Hort Innovation

Final Report

Improving plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing (NGS)

Project leader:

Associate Professor Roberto Barrero

Delivery partner:

Queensland University of Technology

Project code:

MT18005

Project:

Improving plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing (NGS) (MT18005)

Disclaimer:

Horticulture Innovation Australia Limited (Hort Innovation) makes no representations and expressly disclaims all warranties (to the extent permitted by law) about the accuracy, completeness, or currency of information in this Final Report.

Users of this Final Report should take independent action to confirm any information in this Final Report before relying on that information in any way.

Reliance on any information provided by Hort Innovation is entirely at your own risk. Hort Innovation is not responsible for, and will not be liable for, any loss, damage, claim, expense, cost (including legal costs) or other liability arising in any way (including from Hort Innovation or any other person's negligence or otherwise) from your use or non-use of the Final Report or from reliance on information contained in the Final Report or that Hort Innovation provides to you by any other means.

Funding statement:

This project has been funded by Hort Innovation, using the citrus, nursery, potato, rubus and table grape research and development levy and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture.

Publishing details:

Published and distributed by: Hort Innovation

Level 7
141 Walker Street
North Sydney NSW 2060

Telephone: (02) 8295 2300

www.horticulture.com.au

© Copyright 2022 Horticulture Innovation Australia

Contents

Public summary 6
Keywords 6
Introduction 7
Methodology 9
Results and discussion 13
Outputs 32
Outcomes 37
Monitoring and evaluation 40
Recommendations 44
Refereed scientific publications 45
Journal article 45
References 45
Intellectual property 48
Acknowledgements 49
Appendices 49
Appendix 1 - Australian and global initiatives on NGS for viruses and bacterial plant pest diagnostics 50
Appendix 2 – Collaborations with national and international groups conducting NGS diagnostics on viruses and bacteria 55
Appendix 3 - Key plant pathogenic bacteria identified and prioritised 56
Appendix 4 – Current best practices and standards for the diagnosis of viruses and viroids using NGS. 58
Appendix 5 – Comparison of PEQ protocols against small RNA-seq and ribosomal depleted RNA-seq approaches to detect viruses and viroids in positive control plants 85
Appendix 6 – National Plant Biosecurity NGS/NGS Workshop 99
Appendix 7 – Draft NGS protocols for key culturable plant pathogenic bacteria 107
Appendix 8 – Identify and establish NGS protocols for detection of key culturable plant pathogenic bacteria 129
Appendix 9 – Lists of genomes used in comparative analysis. 139
Appendix 10: Phase 1 – Large scale side-by-side comparison of PEQ vs small RNA-seq NGS methods 143
Appendix 11: Cultured bacteria high throughput sequencing pipeline validation 163
Appendix 12: High throughput sequencing protocols for key fastidious and unculturable plant pathogenic bacteria 166
Appendix 13: High throughput sequencing metagenomics sensitivity 173
Appendix 14: Hybridization Probe Design for Targeted Sequencing 178
Appendix 15 – Manuscript on side-by-side comparison of PEQ and NGS methods 199
Appendix 16: Custom Plant Viruses and Viroids Database (PVirDB) for diagnosis of quarantinable and non-quarantinable viruses 219
Appendix 17: Improvements and optimisation of the Virus Surveillance and Diagnosis pipeline 225
Appendix 18: NGS service provider 'Ring' test identifies reliable commercial kit for data generation. 237
Appendix 19: Large scale side-by-side PEQ and HTS testing of diverse plant commodities. 254

Appendix 21: Validation of the NGS pipeine for culturable bacteria 301
Discussion 323
Appendix 22: Testing and optimisation of hybridization probes for enrichment and detection of fastidious bacteria 325
Appendix 23: Evaluation of performance of different small RNASeq library kits 351
Appendix 24: HTS testing for viruses and viroids across diverse plant commodities undergoing quarantine testing at PEQ 364
Appendix 25: Draft policy documents for operational adoption of HTS technology for routine plant virus/viroid diagnosis 366
Appendix 26: Submitted NGS protocol for the diagnosis of all known viruses and viroids in horticulture crops to be submitted to the Subcommittee on Plant Health Diagnostics (SPHD) 415
Appendix 27: Hybridization probe evaluation and sensitivity II 436
Appendix 28 – Improvement of RNA extraction and assessing small RNA-seq controls. 473
Appendix 29: Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass) 478
Appendix 30: Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants 483
Appendix 31: HTS testing for viruses and viroids across diverse plant commodities undergoing quarantine testing at PEQ. 504
Appendix 32: End-user hands-on training workshops 510
Appendix 33: End-User VirReport manual 513
Appendix 34 - Policy paper for adoption of HTS for bacteria in PEQ 528
Appendix 35 - Standard Operating Procedure for culturable bacteria using Illumina sequencing (end-to-end) 539

Public summary

Rapid and safe access to new plant genetic stocks is crucial for plant primary industries to remain profitable, sustainable, and internationally competitive. Imported plant species may spend up to three years in Post Entry Quarantine (PEQ) facilities undergoing pathogen testing, which can impact the ability of plant industries to quickly adapt to new global market opportunities and access new varieties. Existing PEQ diagnostic platforms can be time-consuming, resource intensive, and biological tests may produce ambiguous results. A diagnostic platform that is scalable, robust, accurate, and rapid is needed for PEQ to expedite phytosanitary screening and allow imported plants to move through the quarantine process more quickly potentially.

High throughput sequencing (HTS), also referred to as Next Generation Sequencing (NGS), offers a fast, reliable, and cost-effective method to identify all known plant pathogens in a single test. In partnership with quarantine agencies in Australia and New Zealand, the overall aim of this study was to develop an end-to-end quality assurance program and standard operating procedures, which are harmonised with international best practices, to enable the adoption of NGS technology for phytosanitary screening of exotic pests in PEQ facilities. Two major NGS approaches were piloted to identify the most robust approach to conducting large-scale trials. An NGS method that relies on the host plant response against viruses provided the overall best performance and was used to conduct large-scale side-by-side comparisons of NGS and existing PEQ diagnostic tests. Over 500 plant samples were tested, providing a substantial amount of validation data necessary to enable the implementation of a policy regulatory framework. We developed VirReport and GA-VirReport toolkits to train staff at PEQ and facilitated transferring and uptake of the NGS capability by DAFF. In December 2022, the NGS technology was adopted for testing Prunus, Rubus, Fragaria, and ornamental grasses.

Bacterial pests can cause severe impacts on plant commodities. The project team advanced the use of NGS as a powerful diagnostic tool for rapid detection and accurate identification of culturable and fastidious bacteria in imported horticultural species. Two NGS approaches were assessed, one that collects all genetic information present in a pest sample (whole genome) and another that enriches the genetic information from regions of interest derived from known bacterial pests (targeted capture). The team found that bacterial pure culture provides high-quality genome assemblies containing taxonomically informative information that is not confounded by information from other organisms, while the targeted capture method allowed the detection of unculturable/fastidious bacteria. A policy paper has been drafted to facilitate advancing the use of NGS technologies for routine testing of bacterial pests at PEQ.

Overall, the project, in close partnership with the Department of Agriculture, Fisheries, and Forestry, has enabled the adoption and operationalisation of NGS for routine quarantine testing at PEQ. We envisage that this new capability will enhance Australia's biosecurity system safeguarding domestic plant industries from exotic pests and enabling plant industries to gain accelerated access to new plant genetics to facilitate their access to high-value market opportunities.

Keywords

Horticulture crops; exotic pests; next generation sequencing; high throughput sequencing; plant viruses; viroids; culturable bacteria; fastidious bacteria; NGS diagnosis; post-entry quarantine; VirReport; GA-VirReport; bioinformatics; end-to-end pipelines; genomics; metagenomics; small RNA-seq; ribosomal depleted RNA-seq; whole genome sequencing; hybridization capture; quarantinable pests; policy adoption.

Introduction

Increase in global trade represents new opportunities and challenges to the biosecurity system and industries as plant pests can be moved from one region to another (Bulman et al., 2018). The increase of imported plant materials into Australia, and associated pathogen testing, places pressure on limited resources to ensure the detection of pests of biosecurity concern at Post Entry Quarantine (PEQ) facilities are carried out in a timely manner. Exotic plants pests such as viruses, viroids and bacteria can cause substantial economic and ecological damages (Doddaraju et al., 2019; Loiseau et al., 2019; Malandraki et al., 2015). Current, PEQ diagnostic protocols can be time consuming, resource intensive and biological tests may produce ambiguous results. A diagnostic method that is fast, accurate, reproducible, and scalable capable of scrutinizing from a handful of samples to hundreds can expedite quarantine testing at PEQ. The implementation of such method will provide benefits to industry including faster access to new genetics, lower costs for quarantine testing, option to import a larger volume of plants, ability to quickly respond to emerging market opportunities and increase in profitability.

Preventing the incursion of exotic plant pests is of primary concern [1,2,3]. Traditional diagnostic methods such as serological or molecular assays (i.e., ELISA, PCR) require a priori knowledge of the pathogens' genetic blueprint such as their genome information. Other assays such as electron microscopy, woody indexing, or visual inspection do not rely on such knowledge [4,5,6]. However, they can result in ambiguous detections and generally provide limited diagnosis resolution, for example, morphological features can be used to assign an organism to a high taxonomic category (i.e., genus). Most of these traditional tests are also unable to detect multiple pests in a single assay [4,5,6].

NGS technologies provide an unbiased high-resolution detection capability of highly diverse plant pests without needing prior characterization [5,10,11,12]. NGS also enables the identification of viruses that can cause either non-specific or no symptoms and thus be problematic for traditional diagnostics approaches [13]. With increased affordability and portability of NGS, international groups and regulatory agencies are assessing its viability to rapidly screen for plant pest in quarantine facilities. However, there is a need for harmonization and standardization of NGS-based diagnostics to accelerate quarantine testing. There are several different approaches for nucleic acid extraction, NGS sequencing, and sequence data analysis. The use varying approaches has been reported that can lead to discrepancies between distinct NGS or data analytical methods [14,15,16]. Furthermore, challenges specific to NGS (i.e., adapter dimer formation, barcode misassignment, chimeric sequence assembly) need to be carefully assessed [17].

While some studies have evaluated the sensitivity of specific NGS approaches compared to traditional tests in plant commodities [18,19], NGS technologies still need to pass rigorous tests and repeatedly demonstrate they are equal to, or better than the current sensitivity, specificity, and reproducibility standards of existing methods before being approved for routine use [11,20]. For this reason, NGS is not yet currently considered to be a stand-alone replacement for traditional or molecular assays. Academic and governmental research groups both acknowledge the need for extant protocols to independently confirm the biological significance of detected viruses via NGS. Several international efforts are underway to critically catalogue methods available [17,21,22,23] and provide researchers and diagnostics laboratories with open-source reference datasets to compare and validate their detection pipelines [16,24]. There is also an increase in comparative studies reporting the performance of different approaches [14,15,16,21,25,26]. Together, this will help improve and accelerate harmonization, comparability, and reproducibility of NGS methods.

NGS methods use different nucleic acid fractions for detecting plant pests, with total RNA and small RNA (sRNA) being the two most widely used approaches [17]. Total RNA captures sequences from both the host and pests present in the sample, and strategies to reduce highly abundant hosts sequences such as ribosomal RNAs, can increase the sensitivity of detecting low titre pathogens.

The sRNA method relies on the host response against viruses using Dicer enzymes to break down viral molecules into small pieces, also known as viral derived small interfering RNAs (vsiRNAs) (Aliyari and Ding,

2009; Gouveia et al., 2016; Sabin et al., 2013). Plant species have many Dicer enzymes including Dicer 4 (DCL4), Dicer 2 (DCL2) and Dicer 3 (DCL3) involved in the innate antiviral response that break down viral molecules into sRNAs of 21 nucleotides (nt), 22 nt and 24 nt in length, respectively (Gouveia *et al.*, 2016; Sabin *et al.*, 2013). The size specific cleavage products of these enzymes make it possible to use specific genetic sequences to reconstruct the genomes of viral pathogens. It also enables to show evidence of the active response of the host plant against active viruses and viroids. Overall, it is possible to leverage on the plant immune response against viruses to diagnose one or more viral pests in a single assay (Barrero et al., 2017).

Traditionally, detection of bacterial pathogens has been based on culturing methods, with morphology and biochemical reactions the basis of identification (Gupta et al., 2019). Non-culturable forms of detection and identification have included serology including immunoassays; genotyping which includes amplified fragment length polymorphism (AFLP), PCR-16S rRNA- restriction fragment length polymorphism (RFLP); and amplification methods such as loop-mediated isothermal amplification (LAMP), recombinase polymerase amplification (RPA), PCR coupled with capillary electrophoresis (PCR-CE), and real time PCR (RT-PCR or qPCR) (Doddaraju *et al.*, 2019; Loiseau *et al.*, 2019; Malandraki *et al.*, 2015; Murray et al., 2017; Ordóñez et al., 2019). Some of these protocols can detect pathogen titres as low as 0.1 ng of bacterial genomic/total DNA per reaction (Peňázová et al., 2020). However, running individual tests for the ever-increasing list of both exotic and endemic bacterial pathogens is time consuming, laborious and costly, especially when many plant samples require screening (Anderson et al., 2004; Stukenbrock and McDonald, 2008).

The use of high throughput sequencing (HTS) could provide a rapid and robust approach for detection of multiple bacterial pathogens in a single plant sample in post entry quarantine (PEQ) facilities. The multiplexing capabilities permits sequencing of more than a thousand samples in one run especially when using higher sequencing output platforms such as the Illumina NovaSeq (up to 6 TB per run). Instead of individual tests for each pathogen, with the right technique, HTS would only require one sequencing run to detect all of them. Furthermore, the best outcome from HTS sequencing is its data-rich output providing not only detection, but high taxonomic resolution and origin.

A key component of HTS is the choice of technique used in preparation for sequencing. Some approaches may provide greater sensitivity compared to other bacterial detection methods, via bacterial cell enrichment (Frampton et al., 2018) or through target sequencing (Eichmeier et al., 2019). In this project, the project team assessed, implemented and compared multiple HTS techniques that have been used or could be used for plant bacteria diagnostics in PEQ.

The Hort Innovation funded project “MT18005 Improving plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing (NGS)” conducted large-scale side-by-side comparisons of NGS and existing PEQ diagnostic platforms to provide the substantial amount of validation data necessary to enable the development and policy adoption of the NGS technology for the quarantine testing of plant viruses and viroids in Prunus, Rubus, and Fragaria a the initial set of commodities, expanding the use of the technology that is being used for ornamental grasses. In addition, the project outcomes provide key potential benefits for the industry including rapid access to imported genetics, cost saving for PEQ testing and significantly enable industry to remain competitive domestically and globally.

Methodology

Virology NGS

Pilot - Comparison of two NGS approaches and PEQ protocols for the detection viral pests.

A set of 17 positive control plants (Appendix 5, Table 5.1) encompassing Citrus, Prunus, Rubus, Sweet Potato, Grapevine, Strawberry, and an ornamental grass were selected to conduct a side-by-side comparison of PEQ methods versus small RNA-seq (sRNA-seq) and ribosomal depleted RNA-seq NGS approaches. Total RNA samples were collected at PEQ as described in (Gauthier et al., 2022; Lelwala et al., 2022). Collected total RNA samples were then sent to two HTS service providers, one national (AGRF, Melbourne) and another overseas (BGI, Hong Kong) for NGS data generation using Illumina NextSeq and BGISEQ-500 instruments. Small RNA-seq NGS data was processed using a similar approach as previously described (Barrero *et al.*, 2017), and ribosomal RNA-seq data was processed as described in (Gauthier *et al.*, 2022). Detection of true positives and assessment of index hopping events, cross-sample contamination was evaluated (Gauthier *et al.*, 2022) to define the NGS technology to use for large scale side-by-side PEQ vs. NGS trials.

Phase 1 - Large scale side-by-side comparison of NGS and PEQ protocols for the diagnosis of viruses and viroids.

A total of 271 plant samples were tested in the first quarantine testing season (Spring 2020 to Summer 2021) as part of the MT18005 project and a complementary Biosecurity Innovation Program funding to extend the scope of PEQ testing to a broader range of imported plant commodities. The tested plants included 83 stone fruit, 66 citrus, 25 raspberries, 20 apples, 19 strawberries, 17 blackberries, nine avocados, nine clonal grasses, five pears, and another 18 plants including potato, grapes, and other plant species. Total RNA fractions were collected with an automated Maxwell® Rapid Sample Concentrator (RSC; Promega, Madisson, WI, USA) instrument using Maxwell® RSC SimplyRNA Tissue kit (AS1340, Promega). NGS library prep, sequencing and data analysis was performed as described in (Gauthier *et al.*, 2022) (Appendix 10).

Implementation of VirReport a data analysis pipeline for phytosanitary testing of plant viruses and viroids.

To enable reproducibility, scalability, and portability to run data analysis of small RNA-seq NGS data on diverse compute infrastructure (i.e., on-prem HPC or cloud servers), we developed *VirReport* an end-to-end analytical pipeline implemented in Nextflow (Appendix 15) that incorporated lessons learnt during the project to optimise the diagnosis of plant viruses and viroids (Appendix 17). The pipeline can be accessed at https://github.com/eresearchqut/VirReport. A VirReport user manual is presented in Appendix 33.

Custom plant virus and viroid database (PvirDB)

All available nucleotide sequences (*n* = 17,153) of quarantinable viruses and viroids at PEQ and endemic viruses of interest to Australian plant industries were retrieved in GenBank format using an Entrez text search on the NCBI nucleotide database (Coordinators, 2013). The obtained list was checked manually, and any non-specific sequences removed. To prevent off-target detections, the custom database includes other reference viral sequences derived for higher plant species and abundant host plant sequences (i.e., chloroplast genomes). A nucleotide accession list of 120,945 viruses from higher plants, Embryophyta (taxid 3193), was obtained from the NCBI Virus (Hatcher et al., 2017) (October 2021), and their nucleotide sequences were downloaded in GenBank format using NCBI Batch Entrez. The GenBank-formatted sequences were converted into FASTA format with annotated taxonomic information (Gutierrez et al., 2021). In the quarantinable dataset, sequences with >5% N content were removed using the Seqtk version 1.3 (Shen et al., 2016), as well as sequences >20 Kb, except for complete chloroplast genomes that ranged from 100 to 200 kbp. To avoid ambiguous detection and reporting highly similar sequences (≥99.0%) using CD-HIT version 4.8.1 (Fu et al., 2012) to select a single representative sequence per cluster. A copy of the dataset without reference to quarantine status has made publicly available at Zenodo's data repository (https://doi.org/10.5281/zenodo.6609576) (Appendix 16).

NGS service provider ring test using positive control plants infected with known viruses and viroids.

The COVID19 global pandemic disrupted the supply chain for the provision of NGS services required by the project. To build resilience and flexibility to allow uninterrupted NGS testing, three service providers in

Australia were invited to generate small RNA-seq NGS data using two distinct commercial kits to assess and identify a suitable solution that generates high quality data and is less susceptible to COVID19 supply chain disruptions. A subset of 10 positive control plants used in the pilot trial were selected and sent to service providers for the comparison and evaluation of following commercial kits: 1) NEBNext Multiplex Small RNA Library (NEB); 2) NEXTFLEX Small RNA-Seq Kit v3 plus unique molecular identifiers (UMIs) (Perkin Elmer); 3) NEXTFLEX Small RNA-Seq Kit v3 plus unique dual indexes (UDIs) (Perkin Elmer); 4) SMARTer smRNA-Seq (Clontech, Takara Bio); 5) QIAseq miRNA library kit + UMIs + UDIs; and 6) QIAseq FastSelect kit (ribosomal RNA depletion) + QIAseq miRNA library kit + UMIs + UDIs (Appendix 18).

Phase 2 - Large scale side-by-side comparison of NGS and PEQ protocols for the diagnosis of viruses and viroids.

A total of 206 plant samples were tested during the second quarantine testing season (Spring 2021 to Summer 2022) as part of the MT18005 project and a complementary Biosecurity Innovation Program funding to extend the scope of PEQ testing to a broader range of imported plant commodities. The tested plants included 89 stone fruit, 24 raspberries, 18 grapes, 14 strawberries, 10 citrus, 10 potatoes, eight evergreen shrubs, seven blueberries, seven pears, six apples, four ornamental grasses and one sweet potato. Total RNA fractions were collected with an automated Maxwell® Rapid Sample Concentrator (RSC; Promega, Madisson, WI, USA) instrument using Maxwell® RSC SimplyRNA Tissue kit (AS1340, Promega). NGS library prep, sequencing and data analysis was performed as described in (Gauthier *et al.*, 2022) (Appendix 31).

Bacteriology NGS

DNA extraction

Fig (*Ficus carica*) DNA was extracted from leaf tissues using DNeasy® Plant Mini Kit. Bacterial DNA of *R. solanacearum* and *P. ananatis* was provided by AVR. Concentration of fig DNA was measured using a Quantus™ fluorometer (Promega, Madison, United States) and spiked with 1% *R. solanacearum* and 0.1% *P. ananatis* of total DNA.

Eight *Pantoea* spp (P1-P8) were isolated and grown on Kings B agar media for 2 days at 28 °C A representative Psa3 isolate from Victoria was grown on nutrient sucrose agar (EPPO, 2014) for 2 days at 28 °C. *Pantoea* and Psa3 colonies were scraped from the agar and extracted using Qiagen DNeasy® Blood & Tissue kit (Qiagen, Hilden, Germany).

For MinION sequencing, Psa3 was grown in nutrient broth with no sugar source overnight rotating at 120 RPM at 28 °C. Overnight cultures were centrifuged until a pellet formed. Bacterial pellets were dried in the biosafety cabinet for one hour before proceeding to use Promega Wizard® Genomic DNA Purification Kit (Promega, Madison, United States) for gDNA extraction.

Genomic DNA of 37 cultured, pre-identified bacterial isolates were supplied by PEQ and the SSG community for pipeline validation (Appendix 11).

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, concentration of DNA was measured and 10 – 50 ng DNA for each library preparation. Library preparation kits used on spiked fig DNA included Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with Perkin Elmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes); NEBNext® Ultra™ II FS DNA Library Preparation Kit with NEBNext® Multiplex Oligos for Illumina® (96 UDI Primer Pairs); and NEBNext® Ultra™ II FS DNA Library Preparation Kit with Perkin Elmer NEXTFLEX® UDI Barcodes (384 barcodes). Libraries were sequenced on an Illumina Miseq platform.

Libraries for the *Pantoea* isolates were generated using the NEBNext® Ultra™ II FS DNA Library Preparation Kit (New England Biolabs, Massachusetts, United States). Isolates were then paired end sequenced (2 × 300 bp) using the Illumina MiSeq with a negative control.

Libraries for the Psa3 isolate were generated using the Nextera DNA Flex Library Preparation kit (Illumina Inc., California, United States) and was then paired end sequenced (2 × 150 bp) using the Illumina HiSeq.

ONT library preparation and sequencing

For MinION, approximately 2 µg of DNA was used for library preparation using the ONT ligation kit (LSK109). Psa3 was then sequenced using the flowcell type MIN106 on the MinION for 40 hours.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen et al., 2018) and assembled using SPAdes v3.6.0 (Bankevich et al., 2012). For the spiked fig DNA, Kraken2 analysis was run on trimmed reads and a nucleotide BLAST search (Altschul et al., 1990) was conducted on the assembled reads to check variability of detection limits for each library preparation kit and index combination. The k-mer analysis tool using Kraken2 (Wood et al., 2019), was also used on *Pantoea* trimmed reads.

MLSA and UBCG (Na et al., 2018) analyses were run on the *Pantoea* and Psa3 assemblies. MLSA functions to differentiate species by generating a phylogenetic tree from several concatenated essential protein-encoding gene sequences (housekeeping genes). UBCG does something similar, however, instead of choosing housekeeping genes from literature, it uses 92 core genes for phylogenetic analysis. The MLSA protocol is as described in Basic protocol 9 (Appendix 8) using *infB*, *rpoB*, *atpD* and *gyrB* genes (Brady et al., 2008) for *Pantoea* and *gapA*, *gltA*, *gyrB* and *rpoD* genes (Butler et al., 2013; Flores et al., 2018) for Psa3. MLSA genes used are full genes. The program Insilicoseq was used to generate sequencing reads *in silico* for *Pantoea stewartii* subsp. *stewartii* DC283 and *Pantoea stewartii* subsp. *indologenes* LMG2632. Generated reads were for use for Kraken2 analysis.

Illumina reads were assembled using SPAdes v3.6.0. Assembled contigs were compared to ONT. Raw signal MinION files were basecalled using Guppy v.2.1.3 and subjected to demultiplexing and adapter/barcode removal using Porechop. Raw Illumina reads were subjected to adapter trimming and quality filtering using fastp. Canu v1.9 was used for assembly of for MinION reads and Unicycler v0.4.8 was used for assembly of MinION reads and a hybrid assembly (both MinION and Illumina reads). MinION reads were further polished with two rounds of Racon (Vaser et al., 2017) and Medaka (https://nanoporetech.github.io/medaka/index.html). Pyani v0.2.10 was used to determine the average nucleotide identity (ANI) of polished and unpolished MinION assemblies compared to the hybrid assembly and Prokka v1.14.5 was used to annotate the assembled contigs to identify genes. Both ANI and predicted gene number were used to assess accuracy of polished reads in comparison to unpolished.

Assembled contigs were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990) as described previously (Ashikawa et al., 2018; Hu et al., 2019). Furthermore, all assembled reads, polished and unpolished, were subjected to MLSA and UBCG analysis, as described above.

DNA for test sensitivity comparison

In order to compare the sensitivity of HTS diagnostic techniques to currently used molecular tests like qPRC and LAMP, Olive DNA with various concentrations of spiked *Xylella fastidiosa* (Xf) ST 2 DNA were prepared. Nucleic acids were extracted from olive leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and were shown to be Xf negative using species-specific PCR tests as described in Appendix 13.

Xyllela fastidiosa library preparation, sequencing and analysis

Using the Quantus™ fluorometer, concentration of DNA was measured and ~50 ng DNA was used for each library preparation. Libraries for the spiked Xf serial dilutions were generated in duplicate using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A plant only library and a no template control were also generated. Libraries were then paired end sequenced (2 × 150 bp) using the Illumina NovaSeq at 2 GB, 5 GB and 10 GB sequencing depth.

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen *et al.*, 2018). Trimmed reads were analysed using Kraken 2 (Wood *et al.*, 2019) and assembled using SPAdes v3.6.0 (Bankevich *et al.*, 2012). A nucleotide BLAST search (BLASTn) (Altschul *et al.*, 1990) was conducted on the assembled contigs. Bracken analysis was conducted on Kraken2 results to obtain fraction of reads assigned as Xf. Genome coverage was calculated by determining the total length of assembled contigs and dividing it by the

average Xf genome size. A k-mer based mapping tool, StringMLST v. 0.6.3 (Gupta et al., 2017), was used to determine if the sequence type (ST) could be assigned for each Xf dilution at the varying sequence depths.

Hybridization Capture Targets

Target identification

Genes used as targets for hybridisation probe design were identified two ways: by pan-genome analysis where multiple genomes were available, to identify core genes amongst strains of the same species that are not present in closely related species, and by review of literature and sequence databases to identify genes for species where limited sequence data was available (e.g., phytoplasma species) (Appendix 14).

Unique genes from the pan-genome analysis of each bacterial species that were larger than 1kbp and sequences of genes chosen from literature were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul *et al.*, 1990) to identify any potential cross reaction with non-target species.

Custom probe design

Target regions identified for probe design were supplied to Twist Bioscience for custom capture library tiling design and synthesis. Probes were designed removing regions of low complex with 4 X tiling to increase capture rate of low titre pathogens. Each probe consists of a biotinylated 120 mer dsDNA so once bound to targeted regions of sample DNA, it can be captured using Streptavidin bound to magnetic beads.

Validation of hybridization probes

Nucleic acids were extracted from olive, lemon and apple leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Olive, lemon and apple trees are known to be plant hosts of Xf, Xcc and Ea respectively, hence, used as background DNA for positive samples used in validation experiments. Positive control samples for hybridization are described in Appendix 11. Additionally, quantitative PCR and PCR assays were used to validate hybridization probes as described in Appendix 13.

Hybridization probe sensitivity test and metagenomic comparison

DNA samples extracted from host samples for the *Ca.* Liberibacter and phytoplasma species were sourced from other projects. CLaf was provided by Biosecurity Queensland (Department of Agriculture and Fisheries) which was sourced from CRI facilities, Nelspruit, South Africa. Phytoplasmas were provided by AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria). CLso samples were provided by Plant and Food Research, New Zealand. *Ca.* Liberibacter brunswickensis samples were provided by an AVR PhD project. Bacterial spiked plant DNA (8 dilutions of Ea, Xf and Xcc spiked in plant DNA), 3 phytoplasma samples and five *Ca.* Liberibacter samples (CLaf, 2 CLso and 2 *Ca.* Liberibacter brunswickensis) were used for probe and no-probe comparisons in duplicates (Appendix 14).

Target enrichment: library preparation, hybridization, enrichment and sequencing

Approximately 50 ng of Xf, Ea and Xcc DNA serial diluted and spiked in plant DNA at different concentrations and ~3 ng of plant or insect samples naturally infected with phytoplasma and *Ca.* Liberibacter species was used for library preparation with the exception of the *Ca.* Liberibacter solanacearum infected tomato potato psyllid sample TPP6, where 5 µL containing ~ 0.8 ng of DNA was used as the sample volumes was limited. Libraries were generated with ~50 ng starting DNA using half reaction volumes and ~3 ng starting DNA using full reaction volumes of the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit and PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). Four replicates of each library were prepared along with a plant only library and a no template control library.

Using the Quantus™ fluorometer, DNA concentration of each library was quantified. Four libraries per sample were created. Two libraries were further used for probing while the remaining two were used as metagenomic controls. Around ~220 ng DNA of Xf, Ea and Xcc serial dilution duplicates were pooled eight samples per hybridization capture, which was equivalent to 6 hybridization capture reactions. The concentration of *Ca.*

Liberibacter and phytoplasma libraries were low (between 1-3.4 ng/µL) with one of the duplicates for TPP6 being 0.27 ng/µL, so full volume libraries were pooled in eight samples per hybridization capture, which was equivalent to two hybridization capture reactions. Pooled libraries were subjected to hybridization capture using the Twist Target Enrichment Protocol (Twist Bioscience, USA) according to manufacturer's instructions.

Libraries for the non-captured TPP6 duplicates gave 0.164 ng/µL and 0.14 ng/µL, lower than all other libraries. The 2200 TapeStation system (Agilent technologies) was used to assess library quality and size distribution of libraries. Libraries for the non-captured TPP6 duplicates were not detected on the TapeStation, however, they were still added to the sequencing run using 15 µL. Hybridization captured and non-captured libraries were paired end sequenced (2 x 250 bp) using the Illumina NovaSeq on a SP flow cell.

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, the concentration of each DNA sample was measured and ~8 ng DNA was used for each library preparation. For samples with low concentrations such as PEQ-12, PEQ-15 and PEQ-30 (0.226-0.456 ng/µL) and no detection of DNA for PEQ-16, 8 µL of sample was used. Libraries were generated using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A no template control library was also generated. Libraries were then paired end sequenced (2 × 250 base pairs (bp)) using Illumina NovaSeq.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen *et al.*, 2018) and subsequently to the k-mer-based analysis tool, Kraken2 (Wood *et al.*, 2019) for OTU classification using a custom database containing all completed bacterial reference genomes in NCBI (19/03/2021). Reads were *de novo* assembled using SPAdes v3.15.2 (Bankevich *et al.*, 2012). Assembled contigs were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul *et al.*, 1990). Depending on bacterial genera and the quality of data, 35 PEQ samples were chosen for up-to-date bacterial core gene (UBCG) analysis (Na *et al.*, 2018).

Kraken2 and BLASTn results detected mixed bacterial genera in some samples. Sequences of the identified bacterial genus in mixed samples were isolated from the *de novo* assembled contigs using an in-house Perl script and BLASTn results. Isolated sequences were used for UBCG phylogenetic analysis.
N50 is the length (bp) of the shortest contig in the top 50% of total contig length. This number is used to describe the quality of an assembly, where the higher the number, the better the assembly. N50 was calculated using an open source n50 program on miniconda (https://docs.conda.io/en/latest/index.html).

Results and discussion

< Provide results of the work undertaken including applicable tables, graphs and figures generated through the project. All data sets and/or databases should be supplied as appendices to the final report (and marked confidential as appropriate). The discussion should outline what the results mean for use and uptake by horticultural businesses or other industry stakeholders or for future/related research projects. Detailed scientific information (e.g. tables and figures) should be provided in appendices. Consider the target audience of the final report. For more information, refer to refer to *Attachment A3: Final report guide* >

Comparison of two NGS approaches and PEQ protocols for the diagnosis of viruses and viroids.

We used a set of positive control PEQ reference plants including citrus, prunus, sweet potato, raspberry, strawberry, and an ornamental grass infected with known endemic and quarantinable (i.e., of biosecurity concern) viruses and/or viroids (Table 1) We tested these plants using two NGS approaches, ribodepleted RNA-Seq and sRNA-Seq. NGS data for both methods were generated independently by two service providers (SP1 and SP2) using distinct library prep and sequencing approaches.

For ribodepleted RNA-Seq datasets, we obtained, on average, 37.9 and 46.7 million paired-end reads by SP1 and SP2, respectively. Quality filtering removed on average 5% and 6% of the reads produced by SP1 and SP2 samples, respectively. To assess the effectiveness of ribosomal depletion, quality trimmed reads were mapped

onto the Silva database showing varying levels of retained rRNAs across samples, ranging from 13 to 29% for SP1 and 7 to 33% for SP2. On average, rRNAs accounted for 20% and 16% of the paired-end reads produced by SP1 and SP2, respectively.

Table1: Selected PEQ positive control plants infected with known viruses and viroids. Presence of quarantinable viruses detected using molecular (PCR and ELISA) and bioassays (biological and woody indexing).

Sample ID	Commodity	Species	Positive Detections in PEQ
MT001	Citrus	Citrus Troyer × Frost-Lisbon	CEVd
MT002	Prunus	*Prunus persica*	PNRSV
MT003	Citrus	*Citrus* × *aurantiifolia* (Christm.) Swingle	CTV, CVEV, CDVd, HSVd
MT004	Citrus	*Citrus medica* L.	CEVd, CTV, HSVd
MT005	Raspberry	*Rubus idaeus*	RBDV
MT007	Citrus	*Citrus* sp.	CDVd, HSVd
MT008	Citrus	*Citrus sinensis*	CDVd, HSVd
MT010	Ornamental grass	*Miscanthus sinensis* 'Morning light'	Novel potyvirus (MsiMV)
MT011	Citrus	*Citrus medica* L.	CTV, CVd-VI, HSVd
MT012	Iris	*Iris* sp. 'Crimson colossus'	ISMV, TRSV
MT013	Strawberry	*Fragaria vesca* 'Alpine'	SMoV
MT014	Strawberry	*Fragaria vesca* 'UC4'	SMoV
MT015	Strawberry.	*Fragaria* sp.	SMoV
MT016	Sweet potato	*Ipomoea batatas*	SPFMV

CEVd= Citrus exocortis viroid, CTV = Citrus tristeza virus, CVEV = Citrus vein enation virus, CDVd= Citrus dwarfing viroid, CVd-VI = Citrus viroid VI, HSVd = Hop stunt viroid, ISMV = Iris severe mosaic virus, PNRSV = Prunus necrotic ringspot virus, RBDV = Raspberry bushy dwarf virus, SMoV = Strawberry mottle virus, SPFMV = Sweet potato feathery mottle virus, TRSV = Tobacco ringspot virus.

From sRNA-Seq datasets, we produced on average 55.4 and 27.5 million single-end reads by SP1 and SP2, respectively. Quality filtering resulted in the removal of 15% of reads on average for the SP2 samples and 2% for the SP1 samples, the latter value expected for samples that were already quality-filtered by the sequencing provider. The rRNA content was highly variable in the sRNA samples, accounting for 7% to 53% of the quality-filtered reads in the SP1, compared to 15% to 49% in SP2 samples. There was an association between the commodity and the percentage of rRNA detected in the samples, with the grass, sweet potato, and citrus samples being the least affected while the strawberry, peach and raspberry showed a higher level of rRNAs across both the SP1 and the SP2 samples (Figure 1) suggesting there can be commodity specific biases of genetic material collected and downstream analyses need to take into account source of host plant RNAs.

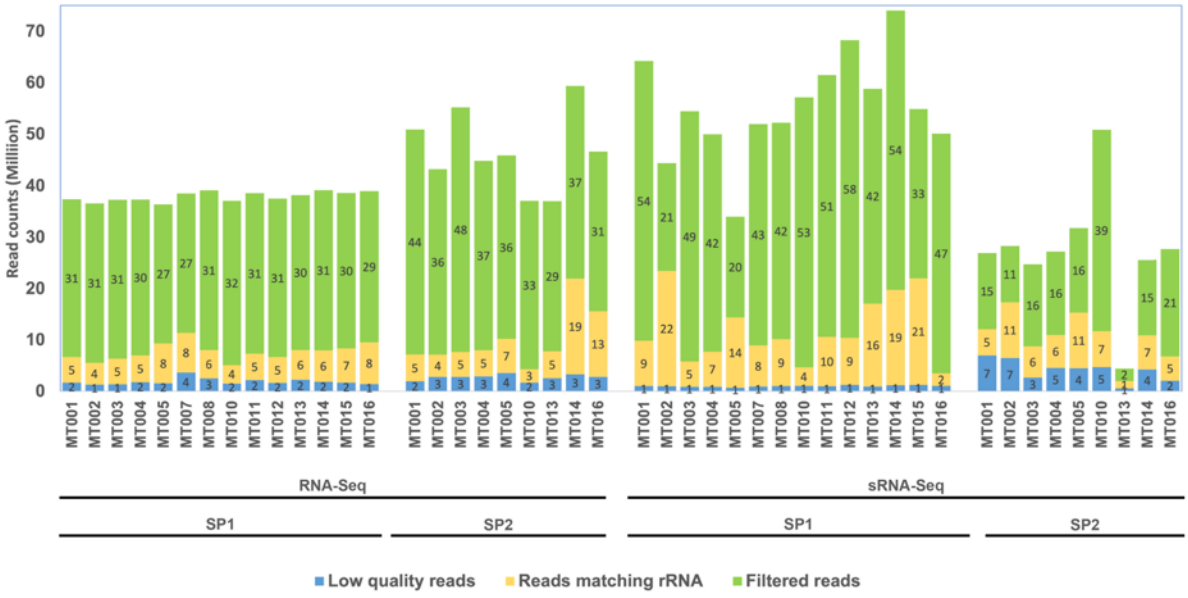


Figure 1: Distribution of RNA reads collected across samples. Bar chart showing number of filtered reads in millions that were left-over (in green) after removing low-quality reads (in blue) and reads matching ribosomal RNA (in yellow) for RNA-Seq (on the left) and sRNA-Seq methods (on the right). The counts were split between sequencing providers (SP1 and SP2).

Evaluating Sensitivity of Detection of Plant Virus and Viroids

We assessed read-based and assembly-based software approaches, the total number of unique viruses and viroids recovered per sample varied significantly. The number of viral candidates detected in the SP1 samples was, on average higher than in the SP2 samples for both sequencing technologies (Figure 2). To measure the sensitivity of each method and HTS technology combination, we checked their ability to detect a total of 38 viruses and viroids across both sequencing platforms. Among the tested viruses, we identified a novel potyvirus in sample MT010 showing high sequence homology to both Sorghum mosaic virus and Maize dwarf mosaic virus, and tentatively named Miscanthus sinensis mosaic virus (MsiMV) (Leblanc et al., 2022) (Table 1).

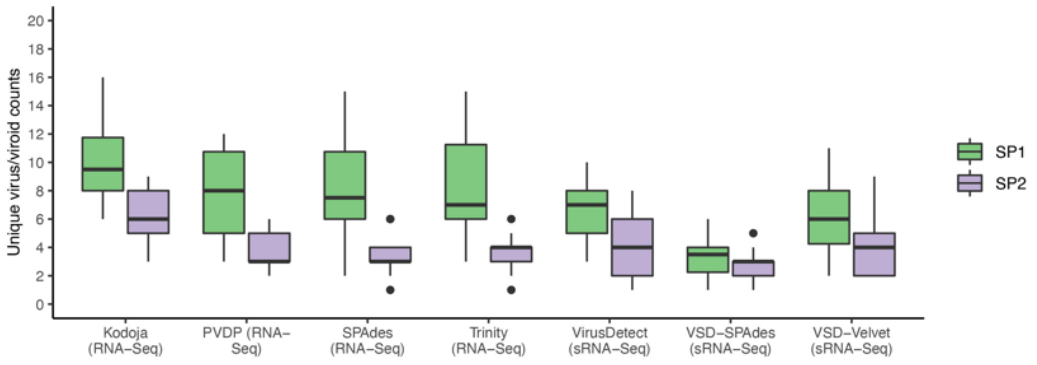


Figure 2: Side-by-side comparison of the total number of unique viruses and/or viroids detected by RNA-Seq and sRNA-Seq across samples by each software. The counts are split between sequencing providers (SP1 in green and SP2 in purple). The boxplot limits indicate the first and third quartiles, with the central line marking the median. Vertical lines extending from each box capture the remaining data that sits within 1.5 times of the interquartile range, while the dots placed past the line edges denote outliers.

For the RNA-Seq samples, only Kodoja could detect all the viruses and viroids in the SP1 samples, while PVDP, SPAdes, and Trinity missed the detection of a Hop stunt viroid (HSVd) present in sample MT003. Kodoja only recovered two pairs of reads matching HSVd that harbored slight nucleotide differences to each other, explaining why the other methods failed to find it. PVDP, Kodoja, and SPAdes were able to detect all the viruses and viroids identified by PEQ in the SP2 samples, while Trinity failed to detect the HSVd present in the sample

MT003. Thus, across the sequencing providers, Trinity achieved overall the lowest sensitivity (36/38, 94.7%), followed by SPAdes and PVDP (37/38, 97.3%), and finally Kodoja (38/38, 100%) (Table 2 and Appendix 15).

For the sRNA data, the contigs generated by VirusDetect and VirReport-Velvet were shorter than the ones generated by VirReport-SPAdes. However, the lower k-mer range (9–21) applied in VirReport-SPAdes meant that we observed many misassembled contigs in the derived assemblies. We could identify that these were caused by shared stretches of up to 20 nucleotides present in the host plant and virus sequences, yielding chimeric plant-virus contigs. SPAdes logs flagged a failure to determine erroneous k-mer threshold and estimate good valley values indicating that the lower bound of the k-mer range was likely yielding suboptimal results. The SPAdes software developers also stated that their tool had not been tested on sRNA reads and recommended not using k-mers below 15 to reduce the risk of misassembly.

Additional viruses detected through NGS include Citrus virus A in MT007 and MT008, Cherry rasp leaf virus in sample MT012, Strawberry vein banding virus in sample MT013, MT014, and MT015, Beet pseudoyellows virus in MT014, Sweet potato symptomless virus 1, and Sweet potato badnavirus in MT016. We also found rubus yellow net virus (RYNV) in samples MT005 and MT006 which are likely endogenous integrations into the host raspberry genome (Ho et al., 2021). These were either not quarantinable pathogens of concern, or not regulated in the given host in which they were found.

Table 2. Average detection sensitivity for the different methods and technology tested in this study.

Sequencing provider	Sequencing technology	Software	Sensitivity (%)
SP1	RNA-Seq	Kodoja	100 (24/24)
		PVDP	95.8 (23/24)
		SPAdes	95.8 (23/24)
		Trinity	95.8 (23/24)
	sRNA-Seq	VirusDetect	100 (24/24)
		VSD-*SPAdes*	100 (24/24)
		VSD-*Velvet*	100 (24/24)
SP2	RNA-Seq	Kodoja	100 (14/14)
		PVDP	100 (14/14)
		SPAdes	100 (14/14)
		Trinity	92.9 (13/14)
	sRNA-Seq	VirusDetect	100 (14/14)
		VSD-*SPAdes*	78.6 (11/14)
		VSD-*Velvet*	100 (14/14)

Detection of Cross-Sample Contamination Event

Some of the viruses and viroids identified in control plants (Table 1) were detected in other multiplexed samples in the same RNA-Seq and sRNA-Seq experiments. These cases were either found in a commodity that is not known to be a host to the virus/viroid or in a known host, but PEQ molecular assays did not confirm its presence. Our samples were processed with Vitis samples on the same sequencing run, but these were not part of our pilot study. Thus, any detection of pathogens specific to grapes across our samples were ignored. Figure 4 depicts identified cross-sample contamination events found for each HTS technology and service provider. For the RNA-Seq samples processed by SP1, each tool detected events across all 14 samples; Kodoja reported a total of 57, PVDP 49, SPAdes 38, and Trinity 38 events, respectively. For those processed by the SP2, Kodoja identified 20 false-positive events across all 9 samples, PVDP 6 events across 5 samples, Trinity 1 event, and SPAdes 1 event. For the sRNA-Seq samples processed by the SP1, VirusDetect, VirReport-Velvet, and VirReport-SPAdes identified 34, 23, and 4 false-positive events, respectively. In comparison to the SP2 samples, VirusDetect, VirReport-Velvet, and VirReport-SPAdes detected 5, 1, and 0 cross-sample contaminations, respectively (Figure 3).

Citrus tristeza virus, ISMV, MsiMV, SMoV, SPFVM, and TRSV accounted for most false positive detections in the

RNA-Seq samples processed by SP1, while PNRSV and RBDV were the most pervasive in SP2. In contrast, CEVd was the most prevalent false positive detection among sRNA-Seq data generated by SP1 (Figure 3A). Interestingly, the plants infected with this viroid showed the highest load of viroid-derived sRNAs (Appendix 15)

A contamination Flag (CF) was implemented for data derived by Kodoja, PVDP, VirusDetect, and VirReport. We found that depending on the bioinformatics tool and the service provider used, the minimum threshold percentage to detect false positives varied to either 0.1% (PVDP), 1% (Kodoja, VirReport-SPAdes, VirReport-Velvet, and VirusDetect for SP2 data), or 10% (VirusDetect for SP1 data). Application of false-positive cut-off thresholds to Kodoja results for the RNA-Seq data resulted in two out of the 57 index hoping events missed for SP1 datasets, while none were missed for the SP2 samples (Appendix 15). We also detected all cross-sample contamination events for data derived by PVDP for the SP1 samples, but one out of the six false positives was missed for the SP2 samples (Appendix 15). For the sRNA-Seq data, the CF identified all contamination events detected by VirReport-SPAdes, VirReport-Velvet, and VirusDetect (Appendix 15).

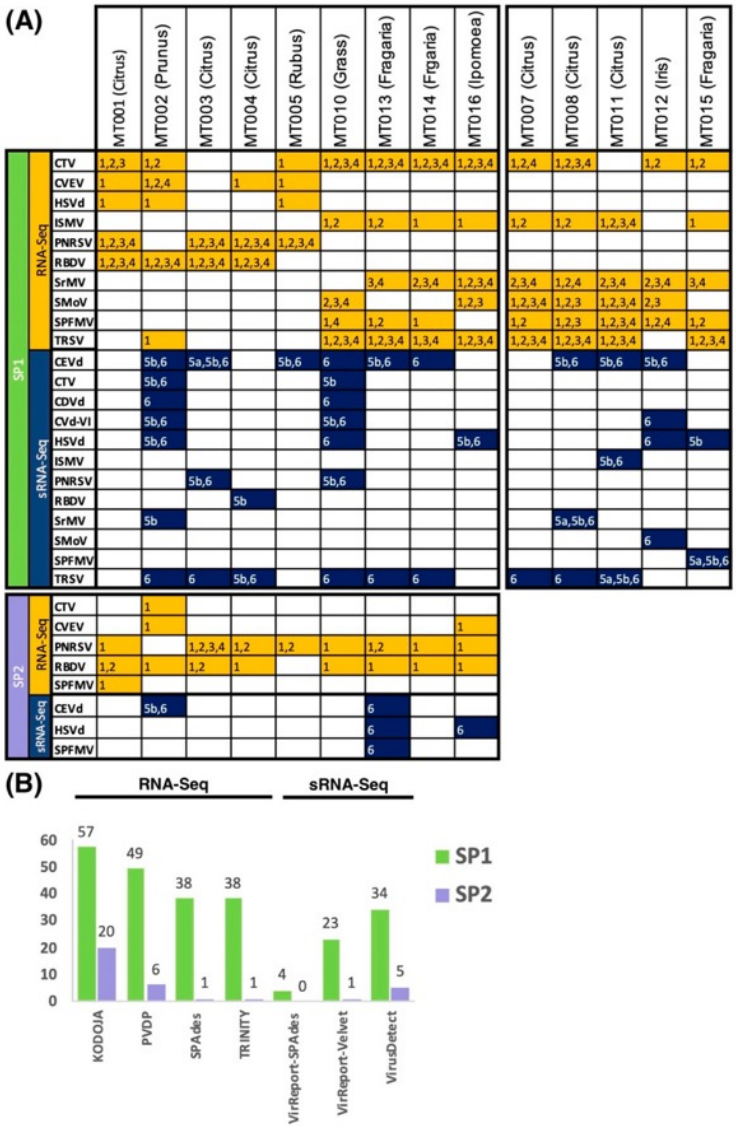


Figure 2: Detection of false positive viruses and viroids across samples. (**A**)**.** Details of all predicted viral contaminants detected in each sample sequenced by SP1 (top) and SP2 (bottom). Each number in the table corresponds to one of the software tested: numbers 1 to 4 refer to methods applied to RNA-Seq data; numbers 5 to 6 refer to methods applied to sRNA-Seq data. 1 = Kodoja, 2 = PVDP, 3 = SPAdes, 4 = Trinity, 5a = VirReport-SPAdes, 5b = VirReport-Velvet, 6 = VirusDetect. (**B**). Total number of false positive events detected per method and sequencing technology tested across 14 plant samples for the SP1 samples and 9 samples for the SP2 samples. The CVD-VI call detected by VirReport-Velvet in sample M010 was only detected in the de novo assembly derived using 24 nt-long reads.

Overall, side-by-side comparison of two NGS methodologies to detect viruses and viroids were tested as compared to PEQ positive control plants. We used two NGS service providers to assess the results of NGS-based diagnosis. We found 100% agreement between small RNA-Seq and PEQ results, while ribominus RNA-seq missed to detect a viroid in a citrus plant. NGS is susceptible to index-switching – cross-contamination between multiplexed samples. We observed two plans showing index switching effects in small RNA-seq, while all the plants tested using ribominus depletion RNA-seq did show varying levels of index switching between SP1 and SP2 approaches. The use of products from the plant antiviral response pathways by the small RNA-Seq method shows promise as it detects the same pest detected by PEQ, detects additional viruses if present in the sample and is less susceptible to cross sample contamination (Figure 3B). Thus, small RNA-seq was selected to perform a large-scale side-by-side PEQ vs NGS comparison in phase 1 and phase 2 trials.

Phase 1 - Large scale side-by-side comparison of NGS and PEQ protocols for the diagnosis of viruses and viroids.

