 1			
Organisation	Flanders Research Institute for Agriculture, Fisheries and Food (ILVO)		
Name of Contact (incl. Title)	Dr. Annelies Haegeman	**Gender**:	F
Job Title	Senior researcher		
Postal Address	Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium		
E-mail	annelies.haegeman@ilvo.vlaanderen.be		
Phone	+32 9 272 28 79		

Partner 2			
Organisation	Canadian Food Inspection Agency – Plant Research & Strategies, Canada		
Name of Contact (incl. Title)	Dr. Michael Rott	**Gender**:	M
Job Title	Research Scientist		
Postal Address	8801 East Saanich Rd, North Saanich, British Columbia, Canada, V8L 1H3		
E-mail	mike.rott@inspection.gc.ca		
Phone	+1 250 363 6650		

Partner 3			
Organisation	Aarhus University, Denmark		
Name of Contact (incl. Title)	Prof. Dr. Mogens Nicolaisen	**Gender**:	M
Job Title	Professor		
Postal Address	Forsøgsvej 1, building 7611, B243, 4200 Slagelse, Denmark		
E-mail	mn@agro.au.dk		
Phone	+4587158137		

Partner 4			
Organisation	Anses, French Agency for Food, Environmental and Occupational Health & Safety, France		
Name of Contact (incl. Title)	Dr. Benoit Remenant	**Gender**:	M
Job Title	Postdoctoral scientist		
Postal Address	ANSES Plant Health Laboratory, 7 Rue Jean Dixméras, CEDEX 01, 49044 Angers, France		
E-mail	benoit.remenant@anses.fr		

Partner 5			
Organisation	National Institute for Agronomic Research, France		
Name of Contact (incl. Title)	Dr. Thierry Candresse	**Gender**:	M
Job Title	Directeur de recherche		
Postal Address	UMR 1332 Biologie du Fruit et Pathologie, INRA, University of Bordeaux, 33140 Villenave d'Ornon, France		
E-mail	thierry.candresse@inra.fr		
Phone	+33 557 12 23 89		

Partner 6			
Organisation	Ministry of Agriculture, Plant Biosecurity, Plant Protection and Inspection Services, Israel		
Name of Contact (incl. Title)	Dr. Noa Sela	**Gender**:	F
Job Title	Research engineer		
Postal Address	Dept. of Plant Pathology Volcani Center-ARO P.O.Box 6 Bet-Dagan 50250, Israel		
E-mail	noa@volcani.agri.gov.il		
Phone	+972 39683986		

Partner 7			
Organisation	Ministry of Agriculture Forestry and Food, Slovenia		
Name of Contact (incl. Title)	Dr. Denis Kutnjak	**Gender**:	M
Job Title	National Institute of Biology (NIB), Department of Biotechnology and Systems Biology, Večna pot 111, SI-1000 Ljubljana, Slovenia		
Postal Address	Research associate		
E-mail	denis.kutnjak@nib.si		
Phone	+386 59 23 28 30		

Partner 8			
Organisation	Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria, Spain		
Name of Contact (incl. Title)	Dr. Antonieta De Cal	**Gender**:	F
Job Title	Researcher		
Postal Address	Ctra. de La Coruña Km. 7,5, 28040 Madrid		
E-mail	cal@inia.es		
Phone	+34 913476839		

Partner 9			
Organisation	Leibniz-Institut DSMZ, Germany		
Name of Contact (incl. Title)	Dr. Paolo Margaria	**Gender**:	M
Job Title	Senior scientist		
Postal Address	Leibniz Institute-DSMZ, Inhoffenstrasse 7b, 38124 Braunschweig, Germany		
E-mail	paolo.margaria@dsmz.de		
Phone	+49 531 2616-0		

Partner 10			
Organisation	National Research Council, Italy		
Name of Contact (incl. Title)	Dr. Laura Miozzi & Dr. Michela Chiumenti	**Gender**:	F; F
Job Title	Researcher		
Postal Address	Institute for Sustainable Plant Protection, National Research Council, Strada delle Cacce, 73, 10135 Torino, Italy Institute for Sustainable Plant Protection, National Research Council, Via Amendola, 122/D, 70126 Bari, Italy		
E-mail	laura.miozzi@ipsp.cnr.it & michela.chiumenti@ipsp.cnr.it		
Phone	+39 113977917 & +39 805443071		

Partner 11			
Organisation	University of Greenwich, National Research Institute, Great Britain		
Name of Contact (incl. Title)	Dr. Steven Okinyi Sewe	**Gender**:	M
Job Title	Researcher		
Postal Address	Natural Resources Institute, University of Greenwich, Central Avenue, Chatham Maritime, Kent ME4 4TB, UK		
E-mail	s.o.sewe@greenwich.ac.uk		