In partnership with the DAFF via the Biosecurity Innovation Program (BIP), we extended the scope of NGS sequencing of the MT18005 project to encompass a broader range of quarantined plants. In addition to VirReport (Gauthier *et al.*, 2022), we implemented a Galaxy Australia VirReport pipeline (GA-VirReport) (Lelwala *et al.*, 2022) to enable the upskilling and training of molecular scientists at PEQ. We also developed a custom virus and viroid database (PVirDB) to enable the unambiguous screening of quarantinable pests using a harmonized, representative, portable and curated PVirDB resource (Lelwala *et al.*, 2022).

Overall, in complementation with MT18005 and this project near two hundred samples encompassing diverse plant species including Citrus, Prunus, Persea, Pyrus, Rosa, Fragaria, Solanum, Olea, Ficus and nursery plants were tested. Overall, side-by-side comparison of PEQ and sRNA-seq HTS methods shows strong correlation detecting 57 (97.3%) out 58 regulated or endemic viruses across 39 plants including 15 positive controls. Over 90% (162/179) of the imported plants did not yield positive hits to regulated viruses, while 3.9% were positive for endemic viruses of interest.

Key Project Achievements:

- Implementation of two automated end-to-end analytical pipelines, namely, VirReport and GA-VirReport for the detection of regulated and endemic viruses in a single assay.
- Development of custom Plant Virus and Viroid Database (PVirDB) that harmonises taxonomic information, collapses highly similar genetic sequences (identity >=99%) and assigs categorical tags (i.e., regulated, or endemic) to facilitate unambiguous diagnosis and HTS reporting.
- Overall, 196 plants including 17 control and 179 imported plants were tested using sRNA-seq NGS approach using the GA-VirReport pipeline against the custom PVirDB (version 20210517).
- PEQ and sRNA-seq NGS protocols showed 100% agreement detecting 23 regulated and 4 endemic viruses/viroids of interest across 15 control plants. Additionally, sRNA-seq HTS detected 8 additional non-regulated viruses in four control plants. Two events of 'index hoping' (cross-sample contamination) were observed in two control plants.
- Out of 179 imported plants, PEQ protocols identified 17 plants (9.5% detection rate) positive for the presence of 24 regulated viruses and viroids. The sRNA-seq NGS method found 23 regulated pests in the same plants. The Citrus bent leaf viroid (CBLVd) was not detected in sample MT177, however, CBLVd was found present in the parental plant MT176 highlighting the necessity to assess viral detections taking into consideration the parental plant's infectious status. Additionally, both PEQ and sRNA-seq NGS methods detected 7 endemic viruses in seven imported plants.
- Overall, across all samples tested sRNA-seq NGS detected 57 (98.27%) out of 58 regulated or endemic viruses and viroids found by PEQ protocols demonstrating a strong correlation with current quarantine testing assays.

Further optimisation of the VirReport and GA-VirReport pipelines is necessary to detect challenging low titre viral pathogens. Here we adopted a dual de novo assembly approach utilising two distinct bioinformatics tools to target mid-high or low titre viruses/viroids. This strategy shows promise to detect nearly all pest of interest in a single assay. Sampling timing and parental relationships among tested

samples are key to comprehensively inform and assess the presence of regulated viruses/viroids, particularly in both, parental and derived early shoot-tip grafted plants.

NGS service provider ring test using positive control plants infected with known viruses and viroids.

Three service providers in Australia were engaged to generate small RNA-seq NGS data using distinct commercial kits including: 1) NEBNext Multiplex Small RNA Library (NEB); 2) NEXTFLEX Small RNA-Seq Kit v3 plus unique molecular identifiers (UMIs) (Perkin Elmer); 3) NEXTFLEX Small RNA-Seq Kit v3 plus unique dual indexes (UDIs) (Perkin Elmer); 4) SMARTer smRNA-Seq (Clontech, Takara Bio); 5) QIAseq miRNA library kit + UMIs + UDIs; and 6) QIAseq FastSelect kit (ribosomal RNA depletion) + QIAseq miRNA library kit + UMIs + UDIs. Overall, the Qiagen miRNA library prep plus UMIs and UDIs outperformed other tested kits (Table 3 and Appendix 18). Interestingly, running a ribosomal depletion prior library prep with the Qiaseq miRNA only marginally improved the depletion of ribosomal RNAs suggesting that further development is necessary to prevent the sequencing of non-informative sequences using FastSelect (QIAGEN).

Table 3: Comparison of mean coverage and genome coverage at 1X depth for all viral targets across all kits

		NEBNext (AGRF)		Nextflex-UMI (AGRF)		Nextflex-UDI (CARF)		Qiaseq rRNA + (Ramaciotti)		Qiaseq rRNA - (Ramaciotti)	
Sample	Virus/Viroid	Mean cov	% bases 1X	Mean cov	% bases 1X	Mean cov	% bases 1X	Mean cov	% bases 1X	Mean cov	% bases 1X
MT000	endoRYNV	46.5	64.0	38.2	34.0	14.2	19.0	50.2	54.0	48.1	53.0
	RBDV-RNA1	97.0	93.0	75.6	74.0	45.9	67.0	109.4	94.0	99.4	94.0
	RBDV-RNA2	141.7	98.0	115.5	85.0	84.9	80.1	157.2	99.0	147.9	99.0
MT001	CEVd	238.8	100.0					239.5	96.0	236.7	100.0
MT002	PNRSV-RNA1	8.9	80.0			22.8	95.0	11	88.0	17.9	92.0
	PNRSV-RNA2	8.9	79.0			23.1	93.0	11.9	88.0	19.4	91.0
	PNRSV-RNA3	102.1	99.0			184.7	100.0	129	99.0	159.5	100.0
MT003	CTV	239.8	100.0	220.9	100.0	247.6	100.0	236.2	100.0	235.9	100.0
	CVEV	2.1	52.6	1.6	34.8	6.9	47.2	3.0	60.0	2.8	65.0
	CVd-III	224.0	97.0	167.7	81.0	183.2	81.0	232.3	98.0	233.2	97.0
	HSVd	171.9	99.0	182.5	99.0	242.3	100.0	205.3	99.0	204.7	100.0
MT005	endoRYNV	199.2	95.0			198.3	87.0	202.5	95.0	201	95.0
	RBDV-RNA1	195.9	96.0			205.5	93.0	195	97.0	194.3	97.0
	RBDV-RNA2	214.5	100.0			222.7	97.0	214.4	99.0	211.4	99.0
MT007	CVd-III	186.0	97.0			193.4	97.0	198.6	97.0	201.6	100.0
	HSVd	150.4	100.0			189.3	98.0	172.8	99.0	143.4	94.0
MT008	CVd-III	220.9	100.0			193.2	97.0	224.1	98.0	223.5	98.0
	HSVd	222.4	100.0			138.1	100.0	164.9	88.0	217.8	98.0
MT010	SgMV	177.5	79.0	46.7	31.0	42	28.0	40.6	30.0	41.4	29.0
MT012	ISMV	166.3	99.0			166.2	100	188.4	100.0	196.1	100.0
	TRSV-RNA1	187.7	96.0					187.5	94.0	199.7	93.0
	TRSV-RNA2	219.0	98.0					222.2	98.0	229.8	98.0
MT016	SPFMV	247.3	100.0					217.3	100.0	235.8	100.0

Phase 2 - Large scale side-by-side comparison of NGS and PEQ protocols for the diagnosis of viruses and viroids.

A total of 206 plant samples were tested during the second quarantine testing season (Spring 2021 to Summer 2022) as part of the MT18005 project and a complementary Biosecurity Innovation Program funding to extend the scope of PEQ testing to a broader range of imported plant commodities. Raw reads for each sample were

analysed using the VirReport pipeline. VirReport specifically reported quarantinable and endemic viruses and viroids of interest from screened plants with some exceptions: clonal grasses, border enforcement samples and special requests from clients (e.g., garlic samples) (Appendix 31). Major results are highlighted below.

- 16 imported plants tested positive for the presence of quarantinable pests (15 viruses and one viroid). Three imported plants tested positive for the presence of endemic virus of interest (6 viruses). These detections were all in agreement with molecular testing (ELISA or Sanger of PCR product) that was performed at SSG either on the same or a different extract from the same plant (**Table 31.1 and 31.2**).
- Contigs matching to Nectarine marafivirus M were detected in 5 prunus samples. This virus was not initially part of the prescribed testing at SSG and after review by PIO, its status was changed to quarantinable and independently validated by PCR and Sanger sequencing (**Table 31.1 and 31.2**).
- NGS failed to detect a cherry virus F (CVF) present at low titer in a prunus sample (MT258) sequenced using AGRF but this virus was detected by another bioinformatics tool, VirusDetect, which uses reference-guided assembly prior to de novo assembly to identify viruses. MT258 is a re-sample of plant MT094 and MT098 and CVF presence was molecularly validated in all extracts using Nepovirus B PCR followed by Sanger sequencing. This suggests that the read abundance was too low in this sample so that our de novo assembly step could successfully assemble them into a contig (**Table 31.1 and 31.2**).
- NGS failed to detect the endemic virus, cherry virus A (CVA) present at low titer in a prunus sample (MT399). This was molecularly validated at SSG using a nested generic Betaflexiridae PCR as well as CVA-specific PCR followed by sander sequencing. The fact that a nested PCR was required to get a PCR product with the generic Betaflexiridae method and that only a faint band was obtained with the CVA-specific PCR method confirm that virus particles were present at low concentration, which might explain why this virus was not detected by NGS (we only recovered one read matching this virus in the sRNA-Seq data) (**Table 31.1 and 31.2**).
- Short contigs matched at low depth to a candidate novel virus with closest match to grapevine red globe virus were also detected in three prunus persica var. nectarina and one prunus avium samples. This virus is not known to infect prunus plants. Sanger sequencing returned chimeric sequences matching both plants and virus. Subsequent sequencing using MinION nanopore sequencing, performed in collaboration with PIC@PEQ, was able to resolve around 5000 bp using nine overlapping amplicons (**Table 31.1 and 31.2**).
- Fourteen rubus plants tested positive for rubus yellow net virus (RYNV) (MT306-MT317, MT320 and MT321). In our previous report, we already described how raspberries are known to harbour up to six endogenous RYNV integrated sites in their genome and thus most RYNV detection in rubus plants are likely to be matches to these integrated sequences. By mapping our contigs to a local database incorporating endogenous RYNVs, which have been provided by Dr Thien Ho, we found that the contigs showed highest homology to these endogenous RYNVs. Thus, we interpret these detections as endogenous RYNVs integrated in the genome of the rubus plant sampled (**Table 31.2**). Therefore, these do not contribute to our episomal viral counts in **Table 31.1**.

Standard operating procedures and policy adoption of NGS for viruses and viroids.

In collaboration with policy regulators from DAFF and the Science and Surveillance Group (SSG) at PEQ. Best practices and standards for NGS applied to viral pest was reviewed (Appendix 4). We also co-organised a NGS workshop to define opportunities and challenges to adopt the technology for diagnosis (Appendix 6). Experts from Australia and New Zealand including managers of pathogen testing schemes, diagnosticians, policy regulators and industry leaders from plant, animal and clinical background convened to identify and discuss the requirements of a framework that may be applied to different commodities and industries to provide the basis of national standards in the detection of pathogens. The workshop attendees recommended:

- An NGS analysis pipeline must be validated as fit-for-purpose. Early in the adoption of NGS protocols, these should be performed in tandem with confirmatory current methods, and with the confidence of successful NGS practices where results accord with traditional assays, NGS may become the primary analysis method.
- Bioinformatics is growing so quickly that organisations should place more resources in getting resident bioinformaticians. It may be desirable to up-skill current diagnosticians and lab operators, however bioinformatics specialists are indispensable and greater focus is required in bringing more into the diagnostic agencies.

- Considering that we are developing a framework, it must be considered that there will have to be different tests for different targets and purposes, and not aim for prescriptive protocols that may be applied to specific commodities or industries.

A draft NGS policy framework and policy principles' report was prepared and proposed to Department of Agriculture, Fisheries and Forestry (DAFF) (Appendix 20) to advance and facilitate the adoption of NGS at PEQ. The draft report considers best practice policy guidelines for the development, validation, and adoption of High-throughput Sequencing (HTS) diagnostic technologies across human, animal and plant health sectors and proposes a hierarchical policy framework and principles for the modular adoption of HTS technologies for plant pests of biosecurity concern. This framework provides a platform for standard operating procedures (SOPs), quality management and assurance, and guidelines for interpreting HTS diagnostic results, which provides confidence in the diagnostic results produced. Standards and guidelines will be informed by the research program and aligned with relevant best practices as they currently exist across the health sector (Appendix 20).

Engagement and leadership by DAFF policy regulators, SSG scientists and the project team facilitated the scoping of the operational adoption of NGS for the diagnosis of viruses and viroids. Standard Operation procedures and updates to instruction manuals wad facilitated by this collaboration (Appendices 25 and 28). The project team also facilitated the transfer of the analytical capability to DAFF by partnering with DAFF's Technology Infrastructure Branch, Information Services Division to implement PEQ-HiTS portal that enables SSG staff at PEQ to automatically process small RNA-seq NGS samples produced by the preferred service provider using both VirReport and GA-VirReport pipelines. NGS has been formally adopted for the testing of Prunus, Rubus and Fragaria on 16 December 2022 as the first tier of plant commodities to benefit from the new technology. DAFF will continue to assess the NGS technology to broaden the scope to other plant commodities over the coming years.

Bacteriology NGS

Comparison of NGS and molecular assays sensitivity test for culturable bacteria

The LAMP and qPCR assays detected Xf at femtogram levels (Table 4 and Appendix 13), 2.4×10^{-05} and 2.4×10^{-06} ng/µL, respectively. Using Kraken2 software to analyse individual reads from the metagenomic sequencing results, Xf was detected up to 2.4×10^{-04} ng/µL of DNA for all sequencing depths assessed, 10 X and 100 X less sensitive than LAMP and qPCR respectively.

Using StringMLST on trimmed reads, the sequence type (ST) was able to be identified in all sequencing depths of the first dilution, 0.24 ng/µL, however, ST was only identified at concentration 2.4×10^{-02} ng/µL when sequenced at 10 GB depth. StringMLST was also able to identify some MLST genes for 2 and 5 GB sequencing depth at concentration 2.4×10^{-02} ng/µL and 10 GB at concentration 2.4×10^{-03} ng/µL, however, it was insufficient to allocate a ST.

As expected, total *Xylella* reads increased as sequencing depth increased and decreased as *Xylella* concentration decreased. Results for largest *Xylella* contig and coverage were linear with total *Xylella* reads with sequencing depth of 10 GB at the highest *Xylella* concentration giving almost full genome coverage at 99.67%.

We showed that metagenomic NGS is not as sensitive for detection of low titre bacterial pathogens as qPCR and LAMP that are currently used in laboratory testing. Results from the sensitivity test using LAMP, qPCR and metagenomics, showed that qPCR was 100 times more sensitive than metagenomic sequencing (using 10 GB of sequence data). However, it should be noted than the qPCR and LAMP used in this analysis could only identify Xf to the species level and metagenomic sequencing, though less sensitive, could provide species, subspecies and even ST level identification at some concentrations. The knowledge that a sample with a qPCR Ct of <24.5 sequenced to a depth of 10 GB could provide enough sequence data to identify an Xf isolate to the ST level is very useful baseline information for use in a biosecurity situation, although for PEQ it may be enough to know Xf is present. Being able to differentiate sequence types of other bacteria could be useful when some types are considered quarantine pathogens, but other types are present.

Table 4: LAMP, qPCR and metagenomic results for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to calculate fraction of reads while coverage was calculated using total length of combined contigs and total genome size. ST2 = sequence type of *Xylella* isolate used in this experiment, UD = Undetermined, Ct = threshold cycle of amplification in qPCR.

Concentration of *Xylella* DNA (ng/µL)	qPCR (Ct)	LAMP (amp time)	StringMLST			Total *Xylella* reads			Total *Xylella* contig			Largest *Xylella* contig			Genome coverage (%)		
			2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB
0.24	21.09	9.9	ST2	ST2	ST2[1]	38162	66031	147981[1]	2561	1486	318[1]	12785	33334	129197[1]	77.5	93.62	99.67[1]
2.4×10^{-02}	24.52	11.2	XF	XF	ST2	2089	4686	9542	133	488	1233	690	1570	3106	1.84	7.23	20.59
2.4×10^{-03}	27.87	12.9	UD	UD	XF	190	501	1155	3	11	45	523	439	575	0.05	0.142	0.619
2.4×10^{-04}	31.17	14.7	UD	UD	UD	13	57	96	1	1	1	317	281	333	0.012	0.011	0.013
2.4×10^{-05}	34.21	20.4	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-06}	37.45	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-07}	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-08}	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD

[1] One replicate was contaminated and was removed from analysis.

As expected, an increase in total sequencing depth generated more Xf reads, increased the Xf genome coverage and improved the resolution of Xf taxonomic identification. This experiment did not include sequencing of a single sample any deeper than 10 GB per sample. It is of interest to the project team to explore whether the sensitivity of qPCR detection of Xf can be achieved with a greater sequencing depth of metagenomic data, however, based on current sequencing costs this exercise would likely be an academic one. Even with the cheapest sequencing on the NovaSeq being approx. $22 per GB, an increased depth of 20 GB would add up to $440 per sample. It is unlikely that this would be a viable option for PEQ diagnostics. For this reason, more cost-effective techniques that facilitate the selective enrichment and sequencing of multiple target pathogens from complex samples were explored in this project. As an example, the hybridisation probe assay designed (Appendix 14) is expected to add approximately $50 to the cost of processing each sample but will drastically reduce the amount of sequencing required.

Targeted NGS has the potential to be more sensitive that, metagenomics for pathogen detection, and potentially as sensitive as polymerase chain reaction (PCR) assays whilst providing core genome information for accurate pathogen identification. The use of hybridisation probes prior to HTS enables capture of target nucleic acids from a broader pool, essentially enriching the sample for a specific target and reducing the amount of sequencing required. Hybridisation probes can be used for a range of applications including diversity studies, molecular diagnostics and breed/cultivar improvement (Donaldson et al., 2017; Mascher et al., 2013; Muraya et al., 2015; Pankin et al., 2014; Whiston et al., 2017). In Milestone 105, we designed and synthesized a hybridization probe capture assay for the detection of pathogenic bacteria within the National Priority Plant Pests (NPPP) list (2019), many of which can be present in low titre in their plant hosts and difficult to accurately identify with routine diagnostics (Table 5 and Appendix 22). Multiple genomic targets are included for each bacterial pathogen, to facilitate accurate identification, but the design strategy varies depending on genome sequence availability. They were also chosen based on their use for species, sub-species or sub-group differentiation for taxonomic purposes. Consequently, a total of 5,901 dsDNA hybridization probes were designed for *Xylella fastidiosa* (Xf), *X. taiwanensis* (Xt)*, Xanthomonas citri* subsp. *citri* (Xcc), *Erwinia amylovora* (Ea), '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso), and 16SrI group phytoplasmas genomes (Table 5). Probes were designed with 4 X tiling to increase capture rate of low titre pathogens and low complex regions were removed to reduce off-target binding. Each probe consists of a 120 mer dsDNA oligonucleotide with a biotin molecule at one end to capture targeted genes using Streptavidin bound to magnetic beads.

Exemplar 1: Xylella fastidiosa

Kraken2 and Bracken software were used to classify individual reads to an operational taxonomic unit for probed and metagenomic serial diluted Xf samples using full genome databases. As expected, total Xf reads and percentage of Xf reads decreased as Xf concentration decreased for both probed and metagenomic samples (Table 6 and Appendix 22, Table 22.6). This was also observed for reads mapped to probed Xf regions. Mapping of Xf probed regions showed that all Xf genes were mapped to.

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples up to Xf concentration of 2.4 x 10^{-06} ng/µL. This was also observed for reads mapped to probed Xf regions. Xf concentration of 2.4 x 10^{-03} ng/µL gave the highest fold increase; 5,736.3 X more when using Bracken with full genome databases and 122,206.3 X more when mapped to probed regions. Xf concentration of 2.4 x 10^{-06} ng/µL gave the lowest fold increase; 93.2X more when using Bracken and 1,963.5X more when mapped to probed regions (Appendix 22, Table 22.6).

Regarding sensitivity, the LAMP and qPCR assays detected Xf at femtogram levels, 2.4 x 10^{-05} and 2.4 x 10^{-06} ng/µL, respectively (Table 7). Using Bracken, Xf was detected confidently up to 2.4 x 10^{-02} ng/µL for metagenomic Xf samples, 1,000 X and 10,000 X less sensitive than LAMP and qPCR respectively. Xf was detected confidently up to 2.4 x 10^{-06} ng/µL for probed Xf samples, 10 X more sensitive than LAMP and equally as sensitive as qPCR.

Using StringMLST on trimmed reads, the sequence type (ST) was able to be identified in the first dilution, 0.24 ng/µL, of metagenomic Xf samples. StringMLST was also able to identify some MLST genes at Xf concentration 2.4 x 10^{-02} ng/µL, however, it was insufficient to allocate a ST. StringMLST was unable to allocate a ST for all probed Xf samples due to the absence of hybridization probes targeting the gene *petC*. Nevertheless, StringMLST was still able to detect the rest of the six MLST genes up to Xf concentration 2.4 x 10^{-05} ng/µL. StringMLST was also able to identify some MLST genes at Xf concentration 2.4 x 10^{-06} ng/µL, however, not all six genes were detected.

Table 5: The genes that were chosen for hybridization capture for detection and identification of *Erwinia amylovora, Xanthomonas citri* subsp. *citri, Xylella fastidiosa, Xylella taiwanensis*, *Candidatus* Liberibacter species and phytoplasmas in the 16SrI group.

Unique regions				**MLST genes+ unique regions**		**MLST genes**			
Erwinia amylovora		***Xanthomonas citri* subsp. *citri***		***Xylella fastidiosa* / *Xylella taiwanensis*** [1]		***Candidatus* Liberibacter sp.**[2]		**Phytoplasma 16SrI**	
Gene annotation	Gene length (bp)	Gene annotation	Gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)
eop2	1991	AMDFNCPH_00526	1518	group_806	2257	*rplL*	396	rbfA	342
WP_004154715.1_7	6255	AMDFNCPH_00529	1185	group_1256	5311	*rplJ*	519	nusA	1074
WP_004155245.1_424	3285	AMDFNCPH_01080	2025	*mshA*	2568	*metG*	1536	rpsC-rpsS	1434
amsE	801	AMDFNCPH_01855	1185	group_7658 (subsp. *pauca* only)	2093	*adk*	606	rplF	555
amsD	1056	AMDFNCPH_01857	1380	*copB*	933	*mutS*	2703	secY	1242
Total length	13388	AMDFNCPH_01859	2988	group 1350	1378	*grpE*	636	map	744
		AMDFNCPH_02036	2109	*dacC*	1446	*dnaG*	1935	secA	2508
		AMDFNCPH_02047	2118	group_852	2217	*glnA*	1404	Total length	7899
		AMDFNCPH_02358	10173	*gltT*	1313	Total length	9735		
		AMDFNCPH_02724	1911	*lacF*	881				
		AMDFNCPH_03192	1218	*nuoL*	2142				
		Total length	27810	*holC*	425				
				leuA	1559				
				cysG	1429				
				Total length	25951				

[1] Includes *Xylella taiwanensis*, *X. fastidiosa* subsp. *fastidiosa*, *X. fastidiosa* subsp. *multiplex*, *X. fastidiosa* subsp. *pauca* and *X. fastidiosa* subsp. *sandyi*
[2] Includes *Candidatus* Liberibacter solanacearum, *Ca.* Liberibacter africanus, *Ca.* Liberibacter americanus and *Ca.* Liberibacter asiaticus

Exemplar 2: *Candidatus* Liberibacters

Mapping of *Candidatus* Liberibacter probed regions showed that reads were mapped to all *Candidatus* Liberibacter genes for both CLaf and CLso samples (Appendix 22, Figure 22.3). As CLbr genes were not added to the probing list, CLbr genes did not show full gene coverage for probed CLbr samples. Probed CLbr samples also showed reduced mapped reads for the *adk* gene and minor increase for the *glnA* gene when compared to metagenomic samples (Figure 3 and Appendix 22).

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples for all *Candidatus* Liberibacter samples. This was also observed for reads mapped to probed *Candidatus* Liberibacter regions (Table 8 and Appendix 22, Table 22.10). CLaf gave the highest fold increase; 470.6 X more when using Bracken with full genome databases and 22,481.36 X more when mapped to probed *Candidatus* Liberibacter regions.

Mapped reads of probed and metagenomic *Candidatus* Liberibacter samples were extracted from BAM files and was *denovo* assembled using SPAdes. Samples that produced full *Candidatus* Liberibacter gene datasets were: probed CLaf (P-Laf), probed CLso-Insect (P-CLso-I), probed CLso-Plant (P-CLso) and metagenomic CLso-Plant (CLso). Choosing one replicate, P-Laf, P-CLso and CLso were, all eight genes (*rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA*) were extracted from assembled reads, concatenated, and aligned against multiple *Candidatus* Liberibacter species. Phylogenetic tree was then constructed (Figure 4 and Appendix 22, Figure 22.5). From the phylogenetic tree, P-CLso and CLso were classified as *Candidatus* Liberibacter solanacearum haplotype A and P-Laf as *Candidatus* Liberibacter africanus closely related to the CLaf isolate PTSAPSY.

The hybridization probe protocol required samples to be pooled before probing and an amplification step post probing. Due to this, sequencing output per sample cannot be controlled in each pool leading to higher amounts of sequencing data for samples with higher pathogenic DNA, which was observed in this report. This could interfere with pathogen detection when low titre pathogens are mixed with high titre pathogens in the same pool. This potential problem will be addressed in future experimentation.

Results from this report showed that probes can increase HTS pathogen detection sensitivity to qPCR/cPCR levels. However, if possible, options to increase sensitivity should be trialled. This report has indicated that longer probed regions might increase sensitivity as seen for Xcc. More probed regions were also indicated to increase fold enrichment, which would lead to better identification due to the extra reads and better gene assembly for phylogenetic assembly. Another method would be to modify the hybridization enrichment protocol. After probe capture using streptavidin-coated magnetic beads, a PCR amplification is done to increase concentration of targeted DNA (Figure 1). Due to the protocols being designed for accumulatively longer probed regions such as whole exome capture, there is a probability that increasing the cycles of the final PCR may increase pathogen detection.

In conclusion, this protocol shows potential for pathogen detection using probe based targeted NGS. Although further improvements and validations are necessary. This includes increasing the number of reference genomes representing a broad range of strains of all pathogens, for hybridization probe design. These draft protocols need to be tested with PEQ to further validify the pipeline and identify any improvements needed.

Hybridization capture HTS uses biotinylated oligonucleotide probes to capture targeted DNA regions within a genome. The use of hybridization capture prior to HTS enables capture of target nucleic acids from a broader pool, essentially enriching the sample for a specific target and reducing the amount of sequencing required. Hybridization capture HTS is normally performed after library preparation enabling indexed libraries to be multiplexed before hybridization capture. For bacterial diagnostic purposes, these probes could either target specific bacterial pathogens for direct identification (Gasc et al., 2017) or target a barcoding region such as the 16s rRNA (Gasc and Peyret, 2018).

Table 6: Bracken and mapping result average for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to assign Xf reads based on full genome databases and Bowtie2 was used for mapping Xf reads to probed Xf regions. Fold enrichment = fold increase of Xf percentage of total reads from metagenomic to probed samples.

			Metagenomic			Probed			
	Sample	Concentration of *Xylella* DNA (ng/µL)	Xf reads	Total paired reads	Xf percentage of total reads (%)	Xf reads	Total paired reads	Xf percentage of total reads (%)	Fold enrichment
K-mer (Bracken)-full genome	XF-2-0	0.24	18030.5	4559165	0.39548	29163112	30223678.5	96.49094	244
	XF-2-1	2.4×10^{-02}	982.5	4084508	0.02405	7550430.5	10432767	72.37227	3,008.7
	XF-2-2	2.4×10^{-03}	134	3788078.5	0.00354	943358.5	4649001	20.29164	5,736.3
	XF-2-3	2.4×10^{-04}	45.5	4147394	0.00110	99788.5	3590935.5	2.77890	2,533
	XF-2-4	2.4×10^{-05}	21.5	3741020.5	0.00057	10622	3138178	0.33848	589
	XF-2-5	2.4×10^{-06}	22.5	4005279.5	0.00056	1745	3332061	0.05237	93.2
	XF-2-6	2.4×10^{-07}	18	3776454.5	0.00048	35.5	2688787.5	0.00132	2.8
	XF-2-7	2.4×10^{-08}	22	4027120	0.00055	51	2229032.5	0.00229	4.2
	Sample	Concentration of *Xylella* DNA (ng/µL)	On-target reads	Total reads	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
Mapping (Bowtie2)-probed region	XF-2-0	0.24	1430.5	9118330	0.01569	53180973	60447357	87.98	5,608
	XF-2-1	2.4×10^{-02}	64	8169016	0.00078	13792088	20865534	66.1	84,370.4
	XF-2-2	2.4×10^{-03}	11.5	7576157	0.00015	1724774	9298002	18.55	122,206.3
	XF-2-3	2.4×10^{-04}	9	8294788	0.00011	183880	7181871	2.56	23,597.2
	XF-2-4	2.4×10^{-05}	0.5	7482041	6.68E-06	19208.5	6276356	0.306	45,796.9
	XF-2-5	2.4×10^{-06}	2	8010559	2.5E-05	3267	6664122	0.049	1,963.5
	XF-2-6	2.4×10^{-07}	0	7552909	0	0.5	5377575	9.3E-06	*
	XF-2-7	2.4×10^{-08}	1.5	8054240	1.86E-05	0.5	4458065	1.12E-05	0.6

* Can't be calculated as fold enrichment is not divisible by zero

Table 7: LAMP and qPCR results from Milestone 105, and sequencing results for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to assign Xf reads and calculate fraction of reads. ST2 = sequence type of *Xylella* isolate used in this experiment, UD = Undetermined, Ct = threshold cycle of amplification in qPCR.

Sample	Concentration of *Xylella* DNA (ng/µL)	qPCR (Ct)	LAMP (amp time)	StringMLST		Xf reads		Xf percentage of total reads (%)	
				Metagenomic	Probed[2]	Metagenomic	Probed	Metagenomic	Probed
XF-2-0	0.24	21.09	9.9	ST2	XF	18030.5	29163112	0.39548	96.49
XF-2-1	2.4×10^{-02}	24.52	11.2	XF[1]	XF	982.5	7550430.5	0.02405	72.37
XF-2-2	2.4×10^{-03}	27.87	12.9	-	XF	134	943358.5	0.00354	20.29
XF-2-3	2.4×10^{-04}	31.17	14.7	-	XF	45.5	99788.5	0.00110	2.78
XF-2-4	2.4×10^{-05}	34.21	20.4	-	XF	21.5	10622	0.00057	0.338
XF-2-5	2.4×10^{-06}	37.45	UD	-	XF[1]	22.5	1745	0.00056	0.0524
XF-2-6	2.4×10^{-07}	UD	UD	-	-	18	35.5	0.00048	0.00132
XF-2-7	2.4×10^{-08}	UD	UD	-	-	22	51	0.00055	0.00229
Negative	0	UD	UD	-	-	35		0.00011	

[1] Incomplete gene detection/typing

[2] *petC* was not included in the probe set

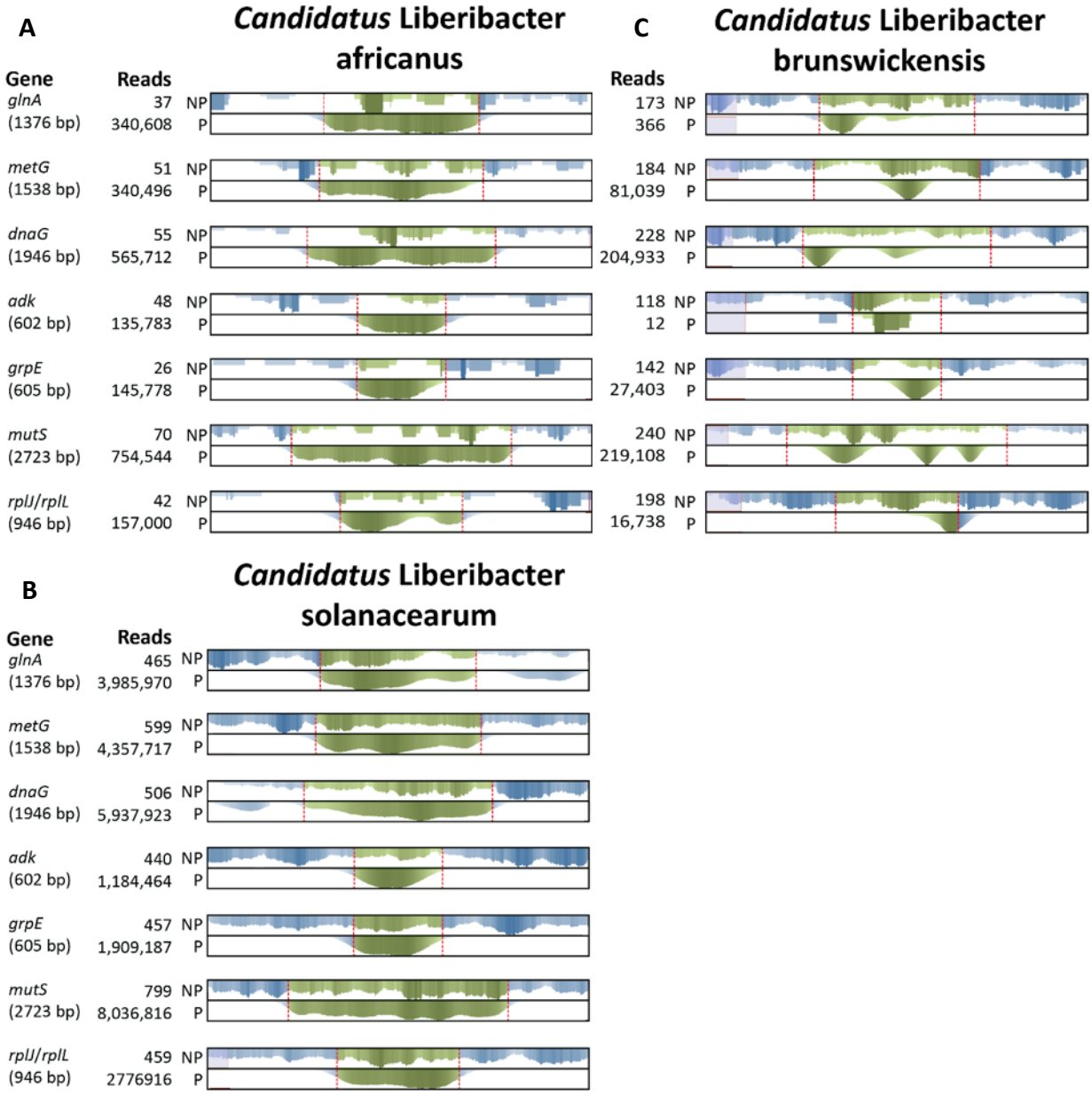


Figure 3: Mapping of metagenomic (NP) and probed (P) reads to *Candicatus* Liberibacter genes for (A) *Ca.* Liberibacter aricanus, (B) *Ca.* Liberibacter solanacearum (plant host) and (C) *Ca.* Liberibacter brunswickensis (insect host). Green regions within the red dotted lines indicate gene location while blue regions are 1 kb sequences 5' and 3' of the gene.

Table 8: Bracken and mapping result average for *Candidatus* Liberibacter brunswickensis (CLbr) in host insect and plant DNA, *Ca*. Liberibacter solanacearum (CLso) in host insect and plant DNA, and *Ca*. Liberibacter africanus (CLaf) in host plant DNA. Bracken was used to assign *Ca*. Liberibacter reads based on full genome databases and Bowtie2 was used for mapping *Ca*. Liberibacter reads to probed *Ca*. Liberibacter regions. Fold enrichment = fold increase of *Ca*. Liberibacter percentage of total reads from metagenomic to probed samples.

		Metagenomic			Probed			
K-mer (Bracken)-full genome	Sample	*Ca.* Liberibacter reads	Total paired reads	*Ca.* Liberibacter percentage of total reads (%)	*Ca.* Liberibacter reads	Total paired reads	*Ca.* Liberibacter percentage of total reads (%)	Fold enrichment
	Clbr-Insect	13106.5	8157829	0.1607	216498	1824880	11.8637	73.84
	Clbr-Plant	168	9217579	0.0018	517	438745	0.1178	64.65
	Clso-Insect	29	1040	2.7885	144866.5	408381	35.4734	12.72
	Clso-Plant	100018.5	8545992	1.1704	14923358	30269491	49.3016	42.13
	CLaf	8031	8671376	0.0926	1110086	2546838	43.5868	470.62
Mapping (Bowtie2)-probed region	Sample	On-target reads	Total yield (reads)	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
	Clbr-Insect	1377	8157829	0.0169	538619	1824880	29.5153	1,748.59
	Clbr-Plant	2	9217579	0.00002	1232	438745	0.2808	12,941.52
	Clso-Insect	0	1040	0	288231	408381	70.5789	*
	Clso-Plant	3638	8545992	0.0426	29505607	30269491	97.4764	2,289.81
	CLaf	334.5	8671376	0.0039	2208676	2546838	86.7223	22,481.36

* Can't be calculated as fold enrichment is not divisible by zero

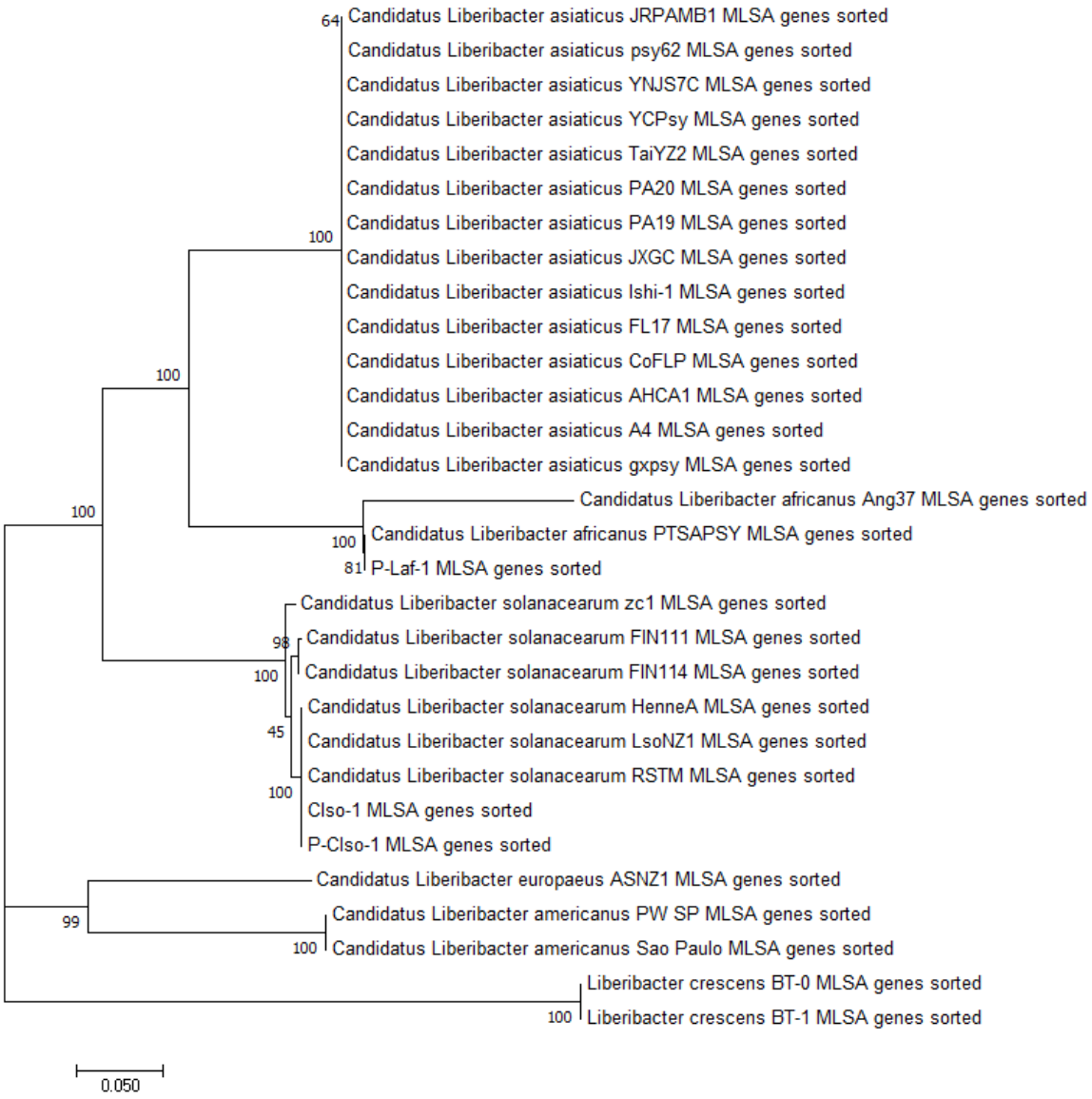


Figure 4: Maximum likelihood tree based on *rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA* gene sequences of probed CLaf (P-Laf), probed CLso-Plant (P-CLso) and metagenomic CLso-Plant (CLso) and their closest phylogenetic neighbours.

Hybridization capture HTS has been used for multiple bacteria detection/identification and as a diagnostic tool. Hybridization capture has been used to sequence full bacterial genomes of species such as *Candidatus* Liberibacter asiaticus (Cai et al., 2019), *Mycobacterium tuberculosis* (Nimmo et al., 2017), *Chlamydia psittaci* (Branley et al., 2016) and *Chlamydia pecorum* (Bachmann et al., 2015), without the need for culturing, to analyse diversity of bacterial strains. Hybridization capture HTS has also been used for 16S sequencing to avoid common problems that have risen from primers as described above. These 16S studies found that hybridization capture HTS reduces the bias in 16S rRNA amplicon sequencing of microbial communities (Cariou et al., 2018; Rassoulian Barrett et al., 2020). Capture of the 16S region has also been designed to specifically target the *Ca*. phytoplasma spp. (Trivellone et al., 2022), giving better resolution than amplicon Sanger sequencing.

Capturing full genomes using hybridization capture HTS has been used for viruses, with some research targeting the whole genomes of up to 356 viral species (Metsky et al., 2019). However, application to full

bacterial genomes can be costly due to their large genome size compared to many viruses. Hence, to target and identify multiple genera of bacterial pathogens in a sample, only selected gene regions within the genome should be captured and could include the 16S region (Trivellone *et al.*, 2022) as well as other taxonomically informative genes. For instance, using five vertebrate model species, >500 loci in highly conserved regions were developed (Lemmon et al., 2012) which could aid in population genomics and phylogenetic studies. Also referred to as Anchored Hybrid Enrichment, hybridization capture HTS has also been used to successfully investigate loci diversity in an entire superfamily of insects (Membracoidea) (Dietrich et al., 2017).

Pros

- Unlimited target capture regions.
- Easy to add more probes (to target more genes/genomes) to original probe sets making it a flexible detection/identification system.
- Able to give high taxonomic resolution.
- Can detect multiple bacterial genera and species in one test.
- Can run on multiplexed samples to reduce total hybridization capture HTS reactions.

Cons:

- Can be costly if targeting full genomes.
- Potential to not be as sensitive as amplicon sequencing and qPCR.
- Requires the availability of the genomic data for target identification.

Conclusion

There are several HTS approaches that can be used to detect and identify bacterial pathogens in a PEQ setting. Metagenomic HTS could be applied when discovery of an unknown pathogen is required. For culturable bacteria, isolation and WGS of pure colonies may be beneficial for confirmatory testing and may be more cost effective when a single culturable bacteria is the target. However, hybridization capture HTS is a targeted approach which overall provides PEQ with the greatest benefits:

- With all probes for every genome in one reaction, hybridization capture HTS is ideal for broad pathogen screening.
- It can be developed for the specific targets that require routine testing in PEQ, reducing the need for policy considerations around the risk of application because novel or unknown pathogens are unlikely to be detected.
- It does not require bacterial isolation and culturing and can be used to detect culturable and non-culturable bacteria.
- It is highly sensitive.
- Interpretation is simple because the analysis pipeline for unique and MLSA/MLST regions is straight forward.
- It is a faster solution than Sanger sequencing multiple amplicons.

The limitation of hybridization capture HTS is the need for genomic information or sequence data of taxonomically informative unique and MLSA/MLST regions, which may not be available for some targets. But this data should be simple to produce in collaboration with international colleagues.
Development, from design to application, would also be faster once quality control and assurance of accurate, reproducible, and consistent results are set in place.

Outputs

Table 9. Output summary

Output	Description	Detail
< List the output (e.g. best practice guide, study tour, new variety, technology or growing protocols)>	< Describe and define the output, including the intended audience. For example, for publication type outputs describe the content, number and reach >	< Include a brief explanation of how and where the output was made available (or next steps required to make it available), and evidence of engagement by the intended audience, e.g. survey data of usefulness, adoption rates, commercialisation agreements. Include a hyperlink for publicly available outputs >
NGS initiatives register	Listing of key NGS initiatives and expert groups on plant viruses, plant bacteria and human bacteria model systems implementing diverse approaches to diagnose these pathogens.	Contact information of groups in Australia and overseas with expertise in the detection of plant viruses, and bacterial pests using specific NGS platforms and NGS methods (**Appendix 1**).
List of priority bacterial pathogens	Key plant pathogenic bacteria identified and prioritised in consultation with the reference group.	Bacterial pathogens identified and prioritized using references obtained from the National Priority Plant Pests (NPPP) (2019) list; Plant Health Australia and the Victorian post-entry quarantine target pathogens; The Emergency Plant Pest Deed (EPPRD) (**Appendix 3)**.
NGS best practices report	Current best practices and standards for the diagnosis of viruses and viroids using NGS. Provides information from sample collection, samples processing and data analysis	High throughput sequencing (HTS), also known as next generation sequencing (NGS), technologies are becoming more commonly used in different fields of virus research including phytosanitary diagnosis. Availability of distinct sequencing platforms and sample preparation approaches contributed to a revolution in detection and discovery of plant viruses and viroids. Adoption of end-to-end standard operating procedures (SOPs) and best practices are key to reliably detect pests in distinct laboratory settings. In this review, we revise the current state of best practices of the use of small RNA (sRNA) sequencing for the diagnosis of viruses and viroids (**Appendix 4**).
Pilot trial report	Comparison of PEQ protocols against small RNA-seq and ribosomal depleted RNA-seq approaches to detect viruses and viroids in positive control plants	Side-by-side comparison of positive control imported plants using PEQ and NGS methods for the diagnosis of viruses and viroids in diverse horticulture commodities. Results and discussion on the performance of two NGS approaches as compared to PEQ exiting protocols is presented (**Appendix 5**).
NGS end-user workshop report	To provide high quality NGS diagnostic testing in biosecurity across Australia and New Zealand, this workshop engaged managers of pathogen testing schemes, diagnosticians, policy regulators and industry leaders in Australia and New Zealand to identify and discuss the requirements of a	Representatives from different research and diagnostic organisations presented and discussed their workflows, objectives, and challenges. This provided a basis of considerations that are common between different users of NGS sequencing and analysis with which to discuss standards required to make a framework, not prescriptive protocols

	framework that may be applied to different commodities and industries and provide the basis of national standards.	that may be applied specific commodities or industries (**Appendix 6**).
NGS protocol	NGS protocols for detection of key culturable plant pathogenic bacteria	Description an NHS pipeline to identify plant pathogenic bacteria from acquisition of bacterial cultures to bioinformatics data analysis. Details on DNA extraction, library preparation (Illuminia and ONT), sequencing and data analysis are presented to analysed gram-negative bacteria, including isolates from the *Pantoea* genus and isolates of *Pseudomonas syringae* pv. *actinidiae* biovar 3 (Psa3) (**Appendices 7 and 8**). Additionally, a list of key bacterial genomes used in the comparative analysis of key bacterial pests are listed (**Appendix 9**).
Technical report	Phase 1 – Large scale side-by-side comparison of PEQ vs small RNA-seq NGS methods	A total of 186 samples were identified processed and evaluated using small RNA-seq NGS. Lessons learnt were used to improve the data analysis pipeline (**Appendix 10**).
Technical report	Cultured bacteria high throughput sequencing pipeline validation.	Validation and optimisation of NGS pipeline by assessing 37 pure culturable bacterial isolates provided by DAFF's SSG team. PEQ samples were collected from various infected plants processed through the NGS pipeline (**Appendix 11**).
NGS protocol	Fastidious and unculturable bacteria pose a problem for post entry quarantine (PEQ) testing as confirmation through isolation is not an option, pathogens can exist in low titre and numerous specific molecular tests are relied upon for detection of a range of exotic pathogens in the one host/sample. Streamlining of this testing into one simplified test to reduce resources and turnaround time is highly desirable and multiple high throughput sequencing (HTS) approaches hold promise to aid in achieving this goal.	The following fastidious bacterial pathogens as case studies (**Appendix 12**); • three phytoplasmas *in planta*, *Candidatus* Liberibacter africanus, *Ca*. L. solanacaereum and *Ca*. L. brunswickensis *in planta* and in their respective vectors; • the following pathogens spiked into host DNA at varying concentrations; ○ *Xanthomonas citri* subsp. *citri* in lemon leaf DNA, ○ *Xylella fastidiosa* in olive leaf DNA; and ○ *Erwinia amylovora* in apple leaf DNA
Pilot trial report	Metagenomic sequencing has the capability to perform unbiased identification of known and unknown plant pathogens as total DNA samples are sequenced. This permits the identification of essential genes that may further clarify genus and species simultaneously.	In this report, the detection limit of metagenomic sequencing of spiked samples at three sequencing depths (2 GB, 5 GB and 10 GB) is compared to current diagnostic approaches including qPCR and LAMP with the aim of getting the baseline for using HTS sequencing without enrichment. Additionally, the level of taxonomic identification (e.g., species, subspecies and sequencing type (ST)) that can be determined from the metagenomic data was also assessed (**Appendix 13**).
Pilot trial report	The simplest and least biased option for pathogen detection using high throughput sequencing (HTS)	An initial trial comparing metagenomics, qPCR and LAMP, discussed in this report (Appendix III), indicated that LAMP and qPCR were more

	sequencing is to sequence the plant samples directly, inclusive of host DNA and the entire microbiome (including the pathogen). This is called a metagenomic sequencing approach. However, in cases where the pathogen is present in low titre within the host, metagenomic sequencing is often not sensitive enough to detect the pathogen.	sensitive at detecting Xylella fastidiosa (Xf), than direct sequencing of infected plants when up to 10 GB of data was sequenced per sample. This indicates a need for enrichment of the pathogen or pathogen DNA if HTS sequencing is going to be used for detection of low titre fastidious pathogens at the same sensitivity as qPCR (**Appendix 14**).
Peer reviewed manuscript	Side-by-side comparison of post-entry quarantine and high throughput methods for the diagnosis of plant viruses and viroids.	Comparison of two NGS methods (RNA-Seq and sRNA-Seq) with that of existing PEQ molecular assays in detecting and identifying viruses and viroids from various plant commodities. Several bioinformatics tools were tested which rely on different approaches: direct-read, *de novo* and reference-guided assembly. Report the development of a new portable, scalable, and reproducible end-to-end workflow of Virus Surveillance and Diagnosis (VSD) implemented in a nextflow pipeline framework (**Appendix 15**).
Report on custom curated virus and viroid database	Custom Plant Viruses and Viroids Database (PVirDB) for diagnosis of quarantinable and non-quarantinable viruses.	PVirDB is a custom curated nucleotide database consisting of sequences of quarantinable viruses and viroids of PEQ and non-quarantinable viruses of interest to domestic plant industries. Descriptors include taxonomic information (Genus, Family) as captured by GenBank records. Future improvements include verification of taxonomic information as per guidelines from the International Committee on Taxonomy of Viruses (ICTV) as well as include PEQ assay information for each quarantinable pest. We continue to liaise with the project team and policy regulators in related activities to improve and optimise the scope and design of PVirDB (**Appendix 16**).
Technical report	Improvements and optimisation of the Virus Surveillance and Diagnosis pipeline.	Improvements to quality trimming and assessment of RNA source origin of NGS data are reported along with relevant code. Also a contamination flag has been introduced to report potential cross-sample contamination events (**Appendix 17**).
Technical report	NGS service provider 'Ring' test identifies reliable commercial kit for data generation.	Three NGS service providers were engaged to test novel commercial kits for generation of small RNA NGS data. Results of these side-by-side comparison are presented (**Appendix 18**).
Technical report	A follow up phase 1 – Large scale side-by-side comparison of PEQ vs small RNA-seq NGS methods is presented.	A total of 199 samples were identified processed and evaluated using small RNA-seq NGS. Lessons learnt were used to improve the data analysis pipeline (**Appendix 19**).
Draft policy framework and principles report.	A High Throughput Sequencing (HTS) framework, policies, and principles for plant diagnostics.	The report guides the development of policies and processes for the adoption of NGS technology for routine phytosanitary screening of exotic viruses and viroids in plants in post entry quarantine. Screening for quarantinable pests verifies the health status of the imported plant material and provides assurance that it

		achieves the appropriate level of biosecurity protection applied by Australia before being released for entry into the country (**Appendix 20**).
Technical report	Validation of the NGS pipeline for culturable bacteria using a larger number culturable bacterial isolates.	In this report, we further validate this pipeline by applying it to bacterial isolates from post entry quarantine (PEQ) samples supplied by the Science and Surveillance Group (SSG). To do so, PEQ sourced pure genomic DNA of 44 cultured bacteria isolated from various infected plants were run through the HTS pipeline (**Appendix 21**).
Technical protocol	Testing and optimisation of hybridization probes for enrichment and detection of fastidious bacteria.	This protocol shows potential for pathogen detection using probe based targeted HTS. although further improvements and validations are necessary. This includes increasing the number of reference genomes representing a broad range of strains of all pathogens, for hybridization probe design. These draft protocols need to be tested with PEQ to further validify the pipeline and identify any improvements needed (**Appendix 22**).
Technical report	Large scale evaluation of the performance of different small RNASeq library kits. Comparison of NGS datasets produced using distinct NGS library approaches informed the best solution to produce high quality data and prevent cross-sample contamination events.	Large scale side-by-side comparison of samples processed using either the NEBNext library kit (AGRF) or the QIAseq library kit (Ramaciotti). In partnership with the SSG team and DPI, a total of 48 samples were selected from a range of commodities (**Appendix 23**).
Policy documents (Confidential)	Draft policy documents for operational adoption of HTS technology for routine plant virus/viroid diagnosis.	PEQ instruction manual and SOPs from nucleic acid extraction to interpretation of NGS results have been prepared in partnership with DAFF's SSG team at PEQ. Prepared documents were reviewed by DAFF policy regulators (**Appendix 25**).
Technical protocol	NGS protocol for the diagnosis of all known viruses and viroids in horticulture crops to be submitted to the Subcommittee on Plant Health Diagnostics (SPHD).	A draft national protocol for the use of NGS for the diagnosis of a range of viruses and viroids in a single assay has been sent for review and consideration to the SPHD (**Appendix 26**).
Technical report	Improvements to the hybridization probe evaluation and sensitivity for bacterial pathogens are reported.	Additional improvements to the hybridization probe targets were attempted and the hybridization capture assay are reported for *Xylella* spp. detection panel to enable ST determination. Additional phytoplasma genomes were included in probe design to evaluate detection of a broader range of phytoplasma 16Sr groups (16SrII, III, V, X, XI, XII and 16SrXXXVII). Comparison of the hybridization capture assay with metagenomic sequencing (non-probed) and current laboratory approaches for samples provided by the PEQ, Northern Australia Quarantine Strategy (NAQS) and Agriculture Victoria Research (AVR). Additionally, the level of taxonomic identification (e.g. species,

		subspecies, haplotype and ST) that can be determined from the HTS hybridization capture data was also assessed. Comparison of the hybridization capture assay with metagenomic sequencing (non-probed) on spiked samples with multiple pathogens. Additionally, the level of taxonomic identification (e.g. species, subspecies, haplotype and ST) that can be determined from the HTS hybridization capture data was also assessed (**Appendix 27**).
Technical report	Improvement of RNA extraction and assessing small RNA-seq controls.	A significant improvement of the small RNA profiles after modification of the extraction protocol was achieved, particularly, small RNA peaks at 21, 22 and 24 nt in length. Additionally alien and spike in controls were introduced to assess cross sample contamination events and sensitivity of detection, respectively (**Appendix 28**).
Peer review manuscript	Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass).	We describe the full-length genome sequence of a novel potyvirus, tentatively named "Miscanthus sinensis mosaic virus" (MsiMV), isolated from Miscanthus sinensis (silver grass) held in a post-entry quarantine facility after being imported into Western Australia, Australia. The MsiMV genome is 9604 nucleotides (nt) in length, encoding a 3071-amino-acid (aa) polyprotein with conserved sequence motifs. The MsiMV genome is most closely related to that of sorghum mosaic virus (SrMV), with 74% nt and 78.5% aa sequence identity to the SrMV polyprotein region. Phylogenetic analysis based on the polyprotein grouped MsiMV with SrMV, sugarcane mosaic virus (SCMV), and maize dwarf mosaic virus (MDMV) (**Appendix 29**).
Peer Review manuscript	Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants.	Developed a (1) virus and viroid report web-based bioinformatics workflows on Galaxy Australia called GA-VirReport and GA-VirReport-Stats for detecting viruses and viroids from host plant sRNA extracts and (2) a curated higher plant virus and viroid database (PVirDB). We implemented sRNA sequencing with unique dual indexing on a set of plants with known viruses. Sequencing data were analyzed using GA-VirReport and PVirDB to validate these resources. We detected all known viruses in this pilot study with no cross-sample contamination. We then conducted a large-scale diagnosis of 105 imported plants processed at the post-entry quarantine facility (PEQ), Australia (**Appendix 30**).
Technical report	Phase 2 – large scale NGS testing for viruses and viroids across diverse plant commodities undergoing quarantine testing at PEQ.	A total 206 plant consignments including Citrus, Rubus, Prunus, Grapevines and ornamental grasses were sequenced using small RNA-seq HTS and screened for the presence of quarantinable pests (**Appendix 31**).
End user hands-on	End-user hands-on training workshops	Training material for GA-VirReport and

training workshops	for the use of VirReport and GA-VirReport were organized for SSG staff at PEQ. Additionally, a hands on workshop was delivered as part of the Australasian Diagnostics and Surveillance Workshops in 2022 engaging a broad range of stakeholders from Australia and overseas.	VirReport along with instructional videos and user guides were prepared to train the SSG team, and diagnosticians from Australasian institutions. Training materials used in the workshops have been stored in GoveTeams for long term access by the SSG team (**Appendix 32**).
User guide	A detailed User guide for the use of the VirReport has been prepared to facilitate the training, transferring of capability and maintenance of the tool with DAFF.	VirReport takes small RNA-Seq fastq files as input. These can either be in raw format (currently only samples specifically prepared with the QIAGEN QIAseq miRNA library kit can be processed this way) or quality filtered. For target identification, VirReport uses a hybrid *de novo* assembly approach to build contigs that are then annotated using blast homology searches against either a virus database or/and a local copy of NCBI nr and nt databases (**Appendix 33**).
Policy draft document	Policy paper for adoption of HTS for bacteria in PEQ.	This draft policy paper discussed on the use of several NGS approaches that can be used to detect and identify bacterial pathogens in a PEQ setting. Metagenomic NGS could be applied when discovery of an unknown pathogen is required. For culturable bacteria, isolation and WGS of pure colonies may be beneficial for confirmatory testing and may be more cost effective when a single culturable bacteria is the target. However, hybridization capture NGS is a targeted approach which overall provides PEQ with the greatest benefits for fastidious/ unculturable bacteria (**Appendix 34**).
Technical protocol	Standard Operating Procedure for culturable bacteria using Illumina sequencing (end-to-end).	A user guide for the NGS testing of culturable bacteria is presented incorporating the lesson learnt throughout the duration of the project (**Appendix 35**).

Outcomes

Table 10. Outcome summary

Outcome	Alignment to fund outcome, strategy and KPI	Description	Evidence
< List the outcome (e.g. knowledge, awareness, practice change, commercialisation, availability of new knowledge for next	< Align to the relevant Fund outcome, strategy and KPI >	< Describe and define the outcome in terms of how it was realised by the target stakeholder group(s). Explain how the outcome is relevant	< What forms of evidence were collected to identify and understand the outcome (e.g. survey, observation, feedback)

phase project) >		at the Fund level >	>
Improved knowledge and understanding of NGS technologies • Understanding of current best practices for NGS diagnosis of plant pests • Appraise strengths and challenges of distict NGS methods for phytosanitary testing • Measure correlation of NGS and PEQ protocols • Priority list bacterial pests for NGS testing defined	• Increased knowledge of NGS diagnosis of plant pests by quarantine agencies (DA and NZ MPI), policy regulators and relevant industry stakeholders • Increased engagment with industry and regulatory stakeholders to guide, design and assess side-by-side comparison of both NGS and PEQ approaches	• Closely engaged quarantine agencies from DAFF and NZ MPI by regular project updates, partnership to deliver to project milestones, providing hands-on training workshops and on demand training materials. • Conducted a pilot and large scale PEQ vs NGS trials over the course of the project processing imported plant material from horticulture commodities including Citrus, Rubus, Grape, Potato and Nursery. • Compared the sensitivity and specificity of detection of NGS as compared to existing NGS protocols. • Engaged and hold regular meetings with policy regulators to advance the adoption of NGS at PEQ. • Released industry newsletter media articles and a podcast to communicate advances made in the project. • Key plant	• Partnership with DAFF to undertake a pilot and large-scale trials comparing NGS and PEQ protocols (**Appendices 5, 10, 19, 23, 24, 31**). • Determine the correlation of NGS and PEQ methods (**Appendices 5, 10, 15**) • NGS service provider ring test to assess detection of viruses and viroids (**Appendix 18**) • List of key plant pathogenic bacteria of biosecurity concern prepared (**Appendix 3**). • Determine the sensitivity of detection of NGS compared to molecular assays for culturable and fastidious bacteria (**Appendices 12, 13, 22, 27**).

		pathogenic bacteria identified and prioritized using references obtained from the National Priority Plant Pests (NPPP) (2019) list; Plant Health Australia and the Victorian post-entry quarantine target pathogens; The Emergency Plant Pest Deed (EPPRD).	
Recommendations for policy adoption of NGS technology will fast track the safe and rapid access to new genetics enabling plant industries to be more competitive nationally and overseas	• Completion and availability of NGS protocols for diagnosis of viruses/viroids and selected non-culturable and fastidious bacteria • Horticulture Industry stakeholders invited to science and end user workshops • NGS policy and principles adopted for PEQ testing of plant viruses and viroids in selected commodities. • Representative diagnosticians from all States, Territories and Crown organisations participated in end user training workshops	• Reviewed current best practices and standards for NGS testing to inform the development of Standard Operating Procedures. • Established NGS protocol for the detection of culturable and fastidious bacteria. • Proposed a draft policy framework and policy principles to engage policy regulators to facilitate the adoption of NGS for quarantine testing of viruses and viroids at PEQ. • Implemented SOPs and instructions manual for the operationalization and adoption of small RNA-seq NGS for the diagnosis of viruses and viroids. • On December 2022, small RNA-seq NGS has been adopted for the screening of viruses and viroids in Rubus, Prunus and Fragaria. • Prepared a policy	• Report on fit for purpose, best practices, and standards for NGS testing of plant viruses and viroids prepared (**Appendix 4**). • User manual for diagnosis of viruses and viroids using VirReport (**Appendix 33**). • End-to end NGS diagnosis pipeline for culturable bacteria (**Appendices 7, 11 and 35**). • Implemented a hybridization capture NGS protocol for fastidious/non-culturable bacteria (**Appendices 14, 22**). • Draft policy framework, and principles for NGS diagnosis of viruses and viroids (**Appendix 20**). • SOPs and instruction manual for the adoption of NGS for quarantine testing of plant viruses and viroids (**Appendix 25**).

		paper for the use of NGS for the detection of bacterial pests. • Organised hands-on training workshops for SSG and diagnosticians from diverse institutions.	• Organised and delivered six end used training workshops for the use of small RNA-seq NGS for the diagnosis of viruses and viroids (**Appendix 32**).

Monitoring and evaluation

< Discuss overall project performance, answering the Key Evaluation Questions (KEQs) that were identified within the project's M&E Plan (if relevant). Areas for continuous improvement should be identified and discussed where relevant. For more information, refer to *Attachment A3: Final report guide* >

KEQs usually consider project effectiveness, relevance, appropriateness and efficiency

Table X. Key Evaluation Questions

Key Evaluation Question	Project performance	Continuous improvement opportunities
< Refer to the M&E Plan >	< Identify aspects of project performance that address the Key Evaluation Questions >	< List opportunities for improvement and future development >
Effectiveness: The project aimed at improving industry access to new genetics through faster and more robust diagnostics methods		
Does NGS methods provide equivalent or better diagnostic capability than existing PEQ assays?	• Diagnosis of plant viruses and viroids using small RNA-seq NGS method detects all pathogens detected by PEQ methods. Furthermore, it provides information of additional pathogens if these are present. This is a significant benefit as list of quarantinable pests are updated overtime and the NGS approach is resilient to cope with changes in regulations, making it a future proof solution to accelerate quarantine testing. • NGS is used as a 'first pass' screening approach detecting relevant quarantinable pests if present, and subsequent PEQ molecular assays are run as a follow up to independently validate detections.	• The small RNA-seq NGS method relies on the host small RNA fraction that contains viral fragments and other host plant sequences. Further R&D is required to deplete non-informative host sequences (i.e., rRNAs, tRNA, chloroplast RNAs) to strengthen the signal detected for viral sequences. • Alternatively, new R&D methods that specifically enriches the small RNA fraction (21nt to 24nt) derived from the host response to viruses will enable to remove non-informative sequences. • Increasing the signal of viruses in the NGS data will provide cost-benefits to industry by reducing the cost of NGS

		testing.
Can NGS methods accelerate the quarantine testing of imported plants?	• A single NGS assay can detect all quarantinable viruses and viroids making it possible to efficiently allocated PEQ resources to validate NGS detections. • The project team in partnership with DAFF's Technology Information Branch has fully automated the NGS diagnostic capability at PEQ. Namely, once NGS data is generated by the accredited service provider, the data is automatically downloaded and processed using VirReport and GA-VirReport pipelines without requiring end user intervention. Thus, maximising the efficiency to process complex data and enable end users at PEQ to run validation tests to inform decision making. • It is anticipated that NGS methods will accelerate the turnaround time of PEQ testing providing fasted access to new genetics to industry as well as saving in importing new plant materials.	
Can NGS methods be used effectively for detecting plant pests?	• A new promising hybridization capture approach coupled with NGS sequencing has the potential to detect both culturable and fastidious bacteria in the same assay, making it possible to create custom panels against bacterial pest of interest.	• Further investment in the hybridization capture technology will enable to explore the upper limit of simultaneous pathogens that can be detected using this method. • Develop commodity specific hybridization capture panels to accelerate the screening of bacterial pest of biosecurity concern.
Relevance: Does the intended end user beneficiaries acquired the skills, resources, and capability to use the NGS technology for routine testing of plant pests?		
How did the project transferred technical skills and resources to end users?	• Six hands-on training workshops were organized to up skill the SSG team at PEQ. • On demand training material	• Training in bioinformatics of diagnosticians in Australia and overseas remains challenge and a gap. Enabling access to

	was prepared using the end user feedback. • The project team supported the development and deployment of PEQ-HiTS an in-house DAFF portal to enable staff to perform routine NGS diagnosis of plant viruses and viroids. • The PEQ team has the capability to run VirReport and GA-VirReport using the PEQ-HiTS portal. Additionally, the Plant virus and viroid database (PVirDB) can be updated using the portal to make sure the most updated list of viral pests deposited in public databases are available to perform NGS quarantine testing. • Hands-on training workshops of diagnosticians from a broad range of institutions in Australia, and overseas were conducted to disseminate the resources generated in the project.	both in person and on demand virtual training opportunities will enable the broad diagnostics community to benefit from using NGS for diverse applications.
Appropriateness: How was the engagement of diverse stakeholders throughout the duration of the project?		
How the COVID pandemic impacted the project?	• The project team was not able to organize end user science workshops in 2020 and 2021 as the global community adapted to live with COVID. • Virtual regular meetings with DAFF colleagues were held throughout the project to deliver to project milestones and engagement with policy regulator colleagues. • COVID impacted the supply chain for NGS services, requiring the team to engage a broader group of providers and commercial solution to increase resilience and flexibility to sustain NGS testing. • Industry communications were disseminated via Reference	

	Group meetings and industry newsletters.	
Efficiency: How does the project contributed to increase the efficiency in PEQ operations		
What processes were improved to increase PEQ testing efficiency?	• The project team and DAFF collaborated closely throughout the project to improve the efficiency of NGS testing from nucleic acids collection to interpretation of NGS results. • DAFF's access to in house cloud infrastructure capability facilitated initial R&D testing and later formal operational adoption of NGS for routine phytosanitary testing.	

Recommendations

- Continued support to PEQ is necessary to expand further the adoption of NGS to a broader range of plant commodities.
- Enhanced R&D methods for small RNA-Seq NGS testing have the potential to continue to improve savings and benefits to horticulture industries.
- Assessment of NGS for diagnosis of Citrus budwood imports can accelerate the access to new genetics by industry.
- Cleaning infected budwood material can be challenging and new methods are required to increase the shoot tip grafting efficiency of virus-free plants for pests detected using NGS.
- Additional investment is required to upscale the use of hybridization capture NGS method to diagnose multiple culturable and non-culturable bacteria in the same assay.

Refereed scientific publications

Journal article

Gauthier, M.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. Biology (Basel) 11:263. doi: 10.3390/biology11020263. PMID: 35205129; PMCID: PMC8868628.

Leblanc, Z., Gauthier, M.E., Lelwala, R., Elliott, C., McMaster, C., Eichner, R., Davis, K., Liefting, L., Thompson, J., Dinsdale, A., et al. (2022). Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass). Archives of Virology 167, 1701-1705. 10.1007/s00705-022-05445-3.

Lelwala, R.V., LeBlanc, Z., Gauthier, M.A., Elliott, C.E., Constable, F.E., Murphy, G., Tyle, C., Dinsdale, A., Whattam, M., Pattemore, J., and Barrero, R.A. (2022). Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants. Viruses 14. 10.3390/v14071480.

Gauthier, M.A., Abeyanake, S.W., Lelwala, R.V., McMaster, C., Eichner, R., Morrison, J., Elliott, C.E., Fiorito, S., Dinsdale, A., Pattemore, J., Barrero, R.A., 2022. First detection and complete genome sequence of a new potexvirus naturally infecting Adenium obesum. Archives of Virology (*submitted, June 2023)*.

References

Aliyari, R., and Ding, S.W. (2009). RNA-based viral immunity initiated by the Dicer family of host immune receptors. Immunol Rev *227*, 176-188. 10.1111/j.1600-065X.2008.00722.x.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410. 10.1016/s0022-2836(05)80360-2.

Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., and Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in Ecology & Evolution *19*, 535-544. https://doi.org/10.1016/j.tree.2004.07.021.

Ashikawa, S., Tarumoto, N., Imai, K., Sakai, J., Kodana, M., Kawamura, T., Ikebuchi, K., Murakami, T., Mitsutake, K., Maesaki, S., and Maeda, T. (2018). Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. Journal of Medical Microbiology *67*, 1589-1595. https://doi.org/10.1099/jmm.0.000855.

Bachmann, N.L., Sullivan, M.J., Jelocnik, M., Myers, G.S.A., Timms, P., and Polkinghorne, A. (2015). Culture-Independent Genome Sequencing of Clinical Samples Reveals an Unexpected Heterogeneity of Infections by Chlamydia pecorum. Journal of Clinical Microbiology *53*, 1573-1581. doi:10.1128/JCM.03534-14.

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., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455-477. 10.1089/cmb.2012.0021.

Barrero, R.A., Napier, K.R., Cunnington, J., Liefting, L., Keenan, S., Frampton, R.A., Szabo, T., Bulman, S., Hunter, A., Ward, L., et al. (2017). An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. BMC Bioinformatics *18*, 26. 10.1186/s12859-016-1428-4.

Brady, C., Cleenwerck, I., Venter, S., Vancanneyt, M., Swings, J., and Coutinho, T. (2008). Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). Syst Appl Microbiol *31*, 447-460. 10.1016/j.syapm.2008.09.004.

Branley, J., Bachmann, N.L., Jelocnik, M., Myers, G.S.A., and Polkinghorne, A. (2016). Australian human and parrot Chlamydia psittaci strains cluster within the highly virulent 6BC clade of this important zoonotic pathogen. Scientific Reports *6*, 30019. 10.1038/srep30019.

Bulman, S.R., McDougal, R.L., Hill, K., and Lear, G. (2018). Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. Australasian Plant Pathology *47*, 467-474. 10.1007/s13313-018-0579-3.

Butler, M.I., Stockwell, P.A., Black, M.A., Day, R.C., Lamont, I.L., and Poulter, R.T.M. (2013). Pseudomonas syringae pv. actinidiae from Recent Outbreaks of Kiwifruit Bacterial Canker Belong to Different Clones That Originated in China. PLOS ONE *8*, e57464. 10.1371/journal.pone.0057464.

Cai, W., Nunziata, S., Rascoe, J., and Stulberg, M.J. (2019). SureSelect targeted enrichment, a new cost effective method for the whole genome sequencing of Candidatus Liberibacter asiaticus. Scientific Reports *9*, 18962. 10.1038/s41598-019-55144-4.

Cariou, M., Ribière, C., Morlière, S., Gauthier, J.-P., Simon, J.-C., Peyret, P., and Charlat, S. (2018). Comparing 16S rDNA amplicon sequencing and hybridization capture for pea aphid microbiota diversity analysis. BMC Res Notes *11*, 461-461. 10.1186/s13104-018-3559-3.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884-i890. 10.1093/bioinformatics/bty560.

Coordinators, N.R. (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *41*, D8-D20. 10.1093/nar/gks1189.

Dietrich, C.H., Allen, J.M., Lemmon, A.R., Lemmon, E.M., Takiya, D.M., Evangelista, O., Walden, K.K.O., Grady, P.G.S., and Johnson, K.P. (2017). Anchored Hybrid Enrichment-Based Phylogenomics of Leafhoppers and Treehoppers (Hemiptera: Cicadomorpha: Membracoidea). Insect Systematics and Diversity *1*, 57-72. 10.1093/isd/ixx003.

Doddaraju, P., Kumar, P., Gunnaiah, R., Gowda, A.A., Lokesh, V., Pujer, P., and Manjunatha, G. (2019). Reliable and early diagnosis of bacterial blight in pomegranate caused by Xanthomonas axonopodis pv. punicae using sensitive PCR techniques. Scientific Reports *9*, 10097. 10.1038/s41598-019-46588-9.

Donaldson, M.E., Rico, Y., Hueffer, K., Rando, H.M., Kukekova, A.V., and Kyle, C.J. (2017). Development of a genotype-by-sequencing immunogenetic assay as exemplified by screening for variation in red fox with and without endemic rabies exposure. Ecol Evol *8*, 572-583. 10.1002/ece3.3583.

Eichmeier, A., Kiss, T., Necas, T., Penazova, E., Tekielska, D., Bohunicka, M., Valentova, L., Cmejla, R., Morais, D., and Baldrian, P. (2019). High-Throughput Sequencing Analysis of the Bacterial Community in Stone Fruit Phloem Tissues Infected by “Candidatus Phytoplasma prunorum”. Microbial Ecology *77*, 664-675. 10.1007/s00248-018-1250-9.

EPPO (2014). PM 7/120 (1) Pseudomonas syringae pv. actinidiae. EPPO Bulletin *44*, 360-375. 10.1111/epp.12171.

Flores, O., Prince, C., Nuñez, M., Vallejos, A., Mardones, C., Yañez, C., Besoain, X., and Bastías, R. (2018). Genetic and Phenotypic Characterization of Indole-Producing Isolates of Pseudomonas syringae pv. actinidiae Obtained From Chilean Kiwifruit Orchards. Front Microbiol *9*, 1907-1907. 10.3389/fmicb.2018.01907.

Frampton, R.A., Thompson, S.M., Kalamorz, F., David, C., Addison, S.M., and Smith, G.R. (2018). Draft Genome Sequence of a “<i>Candidatus</i> Liberibacter europaeus” Strain Assembled from Broom Psyllids (Arytainilla spartiophila) from New Zealand. Genome Announcements *6*, e00430-00418. doi:10.1128/genomeA.00430-18.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152. 10.1093/bioinformatics/bts565.

Gasc, C., Constantin, A., Jaziri, F., and Peyret, P. (2017). OCaPPI-Db: an oligonucleotide probe database for pathogen identification through hybridization capture. Database *2017*. 10.1093/database/baw172.

Gasc, C., and Peyret, P. (2018). Hybridization capture reveals microbial diversity missed using current profiling methods. Microbiome *6*, 61-61. 10.1186/s40168-018-0442-3.

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

Gouveia, B.C., Calil, I.P., Machado, J.P., Santos, A.A., and Fontes, E.P. (2016). Immune Receptors and Co-receptors in Antiviral Innate Immunity in Plants. Front Microbiol *7*, 2139. 10.3389/fmicb.2016.02139.

Gupta, A., Jordan, I.K., and Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for multilocus sequence typing. Bioinformatics *33*, 119-121. 10.1093/bioinformatics/btw586.

Gupta, S., Mortensen, M.S., Schjørring, S., Trivedi, U., Vestergaard, G., Stokholm, J., Bisgaard, H., Krogfelt, K.A., and Sørensen, S.J. (2019). Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. Commun Biol *2*, 291. 10.1038/s42003-019-0540-1.

Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schaffer, A.A., and Brister, J.R. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res *45*, D482-D490. 10.1093/nar/gkw1065.

Hu, Y., Green, G.S., Milgate, A.W., Stone, E.A., Rathjen, J.P., and Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. Phytobiomes Journal *3*, 92-101. 10.1094/PBIOMES-01-19-0004-R.

Leblanc, Z., Gauthier, M.E., Lelwala, R., Elliott, C., McMaster, C., Eichner, R., Davis, K., Liefting, L., Thompson, J., Dinsdale, A., et al. (2022). Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass). Arch Virol *167*, 1701-1705. 10.1007/s00705-022-05445-3.

Lelwala, R.V., LeBlanc, Z., Gauthier, M.A., Elliott, C.E., Constable, F.E., Murphy, G., Tyle, C., Dinsdale, A., Whattam, M., Pattemore, J., and Barrero, R.A. (2022). Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants. Viruses *14*. 10.3390/v14071480.

Lemmon, A.R., Emme, S.A., and Lemmon, E.M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology *61*, 727-744. 10.1093/sysbio/sys049.

Loiseau, M., Plaire, M., Renaudin, I., Taylor, R., Fujikawa, T., Griffin, R., Mann, R., Pion, A., and Renvoisé, J.-P. (2019). Detecting Spiroplasma citri: a comparison of PCR methods to be used for quarantine diagnostics. European Journal of Plant Pathology *155*, 71-80. 10.1007/s10658-019-01750-x.

Malandraki, I., Varveri, C., Olmos, A., and Vassilakos, N. (2015). One-step multiplex quantitative RT-PCR for the simultaneous detection of viroids and phytoplasmas of pome fruit trees. Journal of Virological Methods *213*, 12-17. https://doi.org/10.1016/j.jviromet.2014.11.010.

Mascher, M., Richmond, T.A., Gerhardt, D.J., Himmelbach, A., Clissold, L., Sampath, D., Ayling, S., Steuernagel, B., Pfeifer, M., D'Ascenzo, M., et al. (2013). Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant J *76*, 494-505. 10.1111/tpj.12294.

Metsky, H.C., Siddle, K.J., Gladden-Young, A., Qu, J., Yang, D.K., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., et al. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nature Biotechnology *37*, 160-168. 10.1038/s41587-018-0006-x.

Muraya, M.M., Schmutzer, T., Ulpinnis, C., Scholz, U., and Altmann, T. (2015). Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition. PLOS ONE *10*, e0132120. 10.1371/journal.pone.0132120.

Murray, T.D., Schroeder, B.K., Schneider, W.L., Luster, D.G., Sechler, A., Rogers, E.E., and Subbotin, S.A. (2017). Rathayibacter toxicus, Other Rathayibacter Species Inducing Bacterial Head Blight of Grasses, and the Potential for Livestock Poisonings. Phytopathology® *107*, 804-815. 10.1094/PHYTO-02-17-0047-RVW.

Na, S.-I., Kim, Y.O., Yoon, S.-H., Ha, S.-m., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. Journal of Microbiology *56*, 280-285. 10.1007/s12275-018-8014-6.

Nimmo, C., Doyle, R., Burgess, C., Williams, R., Gorton, R., McHugh, T.D., Brown, M., Morris-Jones, S., Booth, H., and Breuer, J. (2017). Rapid identification of a Mycobacterium tuberculosis full genetic drug resistance

profile through whole genome sequencing directly from sputum. International Journal of Infectious Diseases *62*, 44-46. https://doi.org/10.1016/j.ijid.2017.07.007.

Ordóñez, N., Salacinas, M., Mendes, O., Seidl, M.F., Meijer, H.J.G., Schoen, C.D., and Kema, G.H.J. (2019). A loop-mediated isothermal amplification (LAMP) assay based on unique markers derived from genotyping by sequencing data for rapid in planta diagnosis of Panama disease caused by Tropical Race 4 in banana. Plant Pathology *68*, 1682-1693. https://doi.org/10.1111/ppa.13093.

Pankin, A., Campoli, C., Dong, X., Kilian, B., Sharma, R., Himmelbach, A., Saini, R., Davis, S.J., Stein, N., Schneeberger, K., and von Korff, M. (2014). Mapping-by-Sequencing Identifies HvPHYTOCHROME C as a Candidate Gene for the early maturity 5 Locus Modulating the Circadian Clock and Photoperiodic Flowering in Barley. Genetics *198*, 383. 10.1534/genetics.114.165613.

Peňázová, E., Dvořák, M., Ragasová, L., Kiss, T., Pečenka, J., Čechová, J., and Eichmeier, A. (2020). Multiplex real-time PCR for the detection of Clavibacter michiganensis subsp. michiganensis, Pseudomonas syringae pv. tomato and pathogenic Xanthomonas species on tomato plants. PLOS ONE *15*, e0227559. 10.1371/journal.pone.0227559.

Rassoulian Barrett, S., Hoffman, N.G., Rosenthal, C., Bryan, A., Marshall, D.A., Lieberman, J., Greninger, A.L., Peddu, V., Cookson, B.T., and Salipante, S.J. (2020). Sensitive Identification of Bacterial DNA in Clinical Specimens by Broad-Range 16S rRNA Gene Enrichment. Journal of Clinical Microbiology *58*, e01605-01620. 10.1128/JCM.01605-20.

Sabin, L.R., Zheng, Q., Thekkat, P., Yang, J., Hannon, G.J., Gregory, B.D., Tudor, M., and Cherry, S. (2013). Dicer-2 processes diverse viral RNA species. PLoS One *8*, e55458. 10.1371/journal.pone.0055458.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962. 10.1371/journal.pone.0163962.

Stukenbrock, E.H., and McDonald, B.A. (2008). The Origins of Plant Pathogens in Agro-Ecosystems. Annual Review of Phytopathology *46*, 75-100. 10.1146/annurev.phyto.010708.154114.

Trivellone, V., Cao, Y., and Dietrich, C.H. (2022). Comparison of Traditional and Next-Generation Approaches for Uncovering Phytoplasma Diversity, with Discovery of New Groups, Subgroups and Potential Vectors. Biology *11*, 977.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome research *27*, 737-746.

Whiston, R., Finlay, E.K., McCabe, M.S., Cormican, P., Flynn, P., Cromie, A., Hansen, P.J., Lyons, A., Fair, S., Lonergan, P., et al. (2017). A dual targeted β-defensin and exome sequencing approach to identify, validate and functionally characterise genes associated with bull fertility. Scientific Reports *7*, 12287. 10.1038/s41598-017-12498-x.

Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology *20*, 257. 10.1186/s13059-019-1891-0.

Gutiérrez, P.; Rivillas, A.; Tejada, D.; Giraldo, S.; Restrepo, A.; Ospina, M.; Marín, M. PVDP: A portable open source pipeline for detection of plant viruses in RNAseq data. A case study on potato viruses in Antioquia (Colombia). *Physiol. Mol. Plant Pathol.* **2021**, *113*, 101604.

Ho T., Broome J.C., Buhler J.P., O'Donovan W., Tian T., Diaz-Lara A., Martin R.R., Tzanetakis I.E. Characterization of endogenous Rubus yellow net virus in raspberries. bioRxiv. 2021 doi: 10.1101/2021.06.17.448838.

Intellectual property

No project IP or commercialisation to report.

Acknowledgements

The project team acknowledges the Science and Surveillance Group at DAFF for their leadership, dedication, and contributions to make possible to deliver to project milestones and the adoption of NGS for quarantine testing at PEQ. We acknowledge the DAFF policy regulatory colleagues for their leadership, contributions, and guide to make possible to adopt NGS. We appreciate the tireless and dedicated effort from DAFF's Technology and Infrastructure Branch, Information Services Division for facilitating the automatization of NGS testing. We also acknowledge QUT's eResearch team for their day-to-day support and access to on-prem compute infrastructure to enable the delivery to project activities and deployment of an end-to-end analytical pipeline in a nextflow platform to make it accessible, reproducible, and scalable. The project team also acknowledges the Agriculture Victoria team for their support and assistance to assess and develop new NGS approaches for bacterial testing. We also appreciate the help received from NSW Department of Primary Industries to access biological resources to test NGS methods.

Appendices

< Please note that all appendices are published with the final report. If an appendix includes information that should not be published, such as commercially or institutionally-sensitive material such as sensitive meeting minutes, provide it to Hort Innovation separately and clearly mark it confidential. Appendices should include outputs and key project documents such as an IP Register. >

TBA

Appendix 1 - Australian and global initiatives on NGS for viruses and bacterial plant pest diagnostics

Literature review identified key NGS initiatives on plant viruses (Table 1.1), plant bacteria (Table 1.2) and human (Table 1.3) model systems implementing diverse approaches to diagnose viral and bacterial pathogens.

Table 1.1: Initial list of national and international groups with expertise in virus diagnosis using NGS methods

Country	Contact(s)	Organisation	Species*	HTS methods
Australia	Roberto Barrero, Marie-Emilie Gauthier, Ruvini Lelwala, Doug Beattie Mark Whattam, Adrian Dinsdale, Julie Pattemore, Candace Elliot, SSG team	Queensland University of Technology, Brisbane, Australia Department of Agriculture, Fisheries and Forestry; Post Entry Quarantine, Mickleham, Victoria, Australia	PEQ	Small RNA (sRNA), Oxford Nanopore Technology (ONT)
	Brendan Rodoni, Fiona Constable, Jo Mackie	Agriculture Victoria Research, Agriculture Victoria, AgriBio, The Centre for AgriBioscience, Bundoora, VIC, Australia	Several plant commodities	PCR, HTS, LAMP
	Laura Boykin	The University of Western Australia, ARC Centre of Excellence in Plant Energy Biology and School of Chemistry and Biochemistry, Crawley, Perth 6009, Western Australia, Australia	Cassava, Papaya	total RNA
	Monica Kehoe	Crop Protection Branch, Department of Agriculture and Food, Western Australia, Bentley Delivery Centre, Perth, 6983, Western Australia, Australia	Cassava	total RNA
	Brenda Coutts	Department of Agriculture and Food Western Australia, Crop Protection Branch, Bentley Delivery Centre, Perth, Australia	Zucchini	total RNA
	Steve Wylie	Plant Biotechnology Research Group - Virology, Western Australian State Agricultural Biotechnology Centre, Murdoch University, 90 South Street, Murdoch, 6150, Australia	Ornamental flowers	total RNA
	Nerida Donovan	Elizabeth Macarthur Agricultural Institute, NSW Department of Primary Industries, Menangle, NSW, 2568, Australia	Citrus	total RNA
	Nitin Mantri	School of Science, The Pangenomics Group, RMIT University, Melbourne, VIC, 3083, Australia	Cheakpea	total RNA
	Ralf Dietzgen	The University of Queensland, QAAFI	Maize, capsicum	viral RNA
	Andrew Geering	The University of Queensland, QAAFI	Avocado	total RNA, sRNA
	Anthony James	Centre for Tropical Crops and Biocommodities, Queensland University of Technology	tropical crops	Rolling cycle amplification
	Kathy Parmenter	Queensland Department of Agriculture, Fisheries and Forestry	tropical fruit crops	Molecular diagnostics

	Denis Persley	Queensland Department of Agriculture, Fisheries and Forestry	Capsicum	total RNA
	Nicole Thompson	Sugar Research Australia	Sugar	total RNA
Belgium	Kris De Jonghe	Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium	Prunus and solanaceae	total RNA, enrichment, and pooling strategies
	Sebastien Massart	Integrated and Urban Plant Pathology Laboratory, TERRA-Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés, 2, 5030 Gembloux, Belgium	Fruit trees	sRNA, total RNA
	Brostaux Y	Applied Statistics, Computer Science and Mathematics (SIMa), Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés, 2, 5030 Gembloux, Belgium	Grasses	sRNA
	Haegeman A	Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium	Solanum, cassava	sRNA
Canada	Mike Rott	Centre for Plant Health-Sidney Laboratory, Canadian Food Inspection Agency, 8801 East Saanich Road, Sidney, BC, V8L 1H3, Canada	Fruit trees, strawberries	sRNA, total RNA, dsRNA, ribominus RNA
	James Delano	Centre for Plant Health-Sidney Laboratory, Canadian Food Inspection Agency, 8801 East Saanich Road, Sidney, BC, V8L 1H3, Canada	Fruit trees	total RNA, dsRNA
China	Meiguang Lu	State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China	Fruit trees	sRNA
Czech Republic	Koloniuk I	Biology Centre CAS, Branisovska 31, 370 05 Ceske Budejovice, Czech Republic; seventh author: Crop Research Institute, Drnovska 507, 161 06 Prague, Czech Republic	Strawberry, fruit trees, insect-vector	sRNA, dsRNA, total RNA, poly-A enriched RNA
	Safarova D	Department of Cell Biology and Genetics, Faculty of Science, Palacky University in Olomouc, Slechtitelu 27, 783 71 Olomouc, Czech Republic	Fruit trees	sRNA
	Petr Komínek	Division of Crop Protection and Plant Health, Crop Research Institute, Drnovská 507, 16106 Prague 6, Ruzyně, Czech Republic	Grapevine	sRNA
Finland	Vainio EJ	Natural Resources Institute Finland, Latokartanonkaari 9, 00790 Helsinki, Finland	Trees, fungi	sRNA
France	Jean-Philippe Renvoise, Jean-Emmanuel Gerbault	French Agency for Food, Environment and Occupational Health & Safety, France	Fruit trees	total RNA
	Pooggin MM	Institut national de la recherche agronomique (INRA), Unité Mixte de Recherche-Biologie et Génétique des Interactions Plante-Parasite (UMR BGPI), 34398 Montpellier, France	Mungbean, Arabidopsis, Tomato	sRNA
	Thierry Candresse	UMR 1332 BFP, INRA, Université de Bordeaux, CS20032, 33882 Villenave d'Ornon Cedex, France	Fruit trees, grapevines	sRNA, total RNA
	Verdin E	Pathologie Végétale, INRA, F-84140 Montfavet, France		

Germany	Wilhelm Jelkmann	Julius Kühn-Institut, Federal Research Centre for Cultivated Plants, Institute for Plant Protection in Fruit Crops and Viticulture, D-69221 Dossenheim, Germany	Fruit trees	total RNA
Greece	Varvara I. Maliogka	Laboratory of Plant Pathology, School of Agriculture, Faculty of Agriculture, Forestry and Natural Environment, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece	Grapevines, Cucurbit	total RNA
	Lotos L	Laboratory of Plant Pathology, Faculty of Agriculture, Forestry and Natural Environment, School of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece	Fruit trees, grapevines	sRNA
Hungary	Nikoletta Czotter, Eva Varallyay, et al.	National Agricultural Research and Innovation Center, Agricultural Biotechnology Institute, Gödöllo, Hungary	Grapevines	sRNA
	Varallyay E	Diagnostic Group, Department of Genomics, Agricultural Biotechnology and Research Institute, National Agricultural Research and Innovation Center, Szent-Gyorgyi A street 4, H-2100, Gödöllő, Hungary	Fruit trees	sRNA
Israel	Sela N	Department of Plant Pathology and Weed Research, Agricultural Research Organization, The Volcani Center, 50250 Bet Dagan, Israel	Watermelon, mango,cucurbit, cannabis,insects	sRNA, total RNA
Italy	Michela Chiumenti	Institute for Plant Sustainable Protection-National Research Center, Via Amendola, 165/A, 70126 Bari, Italy	Citrus, grapevine, olive trees	sRNA, total RNA
Japan	Hironobu Yanagisawa	NARO Agricultural Research Center, 3-1-1 Kannondai, Tsukuba, Ibaraki 305-0856, Japan	Blueberries, potato, tomato	total RNA
	Takuya Uehara	Yokohama Plant Protection Station, Ministry of Agriculture, Forestry and Fisheries, 1-7 Nagamine, Tsukuba, Ibaraki 305-0052, Japan	Blueberries	total RNA
	Reiko Tomita	Iwate Biotechnology Research Center, 22-174-4 Narita, Kitakami, Iwate 024-0003, Japan	Blueberries	total RNA
	Go Atsumi	National Institute of Advanced Industrial Science and Technology, 2-17-2-1 Tsukisamuhigashi, Toyohira-ku, Sapporo, Hokkaido 062-8517, Japan	Blueberries	total RNA
New Zealand	Lia Liefting, Zoila Perez	Ministry for Primary Industries, Wellington, New Zealand.	PEQ	total RNA, sRNA
	Simon Bullman	The New Zealand Institute for Plant Food and Research Limited, Better Border Biosecurity, Lincoln, 7608, New Zealand	grapevine	total RNA, sRNA
	Robin MacDiarmid	The New Zealand Institute for Plant & Food Research Limited, Private Bag 92169, Auckland, 1142, New Zealand	grapevine	dsRNA, sRNA
Peru	Kreuze J	Crop Protection Division, International Potato Center (CIP), Aveninda La Molina 1895, La Molina, Lima, Peru	Potato	sRNA
Slovenia	Maja Ravnikar	Ministry of Agriculture Forestry and Food, Slovenia	Potato	sRNA, ribominus RNA

	Denis Kutnjak	Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 111, 1000 Ljubljana, Slovenia	Potato	sRNA, ribominus RNA
South Africa	Maree HJ	Agricultural Research Council, Infruitec-Nietvoorbij: Institute for Deciduous Fruit, Vines and Wine, Nietvoorbij Farm R44, 7600 Stellenbosch, South Africa	Citrus, Grapevine	sRNA
Spain	Antonio Olmos	Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias (IVIA), Ctra. Moncada-Náquera km 4.5, 46113 Moncada, Valencia, Spain	Stone fruit trees, grapevines	total RNA
Switzerland	Schneeberger PHH	Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, University of Basel, Socinstrasse 57, 4051 Basel, Switzerland	Human viruses, bioinformatics	sRNA*
	Turco S, Reynard JS	University of Basel, Department of Environmental Sciences, Botany, 4056 Basel, Switzerland	Solanum, cassava	sRNA
The Netherlands	Robert Vreeburg, Inge van Duivenbode	The Dutch General Inspection Service	Potato	PCR, (HTS)
	Westenberg M	Dutch National Plant Protection Organization (NPPO-NL), P.O. Box. 9102, 6700 HC, Wageningen, The Netherlands	Strawberry, capsicum, insects	sRNA
UK	Richard O'Hanlon	Agri-Food and Biosciences Institute, United Kingdom	Potato	PCR, (HTS)
	Rachel Glover	Fera Science Ltd., National Agri-Food Innovation Campus, Sand Hutton, York, YO41 1LZ, United Kingdom	Sweet potato, potato, oca, insect-vector	sRNA, total RNA
USA	Ioannis E. Tzanetakis	University of Arkansas, USA	Berries, soybean	sRNA, total RNA
	Bob Martin	United States Department of Agriculture–Agricultural Research Service, Horticultural Crops Research Unit, Corvallis, OR 97331, USA	Blueberries	total RNA
	Schilder A	Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, United States	Blueberries	total RNA
	Al Rwahnih M	Department of Plant Pathology, University of California, Davis 95616, USA	Fruit trees, grapevines, rasberry	total RNA
	Diego F.Quito-Avila	Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA	Rasberry	total RNA
	Scott Harper	Clean Plant Center Northwest WSU-IAREC, Prosser 99350	Fruit trees	total RNA

*List of plants species where groups have published over the last few years.

Table 1.2: Initial list of groups in Australia and globally with initiatives in plant bacterial diagnosis using NGS methods

Country	Contact(s)	Organisation	Species*	Platform	NGS methods
Australia	Benjamin Schwessinger	Australian National University, Canberra	Various plant pathogens	MinION	Total nucleic acid
	Brendan Rodoni Fiona Constable	Agriculture Victoria Research, Agriculture Victoria	Various plant pathogens	Illumina	Total nucleic acid
Canada	Bryan Cassone	Brandon University, Brandon	Various plant pathogens	Illumina	Total RNA
Czech Republic	Petr Baldrian	Academy of Sciences of the Czech Republic, Prague	Phytoplasma	Illumina	Amplicon - 16S
Denmark	Mogens Nicolaisen	Aarhus Universitet, Aarhus	Phytoplasma	Illumina	Amplicon – *tuf* gene
Germany	Olga Makarova	Freie Universität Berlin, Berlin	Phytoplasma	Illumina	Amplicon – *tuf* gene
Israel	Shulamit Manulis-Sasson	Agricultural Research Organization of Israel	Various plant pathogens	Minion	Total nucleic acid
New Zealand	Grant Smith	Plant & Food Research, New Zealand	Liberibacter	Illumina	Enrichment-ultracentrifugation
	Gavin Lear	University of Auckland, School of Biological Sciences, Auckland	Various plant pathogens	Illumina	Metabarcoding
	Andrew Pitman	Better Border Biosecurity, Auckland	Various plant pathogens	Illumina	Multilocus sequence analysis
	Simon Bulman	Plant & Food Research, Auckland	Various plant pathogens	Illumina	Metabarcoding
Pakistan	Muhammad Sajjad Mirza	National Institute for Biotechnology and Genetic Engineering Pakistan	Various rhizobium	Illumina	Amplicon - 16S
Netherlands	Theo van der Lee	Biointeractions & Plant Health, Wageningen Plant Research	Various plant pathogens	Illumina	Total nucleic acid
Serbia	Jelena Jovic	Department of Plant Pests, Institute for Plant Protection and Environment	Phytoplasma	Illumina	Amplicon - 16S
United States	Jonas King	Mississippi State University, Starkville	Various plant pathogens	MinION	Total nucleic acid
	William Schneider	Oklahoma State University - Stillwater	Various plant pathogens	Illumina	Metagenomics

Table 1.3: Initial list of groups globally with expertise in human bacterial diagnosis using NGS methods

Country	Contact(s)	Organisation	Platform	NGS methods
Australia	Arthur Georges	University of Canberra, Canberra	Illumina	Genotyping
	Susan Ballard	Microbiological Diagnostic Unit Public Health Laboratory, University of Melbourne	Illumina	Whole genome sequencing
France	Pierre Peyret	Universite d' Auvergne Clermont-FD 1	Illumina	Enrichment - probes
Japan	Satomi Mitsuhashi	Yokohama City University School of Medicine, Yokohama	MinION & Illumina	16S and Total DNA
	Tadashi Imanishi	Tokai University, Hiratsuka	MinION	16S
	Kiichi Hirota	Kansai Medical University, Moriguchi	MinION	16S
United States	Orchid Allicock	Columbia University Irving Medical Center, New York	Illumina	Enrichment - Probes
	David Alland	Rutgers New Jersey Medical School, Newark, United States	Illumina	Amplicon-16S

Appendix 2 – Collaborations with national and international groups conducting NGS diagnostics on viruses and bacteria

We have established collaborations with both national and international groups working on NGS-based bacterial diagnosis approaches.

National

- Dr. Nerida Donovan – New South Wales Department of Primary Industries
- Dr. Toni Chapman – New South Wales Department of Primary Industries
- Dr. Cherie Gambley & Dr. Rebecca Roach – Queensland Department of Agriculture and Fisheries
- Dr. Monica Kehoe - Western Australia Department of Primary Industries & Regional Development

International

- Dr. Jeremy Thompson – New Zealand Ministry of Primary Industries
- Dr. Thien Ho – Driscoll, USA
- Dr. Robert Taylor – Ministry for Primary Industries New Zealand
- Dr. Grant Smith & Dr. Rebekah Frampton – New Zealand Institute for Plant & Food Research Limited
- Dr. George Sundin – Michigan State University
- Prof. Jim Stack – Director Great Plains Diagnostic Network, Kansas State University
- Dr. Mohammad Arif - C-Maiki Institute (Center for Microbiome Analysis through Island Knowledge and Investigation), University of Hawaii

Dr. Chih-Horng Kuo - Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan

Appendix 3 - Key plant pathogenic bacteria identified and prioritised

Key plant pathogenic bacteria identified and prioritized using references obtained from the National Priority Plant Pests (NPPP) (2019) list; Plant Health Australia and the Victorian post-entry quarantine target pathogens; The Emergency Plant Pest Deed (EPPRD).