Partner 12			
Organisation	NAK Services, the Netherlands		
Name of Contact (incl. Title)	Dr. Inge van Duivenbode	**Gender**:	F
Job Title	Researcher		
Postal Address	Dutch General Inspection Service for Agricultural Seed and Seed potatoes (NAK), Randweg 14, 8304 AS Emmeloord, Netherlands		
E-mail	i.vanduivenbode@nak.nl		
Phone	+31 615908730		

Partner 13			
Organisation	Naktuinbouw, the Netherlands		
Name of Contact (incl. Title)	Dr. Thomas van Gurp	**Gender**:	M
Job Title	Senior bioinformatics scientist		
Postal Address	NAKtuinbouw Sotaweg 22, 2371 GD Roelofarendsveen, Netherlands		
E-mail	t.v.gurp@naktuinbouw.nl		
Phone	+31 71 332 62 62		

Partner 14			
Organisation	Agroscope, Switzerland		
Name of Contact (incl. Title)	Dr. Olivier Schumpp	**Gender**:	M
Job Title	Head of Group Virology, Bacteriology & Phytoplasmology		
Postal Address	Agroscope, Plant Protection Department, Route de Duillier 50, Case Postale 1012, CH-1260 Nyon 1, Switzerland		
E-mail	olivier.schumpp@agroscope.admin.ch		
Phone	+41 58 460 43 71		

2. Short project report

2.1. Short executive summary

Plant disease detection by high-throughput sequencing (HTS) is a relatively new and fast developing discipline, with very variable levels of expertise in phytopathology and diagnostics laboratories across the world. The overall goal of the “Plant Health Bioinformatics Network” project (PHBN) was to join different laboratories working with HTS applied to plant disease diagnostics problems and stimulate the exchange of information regarding HTS data analysis, as well as on the interpretation of the results of HTS data in a plant diagnostic context.

In a collaborative effort of > 20 scientists from 11 different countries, **open source training materials were developed**. This resulted in the guide ‘*A primer on the analysis of high-throughput sequencing data for detection of plant viruses*‘, which is useful for both beginners and experts. This guide includes a glossary of terms, a flowchart (showing the typical workflow of an analysis), a checklist with things to keep in mind during data processing, a checklist with points of consideration during taxonomic classification and a quick-start guide. Data analysis pipelines were converted to training materials and compiled with other already well-documented pipelines to make them publicly available.

Training people alone is not sufficient to develop good bioinformatics skills. People may follow a tutorial meticulously, but if the steps/parameters used are not suited for the specific case they investigate, they might misinterpret the results. In order to make virologists and bioinformaticians more aware of the strengths and weaknesses of their pipeline, **(semi-)artificial datasets** were designed and tested. Nine challenges in data analysis that can occur when analyzing HTS datasets for the detection and identification of plant viruses were identified. Based on these challenges, several plant-derived Illumina RNA-seq datasets were selected from different international partners. Three of them showed already one of the challenges and were not modified. For 7 other datasets, artificial reads were added as spike-in, with known read numbers, to mimic one of the challenges. Finally, 8 completely artificial datasets were made for haplotype reconstruction. The (semi-)artificial datasets were made publicly available and recommended by “Peer Community in Genomics”. A **VIROMOCK challenge** was then launched to encourage scientists to analyze the data and upload their results. Although only 29 reports were received (i.e. on average 3 per dataset), we were able to observe that most differences between the participants were due to mapping settings and the choice of the reference genome(s).

Finally, we wanted to demonstrate the potential of HTS in the detection of (non-viral) plant pathogens and pests by re-analyzing existing RNA-seq datasets in an **RNA-seq screening effort**. More specifically, we asked the plant virology community to re-analyze some of their existing datasets, in order to check if traces could be found of non-viral pathogens. This is often overlooked since most virologists only compare the reads or contigs with plant virus sequence databases. In total 15 scientists participated in the screening, together analyzing 101 datasets of which 37 datasets were selected for detailed analysis at ILVO (BE). 29 of the 37 datasets revealed the potential presence of non-viral plant pathogens, with fungi, insects and mites the most observed organism categories. These results show that RNA-seq data generated by virologists can be used to investigate the potential presence of other potentially harmful organisms or potential virus vectors.

2.2. Project aims

High-throughput sequencing (HTS), also referred to as Next Generation Sequencing (NGS), has revolutionized biology and medicine during the past decade. The technique allows the sequencing of millions of DNA molecules in parallel at a low cost. As a consequence, the throughput of molecular analyses has drastically changed, because many samples can be pooled and many genes and/or genomes can be analyzed at once. HTS-based disease diagnostics is beginning to find its way to the clinic for human pathogens (Goldberg *et al*., 2015), and the same trend is expected for plant disease diagnostics (Massart *et al*., 2014). Unlike previous diagnostic sequencing, HTS can deliver a full qualitative and quantitative analysis of the DNA or RNA sequences within a sample in a single test, and thereby promises improved diagnostic yield (Hardwick *et al*., 2017). However, we are still far from wide adoption in plant health diagnostic laboratories since the implementation of HTS still faces major challenges, for example the lack of standards and the varying levels of expertise across the different laboratories. The resulting huge amounts of HTS data – millions of sequences – caused the blooming of the bioinformatics and computational biology scientific fields. Bioinformaticians are seen as the 'missing link' required for improving multidisciplinary research since they can bridge biological sciences, informatics, and mathematics (Vincent and Charette, 2015). Indeed, trained and experienced bioinformaticians are scarce, and on top of that, the available techniques and analyses methods tend to vary according to the discipline (bacteriology, virology, mycology, nematology, entomology) which makes it harder for bioinformaticians working in several disciplines to keep up with all the developments in each field. On the other hand, plant pathologists are often not trained to do bioinformatics analyses, which can have a steep learning curve. Lack of staff and/or expertise are therefore the main reasons why plant diagnostic laboratories do not use HTS (as assessed by an anonymous questionnaire sent to diagnostic laboratories in January 2018).

In this project we wanted to encourage diagnostic laboratories to start using HTS, or get more out of their data, by focusing on the data analysis or bioinformatics part, more specifically on the following aspects: 1) promoting the exchange of expertise among different laboratories by developing training materials, 2) developing tools that can help the comparison and validation of bioinformatics pipelines, and 3) raising awareness of the potential of HTS in a plant diagnostic context.

Since many applications and research goals across plant pathology laboratories in Europe and beyond are very similar, it made sense to build a community network across bioinformaticians and/or plant pathologists to exchange knowledge and hence avoid developing different bioinformatics pipelines. Sixteen plant pathologists / bioinformaticians from 9 different countries with different backgrounds (bacteriology, mycology, virology) and different levels of experience with HTS data analysis discussed the idea. The outcome of this meeting was the identification of the largest needs within the community, and possible ways to improve these. From this, the main goals of this project were extracted, which were 4-fold:

1) Develop training materials to help unexperienced laboratories get started;
2) Develop complex artificial datasets for pipeline testing and validation, including comparing pipelines between laboratories;
3) Transfer some of the knowledge built in the virology community to other disciplines in plant pathology;
4) Improve communication by sharing pipelines and workflows as well as outreach to stakeholders.

2.3. Description of the main activities

The project was divided into 5 work packages, which are presented in Figure 1. For each work package (except WP1 which deals with the management of the project), the main activities are described below.

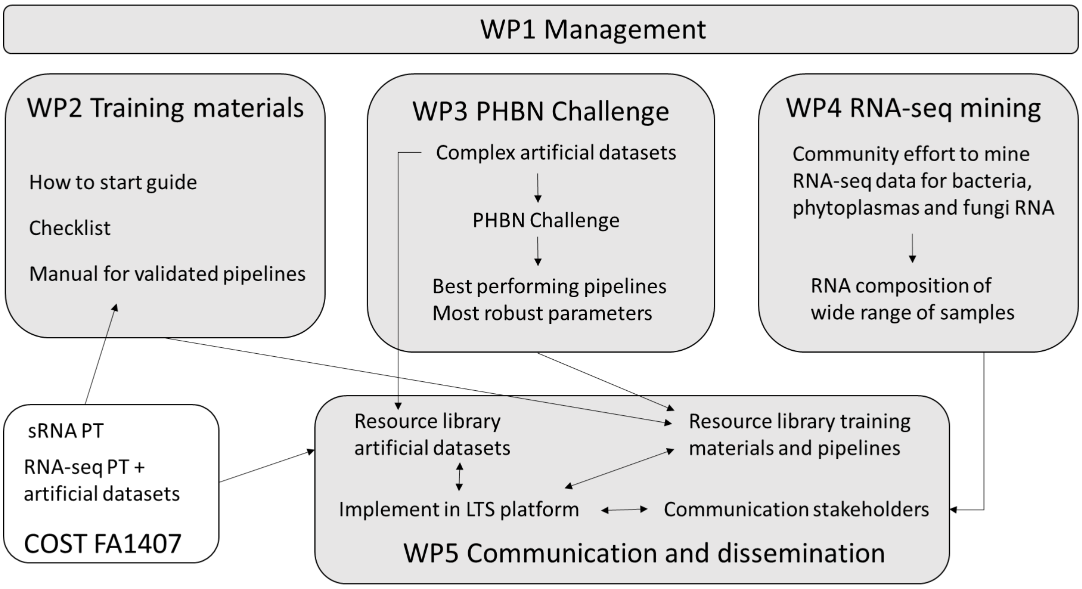
WP1 Management
WP2 Training materials
How to start guide
Checklist
Manual for validated pipelines
WP3 PHBN Challenge
Complex artificial datasets
PHBN Challenge
Best performing pipelines
Most robust parameters
WP4 RNA-seq mining
Community effort to mine RNA-seq data for bacteria, phytoplasmas and fungi RNA
RNA composition of wide range of samples
sRNA PT
RNA-seq PT + artificial datasets
COST FA1407
Resource library artificial datasets
Resource library training materials and pipelines
Implement in LTS platform
Communication stakeholders
WP5 Communication and dissemination

Figure 1 Overview of the different work packages of the PHBN project

Work package 2 had as main goal the development of open-source training materials:

- A beginner's guide to set the context for people with little experience in HTS sequencing. This beginner's guide includes a glossary to explain basic terminology, and a flowchart on how to start working with HTS in a diagnostic context;
- A checklist with critical points when analyzing HTS data;
- Development of training materials from selected data analysis pipelines.