Pathogen	NPPP Ranking	EPPRD Category
Xylella fastidiosa	1	2
Candidatus Liberibacter asiaticus (and other strains) complex	6	2
Candidatus Liberibacter solanacearum	13	2
Xanthomonas citri subsp. *citri*	16	2
Erwinia amylovora	18	2
Phytoplasmas 16SrI group	30	
Candidatus Phytoplasma asteris	42	-
Candidatus Phytoplasma novoguineense		
Ralstonia solanacearum phylotype IIB (banana infecting strains)	-	2
Ralstonia syzygii subsp. *celebesensis*	-	2
Candidatus Phytoplasma pruni	-	3
Candidatus Phytoplasma prunorum	-	3
Clavibacter michiganensis subsp. *Sepedonicus*	-	3
Sugarcane white leaf phytoplasma	-	3
Xanthomonas citri subsp. *Malvacearum*	-	3
Xanthomonas fragariae	-	3
Acidovorax avenae subsp. *citrulli*	-	-
Agrobacterium rubri	-	-
Apple proliferation phytoplasma	-	-
Apple yellows phytoplasma	-	-
Argentinian alfalfa witches' broom phytoplasma	-	-
Aster yellows phytoplasma	-	-
Bermuda grass grassy shoot phytoplasma	-	-
Bermuda grass spiroplasma	-	-
Bermuda grass yellow leaf phytoplasma	-	-
Brachiaria white leaf phytoplasma	-	-
Brenneria rubrifaciens	-	-
Burkholderia glumae	-	-
Candidatus Liberibacter africanus	-	-
Candidatus Liberibacter americanus	-	-
Candidatus Liberibacter asteris	-	-
Candidatus Liberibacter crescens BT-1T gen. Nov	-	-
Candidatus Liberibacter psyllaurous	-	-
Candidatus Phytoplasma americanum	-	-
Candidatus Phytoplasma aurantifolia	-	-
Candidatus Phytoplasma aurantifolia	-	-
Candidatus phytoplasma caricae	-	-
Candidatus Phytoplasma cynodontis	-	-
Candidatus Phytoplasma fragariae	-	-
Candidatus Phytoplasma fraxini	-	-
Candidatus Phytoplasma mali	-	-
Candidatus Phytoplasma phoenicium	-	-
Candidatus Phytoplasma pyri	-	-
Candidatus Phytoplasma solani	-	-
Candidatus Phytoplasma sudamericanum	-	-
Candidatus Phytoplasma trifolii	-	-
Candidatus Phytoplasma ulmi	-	-
Candidatus Phytoplasma vitis	-	-
Candidatus Phytoplasma ziziphi	-	-
Chestnut witches' broom phytoplasma	-	-
Citrus rubbery wood phytoplasma	-	-
Clavibacter iranicus	-	-
Clavibacter michiganensis ssp. *nebraskensis*	-	-
Clavibacter tritici	-	-
Clover yellow edge group	-	-
Clover Yellow Edge Phytoplasma	-	-
Corynebacterium fascians	-	-
Corynebacterium michiganense pv. *tritici*	-	-
Dickeya chrysanthemi	-	-
Dickeya sp syn *Erwinia chrysanthemi* (distinct pathovar)	-	-
Erwinia cypripedii	-	-
Erwinia herbicola	-	-

Erwinia mallotivora	-	-
Erwinia papayae	-	-
Erwinia pyrifoliae	-	-
Erwinia rhapontici	-	-
Erwinia tracheiphila	-	-
European stone fruit yellows	-	-
Grapevine flavescence doree phytoplasma	-	-
Lime witches' broom phytoplasma	-	-
Mexican periwinkle virescence group	-	-
Mulberry dwarf phytoplasma	-	-
Mulberry yellow witches' broom	-	-
Pantoea stewartii subsp. Stewartii	-	-
Peach yellow leafroll phytoplasma	-	-
Pear decline phytoplasma	-	-
Pectobacterium spp	-	-
Phytoplasma 16S – 1B group	-	-
Phytoplasma 16S – VA group	-	-
Phytoplasma 16S – XIIA group	-	-
Phytoplasma asteri HLB-like symptoms	-	-
Polygala phyllody phytoplasma	-	-
Pseudomonas amygaldi	-	-
Pseudomonas avellanae	-	-
Pseudomonas batatas	-	-
Pseudomonas savastanoi pv. *fraxini*	-	-
Pseudomonas syringae pv. actinidiae Psa	-	-
Pseudomonas syringae pv. *avii*	-	-
Pseudomonas syringae pv. *coryli*	-	-
Pseudomonas syringae pv. *lapsa*	-	-
Pseudomonas syringae pv. *morsprunorum*	-	-
Pseudomonas syringae pv. *papulans*	-	-
Pseudomonas syringae pv. *persicae*	-	-
Pseudomonas syringae pv. *syringae* – exotic strain	-	-
Pseudomonas syringae pv. *syringae* Mango strain	-	-
Pseudomonas syringae pv. *tagetis*	-	-
Psuedomonas amygdali	-	-
Ralstonia solanacearum, race 3 (exotic strains)	-	-
Ralstonia solanacearum, race 4 (exotic strains)	-	-
Raspberry witches' broom phytoplasma	-	-
Rathayibacter tritici	-	-
Red currant full blossom phytoplasma	-	-
Rickettsia associated disease	-	-
Rubus stunt phytoplasma	-	-
Spiroplasma citri	-	-
Spiroplasma kunkelii	-	-
Stolbur group phytoplasma	-	-
Strawberry multicipita phytoplasma	-	-
Streptomyces ipomoea	-	-
Streptomyces spp	-	-
Sugarcane grassy shoot phytoplasma	-	-
Sweet potato little leaf phytoplasma	-	-
Unknown (suspected phytoplasma)	-	-
Walnut witches' broom	-	-
Xanthomonas albilineans (exotic strains)	-	-
Xanthomonas arboricola pv. *fragariae*	-	-
Xanthomonas axonopodis pv. *allii*	-	-
Xanthomonas axonopodis pv. *cassavae*	-	-
Xanthomonas axonopodis pv. *citrumelo*	-	-
Xanthomonas axonopodis pv. *dieffenbachiae*	-	-
Xanthomonas axonopodis pv. *malvacearum*	-	-
Xanthomonas axonopodis pv. *manihotis*	-	-
Xanthomonas axonopodis pv. *passiflorae*	-	-
Xanthomonas axonopodis subsp. *citri*	-	-
Xanthomonas batatae	-	-
Xanthomonas campestris – avocado strain	-	-
Xanthomonas campestris pv. *musacearum*	-	-
Xanthomonas campestris pv. *translucens*	-	-
Xanthomonas campestris pv. *viticola*	-	-
Xanthomonas cannabis	-	-
Xanthomonas translucens pv. *translucens*	-	-
Xanthomonas translucens pv. *undulosa*	-	-
Xanthomonas vasicola pv. *musacearum*	-	-
Xylophilus ampelinus	-	-

Appendix 4 – Current best practices and standards for the diagnosis of viruses and viroids using NGS.

4.1 Abstract

High throughput sequencing (HTS), also known as next generation sequencing (NGS), technologies are becoming more commonly used in different fields of virus research including phytosanitary diagnosis. Availability of distinct sequencing platforms and sample preparation approaches contributed to a revolution in detection and discovery of plant viruses and viroids. Adoption of end-to-end standard operating procedures (SOPs) and best practices are key to reliably detect pests in distinct laboratory settings. In this review, we revise the current state of best practices of the use of small RNA (sRNA) sequencing for the diagnosis of viruses and viroids.

4.2 Introduction

Plant viruses and viroids are important plant pathogens, causing economic losses by reducing crop quality and quantity all over the world [1]. Thus, their reliable detection is of a crucial importance for plant protection. Classical methods in plant virus diagnostics can be roughly divided into specific (serological/molecular tests) and non-specific (indicator test plants, electron microscopy) approaches. Specific methods are usually targeted to one or a few viral species and require *a priori* knowledge of the pathogens being tested, whilst non-specific approaches do not require specific knowledge of the pathogens, however, they mostly only classify viruses at a genus level based on the shared physical/biological.

Discovery of new viruses/viroids and new hosts has rapidly increased rapidly after the introduction of HTS technologies. The coupling of HTS with downstream bioinformatics analyses can potentially, without any prior knowledge on a sample infectious status, not only identify known viruses at the species/strain level but also detect yet uncharacterised organisms [2]. Metagenomics refers to HTS data analysis of environmental samples that contain an unknown mixture of typically unculturable organisms. It was first employed for plant virus detection in 2009 [3-5]. Since then, different sample preparation methods have been developed, relying on different nucleic acid inputs.

HTS sequencing of sRNAs and ribominus RNA are the most used input RNAs to detect viruses with different genome types and replication strategies and could be relatively easily integrated in workflows of diagnostic labs as well as downstream bioinformatics analyses. Sequencing and assembly of viral sRNA has been successfully used for detection and identification of several plant viruses and viroids and their complete genome assembly [5]. It has been speculated that this approach could be problematic if used to detect viruses that either do not trigger silencing responses or that express silencing suppressors [6]. On the other hand, the approach is very generic, using the same protocol of sample preparation for many different plant species and does not require high quality of RNA input [7]. More recently, Barrero and colleagues leveraged on the innate plant immune response to detect and reliably assemble viral genomes in a range of plant commodities at post-entry quarantine (PEQ) facilities (**Figure 4.1**) [8].

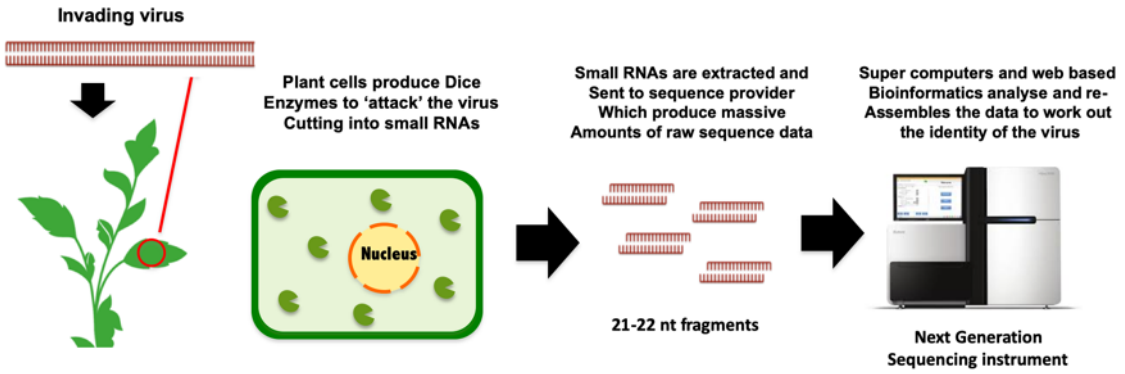


Figure 4.1. Overview of the next generation sRNA sequencing method. Plants harbour intracellular enzymes that recognise and cleave viruses into sRNA pieces which are then sequenced, and the sequence reassembled to determine the identity of viruses present in the plant.

There are several advantages in using either small RNA or ribominus RNA approaches **(Table 4.1)**. Total RNA and ribominus approaches are likely to detect low titre novel viruses and those that potentially do not trigger the RNAi response in the host, while small RNA-Seq approach relies on several internal checkpoint increasing the confidence in the diagnosis of viruses and viroids. Preliminary side by side comparison of PEQ versus HTS approaches showed that ribominus RNA approaches us impacted by index hopping, with the detection of artefact reads derived from other multiplexed samples in the same HTS run. This trend has been reported by other researchers and diagnostic laboratories. On the other hand, sRNA libraries do not seem to suffer from this issue to the same extent. ***Unless a laboratory has validated a strategy to detect reads resulting from index hopping and reliably separate them from true positives present at low titres within a sample, we recommend using sRNA libraries over ribominus RNA libraries.***

Table 4.1: Advantages and challenges of HTS approaches.

	small RNA	**ribominus RNA**
Pros	Lower sequencing depth	Detects all viruses, esp new viruses present at very low titre
	Relies on host immune response for analysis and validation	Detects viruses not triggering RNAi
	Less cost per sample	Assembles full genome
Cons	Misses viruses not triggering RNAi	Requires greater sequencing depth
	Reported to miss some DNA viruses	Higher cost per sample
	Low titre viruses may only assemble into short genomic contigs	

HTS technologies open new possibilities and opportunities in routine diagnostics such as understanding the status of a pest in a region through surveillance programmes, certification of nuclear stock and plant propagation material, post-entry quarantine testing to prevent the introduction of pests into a country or area, and monitoring of imported commodities for new potential risks [9]. The availability of cost-effective platforms and open access bioinformatic tools to analyse sequencing data have driven an increased interest in applying HTS technologies for routine diagnostics. To date, HTS has only been used in plant health diagnostics in specific quarantine situations, such as during the investigation of a plant health outbreak or re-

emerging disease [10, 11]. However, the Commission on Phytosanitary Measures governing body of the International Plant Protection Convention (IPPC) adopted a recommendation on preparing the use of HTS as a diagnostic tool for phytosanitary purposes. This recommendation highlighted the need for robust test design, test validation and quality assurance. One of the main challenges for using HTS as a routine diagnostic test in plant health laboratories is the implementation of guidelines that are internationally recognised for the selection, development, validation, and routine use of HTS-based tests.

This document is intended to be a guideline for any laboratory intending to use sRNA sequencing for the detection and identification of viruses and viroids in plant material.

Defining the aim of the HTS test

Before implementing an HTS test on routine samples, it is important for the laboratory to clearly define what is its main goal, so that the type of test can be selected accordingly, and an adequate experimental design and risk analysis process can be put in place. For instance, the laboratory needs to decide whether it is targeting all currently known viruses and viroids or limiting detection to a specific virus list. This will determine whether public or in-house curated reference databases will be used. It is also important to establish which range of plant commodities, as well as which plant material, are meant to be tested.

HTS validation test

Application of RNA-Seq to detect viruses and viroids in plant samples relies on adequate experimental design that takes into consideration the main goal of the HTS experiment. After setting up an experimental design, a newly developed test should be validated. A test described in an official standard or in a scientific publication with known performance characteristics should be verified. After its validation or verification, the HTS test can be used in routine diagnostic. Performance should be monitored using appropriate controls during its routine use. If changes are to be implemented to the test, it should be verified or validated again to make sure that the changes did not affect performance.

Controls

Laboratories might consider incorporating a negative control for their HTS test to monitor for contaminants. Detection of target(s) in the negative controls would indicate that contamination has occurred during the HTS test. However, a very low amount of contamination by target sequences will often be present in the data generated from negative controls. These traces of contamination might not be detectable by conventional or real-time PCR. This phenomenon will lead to the risk of overestimating the contamination as compared to a sample or control in which trace contaminant DNA is extremely low in sample DNA. For this reason, the use of positive controls containing a DNA quantity similar to the analysed samples could allow a better estimation of contamination in this specific case.

Positive controls are processed alongside the samples to be analysed to assess whether the nucleic acid extraction step was adequate to achieve sufficient sequencing depth for downstream virus detection. Another option is to mix control plant material with tissue from infected plants prior to performing extraction [12]. A commonly used control is Phaseolus vulgaris (cultivar: Black Turtle) infected with the cryptic viruses Phaseolus vulgaris alphaendornavirus 1 and 2 (PvEV1 and PvEV2; JKI ID 31403). An ideal positive control will contain a small but representative fraction of the possible targets the HTS test could detect. It is recommended to use positive controls for which at least some target concentrations are close to the limit of detection. Low level targets are well suited to check the analytical sensitivity of the sequencing run and their low concentration limits the risk of contaminating other samples. Positive controls can also be used to monitor contamination. The detection of an unexpected target in the positive control (in addition to the expected target(s) may be a signal of contamination from another sample.

Alternatively, each sample is added (‘spiked’) with synthetic nucleic acids or a known target not expected to be found in the samples to be analysed. External RNA Control Consortium (ERCC) spike-ins mimic eukaryotic mRNAs and can be added at the same concentration to each sample prior to library construction [13]. Similarly, for normalisation of high-throughput small RNA sequencing (sRNA-Seq) data spike-in oligonucleotides have been developed and added at varying molar concentrations.

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

The following controls are incorporated in our current protocol:

- RNA extraction step (synthetic control)
- cDNA library amplification: positive control (Human XpressRef Universal Total RNA Cat no 338112) and negative control (water)
- Sequencing step (PhiX control)
- Bioinformatics step (control sample run by default by the pipeline)

4.3 Protocol

4.3.1 Sample collection

The laboratory should have a sampling procedure describing the type of material (e.g., expected tissue for plants), the amount of material needed, the number of samples and when relevant, the season of sampling, and the requirements for sampling symptomatic and/or asymptomatic material. The procedure should also define how to deal with samples that do not meet these criteria.

The type of samples (e.g., different plant parts) and the season of sampling can affect the results of the HTS test (e.g., organisms not detected). Virus and viroids titres in host plants vary throughout the year being more abundant during spring and summer seasons, respectively. Hence, ***it is critical to consider the time of the year to collect samples. This may require coordination with plant importers to enable having sufficient plant material for quarantine testing during mid-spring to early summer to increase the efficacy of detecting viruses in imported plant material if present.***

The necessary starting amount of an RNA sample varies between kits and sequencing platforms, and the amount of available RNA is one of the limiting factors for the use of specific HTS test. At least 100 ng of total RNA is typically required for HTS library construction and as much as a few micrograms for RNA-Seq applications, this latter requirement sometimes limiting testing to only high-quality samples. Processing of a broad range of plant commodities at PEQ facilities represents a challenge to consistently achieve high RNA yields. ***Typically, 20-100 mg of input plant material is required in order each sample to yield up to several micrograms of total RNA***. Thus, commercial kits for HTS library construction with the capacity to use low input nucleic acids are critical. However, further development is required to avoid amplifying nucleic acids via polymerase chain reaction (PCR) prior to HTS sequencing. ***We currently recommend using the Maxwell Plant RNA kit for the extraction of total RNA that includes a small RNA fraction***.

A) Materials

- Glove box
- Marker pen
- Surgical scalpel blades
- Resealable plastic bags
- Ice box

B) Methods

Step 1 - Preparations before sampling

- Make a list of plants that you want to sample.
- Ensure all plants are in active growth stage at the time of sampling.
- Write down or print the index numbers and their location in the glasshouse.
- Label the resealable bags with the index numbers, plant species and date of sampling.

- Put adequate gloves, blades, marker pen, labelled bags, and the list of plants you prepared into an ice box.
- Once you are at the glasshouse premises follow the standard protocols before entering each room.
- Locate your plant of interest.

Step 2 - Sampling

- Always sample from the index plant and sample from symptomatic areas of the plant if symptoms are visible.
- Wear a fresh pair of gloves.
- Except in the case of below exceptions, for each plant, pluck three leaves (both blade and petiole) from three parts of the plant from different maturity stages (top, bottom, middle and alternate sides).
    - Exceptions
        - Sessile leaves – pluck leaf blade
        - Monocots and other plants with leaves less than 10 – pluck 2 leaves from different areas
        - For plants with light leaves – sample adequately to make sure you have enough material (>50mg) for nucleic acid extractions.
- Use a scalpel blade where necessary and dispose the blades in between plants (See waste management below)
- Once a plant has been adequately sampled, put the leaf material into the respectively labelled resealable bag and keep in the ice box until you finish sampling.
- Dispose the used gloves properly (see wate management below).
- Use a fresh pair of gloves each time you sample a new plant.
- Once you finish sampling, take the ice box to the extraction room and store the labelled resealable bags with samples at 4 ºC.

4.3.2 RNA extraction

Special precautions and more stringent working practices are required when working with RNA as these molecules are more labile than DNA. To ensure high-quality RNA is isolated, samples need to be processed immediately after harvest. If immediate RNA isolation is not feasible, samples can be stored in an intermediary solution to maintain RNA integrity and allow storage. Commonly used solutions are RNAlater (Thermo Fisher Scientific and Qiagen) and RNAstable (Sigma-Aldrich).

A large range of in-house protocols and commercial kits are currently available for RNA extraction, and the selection of a suitable method depends on the type of nucleic acid to be detected by the HTS test and the type of matrix from which the nucleic acids are extracted (e.g. plant parts: seed, pollen, leaf, stem, purified cultures, soil, water) [12, 14-16]. Some plant hosts are known to be difficult to extract (e.g., grapevine, fruit trees and pollen).

Different extraction methods can differ in their efficiency and can have specific advantages and disadvantages. For example, subtractive hybridisation of the host plant nucleic acids, using tomato (*Solanum lycopersicum*) and *Pepino mosaic virus* (PepMV, RNA +, *Potexvirus, Alphaflexiviridae*) as a model system, resulted in three times more PepMV sequences in subtracted sample [3]. However, it is a time-consuming procedure requiring a healthy plant of the same species, which makes the method unsuitable for high-throughput diagnostics. Another approach targets the sequencing of dsRNA to detect RNA + and RNA +/− viruses, however this procedure typically fails to detect RNA− and DNA viruses [6] but see [17] for an example of a new geminivirus (DNA +) that was found using dsRNA sequencing. RNA isolated from purified viral particles has been successfully used for sequencing plant viruses [6]. **During validation of HTS test in routine testing, several extraction protocols can be evaluated using control samples to ensure they detect the targeted viruses and viroids.**

RNA isolation and extraction protocols can be either manual (e.g., TRIzol–Thermo Fisher Scientific) or automated (e.g., Maxwell Plant RNA Kit—Promega). All these methods broadly consist of three steps, namely, (i) sample solubilisation in the presence of detergent and chaotropic agents, (ii) sample homogenisation for complete cell disruption, and (iii) RNA recovery from the lysate with organic or solid-phase extraction. It is also necessary to have a final RNA free of genomic DNA (gDNA) contaminants. Depending on the downstream application and analysis, gDNA contamination can lead to false-positive signals. Hence, it is recommended to ensure RNA extracts do not contain genomic DNA contamination.

The nucleic acid extraction protocol should deliver nucleic acids of appropriate quality and quantity for HTS testing. The quality (in terms of purity and integrity) and quantity (e.g., ng/µl) of nucleic acids is important as they can affect the results of the HTS test. The quality of extracted nucleic acids can be checked by several methods and minimal thresholds should be defined (e.g., minimum, average, maximum fragment length, minimal acceptable purity indicators) and respected.

RNA quality can be assessed by gel electrophoresis (agarose or polyacrylamide) or through Agilent Bioanalyzer or Agilent TapeStation. RNA quantity can be assessed using spectrophotometer (e.g., Nanodrop), fluorometer (e.g., Qubit), Agilent Bioanalyzer or Agilent TapeStation. No single RNA quantification and quality control method are ideal, and it is necessary to know the limits of each method. **We recommend Nandrop or TapeStation since it measures the RNA integrity and level of degradation by the RNA Integrity Number (RIN) score that allows sample quality comparison by a scale ranging from 1 (most degraded) to 10 (most intact) [18]. There is no consensus about the RIN cut-off for sample inclusion or exclusion in a study, but RIN ≥ 6 are commonly acceptable.** Plants RNA extractions typically yield low RIN scores. Johnson and colleagues [19] recently collected RNA from 629 plant species for HTS application and found that RIN scores are poor predictors of assembly quality, measured by the number of scaffolds (assembled sequences using HTS raw data recovered) because Illumina sequencing is robust to the variation in RNA quality [19].

Note: RNA extraction allows for the removal of inhibitors that can negatively impact the test result. The amount of recovered RNA can be influenced by the plant metabolic state, its stage of growth, type of sample collected (eg. pollen), growth conditions and the presence of plant contaminants such as polyphenolics and polysaccharides. These factors should be considered when low RNA yield is recovered. **The addition of Plant RNA isolation aids is recommended to remove such plant contaminants.**

A) Materials

- Labelled resealable bags with plant material sampled.
- Weighing boats
- Gloves
- surgical scalpel blades
- porcelain tile
- electronic balance
- 2 ml Eppendorf tubes, marker pen
- Stainless steel beads
- Qiagen Tissuelyser LT
- Promega Maxwell RSC Simply RNA Tissue Kit
- Maxwell instrument
- Fridge and freezer
- Laminar flow
- Micropipettes
- Centrifuge
- Vortex
- Kim wipes
- 80% ethanol
- Nuclease free water
- PVP
- DTT

B) Methodology

Step1- Prepare the extraction worksheet

- Create a new extraction worksheet and enter the details of the samples you want to extract RNA from.
- Print out the extraction work sheet and paste in the respective lab book.

Step 2 - Clean the workbench and prepare 2 ml Eppendorf tubes for grinding the samples

- Clean your work bench with 80% ethanol.
- Take one of 2 ml round bottom Eppendorf tube per sample and place in a tube rack.
- Label the lid and the side of with the Eppendorf tubes with the corresponding plant index number in the extraction worksheet using a marker pen.
- Place **7mm stainless steel beads** into the Eppendorf tubes.
- Use one bead per tube for soft plant tissues and 2 beads per tube for petioles and other hard plant tissues.

Step 3 - Weigh 50mg of sample tissue into the tubes and freeze.

- Turn on the laminar flow.
- Wipe the inside of the laminar flow and the balance with 80% ethanol.
- Locate the samples in labelled resealable bags in the extraction room.
- If the samples were temporarily stored in the fridge take them out and move to the laminar flow.
- Wear a fresh pair of gloves.
- Cover a white porcelain tile with a wipe and use as your cutting board.
- Turn on the balance. Take a sterile weighing boat and place it on the balance. Press Tare.
- Dissect a combination of leaf blade and midrib from all leaves present in the bag using a sterile scalpel blade to weigh a maximum of 50 mg per sample.
    - Exceptions – For Rubus spp. use only the petiole and make sure you avoid the leaf blade.
- Cut the weighed tissue into smaller pieces that can fit in the lid of Eppendorf tube.
- Transfer the cut pieces of leaf tissue into the respectively labelled Eppendorf tube with bead/s using the static of the scalpel blade.
- Discard gloves, weighing boat, scalpel and the wipe covering the tile appropriately in between samples.
- Wipe the working space and balance in between samples.
- Once the weighing is completed, transfer the Eppendorf tubes with samples and beads into the Tissuelyser rotor, and then place sample and rotor into the -75°C freezer for 30 minutes.

Step 4 – Prepare Maxwell reagents and elution tubes

- Use either a pre opened **Promega Maxwell RSC Simply RNA Tissue Kit** or open a new kit depending on how many cartridges you need.
- If you are opening a new kit record the batch No. and the date you opened the kit in the logbook.
- **Prepare homogenisation solution**
    - Check whether there is a bottle of prepared homogenisation solution in the fridge.
    - Check the date of preparation and if the homogenisation solution is expired (>1 month old), discard the bottle and prepare a fresh batch.

    - Add 2.5% PVP (0.75g dry PVP), to **Promega Maxwell RSC Simply RNA Tissue Kit** 30mL Homogenization solution Bottle
    - Mix vigorously
    - Add 600uL of 1-thioglycerol to the Homogenisation solution and mix.
    - Write down the date in the bottle if you made a fresh batch.
- **Prepare DNAase 1**
    - DNase 1 is dyed blue and stored at -20°C. Thaw as required.
    - You will need 5µl DNase 1 per sample. Check whether you have enough.
    - To prepare a fresh DNase 1 batch when you open a kit, follow the instructions below.
    - Add 275µl nuclease free water to **Promega Maxwell RSC Simply RNA Tissue Kit** DNase I. Mix/Invert/turn tube to dissolve, do not vortex.
    - Add 5µl of **Promega Maxwell RSC Simply RNA Tissue Kit** blue dye.
    - Store solution at -20°C.
    - Thaw as required, and then refreeze.
- **Label 0.5 ml elution tubes**
    - Label one elution tube per sample from the **Promega Maxwell RSC Simply RNA Tissue Kit** with the sample index number on both the top and the side of the tube.

Step 5 – Prepare the frozen plant tissue for RNA extraction.

- Get the cold Rotor with the frozen tissues from the -75°C freezer and insert into the **Tissuelyser**.
- Operate Tissuelyser for 2 minutes at 50 Hz. Check that a powder has formed, if not, repeat the freezing step and lyser step.
- Working as quickly as possible, add 600µl of **Homogenisation solution preparation** (in 4°C fridge), into the Eppendorf tube with ground tissue.
- Add 300µl **lysis buffer** from **Promega Maxwell RSC Simply RNA Tissue Kit**
- Mix by vortexing.
- Spin sample tube for 5 min at 20 000g.
- If gluggy take off supernatant and spin supernatant a second time.

Step 6 – Prepare the Maxwell instrument for RNA extraction.

- Turn on the touch screen, button top right and turn on the Power at the back of the instrument.
- Press the Maxwell icon on the screen.
- Press the Maxwell icon again on the next screen, (A self-test will occur here)
- Press the start button.
- Choose the extraction program, **AS1340** for RNA extraction.
- Press proceed.
- Press proceed again.
- Then ‘ok’, the door will open.
- Take out the tray and rest it on the bench.
- Load the tray with the number of Maxwell cartridges required, one per sample, press down until the cartridges click in.
- Remove the sealing foil from each of the cartridges aseptically.
- Aseptically insert the **pipette plungers** into the **front well** of each cartridge, (the well closest to the tray holes).

- Aseptically place **elution tubes** into the holes in the **tray** for each cartridge. The lids must be open, opening away from the cartridges.
- Aliquot **100µl** of **nuclease free H_20** from the Promega Maxwell RSC Simply RNA Tissue Kit into each elution tube.
- Add to **5µl** of thawed **Blue DNase I** to the **well with the yellow reagent** in each cartridge, (well #4).
- Finally add **500µl** of **supernatant of your sample preparation** to the **back well**, (the big well), of the corresponding cartridge.
- The tray is ready, it can be placed back onto the Maxwell. Orient the tray with the elution tubes at the front.
- Do a visual scan of your tray as you put it in and check whether
    - you have H_20 in the elution tubes
    - the tubes open
    - your sample is aliquoted in the back well
    - you add blue DNase I into yellow reagent well
    - you put in the plunger
- If all is ok, Press start. (The instrument door will close; the instrument will initialise and the run will start. The progress of the run can be seen on the bar on the computer screen).

Step 7- Post-RNA extraction procedure

- Once the instrument has completed, the bar on the screen will read 100%
- Press 'Open'
- Lift the tray out of the Maxwell and place on the bench.
- Close the lids of the elution and move to a storage box labelled appropriately.
- Store the extracts
    - If you are performing quantification on the same day, store the box in the fridge (4°C) temporally.
    - If the quantification will be done on a later day store the box in the -75°C freezer in the extraction room.
    - Note: if extracts contain too many beads, your extract will look murky, a strong magnet or centrifuging the extract, will help. Separate the supernatant from the beads and store
- Unclick each Maxwell cartridge out of the tray gently and place into a biohazard waste bin.
- Wipe the tray with 80% ethanol using a Kimwipe.
- Place tray back into Maxwell.
- Exit the Extraction program by arrowing, on the screen, back twice.
- You will reach a screen that has a Sanitise option. Press 'Sanitise'.
- Sanitation will take 1 minute.
- On completion of sanitation, exit the Maxwell program with the Arrows and return to the home screen.
- Turn off instrument computer (Top right button)
- Turn off the instrument (button at the back of the Maxwell)

Step 8 – Batching samples for third party sequencing

- Obtain quote from third-party sequencing provider
- Organise purchase order number
- Organise courier shipment of nucleid acid samples on dry ice to the sequencing provider

- Prepare samples in appropriate format for shipment on dry ice
- Submit samples into the sequencing provider's LIMS
- Enter relevant information associated with sequencing run into the local LIMS

Lift the tray out of the Maxwell a

Step 10 – Check quality and quantity of RNA

- Check quality and quantity of RNA via Nanodrop to measure concentration and purity (260/230 and 260/280 ratios). This step will soon be replaced by TapeStation at PEQ.
- Sequencing provider independently checks concentration and purity (260/230 and 260/280 ratios) using Epoch Spectrophotometer. Sample are also run on the Agilent 4200 TapeStation using the RNA ScreenTape system to check integrity.
- Review QC report summary provided by sequencing provider to confirm whether they can proceed to cDNA library preparation step. If sample fails QC check, peform a new RNA exatraction and resubmit sample. If volume recovered in tube is unexpectedly lower than expected or RNA extraction shows unexpected precipitate, do not proceed, and resubmit sample.

4.3.3 cDNA library preparation

cDNA library construction is prepared in accordance with the requirements of the selected sequencing technology to produce a sufficient amount of nucleic acids of appropriate size that are flanked with adapters (oligonucleotide sequences) required for sequencing. Several protocols are available for microRNA sequencing and are often obtainable as a kit with all the reagents included [20].

During the library preparation protocol, an **index** (also known as barcode) is added to each sample to allow **multiplexing** (i.e., the sequencing of a pool of samples). Analysis of pooled samples in a single sequencing run allows a reduction of the costs of the analysis per sample. The index is a short sequence of oligonucleotides flanked with the adapter added during the library preparation and is unique to each sample. This index is sequenced along with the target and allows each sequence to be traced back to the original sample in the pool. Indices can be added to one or both sides (dual indexing) of the fragment. Pooling requires that the amount of nucleic acid of each library in the pool be normalised to minimise the pooling bias that causes the generation of uneven numbers of sequences between samples.

The process of pooling increases the risk of misassignment of reads to a sample due to index cross contamination that can occur during library preparation and sequencing (i.e. index-hopping) or between sequencing runs (i.e. when identical indices are used in successive runs) [21]. The risk is increased when using the latest Illumina machines (e.g., NovaSeq) and when samples are sequenced at very high depth. The laboratory should be aware of the risk associated with pooling and test that the pooling strategy used does not cause contamination and thus affect the test performance by lowering its sensitivity. The pooling method depends on the desired read depth of the targets to be sequenced and should be optimised to ensure that the HTS test meets the criteria of its intended use. Sample misassignments can be reduced on the Illumina platform by using dual indexes [21] and almost abolished by using **unique dual indexes (UDIs)** which increases the bioinformatic power in catching index-hopping [24]. The sequences of sets of indices included in each run should be recorded for trace back purposes and to plan properly the succession of sequencing runs.

Unique molecular identifiers (UMIs) can also be incorporated to mitigate PCR duplicates. These are molecular barcodes used to uniquely tag each molecule in a sample library [25]. By incorporating individual barcodes on each original DNA fragment, variant alleles present in the original sample (true variants) can be distinguished from errors introduced during library preparation, target enrichment, or sequencing. Thus they provide error correction and increased accuracy during sequencing.

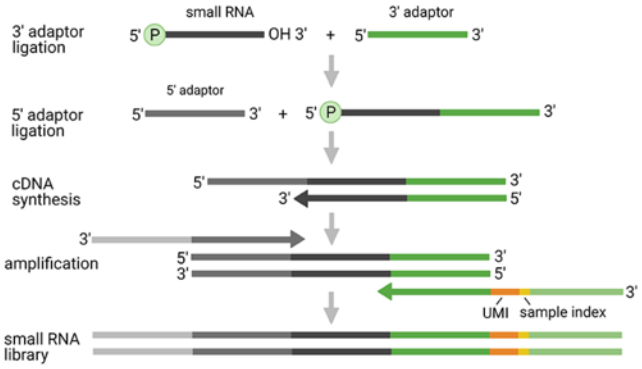


Figure 4.2: Overview of small RNA HTS library prep using a two-adaptor ligation method.

We recommend the use of miRNA library preparation that incorporates unique dual indices and currently outsource library preparation and sequencing to the Ramaciotti Centre for Genomics (Sydney). **QIAseq miRNA** enables HTS of mature miRNAs on Illumina NGS instruments. It uses a ***two-adaptor ligation-based method with unique molecular indices (UMIs) and unique dual indices*** that enable multiplexing of up to 96 samples (**Figure 4.2**). Adapters ligate sequentially to the 3' hydroxyl group and 5' phosphate group of mature miRNAs. To minimise side product formation such a primer dimer, the 5' end of the 3' adapter is pre-adenylated so that ligation can only occur between the 5' adapter end and the 3' hydroxyl group of the sample RNA. In the complementary DNA (cDNA) synthesis step, RNA is then converted to cDNA via reverse transcriptase (RT), generating a cDNA:RNA heteroduplex molecule. The RT primer contains an integrated UMI. The RT primer binds to a region of the 3' adapter and facilitates conversion of the 3'/5' ligated miRNAs into cDNA while **assigning a UMI to every miRNA molecule**. During reverse transcription, a universal sequence is also added that is recognised by the sample indexing primers during library amplification.

After reverse transcription, a cleanup of the cDNA is performed using a streamlined magnetic bead-based method. The first-strand cDNA is then used as a template to generate double-stranded cDNA. Second-strand cDNA synthesis uses an oligonucleotide that is complementary to an adapter sequence located in the 5' end of the RNA template. During library amplification, a dried **universal forward primer** from a plate is paired with 1 of 96 dried reverse primers in the same plate (cat. no. 331565) to assign each sample a **unique custom index**. Library amplification reactions occur directly in the index plate, providing a convenient HTS indexing solution. After library amplification, another cleanup of the miRNA library is performed during which size selection is carried out using a streamlined magnetic bead-based method**. The recommended starting RNA amount is 100 ng** but the amount of template for a single QIAseq miRNA sequencing reaction can range from 500 ng to as little as 1 ng of purified total RNA.

To maximise the yield and quality of sRNA sequencing, precise quantification of HTS libraries is critical. Inaccurate quantification may lead to lower throughput, lower sequences qualities, and non-uniform read depth of multiplexed samples. Several quantification methods are used to evaluate HTS libraries. These include quantitative real-time PCR, spectrophotometer (e.g., Nanodrop), fluorometer (e.g. Qubit), and Agilent Bioanalyzer. It is also recommended to verify the fragment size distribution of libraries using methods such as Agilent Bioanalyzer. Bioanalyzer electropherogram should show a narrow distribution with a peak height of the average size fragmentation value.

A) Materials

- QIAseq miRNA library kit (with UMIs)
- QIAseq miRNA 96 Index Kit IL UDI (96)
- Human XpressRef Universal Total RNA. Cat no 338112
- Quant-iT PicoGreen dsDNA assay kit
- Perkin Elmer LabChip High Sensitivity DNA chip
- Qubit dsDNA HS assay
- Agilent TapeStation High Sensitivity D1000 assay

B) Methods (performed by third-party sequencing provider)

- Follow manufacturer's recommendations and include a positive control (Human XpressRef Universal Total RNA) and a negative control (water) to check if there was any preparation failure.
- Adapters are ligated sequentially to the 3' and 5' ends of miRNAs. Subsequently, universal cDNA synthesis with UMI assignment, cDNA cleanup, library amplification and library cleanup are performed.
- We use a positive control Qiagen XpressRef Universal Total RNA and a negative control (NF water) with each library prep batch.
- We perform QC on individual libraries using the Quant-iT PicoGreen dsDNA assay kit and the Perkin Elmer LabChip High Sensitivity DNA chip.
- Libraries are normalised and pooled in equimolar ratios. A final cleanup is performed with Qiagen QMN beads.
- Final QC on the clean pool is performed using Qubit dsDNA HS assay and Agilent TapeStation High Sensitivity D1000 assay.

4.3.4 Sequencing

There are currently several commercial sequencing platforms available, including Ion Torrent (Thermo Fisher Scientific), Illumina sequencing by synthesis (Illumina), PacBio single molecule real-time (SMRT) (Pacific Biosciences), BGISEQ-500 (Complete Analytics and BGI) and nanopore sequencing (ONT) [26, 27]. Commonly used approaches are clonal amplification-based sequencing (e.g., Illumina), DNA nanoball-based sequencing (e.g., BGI) or single-molecule-based sequencing (e.g. PacBio and ONT). Illumina and BGI technologies offer low-error rates single pass data (close to 100%), while long single pass technologies have average error rates ranging from 10-15%. In virus diagnosis, Illumina and BGI technologies enable to distinguish between closely related viral strains, particularly, in asymptomatic plant material and where virus titres are low. However, long read technologies are gaining popularity to assemble complex genomes, while the portability of ONT devices makes them suitable for pest diagnosis at the farm gate. The sequencing error rate is still an issue but has been reduced over the last few years so that ONT technology is seen as having significant potential for plant virus diagnostics, although its performance at the sequencing depth it provides remains to be evaluated in detail.

Laboratories should determine the type of sequencing platform and the method of sequencing best suited for the intended use of their HTS test. One first consideration is read length and type (e.g., single, paired, mate-pair). **Short single reads are appropriate for sRNA sequencing**. Another consideration is the minimal required sequencing depth to achieve virus detection. Sequencing depth depends on the scope of the HTS analysis. Partial genomic sequences is generally sufficient for virus discovery but full genome recovery is preferred for evolutionary studies, which would require more sequencing depth. A recent study by Massart and colleagues [28] comparing sRNA inputs from 50K to 2.5 million reads concluded that at least 2.5 million reads are necessary to detect known control viruses. Barrero et al. [8] recommended using a minimum of 10 million small RNA reads for Post-Entry Quarantine testing. **Our current sequencing provider (Ramaciotti) recommends using 30M reads per sample for good quality samples and increase sequencing output up to 45M if sample is of lower quality.** Once the required number of reads per sample is established, the laboratory can work out the number of samples they can send per batch by checking the total number of reads expected to be generated per sequencing run for a given technology and whether they will use a whole or partial run (i.e., one or multiple lanes). For instance, the table below shows the different amount of data produced by different flowcells by the NovaSeq 6000 sequencing machine (Illumina).

	SP	**S1**	**S2**
Samples per lane (maximum)	5	10	26
Single-end reads:	650–800 M	1.3–1.6 B	3.3 B–4.1 B

https://sapac.illumina.com/systems/sequencing-platforms/novaseq/specifications.html)

One final important factor to consider are the required **test turn-around time** as well as the manufacturer's level of **technical support**. In terms of deciding whether to invest in a sequencer or to outsource the sequencing step, it is important to consider the **expenses involved in the operation of a sequencing machine**: initial purchase, reagents, consumables, maintenance and the required personnel and expertise. Sequencing platforms are regularly updated and the laboratory should closely monitor these **updates** and evaluate their potential impact on the test results.

Our validation study established that sRNA data derived using cDNA library preparations such as NEBNext and QIASeq sequenced on Illumina sequencing were more specific than sRNA data derived using BGI's patented sRNA kit and sequencing machine, when detecting viruses and viroids in control plants. Therefore, we are currently using the Illumina NovaSeq 6000 machine for sequencing at the Ramaciotti Centre for Genomics located in Sydney.

A) Materials

- HTS sequencing using an Illumina instrument:
    - Illumina NovaSeq 6000 (testing in this study)
    - Illumina NovaSeqX

B) Methods

- We used the Illumina NovaSeq 6000 instrument with either the SP 100c, S1 100c, or S2 100c kits. All run at 1x100bp sequencing the small RNA plus the ligated UMI.
- We recommend generating 40M-50M small RNA reads per sample.
- Demultiplexing is performed using bcl2fastq followed by delivery of fastq files.

4.3.5 Bioinformatics data analysis

The bioinformatics step generally consists of a combination of tools that analyse the raw sequencing data to accurately detect and identify virus targets present in the host plant. Its correct implementation is crucial to avoid false positive or false negative results. A recent paper by Massart *et al*. [23] highlighted the significant impact that a bioinformatic pipeline can have on the correct identification of targets through a test performance study that included 21 plant virology laboratories analysing 10 sRNA datasets.

Several bioinformatics pipelines have been developed that are either open-source, part of commercial packages (such as Geneious Prime (Biomatters) or CLC Genomics Workbench (Qiagen)) or Galaxy-based software packages. These tools either need to be installed and laboratories need to run in-house pipelines or raw data can be submitted on a web interface (VirFind http://virfind.org and VirusDetect http://virusdetect.feilab.net/cgi-bin/virusdetect/vdo_login.cgi). Some tools offer both options [29]. 'One-click' solutions are becoming popular as they simplify usage and parameterisation tools and they do not require extended bioinformatic knowledge. However, it is paramount that the personnel implementing and using these simplified pipelines understands exactly what analyses are being done in the background and what are their limitations.

Both commercial software packages and publicly available tools are available. While the choice of commercial packages can be attractive, a couple of main considerations are the expense, the lack of transparency as to the details of the analyses (as it is often patented), limited upskilling the laboratory personnel, and the possibility that software may not be maintained long term. Aside from the expense, the online submission tools present the same limitations and may have a shorter accessibility than commercial packages.

Side by side comparison of PEQ versus HTS approaches showed that the design of the bioinformatics pipeline is critical to identify true positive and avoid false positive calls. **We recommend using an approach similar to VirReport that rely on *de novo assembly* of a restricted range of small RNA read sizes that are biologically meaningful (i.e. 21-22nt or 24 nt** small RNA reads**) as** reflecting the native plant host antiviral immune

response or small RNA derived from endogenous heterochromatin pathway, respectively. **This maximises the likelihood of detecting solely viruses and viroids which have activated the immune response of the plant and thus represent a real threat to the plant health [8].**

A typical HTS data processing and analysis includes quality control, adaptor trimming, alignment to a reference or *de novo* assembly and further specialised downstream analysis.

Quality control

End-to-end analytical pipelines for RNA-Seq data require checkpoints to ensure reliable results and removal of noise from the biological samples. The first step is to perform a quality control (QC) of the raw data generated.

The sequence data generated by the HTS platforms are delivered together with their corresponding quality scores, usually in FASTQ format. FASTQ is a standard text-based format used to store the nucleotide sequences along with a per base quality score in Phred log scale. The qualities, typically with scores from 0 to 40, are represented by single letters encoded with pre-defined ranges of characters from the American Standard Code for Information Interchange (ASCII) table [30]. Currently, there are two patterns: Phred + 64, used in initial Illumina versions 1.3 + and 1.5 + and Phred + 33, the default encoding for Sanger and more recent sequencers. The FASTQ is widely accepted and used in most downstream software.

Other important QC parameters to take into consideration are: amount of reads per sample, general read and base qualities, mean read length, read length distribution, G + C content, presence of unclipped adapters and unexpected repetitive sequences. This general overview will indicate if library construction and sequencing were properly performed, or if errors like contaminants, poor ribosomal RNA depletion or low sequencing output will demand a new round of experiments. The most common software used to retrieve basic statistics is FastQC [31]and Fastp [32]. Both programs are available with intuitive graphical user interfaces (GUI), accept input files from various sequencing technologies and generate graphical reports, which are very useful for guiding the choice of filtering thresholds.

Read processing

After performing QC on the raw reads, these need to be trimmed of adaptors, primers and indices, and filtered of poor quality reads/bases. This step can significantly influence downstream data analysis [33]. Typically, reads/bases with a mean quality score over 20 (Q20) are retained for downstream analysis that encompass reads with up to 1% error rate. For applications where nucleotide variations are key to identify mutations impacting a particular trait or phenotype, increasing the mean base quality score to >Q30 reflects a 99.9% accuracy rate. Bioinformatics tools used to trim/filter raw reads include Cutadapt **[34]**, Trimmomatic [35], Trim Galore [36] and Fastp [32]. **We currently use Cutadapt for adapter trimming on reads derived from sRNA sequencing libraries, as it is able to efficiently detect and remove any sequencing artefacts (adapters, indices and primers) in a single step.**

For some sequencing technology, like Illumina, the quality of the sequence tends to diminish towards the end of the reads (Kwon *et al*., 2014; Lambert *et al*., 2018), while the first ~6-10 bases in the 5'end of reads can show nucleotide bias if random hexamer RT was used during library preparation. **It is recommended to filter out reads with low mean quality.** Trimming low-quality ends might also improve RNA-Seq assembly, mapping to reference genome and identification of single nucleotide polymorphisms (SNPs) genome-wide. However, shorter reads have a higher probability of erroneous mapping, therefore the length of 3'end trimming should be kept minimal.

Estimation of the amount of usable reads

The sequenced virus reads are expected to be interspersed with background reads that are not related to the target, such as host sequences (e.g., ribosomal sequences, mitochondria, chloroplasts), phage sequences and/or contaminants (environmental or lab-derived). **Therefore, our quality-trimmed reads are mapped to these different categories of genetic material to evaluate the extent to which they contribute to the sequencing effort in each run.** They are subsequently filtered out of our reads. The amount of leftover reads not mapping to the background permits to give an indirect indication of the actual number of reads that are likely to contain the viral sequences. If background sequences are overrepresented in a sequencing run, virus

reads might be too few to permit successful detection and cause the sequencing run to fail. This overrepresentation might be an indication of inadequate size selection of sequences or RNA degradation.

These background reads can also be removed to facilitate the search of target sequences and to reduce the risk of reporting incorrect results [37]. The removal of background reads can be particularly important when the targets are present in low concentration [38]. However, caution should be taken when dealing with organisms such as plant pararetroviruses that are capable of being completely or partially integrated in their host genome because they may be removed during this process [39]. A large-scale performance evaluation of sequence analysis strategies revealed that some laboratories were unable to identify the viral sequences integrated in the host genome present in the sequence dataset [23]. It should be noted that there may also be a risk of removing the target reads during this process when high sequence identities exist between the host and the target, or if the reference genomes used for the removal of background reads contain themselves contaminant target reads.

Target identification

This step aims to identify viral sequences. Target identification relies on the comparison with existing annotated sequences in a database. It can be performed in different ways: (i) on the individual reads (direct classification of individual reads), (ii) after mapping of the reads on reference sequences (reference-based assembly) [29, 40] or iii) after *de novo* assembly of the reads into contigs [8, 28, 40]. A combination of these methods can also be used to increase the likelihood of identifying targets present in low concentration as the contigs resulting from the different methods are pooled together [29, 41].

- Direct annotation of individual reads

The quality checked reads can be annotated without any assembly, clustering, or mapping. This approach relies on k-mer profiling at the nucleotide level and/or sequence alignment at the protein level for virus classification [38, 42, 43]. Tools like Kodoja can rapidly detect very low levels of virus sequences present in RNA-Seq data and extract reads that were assigned to viruses for downstream assembly and alignment for confirmation. This pipeline is currently not optimised for sRNA data. The specificity and sensitivity of the annotation process will be highly dependent on the database used and misclassification has been reported for viruses with small evolutionary distances between some taxa [38]. Therefore, downstream confirmation of detected targets is paramount.

- Mapping to a reference

Mapping of high-quality sRNA reads onto a reference genome (also called reference-guided assembly), can be performed using a bioinformatic tools such as Bowtie or Bowtie2. If the host sequence is available, a filtering step can be applied to facilitate virus genome reconstruction.

The parameters that need to be set and optimised for mapping include the number of mismatches, gaps allowed, minimal identity percentage. These parameters are critical to avoid incorrect results. If they are set too lenient, non-specific mapping can happen, while too stringent parameters can result in poor mapping onto distant isolate [44, 45] . Important mapping results metrics include average read depth, read distribution on the reference sequence and percentage of identity with reference.

- De novo assembly followed by annotation

The quality checked reads from a shotgun sequencing library can be assembled *de novo* to create longer sequences called contigs. The sRNA reads are assembled when they present similar sequences on a portion or on the totality of their length. A range of tools can perform sRNA *de novo* assembly: Velvet [46], Oases [47] and Trans-ABySS [48] or those integrated in commercial software packages, such as Geneious Prime (Biomatters) or CLC Genomics Workbench (Qiagen). Sequencing depth must be obtained at a higher coverage for *de novo* assembly (estimated around 30× [49]) while reference-guided approach requires about 10× [48] to find full-length transcripts. The higher throughput also increases the processing requirements. Digital normalisation can help remove redundancy without impacting on the assembly outcome [50].

The parameters that need to be set and optimised for *de novo* assembly include percentage of identity between reads, the minimum overlap, the minimal length of contigs, k-mer length and bubble size. *De novo*

assembly is often coupled with a contig scaffolding step (e.g. CAP3 [51]) to collapse the obtained sequences. The quality of the assembly can be evaluated using N_{50} or U_{50} values [52].

The most common approach to annotate assembled contigs/scaffolds is to perform homology searches with tools such as BLASTN/BLASTX [53, 54], DIAMOND [55] or MegaBLAST [56] on nucleotide or protein sequences in public (e.g. NCBI) or custom viral databases, Most bioinformatics pipelines rely on BLASTN to find known viruses and BLASTX/DIAMOND to find novel viruses. There are specialised viral genome reference databases that can accelerate the HTS data analysis or provide curated virus pathogen information to enable decision making. The level of certainty of the similarity searches (e.g., e-value) should always be recorded together, along with the tools and database (version) used. Expert judgement may be needed to evaluate the result of a taxonomic position. When it is possible to retrieve the whole genome of a target, through shotgun sequencing, genome completeness and read depth can support the result of a taxonomic annotation (i.e., the more complete the genome is, the more reliable the taxonomic position). Additional analyses such as phylogenetic analysis may also be required.

- Reference sequence database

Sequence databases can be available publicly or can be developed and maintained in-house. The selection of the reference sequence database depends on the intended use of the HTS test and should be clearly defined during the test selection and development. When the focus of the HTS test is on a limited range of known pests, a curated database can be created with sequences of high quality that are accurately annotated and not redundant. However, when searching for unknown organisms, a more extensive and less curated database might provide a much better chance for their discovery than a well curated database with a limited number of entries (Lambert *et al*., 2018; Piper *et al.,* 2019). Sequence databases should be tested for their ability to identify at least the expected targets of the HST tests. The content of public sequence databases is constantly evolving because of scientific discoveries and changes in taxonomy of pests. Therefore, sequence databases used for the bioinformatics step of the HTS test should be kept up to date and information on the databases used during the analysis should be documented for traceability purposes (e.g., version number, date of download, original source, location).

Duplicate read estimation and removal

Duplicate reads originate from the same amplified fragments. Their characteristics are common coordinates (e.g., the same start and end coordinates after mapping), same sequencing direction and identical sequences. A dataset containing lots of duplicated reads in a shotgun library might be the result of a failed library preparation, where too little input material was available. The high abundance of duplicated reads can limit the sensitivity of the HTS test as they can compete with low abundance targets, despite having a large total read number. It is therefore recommended to evaluate the proportion of duplicate reads during the quality control stage of data analyses. Excess duplicate reads can be removed by using read normalisation tools to facilitate the downstream analysis of sequences. If UMIs were used during sample preparation, then reads sharing the same UMIs can be collapsed after alignment to reference.

Sequencing statistics

Since HTS-based approaches are likely to meet contamination issues, the detection of a single or a few viral reads might therefore not be considered sufficient to consider a sample positive. Ideally, all parts of a viral genome should be represented by sRNAs (and their contigs) to ensure that the virus species is correctly identified. With this in mind, pipelines such as VirReport and VirusDetect provide percentage of genome coverage together with total and normalised sequencing depths as standard outputs to aid users in evaluating the significance of identified contigs. If such tool is to be implemented for an HTS diagnostic test, it is recommended to set an additional filtering step which takes into consideration these parameters during the test validation, in order to flag potential false positives.

A) Materials

VirReport processes uses several bioinformatics tools (Table 4.2).

Table 4.2: Lists the current tools used by the separate VirReport processes. The tool is available at https://github.com/eresearchqut/VirReport

Step	Processing component	Tool
Quality Control	Assess overall quality of raw and quality filtered reads	FastQC/ Fastp
	Remove adaptor sequences and trim poor reads/bases	UMICollapse; Cutadapt
Mapping	Mapping reads onto plant host sequences	Bowtie
De novo Assembly	Assembly of siRNAs or RNA-Seq reads	Velvet
Scaffolding	Merge *de novo* assembled contigs	CAP3
Predict protein sequences	Determine protein sequences encoded in assembled contigs/scaffolds	Emboss (getorf)
Find virus	BLATN against nucleotide (NT) database	Blastn/megablast
	BLASTP againts protein (NR) database	blastp, diamond
Mapping	Mapping reads onto best viral reference hit	Bowtie

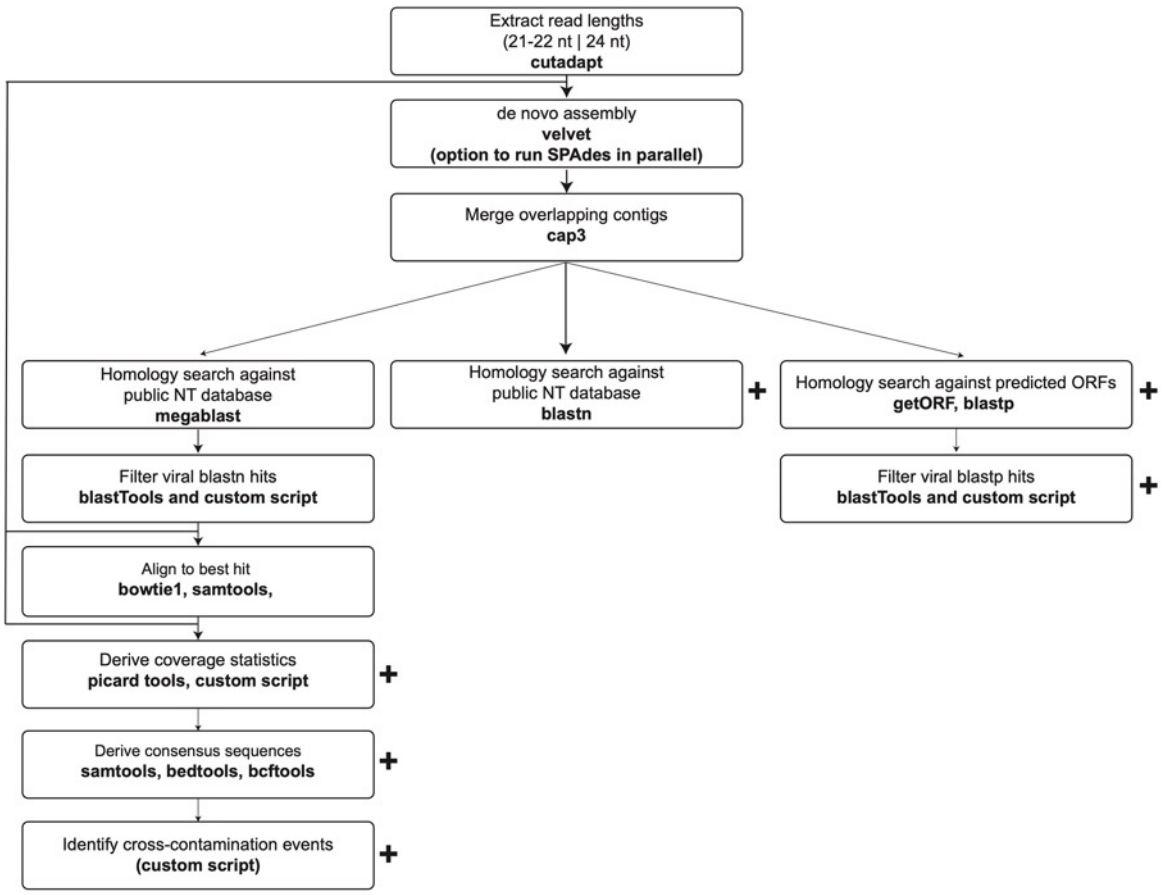


Figure 4.3 VirReport workflow overview. The VirReport pipeline has been implemented in nextflow for its easy access and reproducible and scalable use. The end-to-end VirReport pipeline uses quality filtered small RNA-Seq data as input. Read sequences are selected based on size (e.g. 21-22nt or 24nt) for *de novo* assembly and subsequent scaffolding. Scaffolds are then screened against nucleotide and protein databases. A summary report is generated using BlastTools.jar and custom scripts. Coverage and other statistics are provided to assist with the interpretation of results. Consensus genome sequences, which incorporate the variants of each detected viruses and viroids, can optionally be derived using bedtools (https://bedtools.readthedocs.io/en/latest/content/overview.html) and bcftools [57]. Importantly, VirReport flags potential cross-sample contamination events based on all multiplexed samples in the same HTS experiment. ✚ indicates tools and workflows users can add to the pipeline as by default these are not part of the automated pipeline.

B) Methodology

- Install Nextflow.

VirReport requires Nextflow to be installed in your account on the HPC. Find details on how to install and test Nextflow here prepare a nextflow.config file and run a PBS pro submission script for Nextflow pipelines.

Pull the git repo using:

```
git clone file:///work/pipelines/eresearch/vsd -b vsd-1.0
```

• Install conda3 or miniconda3.

All the required modules are included in the environment.yml file in the pipeline base directory and shows all the tools used in the pipeline:

```
name: vsd-1.0
channels:
  - bioconda
  - conda-forge
  - defaults
  - r
  - anaconda
dependencies:
  - blast=2.11.0
  - cap3=10.2011
  - spades=3.14.0
  - emboss=6.6.0
  - openjdk=8.0.152
  - fastp=0.20.1
  - biopython=1.76
  - numpy=1.16.5
  - matplotlib=2.2.3
  - velvet=1.2.10
  - bowtie=1.3.0
  - samtools=1.12
  - picard==2.25.6
  - bedtools
  - bcftools
  - pandas
  - fastqc=0.11.9
  - fastp=0.20.1
  - cutadapt=3.5
  - umi_tools=1.1.2
```

- **Install a local NCBI blast directory (NT and NR).**

Find detailed info on how to download these databases at
Make sure the taxdb.btd and the taxdb.bti files are also present in the directory.
Create a folder where you will store your NCBI database including the date of download. For instance:

```
#creating a DB on 30 Nov 2022

mkdir blastDB/20221130
```

Run the following PBS script in the newly created folder. Use the update_blastdb.pl script from the blast+ version you will use with your pipeline.

```
#!/bin/bash -l

#PBS -N blastdb_download

#PBS -l walltime=24:00:00

#PBS -l mem=60gb

#PBS -l ncpus=2

cd $PBS_O_WORKDIR

perl update_blastdb.pl --decompress nt [*]

perl update_blastdb.pl --decompress nr [*]

perl update_blastdb.pl taxdb

tar -xzf taxdb.tar.gz
```

- **Install a local NCBI blast directory (NT and NR).**

The VSD workflow will perform the following steps by default:

- Retain reads of a given length (e.g. 21-22 or 24 nt long) from fastq file(s) provided in index.csv file (`readprocessing`)
- De novo assembly using kmer 15 and coverage 3 (`velvet`) -
- Collapse contigs into scaffolds (min length 20) (`cap3`)
- Run megablast homology search against NCBI NT database (`megablast_nt_velvet`)
- Summarise megablast results and restrict to virus and viroid matches (`BlastTools_megablast_velvet`)
- Derive coverage statistics, consensus sequence and VCF matching to top blast hits (`filter_n_cov`)

A number of additional optional steps can be run:

- --blastp: Predict ORF from de novo assembly (derived with Velvet) and run blastP againts NCBI NR (`getorf`, `blastp`, `blastpdbcmd`, `BlastToolsp`) --blastp
- --contamination_detection: Run cross-sample contamination prediction (`contamination_detection`)
- --blastlocaldb: Run blastn and megablast homology search on de novo assembly (derived with Velvet) against local virus and viroid database (`blast_nt_localdb_velvet`, `filter_blast_nt_localdb_velvet`)
- --blastn: Run blastn homology search on de novo assembly (derived with Velvet) against local virus and viroid database (`blastn_nt_velvet`)
- --spades: Run SPAdes 3.14 de novo assembler and perform blastn homology analysis on the derived de novo contigs (`spades, cap3_spades`, `megablast_nt_spades` , `BlastToolsn_megablast_spades`)

A number of additional options are included:

- --targets: A text file with the taxonomy of the viruses/virioids of interest can be provided and only these will be retained in the megablast summary results derived at the `filter_n_cov` step.
- --spadeskmer specifies the range of kmers to use when running spades
- --cap3_len specifies the minimal length of contigs to retain after CAP3 scaffolding
- --blastn_evalue and --blastp_evalue specifies the evalue parameter to use during blast analyses
- --orf_minsize correspond to the minimal open reading frame getorf retains

To enable these options, they can either be included in the nextflow run command provided in the PBS script:

```
nextflow run $PBS_O_WORKDIR/main.nf \
  --indexfile $PBS_O_WORKDIR/index_example.csv \
  --blastlocaldb \
  --spades \
  --contamination_detection \
  -resume
```

or update parameter to **true** in the nextflow.config file. For instance:

```
params {
blastlocaldb = true
spades = true
contamination_detection = true
}
```

- **Preparing the data**
  - Preparing an index.csv file

  You need to create a TAB delimited text file that will be the input for the workflow. By default the pipeline will look for a file called "index.csv" in the base directory but you can specify any file name using the `--indexfile [filename]` in the **nextflow run** command. This text file requires the following columns (which needs to be included as a header): sampleid,samplepath,minlen,maxlen:

  - sampleid will be the sample name that will be given to the files created by the pipeline
  - samplepath is the full path to the quality filtered fastq files that the pipeline requires as starting input
  - minlen and maxlen correspond to the read size that will be retained for downstream analyses.

  An index_example.csv is included in the base directory:
  `sampleid,samplepath,minlen,maxlen`

```
MT212,/work/hia_mt18005/diagnostics/2021/14_RAMACIOTTI_LEL9742-LEL9751/results/06_usable_reads/MT212_21-22bp.fastq,21,22

MT213,/work/hia_mt18005/diagnostics/2021/14_RAMACIOTTI_LEL9742-LEL9751/results/06_usable_reads/MT213_21-22bp.fastq,21,22
```

  You also need to provide the path of your NCBI blast directory in the nextflow.config file. For instance:

```
params {
  blast_db = '/lustre/work-lustre/hia_mt18005/blastDB/30112021'
  }
```

  If you are interested to run a blast analysis against a local database, you also need to specify its path in the nextflow.config file. For example:

```
blastn_local_db = '/work/hia_mt18005/databases/PVirDB/PVriDB_ver2021_11_09/PVirDB_ver20211109.fasta'
```

- **Running the pipeline**

Finally, you need to create a PBS script which includes your nextflow run command. An example of PBS script is included in the base directory and will run the pipeline with default steps:

```
#!/bin/bash -l
#PBS -N NextflowVSD
#PBS -l select=1:ncpus=2:mem=6gb
#PBS -l walltime=24:00:00
cd $PBS_O_WORKDIR
module load java
NXF_OPTS='-Xms1g -Xmx4g'
nextflow run $PBS_O_WORKDIR/main.nf -resume --indexfile $PBS_O_WORKDIR/index_example.csv
```

This following PBS script will additionally run spades, the cross-sample predictor and blast searches against a local database:

```
#!/bin/bash -l

#PBS -N NextflowVSD

#PBS -l select=1:ncpus=2:mem=6gb

#PBS -l walltime=24:00:00

#run spades

#run blast homology against user local database

#run false positive predictor

nextflow run $PBS_O_WORKDIR/main.nf -resume --indexfile
$PBS_O_WORKDIR/index_example.csv --blastlocaldb --spades --
contamination_detection
```

These jobs can be submitted using:

```
qsub nextflow_example.pbs
```

- Monitoring the run
  You can use the command

  ```
  qstat -u $USER
  ```

  Alternatively use the following command:

  ```
  qjobs
  ```

  To check on the jobs you are running. Nextflow will launch additional jobs during the run.
  You can also check the **.nextflow.log** file for details on what is going on.
  Finally, if you have configured the connection to the NFTower you can logon and check your run.

- **Outputs**
  Under the results folder, the pipeline will populate outputs under separate folders for each step. These will be stored in subfolders for each sample:

  results → 01_read_size_selection → sample1
  → sample2
  → sample 3
  → 02_velvet → sample1
  → sample2
  → sample 3
  etc…

The folders are structures as follows (examples of outputs are provided in italics):

- 01_read_size_selection (cutadapt log file and fastq file including reads only matching the size specified in the index.csv file) *MT020_21-22nt_cutadapt.log & MT020_21-22nt.fastq*
- 02_velvet (velvet results and the fasta file which includes the velvet assembled contigs *MT020_velvet_assembly_21-22nt.fasta*
- 02a_spades (if spades is additionally run)
- 03_cap3 (fasta file of the scaffolds produced by CAP3 as well as the singletons) *MT020_velvet_cap3_21-22nt_rename.fasta*
- 04_blastn (all blastn results, filtered results limited to only viruses and viroid top 5 hit matches and their taxonomy) *MT020_velvet_21-22nt_megablast_vs_NT.bls, MT020_velvet_21-22nt_megablast_vs_NT_top5Hits.txt, MT020_velvet_21-22nt_megablast_vs_NT_top5Hits_virus_viroids_final.txt MT020_velvet_21-22nt_megablast_vs_NT_top5Hits_virus_viroids_seq_ids_taxonomy.txt*

- 05_blastoutputs (BlastTools.jar summary output which clusters all the contigs matching to a specific hit. *summary_MT029_velvet_21-22nt_megablast_vs_NT_top5Hits_virus_viroids_final.txt*
- 06_blastp (blastp outputs) *MT020_velvet_21-22nt_getorf.min50aa.fasta, MT020_velvet_21-22nt_getorf.min50aa_blastp_vs_NR_out_virus_viroid.txt*
- 07_filternstats (filtered blast summary with various coverage statistics for each virus and viroid hit, and associated consensus fasta file and vcf file) *MT020_21-22nt_top_scoring_targets_with_cov_stats.txt, MT020_21-22nt_MK929590_Peach_latent_mosaic_viroid.consensus.fasta, MT020_21-22nt_MK929590_Peach_latent_mosaic_viroid_sequence_variants.vcf.gz*
- 08_summary (summary of results for all samples included in the index.csv file. This includes a cross-contamination prediction) *run_top_scoring_targets_with_cov_stats_with_cont_flag_21-22nt_0.01.txt*

Future potential additional features:

- Include a deduplication step for fastq files that have UMIs incorporated.
- Incorporate the fastq file initial filtering steps from sRNAqc as option.
- Work on automation of final summary report.
- Add coverage statistics and cross contamination flag logic to local db blast results.
- Incorporate VirusDetect in the pipeline and derive a summary of results from both pipelines.
- Perform automatically 21-22nt and 24nt analyses by default.

Target confirmation, interpretation, and reporting.

The last step of the HTS test consists of confirming the identity of the targets detected in the sample, interpreting the biological relevance of the targets identified (especially for organisms with a sequence identity close to the threshold of species demarcation) and reporting of the HTS test results.

Several graphical browsers have been developed to inspect HTS sequencing data at the nucleotide level. We recommend Integrative Genome Viewer (IGV) [58] to visualise the distribution of alignments of virus-derived reads onto reference sequences but also to validate host novel transcripts and gene junctions, checking the coverage support for genomic variants and spot read piles, which may represent repetitive regions.

Laboratories might also consider running an independent PCR test to confirm the findings especially if a virus has never been observed in each commodity before.

Conclusion

Development and adoption of end-to-end best practices and standards are key to detect at the border incoming exotic pests. Challenges remain for cases where novel viruses are found and there is little or no biological information on their possible impact on plant hosts and whether they represent an emerging pest with potential to disrupt the biosecurity system and impact plant industries. Further research is necessary to facilitate the development of policy framework that biosecurity officer, plant pathologists and decision makers to safeguard domestic industries and access to high value markets.

References

1. Soliman, T.; Mourits, M. C. M.; Lansink, A. G. J. M. O.; van der Werf, W.; Quantitative economic impact assessment of an invasive plant disease under uncertainty - A case study for potato spindle tuber viroid (PSTVd) invasion into the European Union. *Crop Prot* **2012,** 40, 28-35.

2. Adams, I.; Fox, A., Diagnosis of Plant Viruses Using Next-Generation Sequencing and Metagenomic Analysis. In *Current Research Topics in Plant Virology*, Wang A., Z. X., Ed. 2016; pp 323-335.

3. Adams, I. P.; Glover, R. H.; Monger, W. A.; Mumford, R.; Jackeviciene, E.; Navalinskiene, M.; Samuitiene, M.; Boonham, N.; Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular plant pathology* **2009,** 10, 537-545.

4. Al Rwahnih, M.; Daubert, S.; Golino, D.; Rowhani, A.; Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* **2009,** 387, 395-401.

5. Kreuze, J. F.; Perez, A.; Untiveros, M.; Quispe, D.; Fuentes, S.; Barker, I.; Simon, R.; 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* **2009,** 388, 1-7.

6. Roossinck, M. J.; Martin, D. P.; Roumagnac, P.; Plant Virus Metagenomics: Advances in Virus Discovery. *Phytopathology* **2015,** 105, 716-727.

7. Kutnjak, D.; Elena, S. F.; Ravnikar, M.; Time-Sampled Population Sequencing Reveals the Interplay of Selection and Genetic Drift in Experimental Evolution of Potato Virus Y. *J Virol* **2017,** 91.

8. Barrero, R. A.; Napier, K. R.; Cunnington, J.; Liefting, L.; Keenan, S.; Frampton, R. A.; Szabo, T.; Bulman, S.; Hunter, A.; Ward, L., et al.; An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. *Bmc Bioinformatics* **2017,** 18.

9. Olmos, A.; Boonham, N.; Candresse, T.; Gentit, P.; Giovani, B.; Kutnjak, D.; Liefting, L.; Maree, H. J.; Minafra, A.; Moreira, A., et al.; High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. *EPPO Bulletin* **2018,** 48, 219-224.

10. Candresse, T.; Filloux, D.; Muhire, B.; Julian, C.; Galzi, S.; Fort, G.; Bernardo, P.; Daugrois, J.-H.; Fernandez, E.; Martin, D. P., et al.; Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. *PLOS ONE* **2014,** 9, e102945.

11. Fox, A.; Reconsidering causal association in plant virology. *Plant Pathology* **2020,** 69, 956-961.

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

13. Jiang, L. C.; Schlesinger, F.; Davis, C. A.; Zhang, Y.; Li, R. H.; Salit, M.; Gingeras, T. R.; Oliver, B.; Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **2011,** 21, 1543-1551.

14. Maliogka, V. I.; Minafra, A.; Saldarelli, P.; Ruiz-García, A. B.; Glasa, M.; Katis, N.; Olmos, A.; Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. *Viruses* **2018,** 10.

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

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

17. Al Rwahnih, M.; Dave, A.; Anderson, M. M.; Rowhani, A.; Uyemoto, J. K.; Sudarshana, M. R.; Association of a DNA Virus with Grapevines Affected by Red Blotch Disease in California. *Phytopathology* **2013,** 103, 1069-1076.

18. Schroeder, A.; Mueller, O.; Stocker, S.; Salowsky, R.; Leiber, M.; Gassmann, M.; Lightfoot, S.; Menzel, W.; Granzow, M.; Ragg, T.; The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *Bmc Mol Biol* **2006,** 7.

19. Johnson, M. T. J.; Carpenter, E. J.; Tian, Z. J.; Bruskiewich, R.; Burris, J. N.; Carrigan, C. T.; Chase, M. W.; Clarke, N. D.; Covshoff, S.; dePamphilis, C. W., et al.; Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes. *Plos One* **2012,** 7.

20. Benesova, S.; Kubista, M.; Valihrach, L.; Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. *Diagnostics* **2021,** 11, 964.

21. Kircher, M.; Sawyer, S.; Meyer, M.; Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **2012,** 40, e3.

22. Budowle, B.; Connell, N. D.; Bielecka-Oder, A.; Colwell, R. R.; Corbett, C. R.; Fletcher, J.; Forsman, M.; Kadavy, D. R.; Markotic, A.; Morse, S. A., et al.; Validation of high throughput sequencing and microbial forensics applications. *Investig Genet* **2014,** 5, 9.

23. Massart, S.; Chiumenti, M.; De Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Komínek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L., et al.; Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. *Phytopathology* **2019,** 109, 488-497.

24. MacConaill, L. E.; Burns, R. T.; Nag, A.; Coleman, H. A.; Slevin, M. K.; Giorda, K.; Light, M.; Lai, K.; Jarosz, M.; McNeill, M. S., et al.; Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **2018,** 19, 30.

25. Sena, J. A.; Galotto, G.; Devitt, N. P.; Connick, M. C.; Jacobi, J. L.; Umale, P. E.; Vidali, L.; Bell, C. J.; Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Scientific Reports* **2018,** 8, 13121.

26. Kumar, K. R.; Cowley, M. J.; Davis, R. L.; Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* **2019,** 45, 661-673.

27. Maljkovic Berry, I.; Melendrez, M. C.; Bishop-Lilly, K. A.; Rutvisuttinunt, W.; Pollett, S.; Talundzic, E.; Morton, L.; Jarman, R. G.; 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* **2020,** 221, S292-s307.

28. Massart, S.; Chiumenti, M.; Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Kominek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L., et al.; Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies. *Phytopathology* **2019,** 109, 488-497.

29. Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R. G.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Lin, K. S.; Kreuze, J.; Fei, Z. J.; VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **2017,** 500, 130-138.

30. Cock, P. J.; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M.; The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **2010,** 38, 1767-71.

31. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accesssed on 25 January 2022)

32. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J.; fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018,** 34, i884-i890.

33. Del Fabbro, C.; Scalabrin, S.; Morgante, M.; Giorgi, F. M.; An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *Plos One* **2013,** 8.

34. Martin, M.; Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **2011,** 17, 3.

35. Bolger, A. M.; Lohse, M.; Usadel, B.; Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014,** 30, 2114-20.

36. Krueger, F. Trim_galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 25 January 2022).

37. Lambert, C.; Braxton, C.; Charlebois, R. L.; Deyati, A.; Duncan, P.; La Neve, F.; Malicki, H. D.; Ribrioux, S.; Rozelle, D. K.; Michaels, B., et al.; Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection. *Viruses* **2018,** 10.

38. Baizan-Edge, A.; Cock, P.; MacFarlane, S.; McGavin, W.; Torrance, L.; Jones, S.; Kodoja: A workflow for virus detection in plants using k-mer analysis of RNA-sequencing data. *J Gen Virol* **2019,** 100, 533-542.

39. Hohn, T.; Richert-Pöggeler, K. R.; Staginnus, C.; Harper, G.; Schwarzacher, T.; Teo, C. H.; Teycheney, P.-Y.; Iskra-Caruana, M.-L.; Hull, R., Evolution of Integrated Plant Viruses. In *Plant Virus Evolution*, Roossinck, M. J., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 53-81.

40. Visser, M.; Bester, R.; Burger, J. T.; Maree, H. J.; Next-generation sequencing for virus detection: covering all the bases. *Virol J* **2016,** 13.

41. Ho, T.; Tzanetakis, I. E.; Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **2014,** 471, 54-60.

42. Wood, D. E.; Salzberg, S. L.; Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **2014,** 15, R46.

43. Menzel, P.; Ng, K. L.; Krogh, A.; Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* **2016,** 7, 11257.

44. Roy, S.; Coldren, C.; Karunamurthy, A.; Kip, N. S.; Klee, E. W.; Lincoln, S. E.; Leon, A.; Pullambhatla, M.; Temple-Smolkin, R. L.; Voelkerding, K. V., et al.; Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* **2018,** 20, 4-27.

45. Weiss, M. M.; Van der Zwaag, B.; Jongbloed, J. D. H.; Vogel, M. J.; Brüggenwirth, H. T.; Lekanne Deprez, R. H.; Mook, O.; Ruivenkamp, C. A. L.; van Slegtenhorst, M. A.; van den Wijngaard, A., et al.; Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. *Human Mutation* **2013,** 34, 1313-1321.

46. Zerbino, D. R.; Birney, E.; Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **2008,** 18, 821-829.

47. Schulz, M. H.; Zerbino, D. R.; Vingron, M.; Birney, E.; Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **2012,** 28, 1086-1092.

48. Robertson, G.; Schein, J.; Chiu, R.; Corbett, R.; Field, M.; Jackman, S. D.; Mungall, K.; Lee, S.; Okada, H. M.; Qian, J. Q., et al.; De novo assembly and analysis of RNA-seq data. *Nat Methods* **2010,** 7, 909-U62.

49. Martin, J.; Bruno, V. M.; Fang, Z. D.; Meng, X. D.; Blow, M.; Zhang, T.; Sherlock, G.; Snyder, M.; Wang, Z.; Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *Bmc Genomics* **2010,** 11.

50. Chouvarine, P.; Wiehlmann, L.; Moran Losada, P.; DeLuca, D. S.; Tummler, B.; Filtration and Normalization of Sequencing Read Data in Whole-Metagenome Shotgun Samples. *Plos One* **2016,** 11, e0165015.

51. Huang, X.; Madan, A.; CAP3: A DNA sequence assembly program. *Genome research* **1999,** 9, 868-877.

52. Castro, C. J.; Ng, T. F. F.; U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. *Journal of computational biology : a journal of computational molecular cell biology* **2017,** 24 11, 1071-1080.

53. Altschul, S.; Hot papers - Bioinformatics - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs by S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman - Comments. *Scientist* **1999,** 13, 15-15.

54. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L.; BLAST plus : architecture and applications. *Bmc Bioinformatics* **2009,** 10.

55. Buchfink, B.; Xie, C.; Huson, D. H.; Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **2015,** 12, 59-60.

56. Morgulis, A.; Coulouris, G.; Raytselis, Y.; Madden, T. L.; Agarwala, R.; Schäffer, A. A.; Database indexing for production MegaBLAST searches. *Bioinformatics* **2008,** 24, 1757-64.

57. Danecek, P.; Bonfield, J. K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M. O.; Whitwham, A.; Keane, T.; McCarthy, S. A.; Davies, R. M., et al.; Twelve years of SAMtools and BCFtools. *Gigascience* **2021,** 10.

58. Thorvaldsdottir, H.; Robinson, J. T.; Mesirov, J. P.; Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **2013,** 14, 178-192.

Appendix 5 – Comparison of PEQ protocols against small RNA-seq and ribosomal depleted RNA-seq approaches to detect viruses and viroids in positive control plants

At Post Entry Quarantine (PEQ), molecular assays such as Polymerase Chain Reaction (PCR) and ELISA assays as well as biological indexing and electron microscopy inspections are used for the detection of plant viruses and viroids in imported plant material. In this initial study, we compared two High throughput sequencing (HTS) technologies, also referred a 'next generation sequencing (NGS)' technologies: i) small RNA-seq HTS (sRNA-seq), and ii) ribosomal depleted RNA-seq (rdRNA-seq) HTS against existing PEQ protocols.

A set of 17 positive control plants (Table 5.1) encompassing Citrus, Prunus, Rubus, Potato, Grapevine, Strawberry, and an ornamental grass were selected to conduct a side-by-side comparison of PEQ versus sRNA-seq and rdRNA-seq HTS approaches.

Table 5.1: Selected control plants for PEQ vs HTS comparisons

Sample	Commodity	Virus name
MT001	Citrus	CEVd - Citrus exocortis viroid
MT002	Peach	PNRSV - Prunus necrotic ringspot virus
MT003	Citrus	CTV - Citrus tristeza virus
		CVEV - Citrus vein enation virus
		HSVd - Hop stunt viroid
MT004	Citrus	CTV - Citrus tristeza virus
		CEVd - Citrus exocortis viroid
		HSVd - Hop stunt viroid
MT005	Rubus	RYNV - Rubus yellow net virus
		RBDV - Raspberry bushy dwarf virus
MT006	Rubus	RYNV - Rubus yellow net virus
MT010	Nursery	SrMV - Sorghum mosaic virus
MT013	Strawberry	SMoV - Strawberry mottle virus
MT014	Strawberry	Strawberry vein banding virus
		Beet pseudo-yellows virus
MT016	Sweet potato	SPFMV - Sweet potato feathery mottle virus
MT018	Vitis	GFkV - Grapevine Fleck virus
		GSYSVd-1 - Grapevine yellow speckle viroid 1
		HSVd - Hop stunt viroid
		GRSPaV - Grapevine rupestris stem pitting- associated virus
		GLRaV - Grapevine leafroll associated virus
MT019	Vitis	GFkV - Grapevine fleck virus
		HSVd - Hop stunt viroid
		GLRaV - Grapevine leafroll associated virus

Total RNA samples were collected at PEQ using a SOP described in {Lelwala, 2022 #5;Gauthier, 2022 #6}. Collected total RNA samples were then sent to two HTS service providers, one national (AGRF, Melbourne) and another overseas (BGI, Hong Kong) for HTS data generation.

HTS raw data

- **BGI service provider:**
    - 17 samples were sequenced using small RNA-seq and ribosomal depleted RNA-seq approaches.
    - Small RNA-seq: Generated 979.8 million single-end small RNA sequences with an average of 57.6 million (± 11.7) per sample.
    - Ribosomal depleted RNA-seq: Produced 641.7 million paired-end sequences with an average of 36.6 million (± 0.9) per sample. Most ribosomal RNA was removed from samples averaging only the presence of 9.7% (±3.3) ribosomal RNA in the raw HTS datasets.
- **AGRF service provider:**
    - 12 samples were sequenced using small RNA-seq and ribosomal depleted RNA-seq approaches.
    - Small RNA-seq: Initially, generated 356 million single-end small RNA sequences with an average of 29.7 million (± 7) per sample. Additional small RNA-seq data was collected for five samples where not all expected viruses were detected. A minimal of 40 million reads per sample is required to detect all viruses found in PEQ by HTS data.
    - Ribosomal depleted RNA-seq: Produced 648.8 million paired-end sequences with an average of 58.7 million (± 3.8) per sample.

Bioinformatics pipelines

Overall bioinformatics processing of small RNA-seq and ribosomal depleted RNA-seq raw data is subject to four main steps: i) quality control; ii) genome assembly; iii) sequence comparison of assembled sequences against public databases; and iv) filtering and report.

- **Small RNA-seq**
    - Raw small RNA-seq HTS data was processed using a modified version of the Virus Surveillance and Diagnosis (VSD) pipeline (Barrero et al., 2017) implemented to run on a cloud server and on a local QUT high performance computing (HPC) server.
    - BGI provided adaptor clipped 'clean' reads, while AGRF provided raw HTS data. Prior running the VSD automated pipeline, HTS data was quality assessed using FastQC and then quality trimmed using trimgalore.
    - QCed small RNA reads were then run using the VSD 21-22nt and the 24nt pipelines using SPAdes as the de novo assembler using kmer lengths from 9 to 21, CAP3 was then used to merge and extend assembled sequences, and BLASTN/BLASTP to search sequence similarity against non-redundant nucleotide/protein databases. We use an in-house script to parse and summarise BLAST results.
- **Ribosomal depleted RNA-seq**
    - Raw ribosomal depleted RNA-seq HTS data was QCed using TRIMGALORE or FASTP retaining bases with a quality score of >=30 (equivalent to 99.9% accuracy) and a minimal length of 75 nt.
    - QCed reads were then assembled using METASPADES using kmer lengths of 31, 43 and 57. CAP3 was then used to merge and extend assembled sequences, and BLASTN/BLASTP to search sequence similarity against non-redundant nucleotide/protein databases. We use an in-house script to parse and summarise BLAST results.

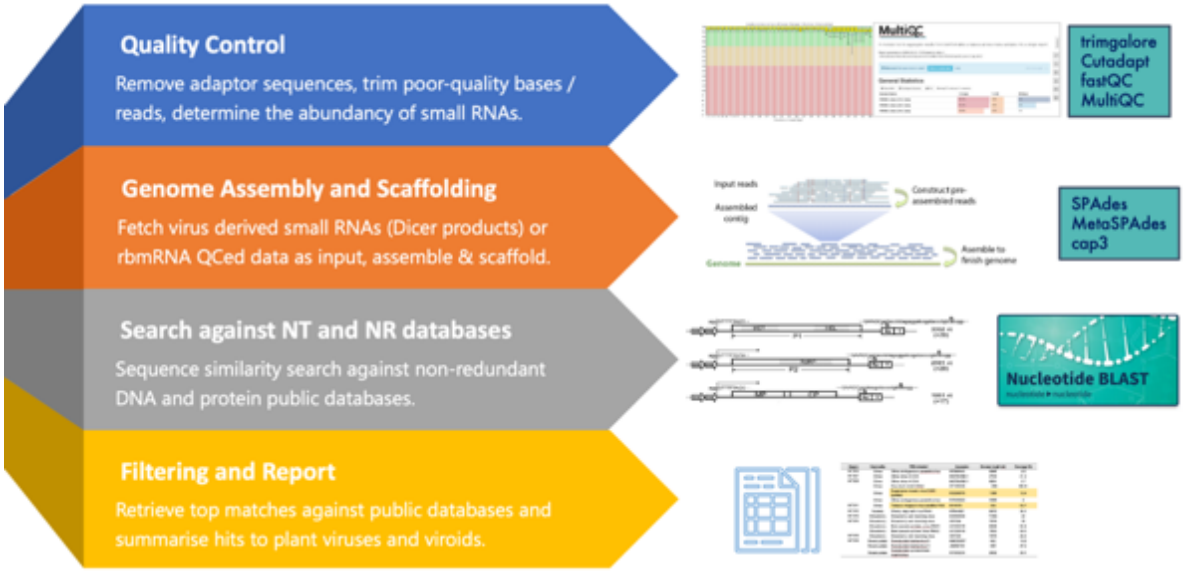


Figure 5.1: Overview of bioinformatics processing of HTS data for the diagnosis of viruses and viroids

Major findings

HTS methods find all PEQ detected viruses and viroids

Small RNA-seq and ribosomal depletion RNA-seq HTS methods detected all know viruses and viroids in control plants. In summary:

- Small RNA-seq HTS data generated by either BGI or AGRF found all known viruses and viroids known to be present in 12 control plants (**Table 5.2**).
- Ribosomal depleted RNA-seq HTS data generated by AGRF detected all known viruses and viroids, while data produced by BGI missed to detect a hop stunt viroid in a Citrus sample - MT003 (**Table 5.3**).
- Based on the 100% agreement obtained using HTS data generated by AGRF, follow up large-scale sequencing will be done with this local service provider.

HTS methods find additional non-quarantinable viruses and cross-sample contamination in control plants

One of the advantages of HTS methods is that these can also identify additional non-biosecurity risk viruses if present in quarantined plants. Also, it is documented that HTS methods can be susceptible to detecting cross-sample contaminations (index hopping) and it is important to design bioinformatics analysis pipelines that minimise the impact of such contaminants.

In summary:

- Small RNA-seq:
  - identified 14 new viruses and viroids in 9 control samples (**Table 5.4**).
  - found 2 index hopping events in small RNA-seq data generated by BGI for 17 samples (**Table 5.4**), while no index hopping events were found in AGRF small RNA-seq HTS data for 12 samples (data not shown).
- Ribosomal depleted RNA-seq:
  - detected 14 new viruses in 10 control plants using BGI data (**Table 5.5**) and 9 new viruses in 6 samples using AGRF data (**Table 5.6**).
  - We also found hits to grape powdery mildew fungus in two grape samples (**Table 5.6**).

- found 34 index hopping events in 14 out of 17 (82.3%) samples using HTS data produced using a rolling cycle sequencing approach (BGI).
    - The most commonly index hopped virus was a Prunus Necrotic Ringspot virus found contaminating six samples
- found 10 index hoping event in 6 out of 12 (50%) samples among HTS data produced by a dual-indexing approach (AGRF).
    - We also found that the most commonly index hopped virus was Prunus Necrotic Ringspot virus found contaminating six samples
- We observed variation in terms of which samples showed cross-contamination among samples sequenced by AGRF and BGI. This is due to a subset of samples sequenced at AGRF.
- Observed index hoping events appear to be aleatory, however PNRSV and a potyvirus are highly prone to cross-contamination in ribosomal depleted RNA-seq HTS samples.

Recommendations:

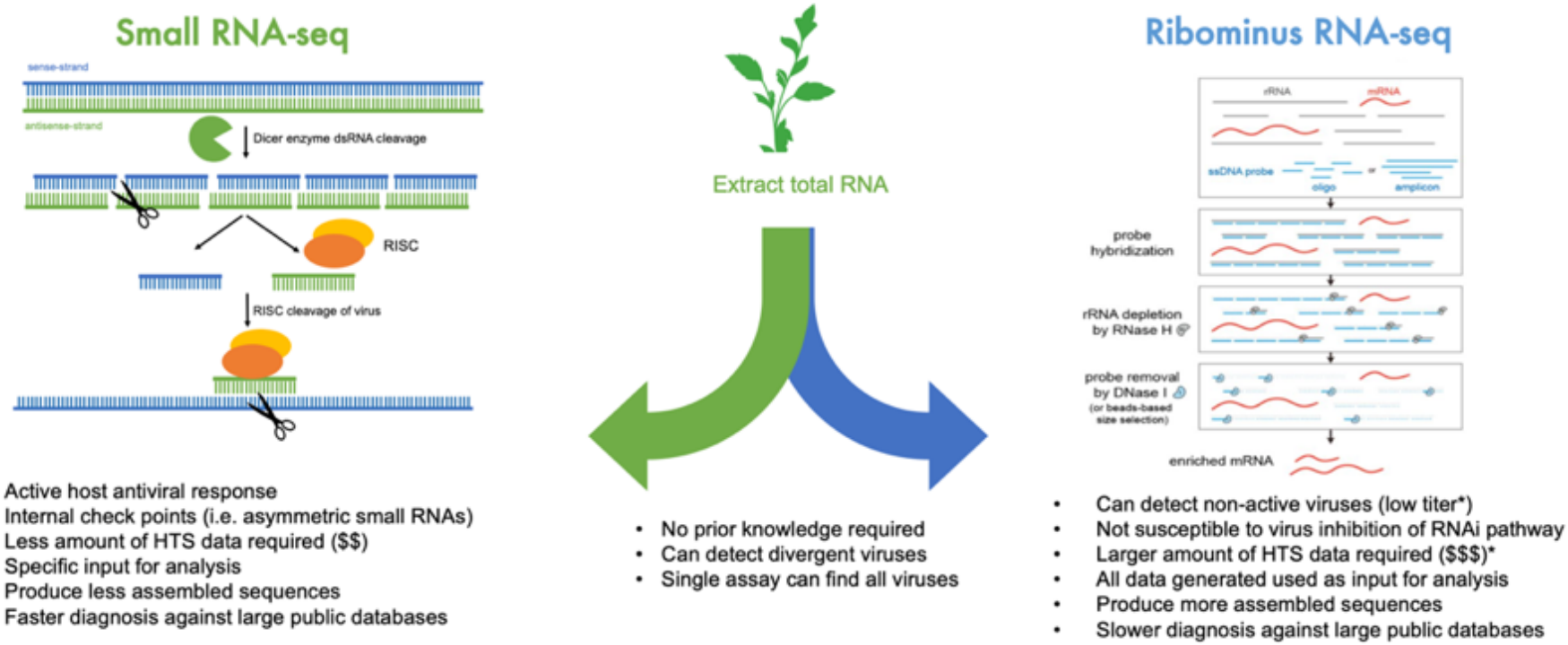
Based on the pilot side-by-side PEQ vs HTS comparisons, the small RNA-seq HTS method is recommended for further large-scale side-by-side studies. The ability to detect all known PEQ viruses and show minimal or no impact of sample cross-contamination provides a competitive advantage over the tested ribosomal depleted RNA-seq HTS method.

Report on additional non- quarantinable viruses if present in quarantined plant samples can assists plant pathologist and industry stakeholders to make more informed regulatory and business decisions.

From a technical point of view, we recommend:

- use a de novo genome assembly approach for HTS diagnosis rather than mapping onto viral genomes as such approach minimises or avoids the impact of index hopping events.
- collect at least 40 million small RNA sequences (also called reads) per sample when conducting virus and viroid diagnosis using HTS data generated by AGRF.
- use small RNA products derived from the host plant immune response against viruses to reconstruct viral genome sequences as not only increases specificity, but also documents if an active virus/viroid is present in quarantined plant samples.

Figure 5.2: Overview of HTS methods. *Small RNA-seq* enables to collect small fragments of genetic information derived from cleavage of virus by host plant enzymes called Dicer. *Ribosomal depleted RNA-seq* (ribominus RNA-seq) discards abundant ribosomal RNA sequences to collect genetic information of interest such as those derived from virus and viroids. Ribosomal RNAs account for ~80% of total RNA fractions, hence, its removal is desirable to improve diagnostic results. Features of both and each individual HTS method are highlighted as dot points.



- Active host antiviral response
- Internal check points (i.e. asymmetric small RNAs)
- Less amount of HTS data required ($$)
- Specific input for analysis
- Produce less assembled sequences
- Faster diagnosis against large public databases

- No prior knowledge required
- Can detect divergent viruses
- Single assay can find all viruses

- Can detect non-active viruses (low titer*)
- Not susceptible to virus inhibition of RNAi pathway
- Larger amount of HTS data required ($$$)*
- All data generated used as input for analysis
- Produce more assembled sequences
- Slower diagnosis against large public databases

Table 5.2: Comparison of PEQ vs small RNA-seq diagnosis for known viruses/viroids in 12 control plants.

Sample	Commodity	Virus	Summary			BGI				AGRF			
			PEQ	AGRF	BGI	accession	Length (bp)	coverage	identity	accession	Length (bp)	coverage	identity
MT001	Citrus	CEVd	YES	YES	YES	KY110721	372	1	98.118	KY110721	372	1	98.473
MT002	Peach	PNRSV RNA1	YES	YES	YES	MH427284.1	1,988	0.774	99.501	MH727235.1	974	0.292	99.179
		PNRSV RNA2	YES	YES	YES	MH427285.1	1,910	0.99	98.75	MF069044.1	915	0.277	99.344
		PNRSV RNA3	YES	YES	YES	MF069044.1	1,647	0.498	99.46	MH427285.1	1755	0.91	98.69
MT003	Citrus	CTV	YES	YES	YES	EU857538	16,232	0.843	99.271	EU857538	9104	0.473	99.265
		CVEV	YES	YES	YES	LC433635	1,595	0.267	98.934	LC360112	99	0.017	100
		HSVd	YES	YES	YES	MG779543.1	161	0.729	99.379	MG779543.1	148	0.67	100
MT004	Citrus	CTV	YES	YES	YES	MF595989.1	1,851	0.096	98.217	KP268199	377	0.885	99.735
		CEVd	YES	YES	YES	KY110721	372	1	97.959	DQ431991	322	0.868	99.069
		HSVd	YES	YES	YES	EU382209	290	0.967	99.31	MF979531.1	297	1	96.584
MT005	Rubus	RYNV	YES	YES	YES	KF241951	2,788	0.351	98.358	KF241951	1261	0.159	98.573
		RBDV	YES	YES	YES	KJ007639	1,417	0.26	97.955	FR687353	249	0.046	94.378
		RBDV	YES	YES	YES	KY417871	1,591	0.734	99.626	KY417868	608	0.28	99.836
MT006	Rubus	RYNV	YES	YES	YES	KM078034	540	0.069	86.111	KM078034	220	0.028	83.71
MT010	Nursery	SrMV	YES	YES	YES	KM025044	5,663	0.588	76.138	KM025044	1372	0.143	80.126
MT013	Strawberry	SMoV RNA1	YES	YES	YES	KU200453	3822	0.545	96.172	KU200453	761	0.108	96.189
		SMoV RNA2	YES	YES	YES	KU200461	3763	0.596	96.769	KU200457	864	0.137	97.338
MT014	Strawberry	SMoV RNA1	YES	YES	YES	KU200453	3590	0.512	95.919	KU200458	280	0.04	96.086
		SMoV RNA2	YES	YES	YES	KU200461	2121	0.336	96.182	KU200457	250	0.04	98.8
MT016	Sweet potato	SPFMV	YES	YES	YES	MG656429.1	3382	0.313	98.587	MG656427.1	971	0.09	98.97
MT018	Vitis	GFkV	n.a.	YES	YES	AJ309022	5024	0.664	92.114	AJ309022	3260	0.431	92.011
		GSYSVd-1	n.a.	YES	YES	KU880712	367	1	98.201	KU880712	367	1	98.201

		HSVd	n.a.	YES	YES	MF979531.1	297	1	96.978	MF774865.1	298	1	99.664
		GRSPaV	n.a.	YES	YES	KC427107	123	0.014	95.122	AY881627	466	0.053	97.21
		GLRaV	n.a.	YES	YES	HQ828135	97	0.12	91.919	FJ467504	6626	0.48	88.329
MT019	Vitis	GFkV	n.a.	YES	YES	AJ309022	6140	0.812	91.149	AJ309022	3726	0.493	91.118
		HSVd	n.a.	YES	YES	MF774873.1	297	1	100	MF774869.1	224	0.754	100
		GLRaV	n.a.	YES	YES	HQ828114	123	0.152	93.496	FJ467504	4047	0.293	86.74

CEVd = Citrus exocortis viroid; PNRSV = Prunus necrotic ringspot virus; CTV = Citrus tristeza virus; CVEV = Citrus vein enation virus; HSVd = Hop stunt viroid; RYNV = Rubus yellow net virus; RBDV = Raspberry bushy dwarf virus; SrMV = Sorghum mosaic virus; SMoV = Strawberry mottle virus; SPFMV = Sweet potato feathery mottle virus; GFkV = Grapevine fleck virus; GSYSVd-1 = Grapevine yellow speckle viroid 1; GRSPaV = Grapevine rupestris stem pitting-associated virus; GLRaV = Grapevine leafroll-associated virus; GFkV = Grapevine fleck virus; n.a. = not tested; coverage = genome coverage where a value of 1.0 equals 100% coverage; identity = percentage nucleotide sequence identity of assembled scaffold when compared to a database entry.

Table 5.3: Comparison of PEQ vs ribosomal depleted RNA-seq diagnosis for known viruses/viroids in 12 control plants.