During a meeting with the project partners, it was decided to combine the first two training materials (beginner's guide and checklist) and write it as an (open access) A1 publication. For the training materials on the data analysis, all transnational partners were asked which of their pipelines would be suited to be transformed into training materials. Next, a selection was made, and this selection was intended to cover multiple approaches: for beginners (using graphical software) to experts (using various pieces of command line software). After the selection, the partners prepared documents with detailed information on their pipelines, and sent them to ILVO (BE). At ILVO (BE), all pipelines were transformed to a uniform format (markdown language) and screenshots were added to clarify the steps in the pipeline.

In **work package 3**, (semi-)artificial datasets for plant viruses were made. These complex datasets containing plant RNA and multiple viruses are either completely artificial, hence with exact known composition; or semi-artificial: consisting of real datasets which were spiked with extra data. In a next step, we launched the VIROMOCK challenge, by inviting virologists to analyze the (semi-)artificial data. People were encouraged to participate via e-mail, social media and presentations. On the online repository, links were added to shared spreadsheets where people were able to submit the results of their analysis. These results were then placed on the corresponding dataset page of the online repository, to enable comparison between participants. The VIROMOCK challenge allowed participants to tweak the parameters of their pipelines as such that they could approximate the real composition of all the datasets as closely as possible. This allowed them to get more insight in the performance of their pipelines in different situations.

Most applications of HTS in plant disease diagnostics are in virology. For other plant pathogens, such as bacteria, phytoplasmas and fungi, there has been much less attention to

HTS for disease detection. For some of these pathogens however (especially the non-cultivable ones), HTS based detection could definitely be an interesting technique to add to the toolbox of the pathologist. Many virology laboratories have RNA-seq datasets available derived from different plant hosts that were used for detection of viruses. In **work package 4**, the community screened their plant RNA-seq datasets for the presence of other (unexpected) pathogens in their data to explore the usefulness of RNA-seq data for other plant pathology disciplines. This community effort was organized as follows:

- In Phase I, the participants needed to download a rRNA reference database which they used to map their dataset(s) against. The mapping report was then sent back to ILVO (BE) for further processing. Next, a visual overview of the rRNA content per organism category (plants, bacteria, fungi, oomycetes, phytoplasmas, insects, spiders and mites and others) was made at ILVO (BE). Finally, a report was returned to the participant including a suggestion of samples that were selected for more detailed analysis.
- In Phase II, participants uploaded the raw data of selected samples to ILVO (BE) where two types of analyses were done. In “Analysis 1”, an RNA assembly was done using rnaSPAdes (Bankevich *et al*., 2012). The resulting contigs were then taxonomically assigned by retaining the top hit from a diamond blastx search against the UniProt protein database. This analysis hence focused on the protein coding genes present in the data. The resulting taxonomic classification was visualized using Krona *(Ondov et al*., 2011), taking the contig length into account as “magnitude”, and adding the %identity of the blast hit as “score”. As an alternative analysis, “Analysis 2”, the reads were directly classified taxonomically using Kraken2 (Wood *et al*., 2019) (with Genbank’s non-redundant Nucleotide database as reference), and visualized using Krona.
- In Phase III, the participants received the Krona reports of the detailed analyses where they could interactively check the detailed taxonomic classification results of their samples. Based on the visual evaluation of these results, they could suggest potentially interesting pathogenic organisms. Furthermore, the Krona plot of “Analysis 2” (Kraken2 analysis) was screened for the presence of pathogens by setting a cut-off of 100 reads per million (rpm). Pathogenic taxa with 100-500 rpm were considered “plausible” to be present, while taxa with >500 rpm were considered “very plausible”. Finally, a questionnaire was sent to the participants, asking for their opinion about the usefulness of these type of metagenomics analyses.

As positive controls, new RNA-seq data from plant tissue infected with a known (non-viral) pathogen was generated. In this way we could confirm that the infected pathogen indeed leaves traces in RNA-seq data.

Work package 5 dealt with the dissemination of the training materials and results from the previous work packages. For each of the scientific work packages (WP2, WP3 and WP4) a scientific publication was envisaged. Furthermore, three publicly available repositories were set up on the platform GitLab and data was shared through platforms Zenodo and Dryad (see point 4). In addition, the planned activities and results of this project were routinely shared on social media (Twitter), and were presented at several scientific events such as an INEXTVIR consortium seminar (11/06/2020), Belgian Scientific Plant Health Symposium (15/10/2020), webinar of the American Association of Phytopathology (03/03/2021), 11th meeting of the EPPO Panel on Diagnostics and Quality Assurance (12/03/2021), AAB International Advances in Plant Virology meeting (21/04/2021), 38th Annual meeting of the Mid Atlantic Plant Molecular Biology Society (16/08/2021), Empowering Biodiversity Research Conference II (24-25/05/2022).

2.4. Main results

In **work package 2**, open source training materials were developed. A beginner’s guide on plant virus diagnostics using HTS was written in the form of a publication (see point 4 for detailed reference). There were already similar publications in other fields, for example on how to analyse bacterial genomes (Edwards and Holt, 2013), or how to do eukaryotic genome annotation (Yandell and Ence, 2012). The paper was divided into six sections that dealt with practical questions such as “what do you need to get started”, “how to prepare the samples and sequence nucleic acids” and “how to analyze the data”. Attention was also given to the different types of similarity searches and taxonomic classification methods. A flowchart (showing the typical workflow of an analysis), a glossary of terms, two checklists and a quick-start guide were made as well.

Next to the beginner’s guide, some analysis pipelines from different partners were selected to be transformed into training materials. We chose to include pipelines that were quite varied: different levels (from beginner to expert) and different types of software (open versus licensed, graphical vs command line). Three of the selected pipelines (Virusdetect, Virtool, Virannot) are well established and widely used pipelines with extensive training materials available (Lefebvre *et al*., 2019; Zheng *et al*., 2017). These pipelines are hence only mentioned on our training repository website, with the link to their own websites. Similarly, for a pipeline called Angua a link has been added to its GitLab page. For other pipelines (CLC_NIB_1, Geneious_DSMZ_1, CL_IPSP-CNR_1 and CL_ILVO_1), training materials were developed in markdown format, and these were shared on the training materials repository (see point 4).

In **work package 3**, (semi-)artificial datasets were created. With the international consortium, a list of challenges that are encountered when analyzing HTS data for virus detection was compiled. To identify / create datasets that are presenting one or more of the identified challenges, the consortium partners were asked to supply data that could be used as a starting point. The data had to be Illumina plant-derived RNA-seq data. Eight datasets were retrieved from partners and were analyzed at ULG. After this analysis, three of the datasets (datasets 7, 8 and 9) were perfect cases for challenges to be tested. Hence, these datasets were not modified or spiked with artificial reads. The other 5 real datasets were used to create 7 semi-artificial datasets, each reflecting one or more of the challenges. In addition 8 completely artificial datasets (datasets 11-18) were constructed consisting of a mix of several strains from the same viral species at different frequencies. An overview of the challenges and datasets is shown in Figure 2.

Once all data was documented, the ’VIROMOCK challenge‘ was launched. The challenge for the VIROMOCK participants was to analyze (some of) the datasets with their own pipeline(s), and approximate the expected results as closely as possible. This helped the participants in understanding their own data analysis pipeline better (for example effect of some parameters, strengths and weaknesses). In total 29 reports were received from analyzed datasets by the community. There were too little reports per dataset to do thorough comparisons, but we were able to infer that the mapping algorithm and settings and also the choice of reference sequences included during the mapping have a large influence on the results.

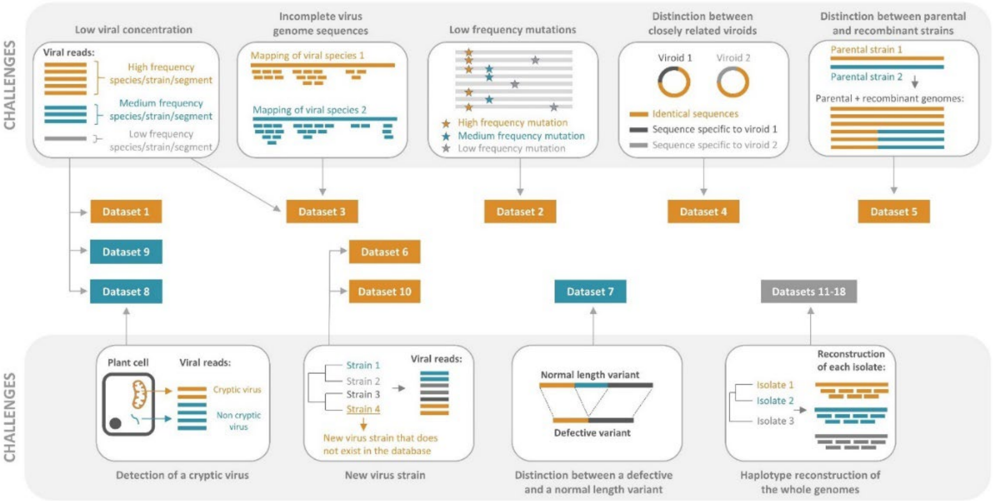
CHALLENGES
Low viral concentration
Viral reads:
High frequency species/strain/segment
Medium frequency species/strain/segment
Low frequency species/strain/segment
Incomplete virus genome sequences
Mapping of viral species 1
Mapping of viral species 2
Low frequency mutations
High frequency mutation
Medium frequency mutation
Low frequency mutation
Distinction between closely related viroids
Viroid 1
Viroid 2
Identical sequences
Sequence specific to viroid 1
Sequence specific to viroid 2
Distinction between parental and recombinant strains
Parental strain 1
Parental strain 2
Parental + recombinant genomes:
Dataset 1
Dataset 3
Dataset 2
Dataset 4
Dataset 5
Dataset 9
Dataset 6
Dataset 8
Dataset 10
Dataset 7
Datasets 11-18
CHALLENGES
Plant cell
Viral reads:
Cryptic virus
Non cryptic virus
Detection of a cryptic virus
Strain 1
Strain 2
Strain 3
Strain 4
Viral reads:
New virus strain that does not exist in the database
New virus strain
Normal length variant
Defective variant
Distinction between a defective and a normal length variant
Reconstruction of each isolate:
Isolate 1
Isolate 2
Isolate 3
Haplotype reconstruction of the whole genomes