Sample	Commodity	PEQ detected	BGI			AGRF			
			Accession	Genome length (nt)	Coverage (%)	Accession	Genome length (nt)	Coverage (%)	Average identity (%)
MT001	Citrus	CEVd	AB054596	372	99.7	JX259396	371	0.782	100
MT002	Prunus	PNRSV RNA1	MF069044.1	3,308	100	MF069044.1	3308	0.874	99.197
		PNRSV RNA2	MH727229.1	2,591	35.5	MH427284.1	2570	0.723	99.785
		PNRSV RNA3	S78312	1,943	83.9	MH730938.1	1943	0.058	99.115
MT003	Citrus	CTV	EU857538	19,251	100	EU857538	19251	0.723	99.175
		CVEV	LC089851	5,983	98.3	LC089851	5983	0.991	98.5
		CVdIIIA	KU861554	294	71.1	S75465	294	0.864	100
		HSVd	n.d.	n.d.	n.d.	KY654683	299	0.987	98.653
MT004	Citrus	CTV	EU937519	19,280	56.4	EU937519	19251	0.618	98.37
		CEVd	DQ431991	371	99.2	EU512994	372	0.831	99.677
		HSVd	EU382209	264	97.7	MF979531.1	297	0.997	96.346
MT005	Rubus	RYNV	KF241951	7,932	68.1	KF241951	7932	0.907	97.952

		RBDV RNA1	KJ007639	5,449	55.5	KJ007639	5449	0.753	96.558
		RBDV RNA2	KY417868	2,169	89.7	KJ007640	2231	0.242	98.804
MT006	Rubus	RYNV	KF241951	7,932	56.9	KF241951	7932	0.494	87.02
MT010	Nursery	SrMV	KJ944752	9,627	44.6	KM025045	9624	0.636	73.597
MT013	Strawberry	SMoV RNA1	KU200453	7,017	88.5	KU200458	7037	0.848	96.332
		SMoV RNA2	KU200457	6,327	87.3	KU200457	6327	0.593	96.618
MT014	Strawberry	SMoV RNA1	KU200458	7,037	84.8	KU200458	7037	0.848	96.215
		SMoV RNA2	KU200457	6,327	55.4	KU200457	6327	0.832	96.753
MT016	Sweet potato	SPFMoV	MF572048.1	10,777	90.2	MF572047.1	10824	0.999	98.262
MT018	Vitis	GRSPaV	KX274277	8724	99.8	KX274277	8724	0.96	93.708
		GLRaV	FJ467504	13807	22	FJ467504	13807	0.48	88.329
		GFkV	AJ309022	7564	71.9	AJ309022	7564	0.648	92.711
		GYSVd	KU880714	366	99.7	KU880714	366	0.995	98.907
		HSVd	MF774873.1	297	99.7	HE575348	264	0.924	91.703
MT019	Vitis	GFkV	AJ309022	7564	89.5	AJ309022	7564	0.854	91.06
		GLRaV	AM494935	4319	11.9	FJ467504	13807	0.293	86.74
		HSVd	MF774873.1	297	99.7	HE575348	264	0.996	99.245
		GRSPaV	KX274277	8724	99.7	KX274277	8724	0.999	93.888

CEVd = Citrus exocortis viroid; PNRSV = Prunus necrotic ringspot virus; CTV = Citrus tristeza virus; CVEV = Citrus vein enation virus; HSVd = Hop stunt viroid; RYNV = Rubus yellow net virus; RBDV = Raspberry bushy dwarf virus; SrMV = Sorghum mosaic virus; SMoV = Strawberry mottle virus; SPFMV = Sweet potato feathery mottle virus; GFkV = Grapevine fleck virus; GSYSVd-1 = Grapevine yellow speckle viroid 1; GRSPaV = Grapevine rupestris stem pitting-associated virus; GLRaV = Grapevine leafroll-associated virus; GFkV = Grapevine fleck virus; n.a. = not tested; coverage = genome coverage where a value of 1.0 equals 100% coverage; identity = percentage nucleotide sequence identity of assembled scaffold when compared to a database entry. n.d. = not detected.

Table 5.4: Detection of additional viruses/viroids and sample cross-contamination by small RNA-seq. Cross-sample contamination, also known as index hopping, was recorded in two samples. * Dr Candace Elliot confirmed that highlighted index hopping events were indeed false positives using PCR assays (data not shown). Samples with no new viruses detected are not shown.

Sample	Commodity	Virus name	Accession	Genome length (nt)	Coverage (%)	Comment
MT003	Citrus	Citrus endogenous para-retrovirus	KF800043	6996	2.9	
MT007	Citrus	Citrus virus A CVA	MG764566.1	2740	21.2	
MT008	Citrus	Citrus virus A CVA	MG764565.1	6691	3.7	
	Citrus	Hop stunt viroid HSVd	EF126046	299	99.34	
	Citrus	Sugarcane mosaic virus SrMV (partial)	KX266876	1380	12.4	index hopping*
	Citrus	Citrus endogenous para-retrovirus	KY609920	4089	2	
MT011	Citrus	Tobacco ringspot virus satellite RNA	M14879	359	55.7	index hopping*
MT012	Nursery	Cherry rasp leaf virus RNA1	KT944021	6872	36.1	
MT013	Strawberry	Strawberry vein banding virus	KX950836	7,854	22	
	Strawberry	Strawberry vein banding virus	X97304	7876	35	
MT014	Strawberry	Beet pseudo-yellows virus RNA1	AY330918	8006	62.4	
	Strawberry	Beet pseudo-yellows virus RNA2	AY330919	7903	68.5	
MT015	Strawberry	Strawberry vein banding virus	X97304	7876	26.4	
	Sweet potato	Sweetpotato badnavirus B	KM009097	642	19.5	
MT016	Sweet potato	Sweetpotato badnavirus C	JQ902104	438	27.2	
	Sweet potato	Sweetpotato symptomless mastervirus	KY565234	2602	28.1	

Table 5.5: Detection of additional viruses/viroids and sample cross-contamination by ribosomal depleted RNA-seq. Index hopping events were detected in 14 out of 17 samples using BGI data. Samples with no new viruses or index hopping events are not shown.

Sample	Commodity	Virus name	Accession	Genome length (nt)	Coverage (%)	Average identity (%)	sRNA-seq?
MT001	Citrus	Citrus endogenous pararetrovirus strain SPBR01 polyprotein gene, complete cds	KY609920	4089	0.293	88.293	YES
		Prunus necrotic ringspot virus isolate Pea4 segment RNA3, complete sequence	MH730939.1	1943	0.362	99.29	
		31.4 kda MP, CP=24.9 kda coat protein {RNA3} [prunus necrotic ringspot ilarvirus PNRSV]	S78312	1943	0.267	99.037	
		Raspberry bushy dwarf virus isolate RR7 coat protein gene, complete cds	KY417879	1033	0.212	99.545	
MT002	Peach	Raspberry bushy dwarf virus strain Can-S coat protein mRNA, complete cds	AF259798	825	0.258	99.533	
MT004	Citrus	Prunus necrotic ringspot virus isolate PNRSV-AL1 movement protein mRNA, complete cds	EU849602	852	0.308	99.62	
MT005	Rubus	Prunus necrotic ringspot virus isolate PNRSV-AL1 movement protein mRNA, complete cds	EU849602	852	0.555	99.375	
MT006	Rubus	Prunus necrotic ringspot virus isolate PNRSV-AL1 movement protein mRNA, complete cds	EU849602	852	0.297	99.606	
MT007	Citrus	Sorghum mosaic virus isolate YN-YZ151, complete genome	KM025039	9624	0.025	76.349	
		Potato virus Y isolate prime-e5, complete genome	MG602675.1	9699	0.019	75.661	
		Citrus virus A isolate W4 segment RNA1, complete sequence	MG764565.1	6691	0.916	95.66	YES
MT008	Citrus	Maize dwarf mosaic virus partial polyprotein, isolate Mv0811, genomic RNA	FM883174	1310	0.239	85.987	
		Citrus virus A isolate W4 segment RNA1, complete sequence	MG764565.1	6691	0.872	95.668	YES
		Citrus endogenous pararetrovirus strain SPBR01 polyprotein gene, complete cds	KY609920	4089	0.121	92.154	

		Tobacco ringspot virus isolate CZ 1104987 capsid protein gene, partial cds	KP144328	1545	0.137	97.183	
		Sorghum mosaic virus isolate YN-YT168	KM025045	9624	0.367	74.101	
		Sugarcane mosaic virus isolate JAL-1	GU474635	9628	0.161	71.573	
MT010	Nursery	Potato virus Y isolate Poha 3 polyprotein gene, complete cds	MF134863.1	9635	0.005	95.652	
MT011	Citrus	Plum pox virus RNA, complete genome, strain: PPV-D, isolate: Ou6	AB576062	9736	0.017	76.786	
		Tobacco Ringspot virus RNA1 polyprotein gene, complete cds	U50869	7514	0.076	95.28	
		Citrus endogenous pararetrovirus strain SPBR01 polyprotein gene, complete cds	KY609920	4089	0.371	88.043	
		Maize dwarf mosaic virus gene for polyprotein, isolate SP, genomic RNA	AM110758	9523	0.028	79.924	
		Tobacco ringspot virus isolate SK segment RNA2, complete sequence	KJ556850	3928	0.394	95.297	
		Potato black ringspot virus isolate PV0056 (2) segment RNA 1	KC832890	7579	0.03	74.348	
		Cherry rasp leaf virus segment RNA-2	AY122330	3318	0.128	87.907	
MT012	Nursery	Mirafiori lettuce big-vein virus isolate LS301-O putative RNA polymerase and 25K protein genes	AF525933	7794	0.217	69.861	
		Cherry rasp leaf virus segment RNA-1	AY764390	7030	0.654	84.06	YES
MT013	Strawberry	Sugarcane mosaic virus isolate CBMungo helper component proteinase gene, partial cds	KX266876	1380	0.246	75.145	
		Potato virus Y isolate prime-e5	MG602675.1	9699	0.019	75.676	
		Sorghum mosaic virus isolate HN-YZ49	KM025051	9627	0.062	77.595	
		Strawberry vein banding virus complete genome	X97304	7876	0.7	95.125	YES
MT014	Strawberry	Beet pseudo-yellows virus segment RNA 1, complete sequence	AY330918	8006	0.753	99.37	YES
		Strawberry vein banding virus complete genome	X97304	7876	0.662	94.87	YES

MT015	Strawberry	Sugarcane mosaic virus genomic RNA for polyprotein gene, isolate Lingpin	AJ310102	9578	0.129	71.598	
		Pennisetum mosaic virus isolate CD1	JX070152	9613	0.023	76.106	
		Tobacco ringspot virus isolate TXJ-TC-5 coat protein gene, partial cds	KY304038	1013	0.208	97.17	
		Sorghum mosaic virus isolate FJ-CI20031	KM025052	9651	0.142	76.394	
		Maize dwarf mosaic virus polyprotein, isolate Sz0605, genomic RNA	FM883211	9515	0.207	74.13	
		Strawberry vein banding virus complete genome	X97304	7876	0.526	95.068	YES
MT016	Sweet potato	Sugarcane mosaic virus isolate CBMungo helper component proteinase gene, partial cds	KX266876	1380	0.279	75.255	
		Pennisetum mosaic virus isolate B, complete genome	AY642590	9611	0.033	74.534	
		Potato virus Y isolate prime-e5, complete genome	MG602675.1	9699	0.02	75.635	
		Maize dwarf mosaic virus partial polyprotein, isolate Sz0807, genomic RNA	FM883187	1298	0.191	85.542	
		Rhabdovirus YHG-2013 isolate IbRNLS coat protein gene, partial cds	JQ902107	981	0.988	99.588	
		Sorghum mosaic virus isolate HN-YZ49, complete genome	KM025051	9627	0.05	76.316	
MT018	Vitis	Prunus necrotic ringspot virus isolate Pea4 segment RNA3, complete sequence	MH730939.1	1943	0.208	99.753	
		Erysiphe necator mitovirus 3 isolate gpm 6 RNA-dependent RNA polymerase gene, complete cds	KY420040	2892	0.937	98.122	
MT019	Vitis	Prunus necrotic ringspot virus isolate PNRSV-AL1 movement protein mRNA, complete cds	EU849602	852	0.386	99.394	

Table 5.6: Detection of additional viruses/viroids and sample cross-contamination by ribosomal depleted RNA-seq. Index hopping events were detected in 6 out of 12 samples using AGRF data. Index hopping events shown in grey background. Putative hits to grape powdery mildew fungus shown in green background. Samples with no new viruses or index hopping events are not shown.

Sample	Commodity	Virus name	Accession	Genome length (nt)	Coverage (%)	Average identity (%)
MT001	Citrus	Citrus endogenous para- retrovirus clone CB17-3	KF800045	6663	0.036	89.583
MT003	Citrus	Prunus necrotic ringspot virus isolate PNRSV-RNA2-SA polymerase protein gene	MH427284.1	2570	0.081	99.524
		Prunus necrotic ringspot virus isolate Pea1 segment RNA3	MH730938.1	1943	0.139	99.262
MT005	Rubus	Grapevine associated narnavirus-1 RdRp gene for RNA dependent RNA polymerase	LN827948	2735	0.593	82.993
		Maize dwarf mosaic virus gene for polyprotein, isolate SP, genomic RNA	AM110758	9523	0.013	90.4
		Prunus necrotic ringspot virus isolate Q15R3N segment RNA3	KY883335	2045	0.083	90.058
MT006	Rubus	Prunus necrotic ringspot virus isolate Q15R3N segment RNA3	KY883335	2045	0.055	90.402
MT010	Nursery	Sorghum mosaic virus isolate YN-YT168, complete genome	KM025045	9624	0.636	73.597
MT013	Strawberry	Strawberry vein banding virus complete genome	X97304	7876	0.96	93.015
		Prunus necrotic ringspot virus isolate Q15R3N segment RNA3	KY883335	2045	0.071	92.466
		Grapevine associated narnavirus-1 RdRp gene for RNA dependent RNA polymerase	LN827948	2735	0.105	83.162
MT014	Strawberry	Strawberry vein banding virus complete genome	X97304	7876	0.96	92.976
		Beet pseudo-yellows virus segment RNA 2, complete sequence	AY330919	7903	0.999	99.311
		Beet pseudo-yellows virus segment RNA 1, complete sequence	AY330918	8006	0.981	99.393
		Prunus necrotic ringspot virus isolate Q15R3N segment RNA3	KY883335	2045	0.079	93.807
MT016	Sweet potato	Sweet potato symptomless virus 1 isolate LVVF-SPSM1-1, complete genome	MG680260.1	2602	0.248	100
		Sweet potato badnavirus B isolate Yanshu1 China 74(2) polyprotein gene, partial cds	KM009092	808	0.146	89.076
		Sweet potato badnavirus C isolate SCCD polyprotein gene, partial cds	JQ902104	438	0.196	95.402

MT018	Vitis	Erysiphe necator mitovirus 3 isolate gpm 6 RNA-dependent RNA polymerase gene	KY420040	2892	0.991	98.047
		Erysiphe necator mitovirus 1 isolate gpm 4 RNA-dependent RNA polymerase gene	KY420038	2571	0.988	99.253
MT019	Vitis	Grapevine Pinot gris virus isolate TI 25 RNA-dependent RNA polymerase gene	KP693445	7172	0.011	98.701
		31.4 kda movement protein, CP=24.9 kda coat protein {RNA3} PNRSV	S78312	1943	0.121	100
		Erysiphe necator mitovirus 1 isolate gpm 4 RNA-dependent RNA polymerase gene	KY420038	2571	0.214	98.732
		Erysiphe necator mitovirus 3 isolate gpm 6 RNA-dependent RNA polymerase gene	KY420040	2892	0.187	97.237

Appendix 6 – National Plant Biosecurity NGS/NGS Workshop

Date: 11-12 February 2020

Venue: Queensland University of Technology

6.1. Executive summary

In order to provide high quality NGS diagnostic testing in biosecurity across Australia and New Zealand, this workshop engaged managers of pathogen testing schemes, diagnosticians, policy regulators and industry leaders in Australia and New Zealand to identify and discuss the requirements of a framework that may be applied to different commodities and industries and provide the basis of national standards. Representatives from different research and diagnostic organisations presented and discussed their workflows, objectives, and challenges. This provided a basis of considerations that are common between different users of NGS sequencing and analysis with which to discuss standards required to make a framework, not prescriptive protocols that may be applied specific commodities or industries.

6.2. Attendees

Name	Location	Area of Expertise	Details
Adrian Dinsdale	Fed Dept Agriculture	PLANT	Operational policy, QUADS rep
Alison Dann	DPIPWE TAS	PLANT	Molecular diagnostician, NGS user
Amy Jennison	Qld Health	HUMAN	NGS user
Brendan Rodoni	AgVic	PLANT	SPHD
Cheryl Jenkins	NSW DPI (Elizabeth Macarthur Ag Institute)	ANIMAL	Bacteriologist, AUSGEM
Craig Smith	QDAF, BQ	ANIMAL	NGS user
David Warrilow	Qld Health	HUMAN	NGS user
Douglas Beattie	QUT	PLANT	HIA NGS project, experienced NGS user
Elizabeth Czislowski	QDAF, BQ		
Gareth Price	QFab & UQ		Facilitator - Head of Computational Biology
James Cunnington	Fed Dept Agriculture	POLICY	Plant Science and Risk Assessments
Jane Oakey	QDAF, BQ	ANIMAL	Entomologist, NGS user
Juliane Henderson	QDAF, BQ	PLANT	Molecular diagnostics plant pathogens
Julie Pattemore	Fed Dept Agriculture		
Kim Halpin	CSIRO	ANIMAL	LEADDR
Louise Jackson	QDAF, BQ		
Lucy Tran-Nguyen	CDU / NT DPIR	ANIMAL	NGS user/policy regulator
Malcolm Letts	QDAF, BQ		
Mark Blacket	AgVic	PLANT	Entomologist, metabarcoding
Mark Whattam	Fed Dept Agriculture	PLANT	Operational policy
Matthew Neave	AAHL CSIRO	ANIMAL	Bioinformatics, Lab manager
Melanie O'Keefe	AGRF		NATA-accredited for some NGS processes
Melanie Blankenberg	QDAF, BQ	ADMIN	Administration support
Monica Kehoe	DPIRD WA	PLANT	Plant pathologist
Paul Campbell	QDAF, BQ	PLANT	NGS diagnostics user
Rikki Graham	Qld Health		
Roberto Barrero	QUT	PLANT	Bioinformatician, NGS Project Leader
Sam Miller	QDAF, BQ	ADMIN	Administration support
Samuel Hair	DPIRD WA	ANIMAL	Microbiologist, NGS user
Shannon Rogers	QDAF, BQ	ADMIN	Administration support
Tegan Ludzioweit	Fed Dept Agriculture	POLICY	Plant Health Policy

6.3. Program and introductory material

6.3.1. Workshop program

Agenda and check list/ deliverables

Day 1			
Start	**Item**	**Presenter**	**Checklist / deliverables**
9.15 AM	Registration		
9:45 AM	Event Opening	Malcom Letts	
9:59 AM	Vote on Dinner	Gareth Price	• Location determined (Gareth to make a slide)
10:00 AM	Introduction and context	Gareth Price	• Attendee's Who's Who • Name / organisation • One question: With reference to NGS: what do you bring? • Max: 30secs each • List definitions for the workshop • Glossary of terms • My experience of NGS in a regulatory environment, and the standards we utilised • Conflict of Interest • Context for the workshop • What's to be achieved • Attendees to sculpt a framework for NGS Standards • Human doesn't equal plant
10:30 AM	What is a NGS Standard?	Brendan Rodoni	• The environment a diagnostic lab works in • What is the logical flow to escalate sample to NGS • Define what is a standard • Write on white board • Return to this throughout the 2 days and refine or modify • Breakdown the steps of NGS pipeline, put it up on a WB and return to it again and again
11:00 AM	NGS Experience: Controls and Workflow Needs	Monica Kehoe (Plants)	• Each presentation is 10-15min to allow for lots of Q&A Past experiences and how improvements were made, with 2 or 3 examples • Outcome: Identify synergies between activities • Outcome: Identify the challenges faced by each sector • Outcome: Quality Assurance 'themes' will start to emerge • Outcome: Draw together a summary of collated experiences • How has NGS impacted on health-related testing? • Are there differences in human and human related pathogen standards and workflows required?
11:20 AM		Matt Neave (Animal)	
11:40 AM		Mark Blackett (Meta Barcoding – Plants)	
12:00 PM	Human Health Experience with NGS	Rikki Graham	
12:45 PM	Lunch		
1:30 PM	NGS User Survey Results and Discussion	Kim Halpin	• Presentation of Survey Results • Once presented: • Does the room agree? • Would they change practice based on what their national colleagues are doing or have access to? • Would a follow-up survey be of value?

2:30 PM	Activity 1: NGS Requirements (e.g. sampling; template, QC; computational and IT challenges; bioinformatic pipelines; reporting)	All	• Context: End of day 2 you ALL will be agreeing on the requirements tabled and debated now • Discuss the minimum requirements we would like to see for each of the following (and therefore what Quality Controls would be put in place at this step) – template (total nucleic acid, amplified) 1: Quality measurement via micro capillary electrophoresis (Tapestation) 2: Positive and Negative standards for data generation tracking from extraction to data output 3: Ancillary testing (mini SNP panels?) 4: Other library prep (use of TapeStation and qPCR) 1: Uniform use of library protocol? 2: Inhouse vs Sequencing supplier 3: As with Template, required positive and negative controls (for example, barcodes with water?) 4: Other sequencing platform 1: Standards for short read, spike-ins, dosage controls, SNV controls 2: Standards for long read, spike-ins, dosage controls, SNV controls 3: Standards for outputs (data generation) 4: Other computational and IT infrastructure • 1: In-house, national or cloud • 2: Standards for security • 3: Is there standard hardware? • 4: Other • bioinformatic pipelines • 1: In-house vs commercial software? • 2: What Standard to use to determine tool accuracy? • 3: When to validate • 4: Other report details (see below) ▪ 1: Pan species? ▪ 2: National, state or organisational templates ▪ 3: Are health templates a good basis to start ▪ 4: Other • Mechanistically: room divided into two working groups led by Monica (Template, Library, Sequencer) and Roberto (IT, Bioinformatics, Reporting) with report back to room • Gareth to draft 3 dot points per topic plus an “other”
3.30 PM	Afternoon tea		
4:00 PM-	Activity 1: NGS Requirements (e.g. sampling; template, QC; computational and IT challenges; bioinformatic pipelines; reporting)		• Continued from above • 4:30pm: Summary by Monica and Roberto back to the room
5:00 PM	Day 1 closeout With Gareth		
Evening	Dinner		
Day 2			

9:00 AM	Dealing with unknown sequencing data	All	• The significance of 'unknowns' might not be realised for years to come, but there is a responsibility of the lab to ensure the finding is not lost. • Reporting structures • Storing data • Sharing data • How do these fit into the framework and requirement for standards? Mediated via white boards, roaming mic (or hand up) Encourage room to use examples from their practice
10:00 AM	What is needed for NGS framework / guidelines (brainstorm)	All	• Gareth to give brief (5min) summary of Day 1 Activity One • Return to Day 1 whiteboard, are we building a framework for NGS standards?
10:45 AM	Morning Tea		
11:15 AM	Activity 2: What should be in a checklist for NGS providers	Melanie O'Keefe All	• AGRF has NATA accredited facilities, let's hear from them • What minimum standards do they promise? • What questions do they want client to ask them? • How within the drafted framework can we assure • The use of outsourced platforms and services for diagnostic services meeting all necessary the standards. • If part of the NGS process is to be outsourced, NATA accredited providers or providers showing full compliance with ISO and/or NATA standards should be used. • Where does the responsibility lie on the submitting laboratory to review, retain and make available for audit all documentation related to NGS testing.
11:30 AM	Agreement on a NGS standard, based on the framework/guidelines	All	• The finer details of the standard will take time to develop. The goal here is to identify what we want to see in the standard. ○ template (total nucleic acid, amplified) ○ library prep ○ sequencing platform ○ computational and IT infrastructure ○ bioinformatic pipelines ○ report details (see below) • Is the Glossary of Terms complete?
12.30 PM	Lunch		
1:15 PM	Australian NGS network: What could it look like - health experience	Rikki Graham	• How did QHealth-FSS establish / build their networks of metadata sharing • The goal of these talks is to show how a networked approach can help create a community of practice and encourage sharing of samples and data. • What practices would you have changed year ago, that made sharing / community more difficult • Reciprocal Reference laboratories
1:45 PM	Australian NGS network: Animal LEADDR network, standards &	Kim Halpin	

	proficiency testing		
2:30 PM	Report Structure, recommendations and next steps	Brendan, Mark, Kim	
3:00 PM	Close		

6.3.2. Key requirements of a Genomics Report:

The minimum suggested content for a report is described below. The following list is not a recommendation for the structure but rather points that should be addressed. Some could be ordered to provide a one page summary of important results relevant to the disease investigation, followed by additional pages detailing the test and any further recommended tests or results that may be appropriate.

Report Details
• Reporting laboratory details
• Title of report
• Report status and report authorisation
• Issue date and time of report
• Submitter's name and address
• Unique laboratory identifier
• Animal or sample identification
• Sample type (blood, tissue and site, fluid)
• Secondary specimen identifier (referring laboratory identifier)

Test description
• Test category
• Purpose of test (e.g to assist in the diagnosis of … or the exclusion of…)
• Methodology used including confirmation by an orthogonal method (such as Sanger sequencing or real time PCR) if performed
• Limitations to test including any remaining uncertainty where it exists; this would include a narrative around any QC failures (i.e. outside acceptance criteria). It may be possible to use a confidence score.

Result summary
• Reference sequences including genome build or reference sequence version
• Variant reporting policy for the reporting laboratory that complies with relevant guidelines

Interpretive comment
• Narrative comment indicating the identified sequence
• If applicable the need for follow up or confirmatory testing should be indicated on the report.
• If this is the first detection of a disease agent, interpret cautiously seeking input and directive from the relevant authority.
• Notification may include taking results to the relevant biosecurity authority (including CVO/CPO).
• The lab should include recommendations for appropriate follow-up in reports. In situations in which further studies may be warranted (e.g. testing of other tissues, or other animals), these recommendations should be included in the test report.

Outputs:

1. Report
2. Horizon scan of infrastructure and software ownership, NGS skills gaps => information obtained through a survey which will be run and analysed before the workshop. Results will be presented at the workshop.
3. Identification of key components of NGS guidelines
4. A checklist of questions to ask NGS providers
5. Proposal for a network model
6. Where to from here

6.4. Issues identified and discussed

- Standards
- Controls
- Replication

- Assay Validation
- Computing
- Data and metadata storage, sharing, and security

6.4.1. Standards

Core standards in diagnostics include reproducibility, reliability, accuracy, sensitivity, and specificity. Diverse industries and commodities were represented at the workshop, and relevant international standards and regulatory groups include:

- International Plant Protection Convention (IPPC)
    - IPPC guide to delivering phytosanitary diagnostic services 2016
    - IPCC subcommittee for plant health diagnostics (IPPC – SPHD).
- WHO
- ISO 17025
- Animal Health Committee (AHC) Subcommittee on Animal Health Laboratory Standards (SCAHLS)
- Laboratories for Emergency Animal Disease Diagnosis and Response (LEADDR)
- Australian and New Zealand Standard Diagnostic Procedure (ANZSDP)
- National association of testing authorities (NATA)

6.4.2. Controls

Biological controls are derived from the initial biological samples, whereas technical controls only monitor the process of the assay. Different controls may be selected and involved in analysis at different stages. e.g. method/extraction positive control ensures all organisms are sufficiently extracted.

Positive controls may be derived from DNA, cells, or individuals, and the availability of reference standards and renewability over time may be a problem. Organism/cell reference standards are available for culturable organisms, and are now commercially available, an example being mock viral community controls such as ATCC MSA-2008. Bioinformatic standards should be available and renewable, but only test that part of the protocol.

Controls should be something that everyone can use, and do not occur naturally in the given context. As an alternative, some groups use surrogate samples as controls e.g. bovine serum.

“Spike-in” positive controls are sample-specific synthetic molecules that allow tracking of samples through entire complex workflow and can resolve sample mix-ups, normalise read depths for quantitative analysis, and quantify index switching. Spike-in controls should have known QC metrics, and comparable metrics to experimental material. It is debatable at which stage the spike-in should take place, e.g. pre- / post-extraction. An example of spike-in material is PhiX bacteriophage material, which is typically used at approx. 1% of a sample for internal controls.

Negative controls identify laboratory / reagent contamination, however removal of all taxa that occur in negative controls to monitor sample cross-contamination is problematic, as real taxa will be removed. Additionally, quantification in low DNA negative controls does not scale in proportion to experimental samples with more DNA, meaning that the results are not comparable. However, some groups sequence negative controls, and the results and the implications to the process are considered almost as important as experimental samples or positive controls themselves.

6.4.3. Replicates

Replicates are useful for identifying cross contamination in lab protocols. The number of replicates used is balanced between the risk of not replicating processes vs the cost. Research in environmental DNA (eDNA) uses several replicates. The temporal and spatial stochasticity of samples is often dramatically higher than between extraction or PCR replications, and the costs to process separate biological replicates are similar to separate technical replicates.

Studies into bulk communities (e.g. environmental DNA) may use fewer replicates, whereas assays with initial low levels of DNA employ relatively more PCR replicates.

6.4.4. Assay Validation

It costs similar amounts in time and money to target a large reportable range of organisms verses a very small range, so an assay should advantage of this and target the broadest range of organisms possible. DNA extraction methods should be confirmed to be non-destructive to provide confidence in experimental data. Assay validation involves proficiency testing, standard QAP programs, interlaboratory QAP’s, and standards in reporting silico results. An example of an international standard in ensuring valid assays for the detection of microbes is the Global Microbial Identifier QAP/UNSGM.

The sensitivity of NGS is proportional to the depth of sequencing, thus subsampling helps work out thresholds of detection. Current human pathogen testing surveillance programs require multiple genotyping methods per organism, and for medically relevant bacteria confirmatory PCR is recommended. Multiple sequencing technologies are beneficial in providing different and comparable results. E.g. large-scale NGS platforms such as Illumina HiSeq or NovaSeq may achieve high coverage, but smaller-scale platforms such as PacBio and Minion can provide portable results with longer read lengths. Combining these methodologies can provide confirmatory analyses and additional information. Combinations of bioinformatic analysis also provide benefits. For example, *de novo* assembly and alignment provides greater phylogenetic information from samples of mixed organisms, whereas mapping is recommended for culturable organisms or pure isolations, and a combination of the two methods can yield greater accuracy and confidence in the results.

6.4.5. Computational considerations

Human interpretation and assessment is not desirable as it is not reproducible, costly, inaccurate or inconsistent between batches or operators. Manual editing is subjective and time consuming, whereas coded parameters provide reproducibility and scalability. Further, manual checking of ambiguities is time consuming, can include subjective editing and steps not easily recorded or transferred.

Options for analysis include command line interfaces dealing with open access software such as that hosted on github. However, this requires greater operator knowledge of operating systems, program dependencies for installation, and low-level programming languages. Alternatively, graphic user interface (GUI) -based programs are easily usable by non-specialists, but require expensive licensing, are not necessarily widely used by other groups for compatibility or interoperability, and there are often limits in the power, versatility or flexibility of non-standardised processes. Many or most free-to-use bioinformatics programs run on linux-based architecture, and it is possible to install some type of linux environment alongside other operating systems on the same machine, or use virtual machines. Alternatives are online pipelines for bioinformatics such as Snakemake and Galaxy, and Nextstrain for visualisation

Cloud based processing involves communication with remote but secure servers to provide high performance computing (HPC) without the requirement of initial investments in hardware and maintainance. Cloud capabilities can be increased and decreased as desired, in order to increase processing power or reduce costs, respectively.

Static data storage formats such as spreadsheets are problematic due to location and sharing issues, and are error prone. Similarly, results such as .pdfs are time consuming to produce and are not interactive. It is preferable to aim for dynamic means of storing, presenting, and sharing data revolve around databases and web platforms that are maintainable, interactive, able to be used ongoing with pipelines, and provide immediate data and context.

6.4.6. Data and metadata sharing

Four principles underpin the guidelines “FAIR”, published in 2016. This was authored by FORCE11: an international group of researchers and funders aiming to improve the research publishing and communication system. These principles are that data and metadata should be:

- Findable: assigned a unique and persistent identifier, and registered in a searchable resource
- Accessible: retrievable by the identifier using a standardized protocol, which is open-source and free.
- Interoperable: stored in a formal, accessible, shared, and broadly applicable language for knowledge representation, and include qualified references to other (meta)data.
- Re-usable: have a plurality of accurate and relevant attributes and are released with a clear and accessible data usage license and meet the standards of the relevant industry.

6.4.7. Version control

Version control in analysis software can be difficult to track, however programs such as Conda are intended to automate the tracking of software versions and program parameters. Github is a resource intended for version control, tracking, and providing reproducible pipelines.

Depending on the type and consequence of software updates and changes to software that occur after pipeline validation, different actions may be taken: For low-level changes where version updates/changes have no impact on analysis or results, the version change is recorded. For mid-level changes where version changes alter some results or analyses, the process is re-verified by comparing results of previous samples to new versions. These outcomes are recorded and decision made if the new pipeline and component software should be retained. For high-level changes where version changes alter all analysis algorithms, results, and data handling, re-verification is performed on all processes by running previously analysed samples on the new pipeline to determine if new version/s should continue to be used.

6.4.8. Data storage and security

Different departments/agencies/states already have standards for data retention times (e.g. the standard for the Department of Agriculture is 7 years), but these differ between states and authorities. These standards may need to be re-evaluated or updated depending on the necessary to revisit samples and reanalyse data. In terms of economy, storing final results and the raw data is the most efficient, whilst also retaining the information for rapid access.

Security for different industries and commodities must be discussed with the relevant internal regulators. E.g. For medical samples, bacterial sequences are not under strict security guidelines and are readily available, whereas the metadata of the samples such as personal details of the human subjects are strictly controlled.

6.5. Conclusions

- An NGS analysis pipeline must be validated as fit-for-purpose. Early in the adoption of NGS protocols, these should be performed in tandem with confirmatory current methods, and with the confidence of successful NGS practices where results accord with traditional assays, NGS may become the primary analysis method.
- Bioinformatics is growing so quickly that organisations should place more resources in getting resident bioinformaticians. It may be desirable to up-skill current diagnosticians and lab operators, however bioinformatics specialists are indispensable and greater focus is required in bringing more into the diagnostic agencies.
- Considering that we are developing a framework, it must be considered that there will have to be different tests for different targets and purposes, and not aim for prescriptive protocols that may be applied to specific commodities or industries.

Appendix 7 – Draft NGS protocols for key culturable plant pathogenic bacteria

Agriculture Victoria - Plant Bacteriology

Introduction

Culturing pathogenic bacteria for whole genome sequencing is a means to isolate pathogen DNA from host and environmental DNA. Shallow short read sequencing (~500 MB) using the Illumina platform or long read sequencing using the Oxford Nanopore Technology (ONT) MinION sequencing of genomic DNA extracted from cultured bacteria is often sufficient for identification of many species and subspecies of bacteria. However, identification of pathogenic variants from closely related taxonomic groupings (e.g., subspecies, pathovar or biovar) may require greater depth of high accuracy sequence data than is generated by these approaches. There are also multiple bioinformatic tools that can analyse individual reads or require the variable or core regions from assembled genomic contigs for identification (Carrie Brady et al., 2008; Glaeser et al., 2015; Gupta et al., 2016).

In the following sections we describe a high throughput sequencing (HTS) pipeline to identify plant pathogenic bacteria from acquisition of bacterial cultures to bioinformatics data analysis as outlined in Figure 7.1. This includes protocols for DNA extraction, library preparation (Illuminia and ONT), sequencing and data analysis. Using these pipelines, we have sequenced and analysed gram-negative bacteria, including isolates from the *Pantoea* genus and isolates of *Pseudomonas syringae* pv. *actinidiae* biovar 3 (Psa3), as case studies. *Pantoea* species were selected to assess the Illumina pipeline because whilst they are easily culturable and sequences readily available online, species of this genus can be difficult to differentiate using standard laboratory tests including polymerase chain reaction (PCR) and often require multi locus sequence analysis (MLSA) for accurate identification, a method that is resource and time intensive for general diagnostics. Whole genome sequencing (WGS) and analysis tools can rapidly identify and differentiate between closely related *Pantoea* species and subspecies. As a part of this project we have assessed multiple bioinformatic pipelines and identified tools that can accurately differentiate and characterise closely related pathogenic strains. A second set of data was generated to assess the use of Illumina and ONT data to accurately identify bacterial species using Psa3 as a case study. Specifically, the taxonomic level that could be identified with this technology was assessed. Analysing Psa3 data sets generated using both ONT and Illumina sequencing, we found that ONT data was accurate to the pathovar level but biovar level detection was only achieved with polished ONT data. Illumina data was accurate to the biovar level with only 300 MB of sequencing.

DNA extractions

Many commercial DNA extraction methods (e.g. Qiagen DNeasy Blood & Tissue Kit, Promega Wizard® Genomic DNA Purification Kit, MagJET Genomic DNA Purification Kits etc) have been used to extract DNA for library preparation in the literature (Ashton et al., 2015; Chalupowicz et al., 2019; Gong et al., 2018). The Qiagen DNeasy Blood & Tissue Kit was used to extract DNA for Illumina library preparation in this work as it is the validated bacterial extraction kit in use at Agriculture Victoria Research (AVR). Variations amongst different commercial kit methodologies have been shown to have little effect on sequencing quality and coverage of gram negative bacterial genomes when using Illumnina sequencing (Becker et al., 2016). The Promega Wizard® Genomic DNA Purification Kit was used for ONT library preparation as this precipitation method causes less shearing of the DNA compared to column-based extraction (Becker et al., 2016) meaning higher molecular weight (HMW) DNA can be yielded. Other DNA extraction kits (including specific HMW extraction kits) can also be used but need to be effectively validated for DNA extraction quality suitable for each sequencing technology. For example, quantity of DNA is important as the amount of DNA can modify library preparation protocols. For this, the Quantus™ or Qubit fluorometer is suggested to measure the concentration of the extracted gDNA. Gel-electrophoresis or a TapeStation can be used to assess the quality of the DNA (e.g. shearing) which is important for long read sequencing such as ONT which works optimally with HMW DNA. Where possible, TapeStation should be the preferred method to assess DNA quality.

Controls to be used during this step is a negative control (blank) to assess cross contamination.

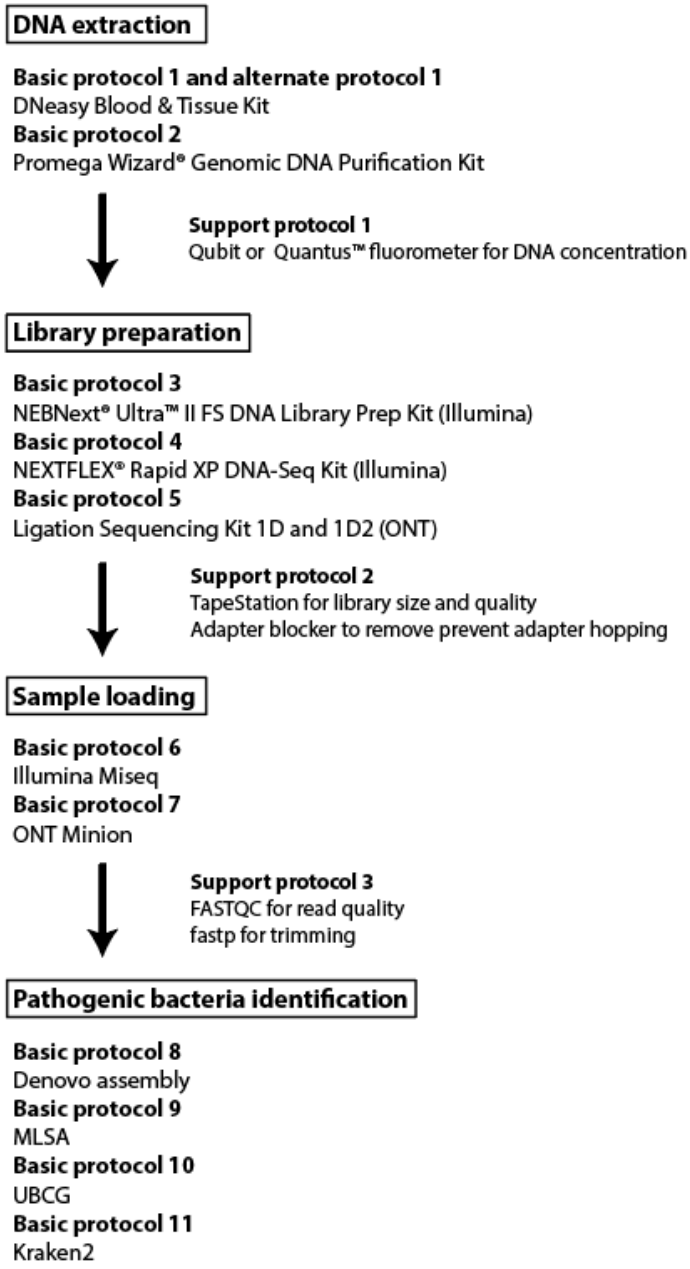


Figure 7.1 Pipeline used to identify culturable plant pathogenic bacteria

Library preparation

Illumina

There are multiple commercial kits that can be used to prepare libraries for Illumina sequencing. Essentially, the overall process begins with fragmentation of extracted DNA to appropriate sizes; between 150 – 300 bp for HiSeq and NovaSeq platforms and 300 – 600 bp for the MiSeq platform if using v3 kits. Most library preparation kits provide a sheering option which can include heat and enzymatic fragmentation. Alternate options include the Covaris g-TUBE; a single-use device that shears genomic DNA into selected fragments sizes ranging from 6 to 20 kb using a centrifuge. For some kits, an additional end repair step is added to accommodate fragmentation to ensure that DNA fragments have blunt ends for sequence adaptor ligation. Fragment size of libraries is an important consideration as insert sizes influence outputs of downstream bioinformatic analysis. For example, larger insert sizes can result in substantial assembly improvements (Huptas et al., 2016). Following sequence adapter ligation individual indexes are added using PCR in a library amplification step for multiplexing multiple samples in the one sequencing run. The workflow of most library preparations is similar, however, some kits have minor

modifications or patented technology in their work flow. For example, enzymes in the Nextera® DNA Library Preparation enable fragmentation and ligation of adapters (which they call tagmentation) in the same step and the NEXTFLEX Rapid XP DNA-Seq Kit which provides the adapter and indexes already together and only requires one ligation step.

Choice of library preparation kit for Illumina sequencing can be influenced by multiple factors. Firstly, the quantity and quality of the DNA can influence kit choice. Library preparation kits such as the NEBNext® Ultra™ II FS DNA Library Preparation Kit can accommodate down to 0.1 ng of DNA input and can be lenient with the quality of DNA. The cost effectiveness of library preparation kits can also influence their use in routine diagnostics. The Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit is known to be one of the cheaper commercial kits available and has potential to be used at half-reaction as shown in this report. Lastly, the sequencing platform being used also influences library kit choice. Due to the smaller throughput of the MiSeq platform (~15 GB), most library preparation kits, including those with limited numbers of indexes (e.g. NEBNext® Ultra™ II FS Library Preparation Kit, Nextera® DNA Library Preparation Kit, NEXTFLEX Rapid XP DNA-Seq Kit, etc.) can be used with this platform. However, as the throughput of the Illumina platforms increases, (e.g. up to 850GB per lane on the NovaSeq), the library preparation kits need to be sourced from companies that have or can adapt to more than 192 compatible indexes for multiplexing of samples. Commercially available indexes such as the Perkin Elmer NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) can accommodate the sample throughout of the Novaseq (up to 300 samples/lane). Perkin Elmer indexes rely on the use of an A-tailing and ligation step that is similar to Truseq and NEBNext® kits making them compatible with a variety of library preparation kits. As such, NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) has been advised to be used when sequencing with the NovaSeq platform.

Indexes are important for differentiating sequence reads of individual samples when multiplexing samples on any of the sequencing platforms. If indexes are not compatible, the demultiplexing procedure will fail and sequencing reads will not be assigned to the corresponding biological samples (Varet et al., 2018). When many samples are multiplexed on larger sequencers, such as HiSeq and NovaSeq platforms, unique dual indexing (UDI) is highly recommended to reduce the phenomenon known as index hopping which is the incorrect assignment of indexes to a sample. When different indexing kits are being used, websites such as https://checkmyindex.pasteur.fr/ can be used to check index compatibility (Varet et al., 2018). Even with the same indexing kit, some companies provide pooling guides to ensure sequence read associated to specific samples can be differentiated. For instance, Figure 7.2 shows the index adapters pooling guide of IDT for Illumina–TruSeq UDI for a four-plex samples which changes for different numbers of samples.

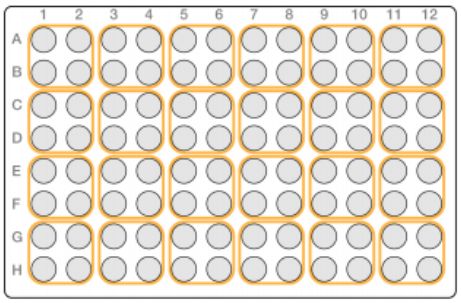


Figure 7.2 Pooling guide for four-plex dual indexing of IDT for Illumina–TruSeq UDI.

There is a need to adapt multiple library preparation kits to one indexing kit due to the different uses of each library preparation kit (e.g., RNA vs DNA library preparation). Most library preparation kits and indexes come with their own adapter sequences and chemistries rendering them incompatible. The Perkin Elmer NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) are compatible with library preparation kits that use A-tailing ligation, of which there are many. However, this means that it is not compatible with the commonly used Nextera® DNA Library Preparation Kits, hence, it is not advised to use the library preparation kit for sequencing on the NovaSeq. Furthermore, Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit is a new library preparation kit introduced to AVR late 2019 with potential to be used in half reactions. Hence, a comparison analysis was done on the Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with Perkin Elmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes); NEBNext® Ultra™ II FS DNA Library Preparation Kit with NEBNext® Multiplex Oligos for Illumina® (96 UDI Primer Pairs); and NEBNext® Ultra™ II FS DNA Library Preparation Kit with Perkin Elmer NEXTFLEX® UDI Barcodes (384 barcodes). The kits were tested using a sample of Fig (*Ficus carica*) DNA spiked with 1% and 0.1% total amount of DNA of *Ralstonia solanacearum* and *Pantoea ananatis* respectively.

ONT

Unlike Illumina, ONT sequencing kits are solely provided by ONT itself. The two core components for a MinION device are the flow cell or pore chemistries and the ligation sequencing kit. Both components need to be bought complementary to one

another. Currently, there are three different pore chemistries available: R9.4.1, R9.5 and R10.3. The most broadly used chemistry is R 9.4.1 which can accommodate the Ligation Sequencing Kit (SQK-LSK109), Rapid Sequencing Kit (SQK-RAD004) and Rapid Barcoding Kit (SQK-RBK004). The latest ONT chemistry, R10.3, which was released January 2020, is said to give higher read count and accuracy and will be compatible with the Ligation Sequencing Kit (SQK-LSK109). The flow cell R.9.5 is used with the $1D^2$ Sequencing Kit (SQK-LSK308).

ONT sequencing accuracy and length is also dependant on the ligation kit. For instance, it has been shown that the $1D^2$ Sequencing Kit (SQK-LSK308) can generate more accurate sequence data than the Ligation Sequencing Kit (Cornelis et al., 2019). However, library preparation is more laborious with low sequencing depth and longer sequencing time (Loit et al., 2019). Hence, it may be preferable to use the Ligation Sequencing Kit (SQK-LSK109) with the new flow cell pore chemistry, R10.3, along with the polishing software to correct reads to improve read accuracy than use the $1D^2$ Sequencing Kit.

Controls

Controls to be used during library preparation include the negative control used in the previous step, no template control to detect cross contamination and a positive control of a known quality DNA. These controls do not need to be sequenced in depth (200-300 MB worth of data is sufficient), however, their use will ensure that contamination can be detected (negative controls) and the library preparation process was successful.

The 2200 TapeStation system (Agilent technologies) should be used to check the quality of libraries. Sheering of libraries, adapter/primer dimers and size of libraries all can be detected on the TapeStation which can affect sequencing quality. For Illumina sequencing, adapter blockers should be used to prevent index hopping.

Bacteria identification

Bacteria can be isolated from a variety of environments and many genera contain species that can survive in a diverse span of these environments (Carrie Brady et al., 2008; Marrero et al., 2013). However, many phytopathogenic bacteria share environmental niches with closely related species and strains, both non-pathogenic strains and pathogenic strains with varying levels of virulence. For example, some bacterial species such as *Pseudomonas syringae* include different pathovars that occupy the same environmental niche. Kiwifruit is a host for both *P. syringae* pv. *actinidiae* and *P. syringae* pv. *actinidifoliorum,* however *P. syringae* pv. *actinidifoliorum* only causes minimal damage to the host in the form of necrotic spots as opposed the vine death that can be seen with *P. syringae* pv. *actinidiae* (Morris et al., 2008). This co-existence in the same environment (Passera et al., 2019) can lead to false positive test results or difficult identification as we try to pry apart pathogenic strains from non-pathogenic strains or to differentiate two closely related species.

Within this report, multiple bioinformatic tools have been trialled for differentiation of closely related pathogens using both Illumina and ONT sequencing data, including to the biovar level using Psa3 as a case study. This includes tools that use trimmed reads such as the k-mer-based analysis tool, Kraken2 (Wood et al., 2019) and the mapping tool, Bowtie2 (Langmead et al., 2012); and tools that use *de novo* assembly such as contig comparison to sequence databases such as GenBank using BLAST (Altschul et al., 1990), MLSA and up-to-date bacterial core gene (UBCG) analysis (Na et al., 2018). UBCG is similar to MLSA, however, instead of specific genes from each species, 92 conserved core-genes are used for MLSA, hence making it a universal method that can supposedly be applied to all bacterial species (Na et al., 2018).

For some of these bioinformatic tools, Conda, a source package management system and environment management system were used. Each bioinformatic tool are dependent on multiple software and programming languages, such as Java and Python, and occasionally these can clash with one another rendering the bioinformatic tool inoperative. In this project, Conda was used to create individual environments for Kraken2, Canu, Unicycler, Medaka and Prokka where they are installed with their dependencies. Within these environments, bioinformatic tools can be easily run and updated. Hence, as a prerequisite, Conda would need to be installed.

Outlined in this report are de novo assembly programs; SPAdes for short reads, Canu for long reads and Unicycler for a hybrid of long and short reads. These assembled reads will then be used to identify variable regions used for MLSA and UBCG. For MLSA, a script was made that encompass Prokka, BLASTn and a few bash and perl scripts to identify, pull out and concatenate MLSA genes.

Controls

FASTQC should be used before analysis to check the quality of reads. Reads would then need to be processed to trim adapters, filter bad quality reads and base correction. Multiple programs can be used such as Trimmomatic, Trim Galore and fastp for Illumina reads, and Porechop for ONT reads.

Positive and negative controls would also need to be analysed. There is a high probability that the read count for the negative controls, if any at all, will be low. Hence, Kraken2 can be used to swiftly analyse contents which can then be compared with samples to confirm contamination. As for the positive control, it would only be necessary to analyse if no other library worked. If the positive control worked and no other library did, it would indicate that the library preparation kit worked, and the problem could leftover inhibitors from DNA extraction.

Bioinformatic tools for pathogenic bacteria identification (*Pantoea* case study)

The genus *Pantoea* belongs within the family *Enterobacteriaceae* and was previously assigned into the *Erwinia herbicola*–*Enterobacter agglomerans* complex (Gavini et al., 1989). *Pantoea* species are a highly diverse group, commonly isolated from plants, humans, insects and the natural environment (Delétoile et al., 2009; Stice et al., 2018; Valiente Moro et al., 2013). Amongst *Pantoea* spp. are phytopathogenic species including *Pantoea ananatis* that are known to be pathogenic to eucalyptus, pineapple, onion, maize and melon; *Pantoea allii* that are known to be pathogenic to onions (Brady et al., 2011) ; and *Pantoea stewartii* subsp. *stewartii* that is known to cause Stewart's disease of maize (Coutinho et al., 2009; Herrera et al., 2008; Stice et al., 2018).

Closely related, plant-associated species of *Pantoea* can be difficult to differentiate using routine diagnostic tests such as PCR. For example, multiple PCR tests for the species-specific detection of *Pantoea stewartii* subsp. *stewartii* generate false positive results with *Pantoea stewartii* subsp. *indologenes* in maize (Nechwatal et al., 2018). Additionally, *Pantoea* species are often isolated during general surveillance in the Crop Health Services diagnostic laboratory at AVR. Specific testing of these isolates can often yield no clear result and there is no single gene recommended for amplification and sequencing for species identification. Instead, MLSA is recommended for identification of species in this genus (Brady 2008), a method that is resource and time intensive when done using PCR and Sanger sequencing. Traditional MLSA can also be biased or hindered by genetic variation in the primer binding regions. Additionally, newly characterised species and newly emerging pathogens often do not have species-specific tests, so a rapid and effective taxonomic identification protocol is required. In this section, closely related *Pantoea* isolates were sequenced on the Illumina Miseq and used as a case study to compare taxonomic classifications generated using a range of different bioinformatic approaches.

ONT vs. Illumina bioinformatic comparisons (Psa3 case study)

Pseudomonas syringae strains are taxonomically grouped into 13 phylogroups using molecular characterisation which have relatively conserved virulence-associated and phenotypic features (Berge et al., 2014) and has previously been separated into more than 60 pathovars on the basis of host of isolation and host range (Xin et al., 2018). These taxonomic classifications are important in a biosecurity setting for a species like *P. syringae* as it is found globally, however, specific pathovars with distinct pathogenic characteristics can be exotic to a region. *P. syringae* pv. *actinidiae* (Psa) has been further split into five biovars of which biovar 3 (Psa3) is known to be a highly virulent causal agent of bacterial canker of kiwifruit and biosecurity presence/absence data is monitored at biovar level (Chapman et al., 2012; Flores et al., 2018; Fujikawa et al., 2019). Due to the complex nature of Psa, it was an ideal test subject for ONT and Illumina sequencing comparisons to assess the taxonomic level that ONT can identify.

Material and Methods

DNA extraction

Fig (*Ficus carica*) DNA was extracted from leaf tissues using DNeasy® Plant Mini Kit. Bacterial DNA of *R. solanacearum* and *P. ananatis* was provided by AVR. Concentration of fig DNA was measured using a Quantus™ fluorometer (Promega, Madison, United States) and spiked with 1% *R. solanacearum* and 0.1% *P. ananatis* of total DNA.

Eight *Pantoea* spp (P1-P8) were isolated and grown on Kings B agar media for 2 days at 28 °C A representative Psa3 isolate from Victoria was grown on nutrient sucrose agar (EPPO, 2014) for 2 days at 28 °C. *Pantoea* and Psa3 colonies were scraped from the agar and extracted using Qiagen DNeasy® Blood & Tissue kit (Qiagen, Hilden, Germany).

For MinION sequencing, Psa3 was grown in nutrient broth with no sugar source overnight rotating at 120 RPM at 28 °C. Overnight cultures were centrifuged until a pellet formed. Bacterial pellets were dried in the biosafety cabinet for one hour before proceeding to use Promega Wizard® Genomic DNA Purification Kit (Promega, Madison, United States) for gDNA extraction.

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, concentration of DNA was measured and 10 – 50 ng DNA for each library preparation. Library preparation kits used on spiked fig DNA included Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with Perkin Elmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes); NEBNext® Ultra™ II FS DNA Library Preparation Kit with NEBNext® Multiplex Oligos for Illumina® (96 UDI Primer Pairs); and NEBNext® Ultra™ II FS DNA Library Preparation Kit with Perkin Elmer NEXTFLEX® UDI Barcodes (384 barcodes). Libraries were sequenced on an Illumina Miseq platform.

Libraries for the *Pantoea* isolates were generated using the NEBNext® Ultra™ II FS DNA Library Preparation Kit (New England Biolabs, Massachusetts, United States). Isolates were then paired end sequenced (2 × 300 bp) using the Illumina MiSeq with a negative control.

Libraries for the Psa3 isolate were generated using the Nextera DNA Flex Library Preparation kit (Illumina Inc., California, United States) and was then paired end sequenced (2 × 150 bp) using the Illumina HiSeq.

ONT library preparation and sequencing

For MinION, approximately 2 µg of DNA was used for library preparation using the ONT ligation kit (LSK109). Psa3 was then sequenced using the flowcell type MIN106 on the MinION for 40 hours.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen et al., 2018) and assembled using SPAdes v3.6.0 (Bankevich et al., 2012). For the spiked fig DNA, Kraken2 analysis was run on trimmed reads and a nucleotide BLAST search (Altschul et al., 1990) was conducted on the assembled reads to check variability of detection limits for each library preparation kit and index combination.

Analysis of *Pantoea* trimmed reads included mapping to diverse *Pantoea* references using Bowtie2. Genome references used are in appendix 3. The k-mer analysis tool using Kraken2 (Wood et al., 2019), was also used on *Pantoea* trimmed reads.

Assembly of Illumina reads was done using SPAdes v3.6.0 (Bankevich et al., 2012). Assembled contigs were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990). MLSA and UBCG (Na et al., 2018) analyses were run on the *Pantoea* and Psa3 assemblies. MLSA functions to differentiate species by generating a phylogenetic tree from several concatenated essential protein-encoding gene sequences (housekeeping genes). UBCG does something similar, however, instead of choosing housekeeping genes from literature, it uses 92 core genes for phylogenetic analysis. The MLSA protocol is as described in Basic protocol 9 (Appendix 8) using *infB*, *rpoB*, *atpD* and *gyrB* genes (C. Brady et al., 2008) for *Pantoea* and *gapA*, *gltA*, *gyrB* and *rpoD* genes (Butler et al., 2013; Flores et al., 2018) for Psa3. MLSA genes used are full genes. The program Insilicoseq was used to generate sequencing reads *in silico* for *Pantoea stewartii* subsp. *stewartii* DC283 and *Pantoea stewartii* subsp. *indologenes* LMG2632. Generated reads were for use for Kraken2 analysis.

Illumina reads were assembled using SPAdes v3.6.0. Assembled contigs were compared to ONT. Raw signal MinION files were basecalled using Guppy v.2.1.3 and subjected to demultiplexing and adapter/barcode removal using Porechop. Raw Illumina reads were subjected to adapter trimming and quality filtering using fastp. Canu v1.9 was used for assembly of for MinION reads and Unicycler v0.4.8 was used for assembly of MinION reads and a hybrid assembly (both MinION and Illumina reads). MinION reads were further polished with two rounds of Racon (Vaser et al., 2017) and Medaka (https://nanoporetech.github.io/medaka/index.html). Pyani v0.2.10 was used to determine the average nucleotide identity (ANI) of polished and unpolished MinION assemblies compared to the hybrid assembly and Prokka v1.14.5 was used to annotate the assembled contigs to identify genes. Both ANI and predicted gene number were used to assess accuracy of polished reads in comparison to unpolished.

Assembled contigs were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990) as described previously (Ashikawa et al., 2018; Hu et al., 2019(Ashikawa et al., 2018; Hu et al., 2019). Furthermore, all assembled reads, polished and unpolished, were subjected to MLSA and UBCG analysis, as described above.

Results

Library preparation kits comparisons

Library preparation kits used for comparisons include Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with Perkin Elmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PEPE); NEBNext® Ultra™ II FS DNA Library Preparation Kit with NEBNext® Multiplex Oligos for Illumina® (96 UDI Primer Pairs) (NEBNEB); and NEBNext® Ultra™ II FS DNA Library Preparation Kit with Perkin Elmer NEXTFLEX® UDI Barcodes (384 barcodes) (NEBPE). All three combinations were sequenced on the Miseq and both Perkin Elmer indexed libraries were sequenced on the NovaSeq.

Sequencing

Library information provided in the pre-sequencing column of Table 7.1 shows that Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit (PEPE) produces longer fragment sizes in comparison to NEBNext® Ultra™ II FS DNA Library Preparation Kit. Even though PEPE produced longer fragments, it produced the shortest largest contig and more than double the number of contigs in comparison to the NEB kit. The BLASTn search results identified the largest contig of PEPE as *Ralstonia solanacearum* and the largest contig generated by the NEB library preparation kits (NEBNEB and NEBPE) had the highest identity with the plant species *Morus notabilis* (mulberry*)*. There were no obvious differences between the NEB library preparation kits using adapters from different companies (NEBNEB and NEBPE).

Detection of spiked genomes

The no-template control used during the library preparation step had four reads that were not proper sequences indicating there was not contamination. The Kraken2 results showed that percentage of *R. solanacearum* reads for PEPE was 7.34%, NEBPE was 4.98 and NEBNEB was 5.02%. PEPE showed to the largest contig for *R. solanacearum* which was 48.7 kbp in comparison to other kits that generated contigs below 10 kbp. There was an increased number of contigs for *P. ananatis* for PEPE preparation kit which was 1439 contigs in comparison to the NEB library preparation kits which gave less than 400 contigs. Majority of the contigs could not be matched to known sequences (between 82 to 94.1%). It is presumed that these unknown contigs were most likely associated with fig DNA however fig genomes were not available in the Kraken2 database to support this hypothesis. Results also show little variation between NEB library preparation kit with NEB adapters and Perkin Elmer adapters.

Table 7. 1 Sequencing results for 3 library kit and index combination; NEB library preparation kit with NEB indexes (NEBNEB), NEB library preparation kit with Perkin Elmer indexes (NEBPE), and Perkin Elmer library preparation kit with Perking Elmer indexes (PEPE).

	PRE-SEQUENCING		POST-SEQUENCING				
			Post trimming		Post assembly		
LIBRARY KIT	**Avg lib size (bp)**	**Lib concentration**	**No. of reads**	**Size (GB)**	**No. of contigs**	**Largest contig size**	**BLASTn**
PEPE	414	18.9	2713806	0.688	95435	48749	*R. solanacearum* strain UY031 plasmid
NEBPE	271	9.86	3835266	0.525	34630	56756	*Morus notabilis* mitochondrion
NEBNEB	298	7.08	3789304	0.599	50551	61448	*Morus notabilis* mitochondrion
NEBNEB (FIG)	292	6.28	2879274	0.428	23761	57390	*Morus notabilis* mitochondrion

Table7. 2 Detection of spiked genomes using BLASTn and Kraken2 for 3 library kit and index combination; NEB library preparation kit with NEB indexes (NEBNEB), NEB library preparation kit with Perkin Elmer indexes (NEBPE), and Perkin Elmer library preparation kit with Perking Elmer indexes (PEPE).

LIBRARY KIT	**SPIKED GENOMES**	**BLASTN**	**LARGEST CONTIG (BP)**	**NO OF CONTIGS**	**UNKNOWN CONTIGS**	**% OF UNKNOWN CONTIGS**	**% READS**	**KRAKEN**
PEPE	*R. solanacearum*	*R. solanacearum*	48749	1825	85419	89.5	7.34	*R. solanacearum*
	P. ananatis	*P. ananatis*	2558	1439			0.88	*P. ananatis*
NEBPE	*R. solanacearum*	*R. solanacearum*	7899	4018	28414	82.1	4.98	*R. solanacearum*
	P. ananatis	*P. ananatis*	1851	259			0.79	*P. ananatis*
NEBNEB	*R. solanacearum*	*R. solanacearum*	9115	4033	43137	85.3	5.02	*R. solanacearum*
	P. ananatis	*P. ananatis* LMG 5342	1264	383			0.75	*P. ananatis*
FIG	*R. solanacearum*	N/A		N/A	22350	94.1	0.00	*R. solanacearum*
	P. ananatis	N/A	N/A	N/A			0.01	*P. vagans*

Bioinformatic tools for pathogenic bacteria identification (*Pantoea* case study)

Sequencing

Pantoea isolates were sequenced through Illumina MiSeq. The MiSeq run had a cluster density of 895 K/mm^2 and clusters passing filter at 95.0% with estimated yield of 12.65 GB. Generated reads were trimmed using fastp and information for each isolate shown in Table 7.1.

Table 7. 3 Post trimming sequencing results for each Pantoea isolate.

	TOTAL NO. OF READS	TOTAL AMOUNT OF DATA (GB)	N50 LENGTH (BP)	MAX LENGTH (BP)
P1	3,196,954	0.44	149	300
P2	3,949,256	0.59	164	300
P3	2,165,542	0.34	172	300
P4	3,036,596	0.46	168	300
P5	3,398,142	0.49	158	300
P6	4,442,318	0.54	128	300
P7	3,000,554	0.47	173	300
P8	2,712,160	0.42	171	300

Mapping

Pantoea isolates were mapped to reference genomes and genomes with the highest number of reads are listed in Table 7.4. Isolate P1 identified as *Pantoea ananatis* while isolate P6 identified as *Pantoea agglomearans*. Isolate P2, P5, P7 and P8 identified as *Pantoea allii*. Isolate P4 was classified as *Pantoea coffeiphila*. Isolate P3 could not be clearly identified as there was more than one possibility presented.

Table 7.4 Genomes with highest number of reads for each isolate.

ISOLATE	GENOME WITH HIGHEST NO. OF READS	% OF READS
P1	*Pantoea ananatis* PA13	19.4
P2	*Pantoea allii* OXWO6B1	20.6
P3	*Pantoea eucalypti* LMG 24197	10.2
	Pantoea vagans FBS135	10.1
	Pantoea agglomerans C410P1	7.6
P4	*Pantoea coffeiphila JUb26*	17.5
P5	*Pantoea allii* OXWO6B1	19.8
P6	*Pantoea agglomerans* C410P1	23.6
P7	*Pantoea allii* PNA 200-10	23
P8	*Pantoea allii* PNA 200-10	23

MLSA

Four genes (*infB*, *rpoB*, *atpD* and *gyrB*) were extracted from assembled reads from each isolate, concatenated and aligned against multiple species of *Pantoea*. Phylogenetic tree was then constructed and shown in Figure 7.6. From the phylogenetic tree, P2, P5, P7 and P8 were classified as *P. allii*, P1 as *P. ananatis*, P6 as *P. agglomerans,* and P4 as a *Pantoea* species closely related to *P. coffeiphila,* and P3 as a *Pantoea* species closely related to *P. eucalypti*.

UBCG

UBCG java script was ran on all eight *Pantoea* isolates to extract 92 core genes. The 92 core genes were aligned to multiple species of *Pantoea*. Phylogenetic tree was then constructed and shown in Figure 7.7. From the phylogenetic tree, P2, P5, P7 and P8 were classified as *P. allii*, P1 as *P. ananatis*, P6 as *P. agglomerans,* P4 as a *Pantoea* species closely related to *P. coffeiphila* and P3 as a *Pantoea* species closely related to *P. eucalypti*.

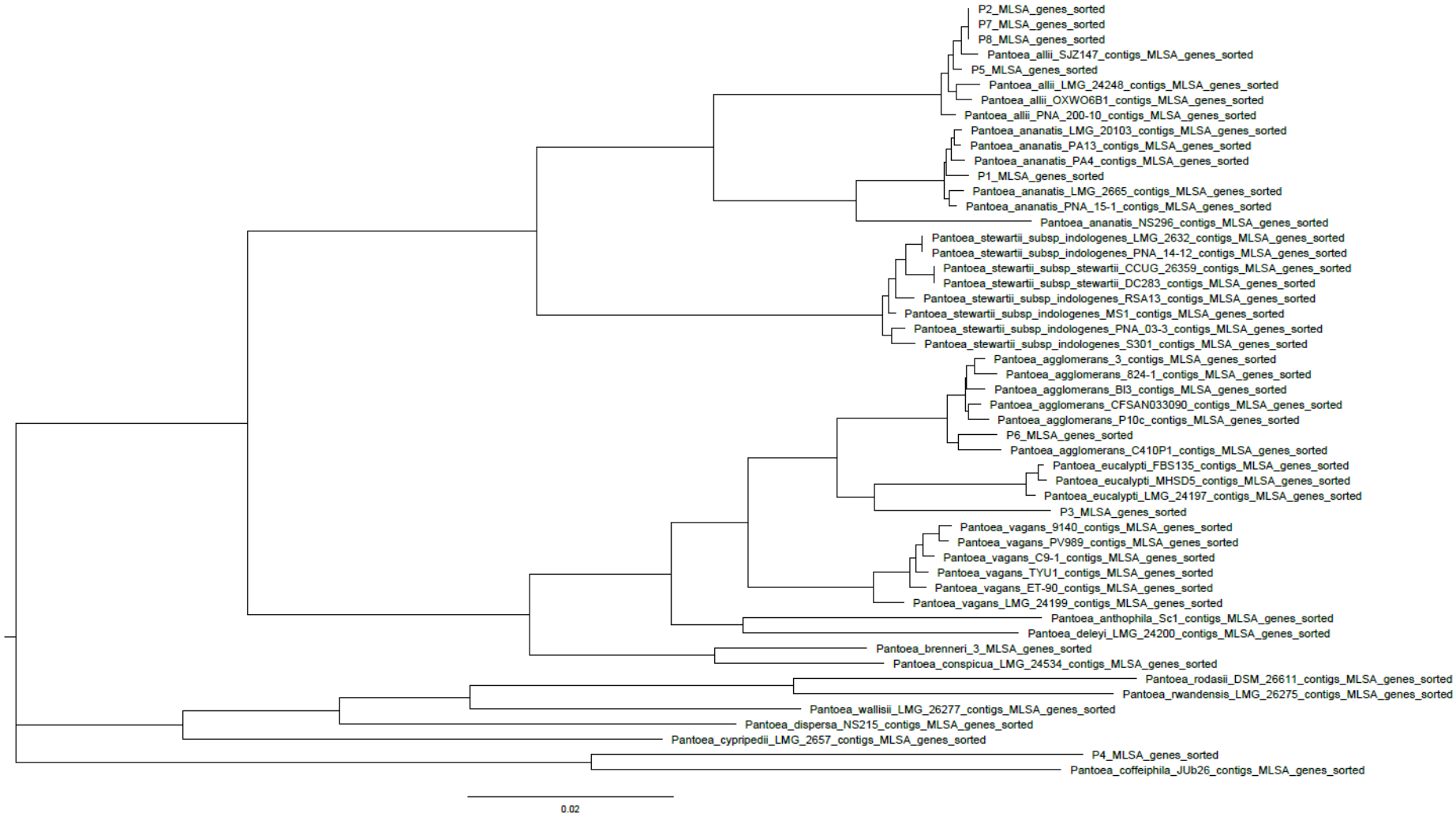


Figure 7.3 Maximum likelihood tree based on infB, rpoB, atpD and gyrB sequences of isolated Pantoea species and their closest phylogenetic neighbours.

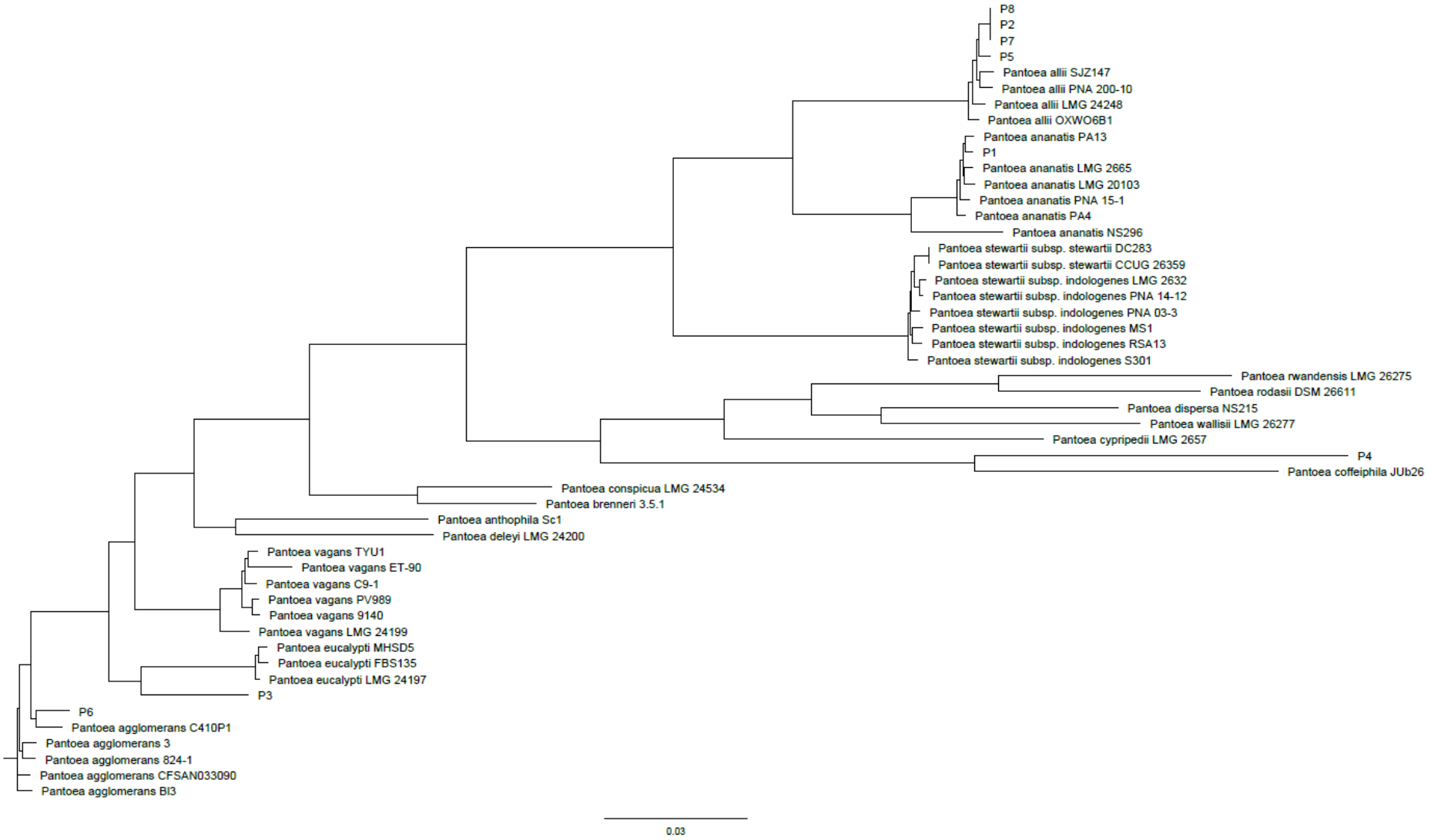


Figure 7. 4 Maximum likelihood tree based on 92 core gene sequences of isolated Pantoea species and their closest phylogenetic neighbours.

Table 7.5 Summary of bioinformatic results.

Samples		BLASTn*	Kraken2**	Mapping	MLSA	UBCG
P1	Hit1 Hit2	*P. ananatis*	*P. ananatis*	*P. ananatis*	*P. ananatis*	*P. ananatis*
P2	Hit1 Hit2	*P. ananatis*	*P. ananatis* *P. stewartii*	*P. allii*	*P. allii*	*P. allii*
P3	Hit1 Hit2	*P.* vagans *P. agglomerans*	*P.* vagans *P. agglomerans*	*P. eucalypti* *P. agglomerans*	*P. eucalypti*	*P. eucalypti*
P4	Hit1 Hit2	*Erwinia persicina* *Erwinia pyrifoliae*	*Erwinia persicina* *Erwinia tasmaniensis*	*P. coffeiphila*	*P. coffeiphila*	*P. coffeiphila*
P5	Hit1 Hit2	*P. ananatis*	*P. ananatis* *P. stewartii*	*P. allii*	*P. allii*	*P. allii*
P6	Hit1 Hit2	*P. agglomerans*	*P. agglomerans*	*P. agglomerans*	*P. agglomerans*	*P. agglomerans*
P7	Hit1 Hit2	*P. ananatis*	*P. ananatis* *P. stewartii*	*P. allii*	*P. allii*	*P. allii*
P8	Hit1 Hit2	*P. ananatis*	*P. ananatis* *P. stewartii*	*P. allii*	*P. allii*	*P. allii*
P. stewartii subsp. stewartii DC283	Hit1	N/A	*P. stewartii subsp. stewartii* DC283	N/A	*P. stewartii subsp. stewartii* DC283	*P. stewartii subsp. stewartii* DC283
P. stewartii subsp. *indologenes* LMG2632	Hit1 Hit2	N/A	*P. stewartii subsp. stewartii* DC283 *P. ananatis* PA13	N/A	*P. stewartii* subsp. *indologenes* LMG2632	*P. stewartii* subsp. *indologenes* LMG2632

*Majority species of the first 5 contigs of BLASTn search results.

** Second hit on Kraken2 was displayed if the percentage of reads were below 80%.

BLASTn and Kraken2

As observed from the UBCG and MLSA phylogenetic tree, *P. stewartii subsp. stewartii* and *P. stewartii* subsp. *indologenes* are very closely related and may be a problem for identification if a new strain was to be detected under post entry quarantine (PEQ). This is important as *P. stewartii subsp. stewartii* is an exotic. Hence, Insilicoseq was used to generate sequencing reads on both strains *in silico* and an extra Kraken2 analysis was conducted with results shown in Table 7.5. Remaining BLASTn searches and Kraken2 results for the *Pantoea* isolates can also be viewed in Table 7.5 in comparison with mapping, MLSA and UBCG results. BLASTn and Kraken2 could only identify *P. agglomerans*, and *P. ananatis* and *P. stewartii* subsp. *stewartii* accurately. Mapping results could not accurately identify isolate P3 among the isolates and identified P4 as *P. coffeiphila*.

ONT vs. Illumina bioinformatic comparisons (Psa3 case study)

Sequencing data

The Psa3 isolate was sequenced on both Illumina HiSeq and ONT MinION. Generated Illumina reads were trimmed using fastp and ONT reads trimmed using porechop. Information of trimmed reads for each sequencer are shown in Table 7.3.

Table 7.6 Post trimming sequencing results for the Psa3 isolate on ONT MinION and Illumina HiSeq.

	TOTAL NO. OF READS	TOTAL AMOUNT OF DATA (GB)	N50 LENGTH (BP)	MAX LENGTH (BP)
MINION	91,440	1.09	27275	215652
ILLUMINA	1,347,586	0.3	250	251

ANI and annotation

To analyse accuracy of polished and unpolished MinION assembly, a comparison study to the hybrid assembly was done comparing number of predicted genes and average nucleotide identity (ANI) with two *de novo* assembly programs, Canu and Unicycler. High numbers of predicted genes indicate fragmented and incomplete genes. Results shown in Table 7.4 indicates that polishing MinION reads does aid in correcting reads.

Table 7. 7 ANI and number of predicted gene summary for raw and polished MinION reads.

Assembler		Raw output	Racon polish round 1	Racon polish round 2	Medaka polish	Unicycler hybrid assembly
Canu	ANI	99.637%	99.889%	99.902%	99.966%	N/A
	No. of predicted genes	9896	N/A	7166	6391	6002
Unicycler	ANI	99.883%	99.907%	99.907%	99.963%	N/A
	No. of predicted genes	7631	N/A	7248	6491	6002

* ANI to the hybrid assembly.

MLSA

Four genes (*gapA*, *gltA*, *gyrB* and *rpoD*) were extracted from each assembled and polished reads of Psa3, concatenated and aligned against multiple biovars of Psa. Phylogenetic tree was then constructed (data not shown) From the phylogenetic tree, all Psa3 assemblies and polishes have been classified as Psa3.

UBCG

UBCG java script was run on all Psa3 assemblies to extract 92 core genes. The 92 core genes were aligned to multiple biovars of Psa. A phylogenetic tree was then constructed and shown in **Figure 7.8.** From the phylogenetic tree, Illumina and hybrid assembled reads sits within the Psa3 clade. However, MinION assembled and polished reads are starting to diverge off, display long branch lengths indicating genetic diversity, with non-polished reads diverging more than their polished counterparts. This indicates that there are enough errors in the reads to alter phylogenetic results at the biovar level.

To confirm whether the same results are observed in the pathovar level, a similar analysis was conducted, however, the Psa3 assemblies were aligned against different pathovar of *Pseudomonas syringae* shown in **Figure 7.9.** Similar to the phylogenetic analysis against Psa biovars, the MinION assembled and polished exhibit increased branch length implying genetic diversity, however, all assemblies clearly cluster within the Psa clade indicating it has potential to be used for bacterial pathogen identification down to the pathovar level.

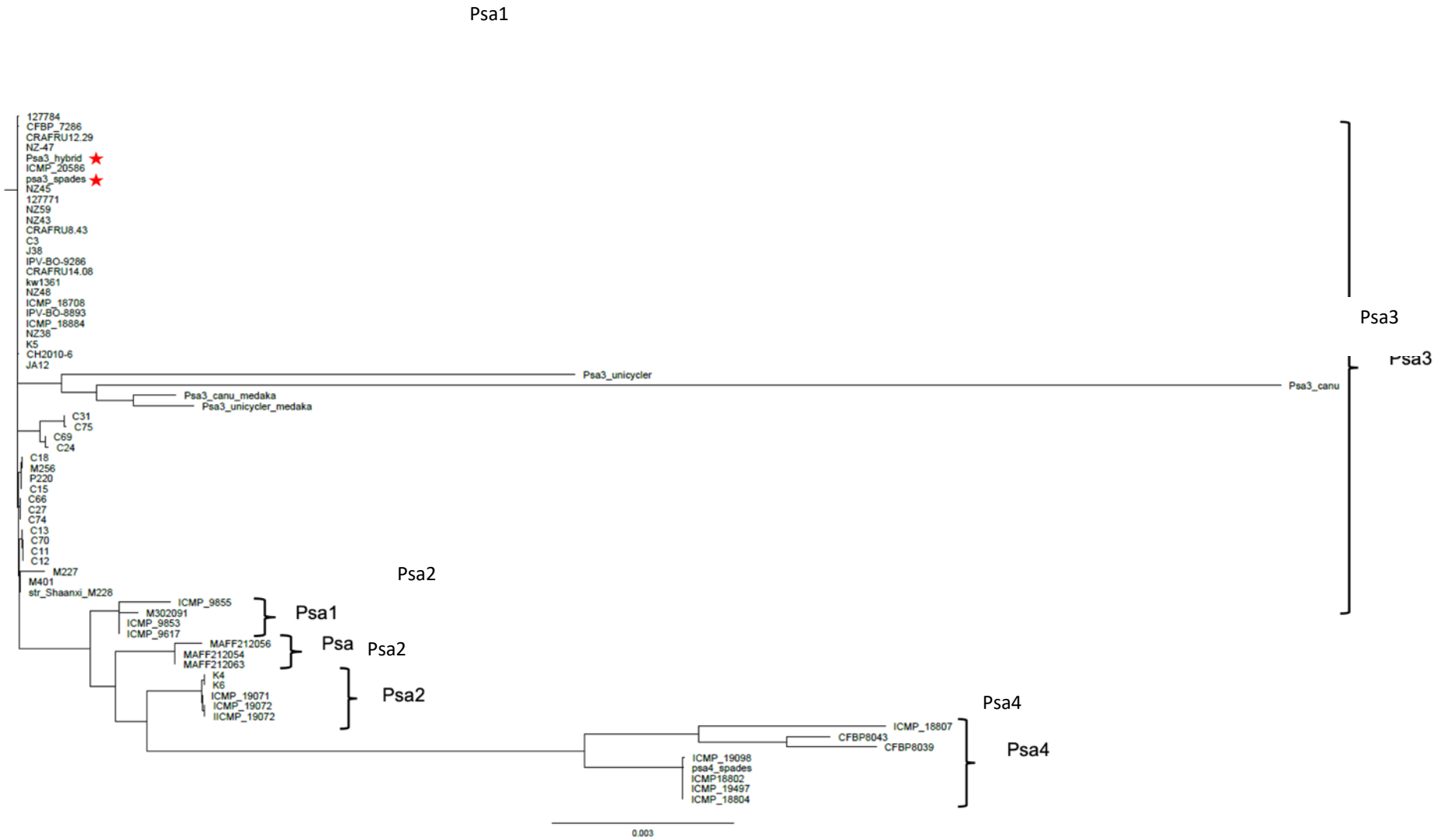


Figure 7.8 Maximum likelihood tree based on 92 core gene sequences of isolated Psa3 assemblies and their closest phylogenetic Psa neighbours. The two red stars shows the Illumina and hybrid assembled reads.

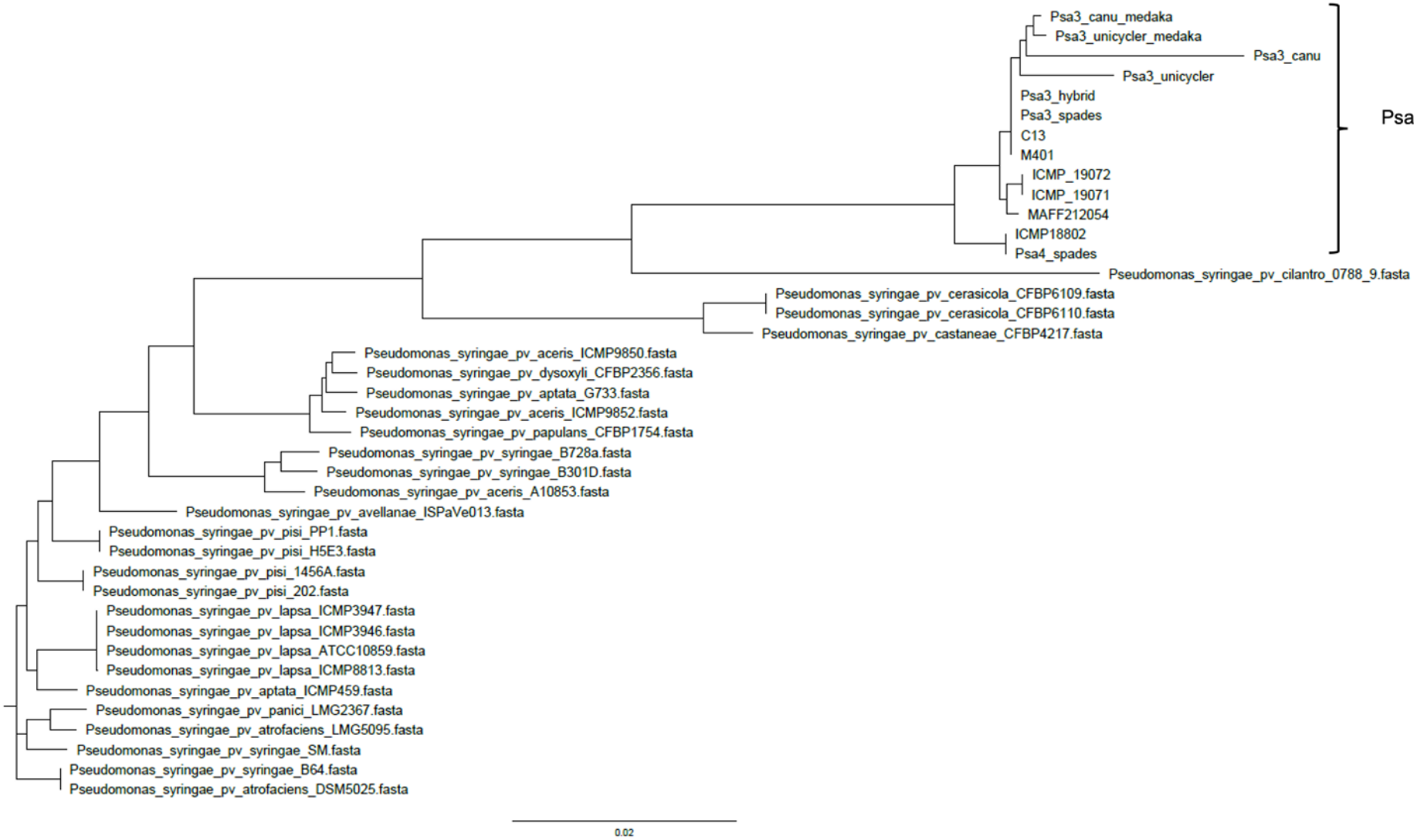


Figure 7.5 Maximum likelihood tree based on 92 core gene sequences of isolated Psa3 assemblies and their closest phylogenetic Pseudomonas syringae neighbours. The two red stars shows the Illumina and hybrid assembled reads.

Discussion

Each step from DNA extraction to data analysis can play a big part in attaining high quality data and accurate identification of bacteria using HTS. The aim of this study was to assess the use of different combinations of library preparation kits and barcodes, sequencing platforms, and bioinformatics tools to identify the most reliable procedure and to develop draft standard operating procedures (appendix 2) for identification of culturable bacteria by HTS. We did not assess different commercial DNA extraction kits as good guidelines are available for DNA quality control as input into both long and short read sequencing procedures. For Illumina sequencing, DNA was extracted from bacterial cultures using the Qiagen DNeasy Blood & Tissue Kit (Basic protocol 1, appendix 2) as it has been a validated bacterial extraction kit at AVR. It can be adapted as a high throughput option where 96 samples can be extracted simultaneously using robotics. For ONT libraries, The Promega Wizard® Genomic DNA Purification Kit (Basic protocol 2) was chosen as it uses a precipitation method that causes less shearing of the DNA compared to column-based extraction with higher yields of high molecular weight (HMW) DNA. This is important for ONT sequencing as HMW DNA is required to attain high quality data. Commercial kits such as these provide assurance of DNA quality and yield due to the manufacturer's quality control processes for critical reagents. Nevertheless, the DNA quality must be measured before HTS can proceed. Although gel electrophoresis can be used to estimate DNA quality, the TapeStation, which was used throughout this study, provides a precise analytical evaluation of DNA size, quantity, and integrity and is recommended as part of a quality assurance process for bacterial HTS. The DNA must also be quantified as the protocols for library preparation kits relies heavily on the amount of input DNA and this is most accurately done using a fluorometer such as Quantus™ or Qubit to measure concentration of DNA. Although the DNA extraction kits used in this study have been validated and are constantly used, controls are still a necessity. Among the controls that could be added are the addition of a negative (blank) control during extraction to measure possible contamination during the extraction process. This is especially necessary when doing high-throughput extractions as probabilities of cross contamination is higher.

Based on the results of this study the NEBNext® Ultra™ II FS DNA Library Preparation Kit first library preparation kit is recommended for Illumina sequencing is (Basic protocol 3). This is due to the low input of DNA (0.1 ng) necessary to make libraries, which is useful for difficult to extract samples, hence is recommended for DNA input <10 ng. The second library preparation kit suggested is the NEXTFLEX Rapid XP DNA-Seq Kit (Basic protocol 4). This kit is fairly new and has only been introduced to AVR late 2019. In this study we trialled this kit using half reactions, which at $13 per sample reaction, represents a cost saving compared to other kits and is recommended for input DNA >10 ng. We observed that the NEXTFLEX Rapid XP DNA-Seq Kit produces longer fragments for sequencing compared to the other kits, which considerably aids genome assembly. In this study, the assembly of spiked *R. solanacearum* from data generated in the library preparation kit comparisons highlighted the impact fragment (or insert) length can have on bacterial genome assembly. The largest assembled *R. solanacearum* contig from PEPE generated sequence reads was 48749 bp, approx. six times longer than the largest assembled contig from the NEB kit generated sequence reads with shorter insert length.

When using high throughput sequencers, such as the NovaSeq and HiSeq, the Perkin Elmer NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) is highly recommended as they can differentiate up to 1,536 samples giving room to incorporate libraries from hundreds of samples simultaneously in a single sequencing run. Costs per sample are also reduced to $5 per barcode which can be reduced by half if using the half reaction of the Perkin Elmer library preparation kit. This study has shown that indexes from Perkin Elmer NEXTFLEX® Unique Dual Index Barcodes (384 UDI barcodes) are also adaptable to the NEB library preparation kit and there were no major differences in results of this combination to the NEB library preparation kit using NEB indexes. Adapting one set of UDI barcodes for multiple library preparation kits can save both time and costs as there would be no need to worry about incompatible indexes and neither the need to purchase multiple sets of UDIs for each library preparation kit. Before loading samples, the quality and quantity of the libraries needs to be analysed on the TapeStation or Agilent Bioanalyzer. Examples in support protocol 2.1 show the presence of adapter and/or primer dimers that was detected on TapeStation in the library. The dimers would both skew the final concentration of a library and will take up a lot of space on the sequencing run as they are smaller fragments. An extra adapter blocker step is highly suggested when using high throughput sequencers to prevent adapter hopping. Controls from this step include a positive control of a known working library to confirm the sequencing run has worked and an extra negative (no-template) control to detect cross contamination.

The final and most variable step in most NGS pipelines is the bioinformatics approach used to identify the bacterial isolates in the sample. First a quality check using FASTQC (Support protocol 3.1) followed by fastp (Support protocol 3.2) which is a trimming, base correction and read filtering program is recommended. Subsequently, as seen from this report, a k-mer analysis tool such as Kraken 2, can quickly identify to the genus level using only the trimmed reads. However, in this analysis Kraken 2 did not accurately identify some *Pantoea* sp. including *P. stewartii* pv. *indologenes*. Similarly, when BLASTn and

a mapping approach was used on assembled contigs to identify bacterial samples, not all isolates were accurately identified to the species level. It is highly probable that this misclassification reflects the content of the databases used rather than the ability of the actual program. However, it does highlight the importance of up-to-date databases and the fact that for newly described or emerging pathogens, identification pipelines require complementation with less biased approaches. Since k-mer analysis tools do not require an assembly of reads, which can be a bottleneck for analysis, it is be worthwhile curating our own database for rapid analysis and further validating this approach.

As seen in this report, using core genes in a phylogenetic analysis as outlined for MLSA and UBCG, enabled identification of the Psa3 isolate to the biovar level and identification of most of the *Pantoea* isolates. The P3 isolate, which was not clearly identified may be a novel *Pantoea* sp. or a genetic variant of *P. eucalyptii* based on its phylogenetic placement in the trees generated using MLSA and UBCG data. This also shows the importance of phylogenetic analysis to detect new species or genetic variants that may give false positives or may not be detected using some of the database comparison methods. However, it should be noted that the accuracy and quality of information derived from using an MLSA core gene phylogenetic approach is also dependant on including appropriate references for comparison further affirming the need for up-to-date databases. Both MLSA and UBCG are included in the pipeline as some reference isolates may not have a publicly available genome (e.g. in the case of newly classified or emerged species or strains). However, MLSA requires the download of MLSA genes for the analysis of a new genus. Furthermore, it would need users to be competent in understanding bash scripts as modifications are necessary to the script. UBCG only requires two lines to be written for a full analysis without the need to download reference MLSA genes for every genus, hence would be the preferred option for bacterial identification.

Based on the abovementioned findings, the draft pipeline for bacterial identification using Illumina sequencing is rapid taxonomic grouping using Kraken 2 (Basic protocol 11), followed by a *de novo* assembly (Basic protocol 8) and UBCG (Basic protocol 10) depending on the available sequence data.

The draft recommended protocol for ONT sequencing uses either the Ligation kit (SQK-LSK109) with the latest nanopore flow cell (R10.3) or the use of $1D^2$ sequencing (Basic protocol 5) as they increase sequence accuracy. As seen from this report, ONT sequencing can be error prone which can lead to misidentification shown by the diverging clade of ONT assembly for the UBCG results of Psa3. However, polishing the assembled reads can alleviate some of the divergence and may become a useful sequencing platform as it is easy and cheap to establish in the lab.

This study showed that ONT technology on its own may not provide sufficient information to identify a bacterium to a pathogenic biovar level but it can be used to rapidly identify bacterial species to the species and pathovar level (though this may vary for some species), whilst there were mild discrepancies observed for the unpolished and polished ONT read assemblies. It was clear from the phylogenetic analysis and ANI to the hybrid assembly, that the polished reads gave better results than their non-polished counterparts, but would it be good enough for basic bacterial identification in most cases. As mentioned above, the new flow cell chemistry may prove to be a game changer for ONT sequencers with the increased accuracy and should be assessed in the future. Furthermore, ONT library preparation and running ONT sequencers are less time and labour intensive compared to their Illimina equivalents making ONT an attractive tool for a swift analysis. The only downside is the minimal multiplexing that can be done on a MinION. However, the processes are adaptable to the PromethION which can run multiple, high-capacity flow cells simultaneously, and can therefore be a high-throughput multiplexing tool.

When using Illumina sequencing, Psa isolates were identified to the biovar 3 level with only 300 MB of Illumina data. Hybrid assembly using Illumina and ONT sequence data not only enabled accurate identification, but it also generated complete genomes. The value of generating the Illumina and/or hybrid assemblies is that both could be used in more in-depth downstream analysis like source tracing. For this reason, we recommend sequencing with both Illumina and ONT for more in depth and descriptive information of the isolate.

Bioinformatic analysis tools for bacterial identification that have not been assessed in this report but will be considered in the future as part of pipelines for fastidious bacterial pathogens include core genome analysis (both genomic and single nucleotide polymorphism (SNP) based), as it would not only identify the pathogen but also aid in tracing source of origin of the bacterial pathogen. Additionally, bioinformatic tools such as the k-mer analysis tool kSNP3.0 (Gardner et al., 2015) and KrakenUBCG (Chalita et al., 2020) that combines kmer analysis with SNP analysis, and Kraken 2 and UBCG subsequently, will also be considered.

In conclusion, recommended draft protocols for bacterial genome identification are outlined in Appendix 3. Multiple options are provided at the different stages to accommodate variation in sample and output requirements (e.g. low input DNA, desire to use data for tracing activities) have their own advantages. Obtaining accurate results is a priority imperative

in biosecurity diagnostics as false positives and/or negatives can have drastic impacts on industry and trade. Hence, it is advised that at the sequence analysis stage multiple approaches should be used for identification to ensure accurate results. Furthermore, like with all diagnostic testing, controls play a major role in preventing false positives and negatives and must be included where necessary, an important step this is sometimes overlooked in NGS testing. These draft protocols need to be tested with PEQ to further validify the pipeline and identify any improvements needed.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Ashikawa, S., Tarumoto, N., Imai, K., Sakai, J., Kodana, M., Kawamura, T., . . . Maeda, T. (2018). Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. *Journal of Medical Microbiology, 67*(11), 1589-1595. doi:https://doi.org/10.1099/jmm.0.000855

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., . . . O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology, 33*(3), 296-300. doi:10.1038/nbt.3103

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Becker, L., Steglich, M., Fuchs, S., Werner, G., & Nübel, U. (2016). Comparison of six commercial kits to extract bacterial chromosome and plasmid DNA for MiSeq sequencing. *Scientific Reports, 6*(1), 28063. doi:10.1038/srep28063

Berge, O., Monteil, C. L., Bartoli, C., Chandeysson, C., Guilbaud, C., Sands, D. C., & Morris, C. E. (2014). A user's guide to a data base of the diversity of Pseudomonas syringae and its application to classifying strains in this phylogenetic complex. *PLOS ONE, 9*(9), e105547-e105547. doi:10.1371/journal.pone.0105547

Brady, C., Cleenwerck, I., Venter, S., Vancanneyt, M., Swings, J., & Coutinho, T. (2008). Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Systematic and Applied Microbiology, 31*(6), 447-460. doi:https://doi.org/10.1016/j.syapm.2008.09.004

Brady, C., Cleenwerck, I., Venter, S., Vancanneyt, M., Swings, J., & Coutinho, T. (2008). Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Syst Appl Microbiol, 31*(6-8), 447-460. doi:10.1016/j.syapm.2008.09.004

Brady, C. L., Goszczynska, T., Venter, S. N., Cleenwerck, I., De Vos, P., Gitaitis, R. D., & Coutinho, T. A. (2011). Pantoea allii sp. nov., isolated from onion plants and seed. *Int J Syst Evol Microbiol, 61*(Pt 4), 932-937. doi:10.1099/ijs.0.022921-0

Butler, M. I., Stockwell, P. A., Black, M. A., Day, R. C., Lamont, I. L., & Poulter, R. T. M. (2013). Pseudomonas syringae pv. actinidiae from Recent Outbreaks of Kiwifruit Bacterial Canker Belong to Different Clones That Originated in China. *PLOS ONE, 8*(2), e57464. doi:10.1371/journal.pone.0057464

Chalita, M., Ha, S.-m., Kim, Y. O., Oh, H.-S., Yoon, S.-H., & Chun, J. (2020). Improved Metagenomic Taxonomic Profiling Using a Curated Core Gene-Based Bacterial Database Reveals Unrecognized Species in the Genus Streptococcus. *Pathogens, 9*(3), 204.

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., . . . Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology, 68*(2), 229-238. doi:10.1111/ppa.12957

Chapman, J. R., Taylor, R. K., Weir, B. S., Romberg, M. K., Vanneste, J. L., Luck, J., & Alexander, B. J. R. (2012). Phylogenetic Relationships Among Global Populations of Pseudomonas syringae pv. actinidiae. *Phytopathology®, 102*(11), 1034-1044. doi:10.1094/PHYTO-03-12-0064-R

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Cornelis, S., Gansemans, Y., Vander Plaetsen, A.-S., Weymaere, J., Willems, S., Deforce, D., & Van Nieuwerburgh, F. (2019). Forensic tri-allelic SNP genotyping using nanopore sequencing. *Forensic Science International: Genetics, 38*, 204-210. doi:https://doi.org/10.1016/j.fsigen.2018.11.012

Coutinho, T. A., & Venter, S. N. (2009). Pantoea ananatis: an unconventional plant pathogen. *Molecular plant pathology, 10*(3), 325-335. doi:10.1111/j.1364-3703.2009.00542.x

Delétoile, A., Decré, D., Courant, S., Passet, V., Audo, J., Grimont, P., . . . Brisse, S. (2009). Phylogeny and Identification of Pantoea Species and Typing of Pantoea agglomerans Strains by Multilocus Gene Sequencing. *Journal of Clinical Microbiology, 47*(2), 300. doi:10.1128/JCM.01916-08

EPPO. (2014). PM 7/120 (1) Pseudomonas syringae pv. actinidiae. *EPPO Bulletin, 44*(3), 360-375. doi:10.1111/epp.12171

Flores, O., Prince, C., Nuñez, M., Vallejos, A., Mardones, C., Yañez, C., . . . Bastías, R. (2018). Genetic and Phenotypic Characterization of Indole-Producing Isolates of Pseudomonas syringae pv. actinidiae Obtained From Chilean Kiwifruit Orchards. *Frontiers in microbiology, 9*, 1907-1907. doi:10.3389/fmicb.2018.01907

Fujikawa, T., & Sawada, H. (2019). Genome analysis of Pseudomonas syringae pv. actinidiae biovar 6, which produces the phytotoxins, phaseolotoxin and coronatine. *Scientific Reports, 9*(1), 3836. doi:10.1038/s41598-019-40754-9

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics, 31*(17), 2877-2878. doi:10.1093/bioinformatics/btv271

Gavini, F., Mergaert, J., Beji, A., Mielcarek, C., Izard, D., Kersters, K., & DE LEY, J. (1989). Transfer of Enterobacter agglomerans (Beijerinck 1888) Ewing and Fife 1972 to Pantoea gen. nov. as Pantoea agglomerans comb. nov. and Description of Pantoea dispersa sp. nov. *International Journal of Systematic and Evolutionary Microbiology, 39*(3), 337-345.