Figure 2 Schematic representation of the bioinformatics challenges presented that could prevent virus detection. Each challenge is addressed by at least one dataset. The datasets are either real (blue), semi-artificial (orange) or completely artificial (grey).

In **work package 4**, the RNA-seq community effort, the goal was to have participants re-analyze their data to see if any non-viral pathogens (which are typically not searched for) are present. In Phase I, we let the participants map their own datasets against a LSU rRNA database to be able to quickly screen if their samples might contain lots of non-plant, non-viral sequences. In Phase II, selected samples were transferred to ILVO (BE) for more detailed taxonomic analysis. In Phase III the results were interpreted and feedback was asked from the community. In total 15 different scientists from 10 different countries participated, with a number of samples ranging from 1 to 20 per participant, with 101 samples in total. After receiving the mapping reports from the participants, the mapping statistics of the rRNA mapped reads were processed at ILVO (BE), and graphs were produced showing the rRNA content of each sample per participant. From these graphs, 37 datasets were selected for further in-depth analysis using two types of analyses as explained above. An overview of these results is shown in Figure 3. In 29 of the 37 samples (78%), traces of non-viral pathogens were found (>100 reads per million). The most observed organism categories were fungi (15/37 samples, 41%), insects (13/37 samples, 35%) and mites (9/37 samples, 24%). Nematodes were not observed and only a few samples showed the presence of plant pathogenic phytoplasmas (1/37 samples, 3%), bacteria (3/37 samples, 8%) and oomycetes (4/37 samples, 11%).

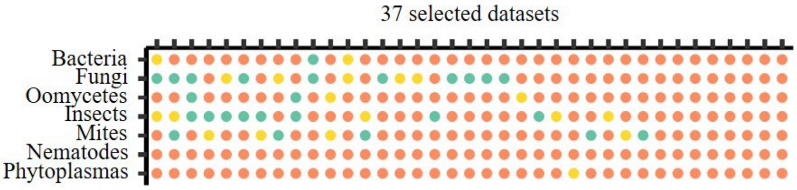

Figure 3 Overview of the selected datasets and the potential presence of non-viral pathogens. Red: presence is not plausible (<100 rpm), yellow: presence is plausible (100-500rpm), green: presence is very plausible (>500rpm).

2.5. Conclusions and recommendations to policy makers

WP2 produced a beginner's guide to detect plant viruses through HTS that was published as A1 paper in the open access journal Microorganisms, including several handy appendices such as a glossary and checklists. Training materials for several pipelines were made and shared online. We received a lot of positive feedback from the community, both from beginners and from experts, who indicated that they learned a lot from the developed materials.

The artificial datasets developed in WP3 provide a useful resource to evaluate the effectiveness of bioinformatics pipelines to cope with typical challenges in plant virus detection using high throughput sequencing. These datasets were made publicly available through GitLab, Zenodo and Dryad, and an accompanying publication was written, published in Peer Community Journal (open access). By checking the results of the different participants, we could conclude that identification of viruses and viroids on species level is typically no problem, also not if there are multiple species present at different abundances. The main lessons learned from the results of the participants is that the different mapping strategies can lead to differences in results, mainly in identifying the closest related strain from the public database. Also relative frequencies of abundance of different species/strains and/or coverage statistics can deviate a lot between participants. This suggests that the mapping algorithm and settings and also the choice of reference sequences included during the mapping have a large influence. Some participants also relied too much on the mapping results rather than putting efforts in assembly or hybrid approaches. Finally, detection of mutations (SNPs and indels) proved no problem for the participants as long as the relative frequency of the mutations was higher than the frequency of the noise. Despite several attempts to motivate scientists in the plant virology community, only a limited number of people participated to the VIROMOCK challenge. It was of course quite some work to read the information on the datasets, analyse them and report the results. Nevertheless, we are confident that the open source datasets will be frequently used by the community in the future during pipeline testing and validation.

In WP4 we organized an RNA-seq community effort to help virologists re-analyze some of their RNA-seq datasets to see if traces from other pathogens could be found. Fifteen persons participated and several unexpected pathogens were presumably found in the data. Although in some cases, it was very clear that the pathogen was present, in many cases interpretation remained difficult. Low numbers of reads are unreliable, and as an arbitrary cut-off for further investigation, we propose 100 reads per million (rpm). However, this cut-off depends on numerous factors (sample type, pathogen type, RNA extraction efficiency, library preparation, etc.). This strengthens our belief that although this tool can be useful to do a full pathogen screen, interpretation of the results is extremely dependent on the type of samples and the nature of the pathogen. Therefore plant pathologists should always interpret the results on a sample-per-sample basis and confirm the presence of putative pathogens by independent confirmation assays. Nevertheless, by organizing the RNA-seq community effort, awareness among virologists was raised that they can also use the data to detect potential other pathogens. The participants completed a questionnaire that asked their opinion about the usefulness of these kind of analyses. The questionnaire revealed that similar methods were almost exclusively being used by people with an expert bioinformatics level, and that regardless of the bioinformatics level, all participants will probably use these metagenomics methods in the future (Figure 4). Recently, the knowledge built from this project was used in an international consortium dealing with a relatively new virus, i.e. tomato fruit blotch virus (ToFBV) for which the transmission route is not known yet. By analyzing the RNA-seq data of different ToFBV positive datasets from different countries, we were able to identify a common organism present in 8/9 datasets, *Aculops lycopersici*, (an Eriophyid mite), making this a good vector candidate.

Finally, in WP5 we tried to pave the way towards broader communication towards scientists and stakeholders, by communicating as much as possible on the results, and sharing everything through publications and (open access) online repositories.

This project has shown that high-throughput sequencing is becoming an undismissible tool in plant pathology / diagnostics, with very interesting applications (cfr. complete disease screening through RNA-sequencing, vector identification for new viruses, etc.). However, the interpretation of the results and training of plant pathologists to keep up with the technology and data analysis remains far from straightforward. This project was useful in capacity building, networking and education, which will be a continuous effort in the future as new sequencing technologies and data processing methods arise. Our project results are in a too early stage to give concrete advise to policy makers.

2.6. Benefits from trans-national cooperation

This project was started because many laboratories were struggling to apply HTS because of the lack of knowledge and experience. The main goal of the project was to share knowledge between partners, also paying attention to laboratories with little experience. Without the transnational cooperation, this project would not have existed nor succeeded.

The different partners in the project had a various background in working with HTS data (beginner to expert), but all had the opportunity to participate in the different work packages. Work package 2 (training material and beginner's guide) was oriented specifically towards the less experienced laboratories, to develop and share materials that could be useful as a starting point in the data analysis. The developed beginner's guide gives a concise and complete overview of the different steps in a HTS workflow for virus detection. The guide contains both basic information as well as rather detailed points that can help in the analysis and interpretation of the results. Although the guide was initially intended towards users with little experience, many experienced users indicated that they actually learned a lot from the interaction with each other during the writing of the beginner's guide. The activities in work package 3 were more oriented towards specialists, since there we developed (semi-)artificial datasets that can be useful to test the behaviour of different pipelines, hence the group of people who worked on this work package was smaller than for the other work packages. Also here, the transnational cooperation was necessary to reach the WP3 goals because the partners decided together which challenges should be addressed in the datasets, and the partners also supplied datasets that were used as a starting point. Finally, in WP4, we encouraged as many laboratories as possible to participate in our RNA-seq community effort, where we looked for the presence of non-viral pathogens in existing RNA-seq datasets (intended to detect viruses). The community effort not only showed that it is indeed possible to detect other pathogens, it also raised the awareness among researchers that they should not only compare their data to virus databases. The fact that researchers could participate with their own datasets made them feel very involved with the results. The questionnaire also showed that some participants considered themselves as beginner, while others thought of their HTS and bioinformatics experience as intermediate. The same questionnaire also revealed that almost no (non-expert) participants used metagenomics techniques before to analyze their RNA-seq data, but after seeing the results everyone considers using these techniques in the future (Figure 4). We believe that these techniques will become more important, and by means of this project, more researchers were made aware of their existence and usefulness.

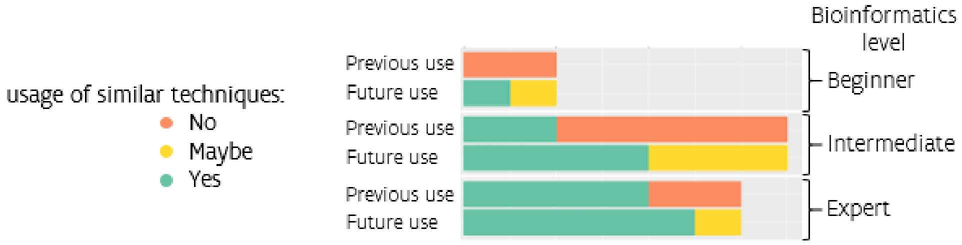

Figure 4 (Partial) result of the questionnaire, where participants were asked whether or not they used similar metagenomics techniques in the past, and if they consider to use them in the future (split by bioinformatics level of the participant).

As this project was called “Plant Health Bioinformatics Network”, one of its main goals was to create a network of bioinformaticians working in plant health. Unfortunately, due to the COVID-19 pandemic, we were unable to organize any physical networking activities (except for the project kick-off meeting in October 2019 in Rome). Despite this restriction, many laboratories and researchers stayed involved in one or more of the work packages. The network kept on growing during the project, and by the end of the project, several bioinformaticians and plant pathologists regularly exchanged e-mails with practical questions regarding analysis of HTS data. We hope that this network and the openness of the scientists to share experiences will continue to exist and grow in the future.

2.7. References

- Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM (2015). Making the Leap from Research Laboratory to Clinic : Challenges and Opportunities for Next-Generation Sequencing in Infectious Disease Diagnostics. MBio 6, e01888-15. https://doi.org/10.1128/mBio.01888-15
- Hardwick S, Deveson IW, Mercer TR (2017). Reference standards for next-generation sequencing Reference standards for next-generation sequencing. Nat. Rev. Genet. 18, 473–484. https://doi.org/10.1038/nrg.2017.44
- Massart S, 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
- Vincent AT, Charette SJ (2015). Who qualifies to be a bioinformatician? Front. Genet. 6, 1–3. https://doi.org/10.3389/fgene.2015.00164

3. Publications

3.1. Article(s) for publication in the EPPO Bulletin

None.

3.2. Article for publication in the EPPO Reporting Service

None.

3.3. Article(s) for publication in other journals

Kutnjak D, Tamisier L, Adams I, Boonham N, Candresse T, Chiumenti M, De Jonghe K, Kreuze JF, Lefebvre M, Silva G, Malapi-Wight M, Margaria P, Mavrič Pleško I, McGreig S, Miozzi L, Remenant B, Reynard JS, 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, 841.
https://doi.org/10.3390/microorganisms9040841

Tamisier L, Haegeman A, Foucart Y, Fouillien N, Al Rwahnih M, Buzkan N, Candresse T, Chiumenti M, De Jonghe K, Lefebvre M, Margaria P, Reynard JS, Stevens K, Kutnjak D & Massart S (2021). Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection.
Original publication: Zenodo, 4584718, ver. 4. https://doi.org/10.5281/zenodo.4584718

Reviewed and recommended by Peer Community In Genomics:
https://doi.org/10.24072/pci.genomics.100007

Publication in Peer Community Journal,1, article no. e53: https://doi.org/10.24072/pcjournal.62

Haegeman A, Foucart Y, Goedefroit T, De Jonghe K, Al Rwahnih M, Boonham N, Candresse T, Gaafar Y, Hurtado-Gonzales O, Kogej Z, Kutnjak D, Lamovšek J, Lefebvre M, Malapi-Wight M, Mavrič Pleško I, Onder S, Reynard J.-S, Salavert Pamblanco F, Schumpp O, Stevens K, Pal C, Tamisier L, Ulubaş Serçe Ç, van Duivenbode I, Waite D, Xiaojun H, Ziebell H and Massart S. Revisiting high throughput sequencing data used for plant virus detection in order to find evidence of non-viral plant pathogens and pests (tentative title). In preparation.

4. Open Euphresco data

For the project, three GitLab repositories were created of which some were transferred to long-term storage data platforms.

The first repository contains all training materials developed (https://gitlab.com/ilvo/phbn-wp2-training) accompanying the publication mentioned above (Kutnjak et al., 2021). The repository was also transferred to the Zenodo data repository platform for long-term storage under a CC BY 4.0 license (https://doi.org/10.5281/zenodo.6390814).

A second repository was set up with more information about the (semi-)artificial datasets, including links to the datasets itself and information on how to participate to the VIROMOCK challenge (https://gitlab.com/ilvo/VIROMOCKchallenge). This repository contains detailed information on how the datasets were made, and also includes forms for the participants to upload their own results. Since the datasets are useful as a resource for future pipeline validation, they were also transferred to the long-term supported data platform Dryad (https://doi.org/10.5061/dryad.0zpc866z8) under a CC0 1.0 license.

A third repository contains all materials regarding the RNA-seq community effort: the rRNA database, instructions on how to participate and reports from all participants. (https://gitlab.com/ilvo/PHBN-WP4-RNAseq_Community_Screening). This repository is meant as supporting information for the publication that is being prepared.