Glaeser, S. P., & Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology, 38*(4), 237-245. doi:https://doi.org/10.1016/j.syapm.2015.03.007

Gong, L., Huang, Y.-T., Wong, C.-H., Chao, W.-C., Wu, Z.-Y., Wei, C.-L., & Liu, P.-Y. (2018). Culture-independent analysis of liver abscess using nanopore sequencing. *PLOS ONE, 13*(1), e0190853. doi:10.1371/journal.pone.0190853

Gupta, A., Jordan, I. K., & Rishishwar, L. (2016). stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics, 33*(1), 119-121. doi:10.1093/bioinformatics/btw586

Herrera, C. M., Koutsoudis, M. D., Wang, X., & von Bodman, S. B. (2008). Pantoea stewartii subsp. stewartii Exhibits Surface Motility, Which is a Critical Aspect of Stewart's Wilt Disease Development on Maize. *Molecular Plant-Microbe Interactions®, 21*(10), 1359-1370. doi:10.1094/mpmi-21-10-1359

Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., & Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. *Phytobiomes Journal, 3*(2), 92-101. doi:10.1094/PBIOMES-01-19-0004-R

Huptas, C., Scherer, S., & Wenning, M. (2016). Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC Research Notes, 9*(1), 269. doi:10.1186/s13104-016-2072-9

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., . . . Tedersoo, L. (2019). Relative performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation sequencing platforms in identification of agricultural and forest pathogens. *bioRxiv*, 592972. doi:10.1101/592972

Marrero, G., Schneider, K. L., Jenkins, D. M., & Alvarez, A. M. (2013). Phylogeny and classification of Dickeya based on multilocus sequence analysis. *International Journal of Systematic and Evolutionary Microbiology, 63*(Pt_9), 3524-3539. doi:https://doi.org/10.1099/ijs.0.046490-0

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffière, A., . . . Thompson, B. M. (2008). The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. *The ISME Journal, 2*(3), 321-334. doi:10.1038/ismej.2007.113

Na, S.-I., Kim, Y. O., Yoon, S.-H., Ha, S.-m., Baek, I., & Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *Journal of Microbiology, 56*(4), 280-285. doi:10.1007/s12275-018-8014-6

Nechwatal, J., Friedrich-Zorn, M., Theil, S., Gebauer, P., & Wensing, A. (2018). Validation of a specific PCR screening test for Pantoea stewartii subsp. stewartii in maize (Zea mays) samples. *EPPO Bulletin, 48*(1), 78-85. doi:10.1111/epp.12448

Passera, A., Compant, S., Casati, P., Maturo, M. G., Battelli, G., Quaglino, F., . . . Mitter, B. (2019). Not Just a Pathogen? Description of a Plant-Beneficial Pseudomonas syringae Strain. *Frontiers in microbiology, 10*, 1409-1409. doi:10.3389/fmicb.2019.01409

Stice, S. P., Stumpf, S. D., Gitaitis, R. D., Kvitko, B. H., & Dutta, B. (2018). Pantoea ananatis Genetic Diversity Analysis Reveals Limited Genomic Diversity as Well as Accessory Genes Correlated with Onion Pathogenicity. *Frontiers in microbiology, 9*, 184-184. doi:10.3389/fmicb.2018.00184

Valiente Moro, C., Tran, F. H., Raharimalala, F. N., Ravelonandro, P., & Mavingui, P. (2013). Diversity of culturable bacteria including Pantoea in wild mosquito Aedes albopictus. *BMC microbiology, 13*, 70-70. doi:10.1186/1471-2180-13-70

Varet, H., & Coppée, J.-Y. (2018). checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes. *Bioinformatics, 35*(5), 901-902. doi:10.1093/bioinformatics/bty706

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research, 27*(5), 737-746.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Xin, X.-F., Kvitko, B., & He, S. Y. (2018). Pseudomonas syringae: what it takes to be a pathogen. *Nature Reviews Microbiology, 16*(5), 316-328. doi:10.1038/nrmicro.2018.17

Appendix 8 – Identify and establish NGS protocols for detection of key culturable plant pathogenic bacteria

Protocols

NGS protocols for detection of key culturable plant pathogenic bacteria are described below. Within most sections, hyperlinks are included to redirect you to the official manufacturer's protocol.

WORKFLOW:

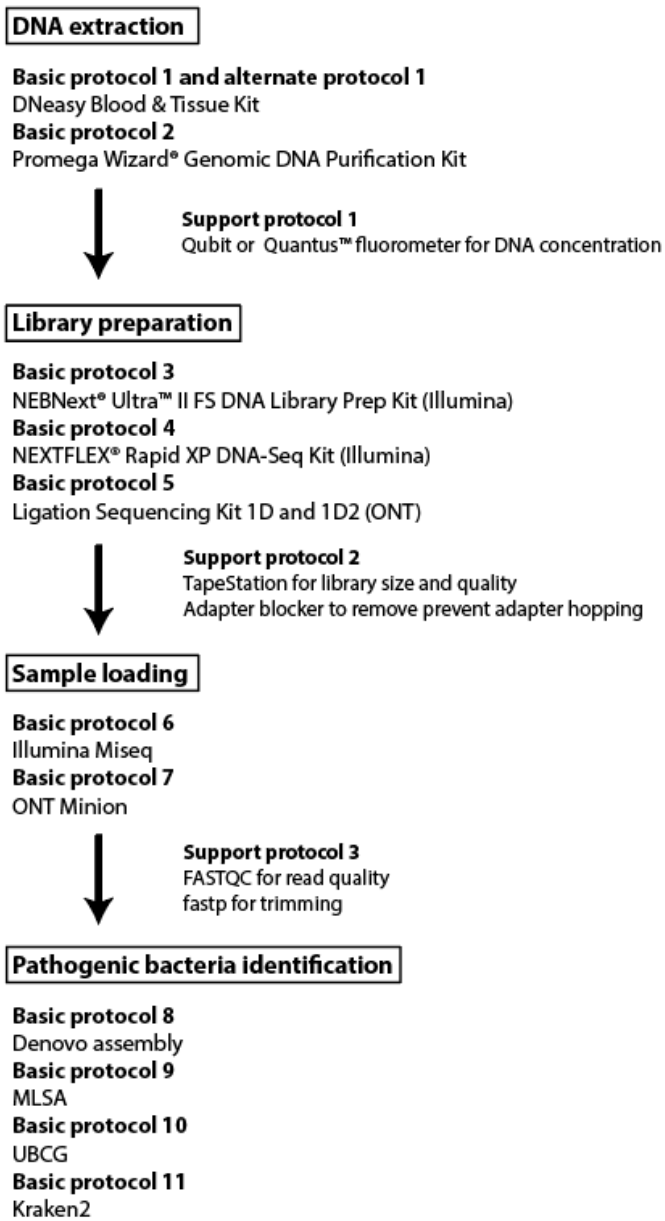


Figure 8.6 Pipeline used to identify culturable plant pathogenic bacteria

DNA extractions

The Qiagen DNeasy Blood & Tissue Kit is recommended for Illumina library preparation as it is the validated bacterial extraction kit in use at Agriculture Victoria Research (AVR). The Promega Wizard® Genomic DNA Purification Kit is recommended for Oxford Nanopore Technology (ONT) library preparation as this precipitation method causes less shearing of the DNA compared to column-based extraction meaning higher molecular weight.

Basic protocol 1: Qiagen DNeasy® Blood & Tissue kit for Illumina sequencing

This kit can be used for bacteria growing in both liquid and solid cultures. Bacterial genomic DNA is extracted using the Qiagen DNeasy Blood & Tissue Kit, following the instructions provided by the manufacturer with minor changes. Options for lysing bacterial cells for 1 hour and RNase A treatment are used. DNA is eluted from the column with half the eluent given volume (100 µl) to increase the concentration of gDNA extracted.

Basic protocol 2: Promega Wizard® Genomic DNA Purification Kit for ONT

Liquid cultures are necessary for this protocol. DNA is extracted according to the manufacturer's instructions with minor modifications. After removing the supernatant in step 2, the pellet is air dried in the open tube in a biosafety cabinet for one hour before continuing the DNA extraction protocol.

Support protocol 1.1 Quantus™ fluorometer

DNA was quantified using the Quantus™ fluorometer (Promega) according to the manufacturer's instructions.

Library preparation

The Illumina NEBNext® Ultra™ II FS DNA Library Preparation Kit is recommended for low input DNA as it can accommodate down to 0.1 ng of DNA input. The Perkin Elmer NEXTFLEX Rapid XP DNA-Seq Kit is recommended for normal use as it is cost effective and it produces longer fragments that aid with genome assembly

Basic protocol 3: NEBNext® Ultra™ II FS DNA Library Prep Kit for low input DNA

Follow the manufacturer's instructions to prepare libraries. Parts of this protocol changes to accommodate the amount of input DNA. Targeted fragment size is 100 bp-250 bp.

This kit can be adapted to the Perkin Elmer NEXTflex® Unique Dual Index Barcodes. Changes include replacing NEBNext Adaptor for Illumina** for NEXTflex® Unique Dual Index Barcodes 1 - 96* (25 µM) in step 1.2.1 while adjusting adapter dilutions according to the amount of input DNA; skipping step 1.2.4 and 1.2.5; and replacing the 10 µl Index/Universal Primer**** with 2 µl NEXTflex® Primer Mix (12.5 µM) (provided with the barcodes) compensating the lesser volume with ultragrade water in step 1.4.1B.

Basic protocol 4: NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) for standard use

Half reactions can be used for sequencing on the Illlumina MiSeq, HiSeq and NovaSeq platforms. Reactions are prepared according to the manufacturer's instructions, except volumes are of the components are halved. This protocol is dependent on the amount of input DNA and should be adjusted as such. Minimum amount of DNA for the half reaction protocol is 10 ng of DNA. Targeted fragment size was 200 bp-300 bp.

Option 2 of the protocol is suggested for DNA above 10 ng as it gives single sharp peaks on the TapeStation indicating lack of primer/adapter dimers and smearing

NOTE: EDTA disrupts the fragmentation step in this protocol. It is suggested to either clean the library with Ampure beads and resuspend with ultragrade water (page 23) or to use their condition solution (page 24).

Oxford Nanopore Technologies Library Preparation
The Ligation Sequencing Kit (SQK-LSK109) is recommended with the new flow cell chemistry, R10.3. The polishing software, Racon and Medaka is recommended to correct reads to improve read accuracy.

Basic protocol 5 1D Genomic DNA sequencing by Ligation kit (SQK-LSK109)

ONT libraries should be generated according to the manufacturer's instructions. The minimum amount of DNA required is 1 - 3 µg of DNA.

Support protocol 2.1 TapeStation

A detailed protocol to use the 2200 TapeStation system (Agilent technologies) for quality checking of libraries can be found here. For Illumina sequencing, HSD1000 tape assays are used for DNA concentrations below 30 ng/µl and D1000 tape assays are used DNA concentrations above 30 ng/µl . When analysing DNA for the ONT MinION device the D5000 tape assays are used.

Figure 8.1 and Figure 8.2 show examples of a bad and good library, respectively. Libraries should peak at a minimum of 280 bp with no smears or peaks below the 200 bp mark.

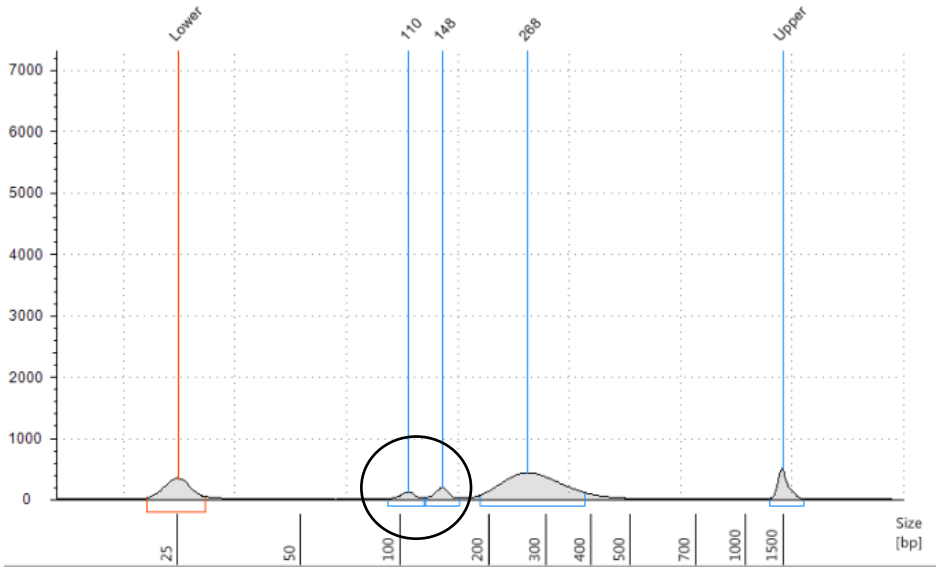


Figure 8. 7 Example of a bad library on a HSD1000 tape assay on the TapeStation. The peaks between 100 – 150 bp (circled) indicates the presence of either primer dimers or excess indexes.

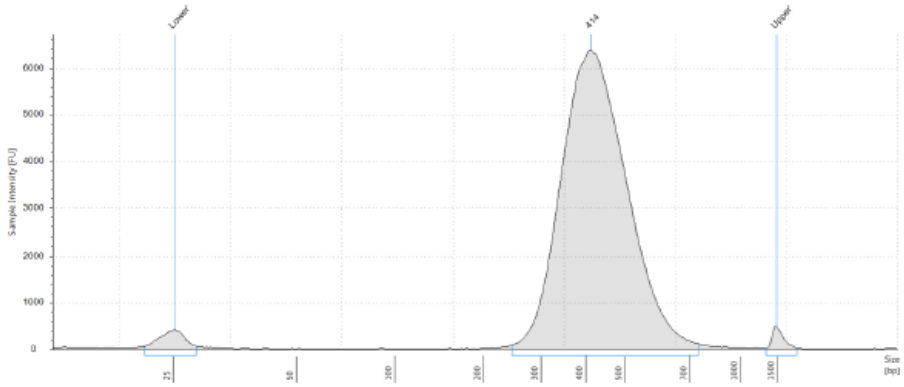


Figure 8. 8 Example of a good library on a HSD1000 tape assay on the TapeStation. One strong peak is observed at 414 bp.

Support protocol 2.2 Adapter blocker

Blocking unligated adapters to prevent index hopping is done following the instructions provided by the manufacturer.

Sample preparation and loading

Basic protocol 6.1: Illumina Miseq (MiSeq Reagent Kit v3) sample preparation for loading, denaturing and diluting libraries

To prepare the PhiX calibration control and denature and dilute libraries for loading onto the Miseq sequencer follow the instructions provided by the manufacturer. At the **Denature and Dilute PhiX Control** step it is suggested to make 20 pM PhiX libraries. PhiX libraries can be kept up to a month at -15°C to -25°C. Aim to make between 8 pM to 10 pM libraries and combine with 1% 20 pM PhiX.

Basic protocol 6.2: Illumina Miseq (MiSeq Reagent Kit v3) Sample loading

Preparing sample sheet

Create the sample sheet in Microsoft Excel and save it as a .csv file. There are three components of the sample sheet. First is the header where details of the run are written. Some of the fields can be changed to suit the current experiment. This includes investigator name, date and assay. In the experiment name, the flowcell ID (Figure 8.4B) is usually incorporated into the name. The two fields that cannot be changed are the workflow field, which needs to be filled with the term GenerateFASTQ, and the chemistry field which needs to be filled with the term amplicon.

In the Reads section the amount of basepairs for read 1 and read 2 sequence is listed. Since this protocol is aimed at using the MiSeq Reagent Kit V3, 300 is written as up to 300 sequencing cycles can be completed. In the Settings section, the adapter sequence can be filled, however, it can remain also empty. The Data section is where information of individual samples is written. Example of a sample sheet is shown below.

[Header]									
IEMFileVersion	4								
Investigator Name	Your name								
Experiment Name	J3KH7-your project								
Date	Date								
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	Truseq								
Description	Description								
Chemistry	Amplicon								
[Reads]									
300									
300									
[Settings]									
Adapter	Adapter sequence								
[Data]									
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
S-14RD	S-14RD	1	A1	S747	TAGGAGCT	S524	GCCTTAAC	Pathogen	
S-17RD	S-17RD	1	B1	S716	CGAATTGC	S559	GTAAGGTG	Pathogen	
S-19RD	S-19RD	1	C1	S770	GTCCTAAG	S540	AGCTACCA	Pathogen	
S-36RD	S-36RD	1	D1	S749	CTTAGGAC	S565	CTTCACTG	Pathogen	
S-39RD	S-39RD	1	E1	S784	TCCACGTT	S541	GGTTGAAC	Pathogen	
S-40RD	S-40RD	1	F1	S780	CAACACAG	S530	GATAGCCA	Pathogen	
20-01805-1	20-01805-1	2	A1	UDI0094	CGTCTGCG	UDI0094	TTCACAAT	Pathogen	
20-01805-2	20-01805-2	2	B1	UDI0095	TCGATATC	UDI0095	ACACGAGT	Pathogen	
20-01801	20-01801	2	C1	UDI0278	ATCGTTAG	UDI0278	TGCGAAGT	Pathogen	

Figure 8. 9 An example of a sample sheet layout.

Once the sample sheet is completed, the name of the file should be saved as the barcode number on the reagent cartridge. As an example, the name of this sample sheet will be MS9857995-600V3.csv as shown from the Figure 4A.

A

B

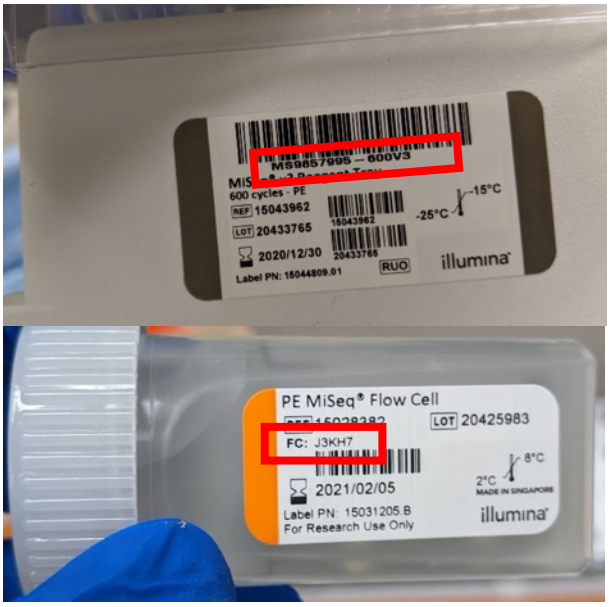


Figure 8.10 Barcodes from the MiSeq kits required to create the sample sheet. A) Reagent cartridge and B) Flow Cell

Sample loading

In depth protocols and troubleshooting can be found here. The sequence of steps are as follows:

1. Inspect the Reagent Cartridge
2. Load Sample Libraries
3. Set Up a Run Using MCS
4. Clean the Flow Cell
5. Load the Flow Cell
6. Load Reagents

a. Load Sample Libraries can be done here if you prefer to leave you libraries on ice

7. Change Sample Sheet

a. Can be added through a USB or through email sent to self as you can log on your email on the MiSeq

8. Post-Run Wash

WARNING: of the set of reagents used in this process contains potentially hazardous chemicals. Personal injury can occur through inhalation, ingestion, skin contact, and eye contact. Wear protective equipment, including eye protection, gloves, and laboratory coat appropriate for risk of exposure. Handle used reagents as chemical waste and discard in accordance with applicable regional, national, and local laws and regulations.

Basic protocol 7: ONT loading Samples

ONT libraries should be loaded according to the manufacturer's instructions.

Bacteria identification

Conda or Miniconda needs to be installed before running most of these scripts. Refer to link on how to install Conda or Miniconda on your computer. Basic bash commands are required for most of the protocols.

Kraken2 is recommended for a swift analysis to identify genus which is then followed by *de novo* assembly to generate contigs. UBCG is the recommended protocol for phylogenetic analysis used for identification.

Basic protocol 8.1: Denovo assembly – SPAdes (short reads)

SPAdes is available to download from GitHub at https://github.com/ablab/spades#sec1.2. It can be installed by going to the directory in which you wish SPAdes to be installed and using the following command on a Linux terminal:

```
wget http://cab.spbu.ru/files/release3.14.1/SPAdes-3.14.1-Linux.tar.gz
tar -xzf SPAdes-3.14.1-Linux.tar.gz
cd SPAdes-3.14.1-Linux/bin/
```

Once installed, you should test the installed software before starting any analysis by running the script:

```
python /SPAdes-3.14.1-Linux/bin/spades.py --test
```

The log would indicate if your test was successful. If it is, you can start assembly using the script:

```
python /SPAdes-3.14.1-Linux/bin/spades.py -k 127,107,87,67,31 --careful --pe1-1 *R1.fastq.gz
--pe1-2 *_R2.fastq.gz -o /output/folder/
```

SPAdes will create your output folder so making one beforehand is not necessary. If you do make one, an error will occur, and you would either add "-force" to the script or remove the folder.

Basic protocol 8.2: Denovo assembly – canu (long reads)

Canu is available for download in Conda. It can be installed by running the script:

```
conda install --name canu canu
```

Once installed, you can run the script:

```
canu -p assembly -d /output/folder/ genomeSize=40k minReadLength=500 stopOnReadQuality=false
useGrid=false corOutCoverage=1000 maxThreads=32 minThreads=30 -fast -nanopore-raw
input.fastq.gz
```

However, you would need to change the genome size on the script to accommodate your sequenced bacteria.

Basic protocol 8.3: Denovo assembly – Unicycler (hybrids)

Unicycler is available for download in Conda. It can be installed by running the script:

```
conda create --name unicycler unicycler
```

Once installed, you can run the script:

```
python unicycler-runner.py --long MinION.fastq.gz --short1
/Illumina_reads/*R1_clean.fastq.gz --short2 /Illumina_reads/*_R2_clean.fastq.gz -o
```

```
/output/folder/ -t 24
```

Basic protocol 9: MLSA

Below is the shell script used to pull out genes and concatenate them for MLSA phylogenetic analysis. Before this script can be run, some setup is necessary. Firstly, is to install Prokka, an annotation tool. Prokka can be installed through Conda using the script:

```
conda create --name prokka prokka
```

This will create an environment named prokka while installing Prokka.

Secondly, all genomes where the genes are to be extracted from should be in one folder and the path is to be filled in "`/path/to/fasta/files/`". Lastly, a sample or two of each gene needs to be compiled in one fasta file which will replace "`for_blast_MLSA.fa`" in the script. Once the script has been run, the cat_MLSA.fasta file will contain all sequences necessary to for alignment and phylogenetic analysis.

```
## Activate conda/prokka ##
eval "$(conda shell.bash hook)"
conda activate prokka


###### Run scripts ########

echo "---Prokka started---"
cd /path/to/fasta/files/
for f in *.fasta; do

        prokka --prefix ${f%%.*} --force --centre X --compliant --addgenes $f --kingdom
Bacteria --gcode 11
        echo "---makedb started---"
        makeblastdb -in ${f%%.*}/*.ffn -dbtype nucl -title ${f%%.*}_db -out ${f%%.*}_db
        echo "---blastn started---"
        blastn -query for_blast_MLSA.fa -db ${f%%.*}_db -out MLSA_blast_${f%%.*}_db.txt -
outfmt "6 sseqid" -evalue 10e-80
        echo "---remove duplicates started---"
        uniq MLSA_blast_${f%%.*}_db.txt MLSA_blast_sorted_${f%%.*}_db.txt
        echo "---extraction started---"
        perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV'
MLSA_blast_sorted_${f%%.*}_db.txt ${f%%.*}/*.ffn > ${f%%.*}_MLSA_genes.fasta
        done


echo "---Gene extraction started---"


for f in *_MLSA_genes.fasta; do


        awk 'BEGIN{RS=">"} NR>1 {gsub("\n", "\t"); print ">"$0}' $f | sort -t $'\t' -k2 |
awk '{sub("\t", "\n"); gsub("\t", ""); print $0}' > ${f%%.*}_sorted.fasta
        sed -i '/>/d' ${f%%.*}_sorted.fasta
```

```
done
        for f in *_sorted.fasta; do echo ">""${f%%.*}";echo;cat "$f"; done > cat_MLSA.fasta
done

echo "=====END===="
```

Basic protocol 10: UBCG

UBCG is available to download from https://www.ezbiocloud.net/tools/ubcg. A zip file will be downloaded which would need to be extracted and moved into your Linux system. Before the java scripts can be run, an environment in Conda can be made by running the script:

```
conda create --name UBCG
conda install java-jdk
conda install prodigal
conda install hmmer
conda install mafft
conda install fasttree
```

This will install all necessary requirements to run the UBCG script. Once ready, fasta files of chosen genomes and samples are run through the first UBCG script:

```
conda activate UBCG
java -jar UBCG.jar extract -bcg_dir /output/file/ -i genome.fasta -label "sample_name"
```

This will create UBCG files. Accumulated UBCG files will need to be in one folder as the next step is to align and create a phylogenetic tree through RAxML through the script:

```
java -jar UBCG.jar align -bcg_dir /output/file/ -prefix psa
```

The output folder will contain both the fasta file and .nwk file of all gene alignment and a combined gene alignment. The .nwk file can then be viewed on FigTree or the Mega software.

Basic protocol 10: Kraken2

Kraken2 is available for download in Conda. It can be installed by running the script:

```
conda create --name kraken2 kraken2
```

Once installed, you can install the database from here. When done, you can run the script:

```
kraken2 --db /path/to/database --gzip-compressed --paired *R1.fastq.gz
*R2.fastq.gz --report name.txt --output report
```

Remove `--gzip-compressed` if not using a compressed file.

Support protocol 3.1 FASTQC

FASTQC is a common method for checking the quality of reads. It can be downloaded in Conda or as a module on Linux. To run FASTQC on your reads, you can run the script:

```
fastqc -o /output/folder -f fastq *R1.fastq.gz *R2.fastq.gz
```

Data obtained from FASTQC is in a .html file. When opened, a summary is given (Figure 5) with details of each error accompanying it.

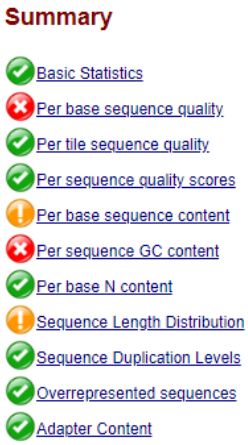


Figure 8.11 Summary as given by FASTQC results.

Support protocol 3.2 fastp

Fastp is available for download in Conda. It can be installed by running the script:

```
conda install --name fastp fastp
```

For paired end reads, you can run the script:

```
fastp -i input_R1_001.fastq.gz -I input_R2_001.fastq.gz -o outpout_R1.fastq.gz -O
outpout__R2.fastq.gz
```

For long reads such as those from ONT, you can run the script:

```
fastp -i input.fastq.gz -o outpout.fastq.gz
```

Appendix 9 – Lists of genomes used in comparative analysis.

Table 9.1. Genomes of Pantoea strains downloaded from EzBioCloud for use in comparative analysis.

***Pantoea* strain**	**DB accession number**
Pantoea agglomerans 3	GCA_001597625.1
Pantoea agglomerans 824-1	GCA_001661985.1
Pantoea agglomerans Bl3	GCA_004793995.1
Pantoea agglomerans C410P1	GCA_001709315.1
Pantoea agglomerans CFSAN033090	GCA_001264395.1
Pantoea agglomerans P10c	GCA_001288285.1
Pantoea allii LMG 24248	GCA_002307475.1
Pantoea allii OXWO6B1	GCA_001641135.1
Pantoea allii PNA 200-10	GCA_003148935.1
Pantoea allii SJZ147	GCA_007828605.1
Pantoea ananatis LMG 20103	GCA_000025405.2
Pantoea ananatis LMG 2665	GCA_000710035.1
Pantoea ananatis NS296	GCA_001475715.1
Pantoea ananatis PA13	GCA_000233595.1
Pantoea ananatis PA4	GCA_000710015.1
Pantoea ananatis PNA 15-1	GCA_002959585.1
Pantoea anthophila Sc1	GCA_000255315.2
Pantoea brenneri 3-5-1	GCA_000731025.2
Pantoea coffeiphila JUb26	GCA_003751505.1
Pantoea conspicua LMG 24534	GCA_002095315.1
Pantoea cypripedii LMG 2657	GCA_002095535.1
Pantoea deleyi LMG 24200	GCA_006494415.1
Pantoea dispersa NS215	GCA_001476715.1
Pantoea eucalypti FBS135	GCA_000731025.2
Pantoea eucalypti LMG 24197	GCA_009646115.1
Pantoea eucalypti MHSD5	GCA_002959725.1
Pantoea rodasii DSM 26611	GCA_002811195.1
Pantoea rwandensis LMG 26275	GCA_002095475.1
Pantoea stewartii subsp *indologenes* LMG 2632	GCA_000757405.1
Pantoea stewartii subsp *indologenes* MS1	GCA_010273335.1
Pantoea stewartii subsp *indologenes* PNA 03-3	GCA_003201175.1
Pantoea stewartii subsp *indologenes* PNA 14-12	GCA_004364735.1
Pantoea stewartii subsp *indologenes* RSA13	GCA_001477215.1
Pantoea stewartii subsp *indologenes* S301	GCA_001310295.1
Pantoea stewartii subsp *stewartii* CCUG 26359	GCA_008801695.1
Pantoea stewartii subsp *stewartii* DC283	GCA_002082215.1
Pantoea vagans 9140	GCA_000745295.1
Pantoea vagans C9-1	GCA_000148935.1
Pantoea vagans ET-90	GCA_008039815.1

Pantoea vagans LMG 24199	GCA_004792415.1
Pantoea vagans PV989	GCA_003032455.1
Pantoea vagans TYU1	GCA_002287745.1
Pantoea wallisii LMG 26277	GCA_002095485.1

Table 9.2. Genomes of Pseudomonas syringae strains downloaded from National Center for Biotechnology Information (NCBI) for use in comparative analysis.

***Pseudomonas syringae* strain**	**GenBank assembly accession**
Pseudomonas syringae pv. *aceris* A10853	GCA_001270465.1
Pseudomonas syringae pv. *aceris* ICMP9850	GCA_003700545.1
Pseudomonas syringae pv. *aceris* ICMP9852	GCA_003700575.1
Pseudomonas syringae pv. *aptata* G733	GCA_003699535.1
Pseudomonas syringae pv. *aptata* ICMP459	GCA_001401335.1
Pseudomonas syringae pv. *atrofaciens* DSM5025	GCA_003699335.1
Pseudomonas syringae pv. *atrofaciens* LMG5095	GCA_003047185.1
Pseudomonas syringae pv. *avellanae* ISPaVe013	GCA_000302795.1
Pseudomonas syringae pv. *castaneae* CFBP4217	GCA_001538245.1
Pseudomonas syringae pv. *cerasicola* CFBP6109	GCA_900235885.1
Pseudomonas syringae pv. *cerasicola* CFBP6110	GCA_900289135.1
Pseudomonas syringae pv. *cilantro* 0788_9	GCA_001293775.1
Pseudomonas syringae pv. *dysoxyli* CFBP2356	GCA_009898305.1
Pseudomonas syringae pv. *lapsa* ATCC10859	GCA_001482725.1
Pseudomonas syringae pv. *lapsa* ICMP3946	GCA_003698895.1
Pseudomonas syringae pv. *lapsa* ICMP3947	GCA_001400495.1
Pseudomonas syringae pv. *lapsa* ICMP8813	GCA_003699285.1
Pseudomonas syringae pv. *panici* LMG2367	GCA_000282735.1
Pseudomonas syringae pv. *papulans* CFBP1754	GCA_001535905.1
Pseudomonas syringae pv. *pisi* 1456A	GCA_003205935.1
Pseudomonas syringae pv. *pisi* 202	GCA_003205965.1
Pseudomonas syringae pv. *pisi* H5E3	GCA_003699035.1
Pseudomonas syringae pv. *pisi* PP1	GCA_000452445.3
Pseudomonas syringae pv. *syringae* B301D	GCA_000988485.1
Pseudomonas syringae pv. *syringae* B64	GCA_000331385.1
Pseudomonas syringae pv. *syringae* B728a	GCA_000012245.1
Pseudomonas syringae pv. *syringae* SM	GCA_000412165.1

Table 9.3. Genomes of Pseudomonas syringae pv actinidiae strains downloaded from National Center for Biotechnology Information (NCBI) for use in comparative analysis.

***Pseudomonas syringae* pv. *actinidiae* strain**	**GenBank assembly accession**
Psa105743	GCA_002750505.1
Psa127771	GCA_002750525.1
Psa127784	GCA_002750515.1

C11	GCA_002174935.1
C12	GCA_002174895.1
C13	GCA_002174925.1
C15	GCA_002174975.1
C18	GCA_002175065.1
C24	GCA_002175085.1
C27	GCA_002175075.1
C3	GCA_002174875.1
C31	GCA_002175185.1
C66	GCA_002111675.1
C69	GCA_002111745.1
C70	GCA_002111695.1
C74	GCA_002111785.1
C75	GCA_002111795.1
CFBP8039	GCA_001497595.1
CFBP8043	GCA_001497495.1
CFBP 7286	GCA_000245415.1
CH2010-6	GCA_000245475.1
CRAFRU12.29	GCA_002024285.1
CRAFRU14.08	GCA_002024305.1
CRAFRU8.43	GCA_000233815.2
ICMP18802	GCA_002003545.1
ICMP 18708	GCA_000344355.2
ICMP 18804	GCA_000416905.2
ICMP 18807	GCA_000416625.1
ICMP 18884	GCA_000648735.3
ICMP 19071	GCA_000416485.1
ICMP 19072	GCA_000416885.1
ICMP 19098	GCA_000416545.1
ICMP 19497	GCA_001467225.1
ICMP 20586	GCA_001913215.1
ICMP 9617	GCA_000658965.1
ICMP 9853	GCA_000344335.2
ICMP 9855	GCA_000416665.1
IPV-BO-8893	GCA_002750445.1
IPV-BO-9286	GCA_002750435.1
J38	GCA_002111805.1
K4	GCA_002111835.1
K5	GCA_002111875.1
K6	GCA_002111865.1
M227	GCA_002890365.1
M256	GCA_002890485.1
M302091	GCA_000145865.1

M401	GCA_002890475.1
MAFF212054	GCA_002762615.1
MAFF212056	GCA_001270005.1
MAFF212063	GCA_002763655.1
NZ-47	GCA_001913235.1
NZ38	GCA_002112095.1
NZ43	GCA_002112045.1
NZ45	GCA_001913215.1
NZ48	GCA_002111955.1
NZ59	GCA_002112165.1
P220	GCA_003665415.1
Shaanxi M228	GCA_000344475.3
kw1361	GCA_013305425.1

Appendix 10: Phase 1 – Large scale side-by-side comparison of PEQ vs small RNA-seq NGS methods

Summary

A total of 90 samples were identified for the first batch (batch 1) comparison of PEQ and NGS methods. In average 48M reads (5Gb) per sample was generated using a small RNA-seq approach. In comparison to a previous pilot study (milestone 104) the content of batch 1 samples showed a drastic increase in ribosomal RNA content. To understand the technical challenges leading to a distinct RNA content profile, we investigated batch 1 samples (**Part 1)** and also conducted a controlled trouble shooting experiment to assess if commercial kits used in the generation of the NGS data may have impacted the outputs (**Part 2**). We are working actively with the service provider to resolve the technical issues to conduct future large scale sequencing experiments. In addition, we compared available PEQ diagnostic results with preliminary NGS-based batch 1 diagnosis results and found a strong correlation despite the lesser amount of NGS data available for virus/viroid detection (**Part 3**).

Sample metadata

To conduct large scale side-by-side comparisons of PEQ vs NGS methods, a total of 186 samples were identified and split into two batches for processing: i) batch 1 encompassing 90 samples (**Table 10.1**), and ii) batch 2 consisting of 96 plants (**Table 10.2**). Total RNA was collected and sent to NGS service providers. NGS data for batch 1 was received (Appendix 6), while for batch 2 are being processed.

Table 10.1: Sample metadata information for PEQ plants encompassing the NGS batch 1 set.

Sample ID	Commodity / Genus	Plant tissue collected	Collection date (Month, Year)	Total RNA extraction kit
MT020	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT021	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT022	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT023	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT024	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT025	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT026	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT027	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT028	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT029	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT030	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT031	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT032	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT033	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT034	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT035	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT036	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT037	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT038	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT039	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT040	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA

MT041	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT042	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT043	Lonicera	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT044	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT045	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT046	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT047	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT048	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT049	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT050	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT051	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT052	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT053	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT054	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT055	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT056	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT057	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT058	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT059	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT060	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT061	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT062	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT063	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT064	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT065	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT066	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT067	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT068	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT069	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT070	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT071	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT072	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT073	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT074	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT075	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT076	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT077	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT078	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT079	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT080	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA

MT081	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT082	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT083	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT084	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT085	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT086	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT087	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT088	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT089	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT090	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT091	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT092	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT093	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT094	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT095	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT096	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT097	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT098	prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT099	prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT100	rosa	50 mg symptomatic leaf	Nov-20	Maxwell RSC Pure RNA
MT101	rosa	50 mg symptomatic leaf	Nov-20	Maxwell RSC Pure RNA
MT102	rosa	50 mg symptomatic leaf	Nov-20	Maxwell RSC Pure RNA
MT103	Citrus	50 mg symptomatic leaf	Jul-20	Maxwell RSC Pure RNA
MT104	Citrus	50 mg leaf cross section	Jul-20	Maxwell RSC Pure RNA
MT105	Citrus	50 mg leaf cross section	Jul-20	Maxwell RSC Pure RNA
MT106	Citrus	50 mg leaf cross section	Jul-20	Maxwell RSC Pure RNA
MT107	Citrus	50 mg leaf cross section	Jul-20	Maxwell RSC Pure RNA
MT108	Citrus	50 mg leaf cross section	Jul-20	Maxwell RSC Pure RNA
MT114	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA

Table 10.2: Sample metadata information for PEQ plants encompassing the NGS batch 2 set

Sample ID	Commodity / Genus	Plant tissue collected	Collection date (Month, Year)	Total RNA extraction kit
MT116	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT117	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT118	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT119	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT120	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT121	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT122	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT123	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA

MT124	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT125	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT126	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT127	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT128	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT129	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT130	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT131	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT132	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT133	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT134	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT135	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT136	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT137	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT138	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT139	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT140	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT141	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT142	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT143	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT144	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT145	Rubus	50 mg leaf/petiole	Sep-20	Maxwell RSC Pure RNA
MT146	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT147	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT148	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT149	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT150	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT151*	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT152	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT153	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT154	Persea	50 mg leaf/midrib	Sep-20	Maxwell RSC Pure RNA
MT155	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT156	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT157	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT158	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT159	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT160	Pyrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT161	Pyrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT162	Pyrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT163	Pyrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA

MT164	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT165	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT166	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT167	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT168	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT169	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT170	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT171	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT172	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT173	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT174	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT175	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT176	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT177	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT178	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT179	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT180	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT181	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT182	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT183	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT184	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT185	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT186	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT187	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT188	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT189	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT190	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT191	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT192	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT193	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT194	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT195	Malus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT196	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT197	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT198	Olea	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT199	Prunus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT200	Pyrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT201	Ficus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT202	Rubus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT203	Citrus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA

MT204	Rubus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT205	Rubus	50 mg leaf/petiole	Nov-20	Maxwell RSC Pure RNA
MT206	Rubus	50 mg leaf/petiole	Nov-20	Maxwell RSC Pure RNA
MT207	Rubus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT208	Rubus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT209	Rubus	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT210	Solanum	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA
MT211	Solanum	50 mg leaf/midrib	Nov-20	Maxwell RSC Pure RNA

Part 1: NGS data for batch 1 plants show an increased amount of ribosomal RNAs

Small RNA-seq HTS data received for batch 1 - CAGRF20093982 (90 samples) showed increased level of ribosomal RNAs as compared to previous control plants (12 samples) tested to define depth coverage for large scale HTS processing of samples. Control plants showed in average 29.8% of reads derived from rRNAs, while the batch 01 samples showed in average to 67.5% of the reads mapped onto rRNAs (**Figure 10.1**). We found reads mapped onto rRNAs ranging in size from 18 nt to over 75 nt, showing an increased number of mapped reads as the size get shorter in length (**Figure 10.2**).

Mapping of small RNA reads onto other potential sources including rRNAs, mitochondrion, chloroplast, plant miRNAs, transposon and others shows mapped reads with distinct profile to those mapped onto rRNAs, in contrast to rRNAs, reads mapped onto other sources show the presence of specific size peaks (**Figure 10.3**).

Further analysis of reads mapped onto rRNAs show that these are mapped onto 5S, 5.8S, 18S and 26S rRNAs. Surprisingly mapped rRNA reads show a significant bias to map onto the 'forward' strand (**Table 10.1** and **Figure 10.4**). In contrast, reads mapped onto a reference virus and the chloroplast genome did not show such strand mapping bias (**Table 10.1**). To illustrate further these observations, find included the genome browser mapping of reads onto 5.8S (**Figure 10.5**), 26S (**Figure 10.6**) and a reference virus (**Figure 10.7**).

In conclusion, we did not observe apparent issues with reads derived from viruses, chloroplast, mitochondria, transposon and plant miRNAs. However, small RNAs derived from ribosomal RNAs account for a significantly large amount of reads in batch 01 (96 samples) HTS data. Furthermore, strand mapping bias onto the forward strand was observed for small RNAs mapped on ribosomal RNAs and chloroplast genome.

Plant samples

Diverse plant species are included in both 'control plants' and 'batch 01' sample sets:

- **Control plants**: 3 citrus, 2 rubus, 2 strawberry, 1 miscanthus, 1 sweet potato, 1 prunus, 2 vitis
- **Batch 1**: 79 Prunus, 6 citrus, 5 strawberry, 3 rosa, 2 rubus and 1 lonicera

Methodology

Total RNA extraction

Tissue samples are collected from leaves, midrib and/or petioles using sterile scalpel (single use) and gloves are changed between samples. Total RNA is extracted and quantified in a dedicated sample preparation environment. New steel balls and tubes are used for homogenisation. Extraction is done with a Promega Maxwell RSC Pure RNA kit (lot number recorded) and the procedure is to use only reagents that come in the kit (this includes all plasticware, tubes and reagents). In the homogenization buffer we are required to add PVP, and we have a dedicated container only used for this purpose. All tips are barrier tips and come pre-racked. Lab benches are wiped down either with 80% ethanol or RNAse AWAY. The same extraction protocol has been used for both 'control plants' and 'batch 01' sample sets.

RNA origin analysis

To assess the origin of small RNAs reads were mapped using bowtie allowing up to 1 mismatch onto rRNAs (https://www.arb-silva.de; SILVA_138.1_LSURef_NR99_tax_silva_trunc.fasta.gz), mitochondria, chloroplast, transposon, PhyX, vector sequences (Univec) and plant miRNAs representative sequences.

rRNA analysis

Representative sequences for Prunus 5S (119nt), 5.8S (159nt), 18S (1,807nt), 26S (2,827nt) and **28S** (3,172nt) were selected to assess the origin of rRNAs reads present in Prunus HTS samples MT020, MT021, MT022 and MT023. Additionally, a representative virus genome found only in sample MT020 and chloroplast genome were also included to assess the mapping of reads.

Results

RNA origin of small RNAs in plant HTS sets

Figure 10.1 depicts the percentage reads mapped onto rRNAs, mitochondria, chloroplast, transposon, plant miRNAs, PhIX, vector sequences and the remaining fraction of unmapped reads that is of interest for diagnostic analysis. We observed a large increase in the percentage of reads that are mapped onto rRNAs in the latest Batch 01 as compared to the initial control plant set.

The percentage of reads mapped onto chloroplast, mitochondria, transposon and plant miRNAs show results for both control and batch 01 sets. However, the fraction of unmapped reads, which are reads of interest for our downstream virus/viroid diagnostic analysis shows considerable reduction in the latest batch 01 as compared to control plants (**Figure 10.1**). We found in average the number of unmapped reads is 21.2 M (21,229,672) and 7.8M (7,816,227) for control and batch 01 sample sets, respectively. This means there is a reduction of ~63% of the reads that can be used for diagnostic purposes which directly impact on our capability to reliably perform the diagnostic of plant viruses and viroids.

rRNA analysis – bias of mapping of small RNAs onto the forward strand of all types of rRNAs

Mapping of small RNA reads onto representative 5S, 5.8S, 18S, 26S and 28S rRNA sequences shows a significant number of reads mapped onto the forward strand, while such strand mapping bias is not observed for reads mapped onto chloroplast genome or a reference viral genome (**Table 10.3** and **Figure 10.4**). We estimate that in average the fold increase of reads mapped onto the forward strand is 4884 (5S), 4015 (5.8S), 1609 (18S), 2453 (26S) and 1267 (28S) in four selected samples. In contrast the fold increase of reads mapped onto the forward strand of the chloroplast genome was in average only 1.5 across the same MT020, MT021, MT022 and MT023 samples.

As an example, to illustrate the mapping bias of rRNAs onto the forward strand, **Figure 10.5** and **Figure 10.6** shows the mapping of reads onto the 5.8S (Accession MT796549.1) and 26S (Accession MT796549.1) representative sequences. The strand bias is not observed for representative virus (Accession KY635989; **Figure 10.7**).

Conclusions (Part 1):

We did not observe apparent issues with reads derived from viruses, chloroplast, mitochondria, transposon and plant miRNAs. However, small RNAs derived from ribosomal RNAs account for a significantly large amount of reads in batch 01 (96 samples) HTS data. Furthermore, strand mapping bias onto the forward strand was only observed for small RNAs mapped on ribosomal RNAs.

The amount of usable data from batch 01 for the diagnostic of plant viruses and viroids limits the reliable detection of true positives and decision making. Based on our preliminary sequencing and analysis of control plants, we estimated that we require in average 21 million reads for virus diagnosis, however, the batch 01 set resulted in average in 7.8 million usable reads for diagnostics. This imply that 3-times the current available amount of reads for the batch 01 set will be necessary to complete our downstream analysis.



Figure 10.1: Percentage of mapped reads onto various RNA sources. NGS reads were mapped onto ribosomal RNAs, chloroplast and mitochondria genomes, microRNAs, transposons and vector sequences.

Figure 10.2: Size distribution of reads mapped onto ribosomal RNAs.

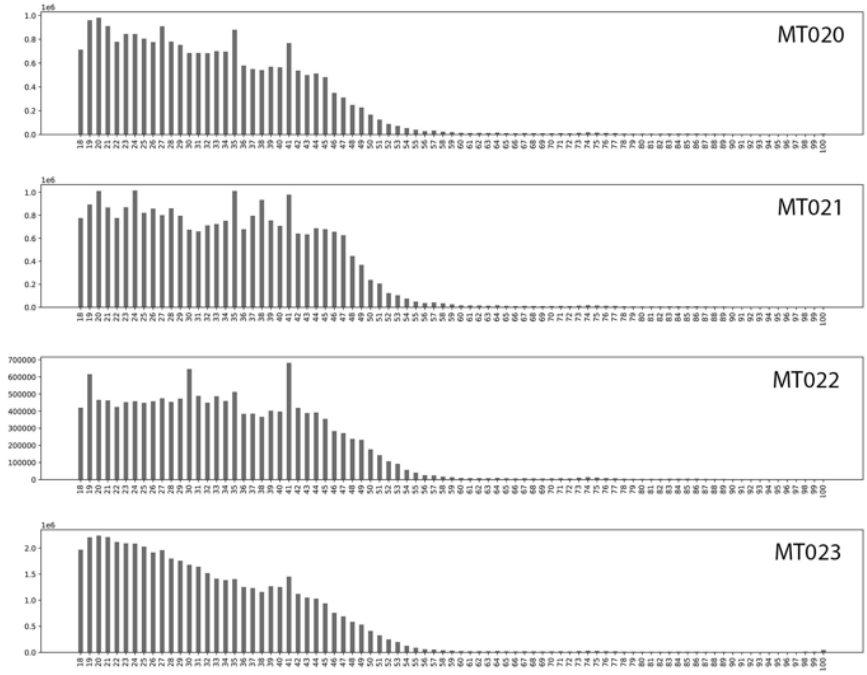


Figure 10.3: Example of the size distribution of mapped reads onto various RNA sources

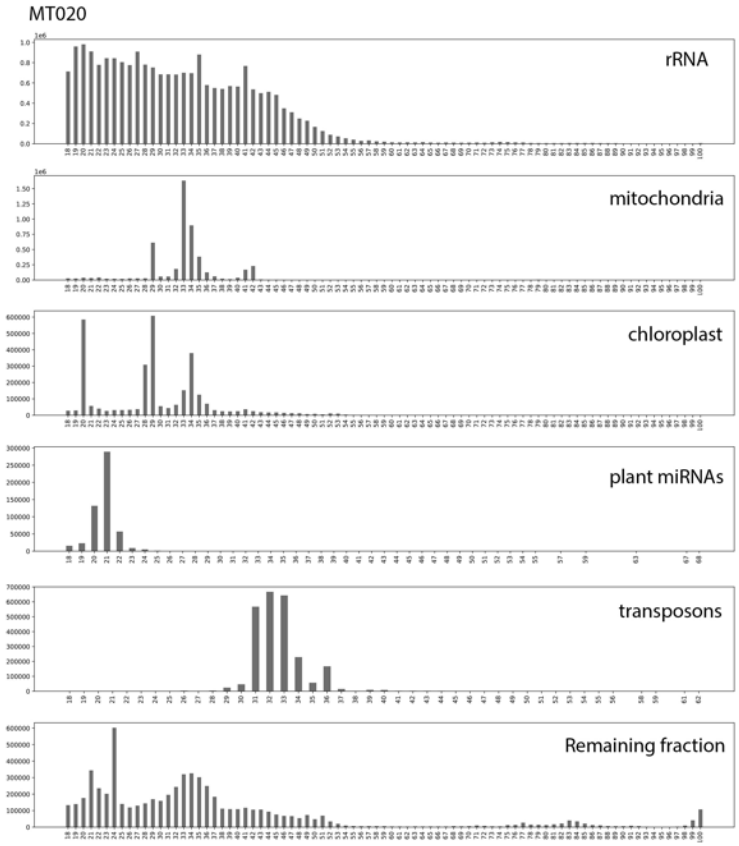


Table 10.3: Statistics of mapped small RNA reads onto reference sequences including a virus (Palv), rRNAs (5S, 5.8S, 26S and 28S) and chloroplast genome. The number of mapped reads on both forward (+) and reverse (-) strands are presented for each representative sequence. Note, all ribosomal RNAs show a significant bias of reads mapped onto the forward strand as compared to the virus and chloroplast derived small RNAs.

Sequence	Length (bp)	Strand	MT020	MT021	MT022	MT023
Palv (virus) +	5814	Forward	9,606	4	9	26
Palv (virus) -	5814	Reverse	7,010	5	13	56
Prunus dulcis (host) 5s +	119	Forward	518,368	617,593	450,450	630,717
Prunus dulcis (host) 5s -	119	Reverse	436	191	79	67
Prunus serotina (host) 5.8s +	159	Forward	394,348	287,826	261,669	974,554
Prunus serotina (host) 5.8s -	159	Reverse	132	64	157	141
Prunus dulcis (host) 18s +	1807	Forward	3,514,043	2,726,001	2,514,825	10,269,189
Prunus dulcis (host) 18s -	1807	Reverse	5,440	3,062	2,621	2,606
Prunus serotina (host) 26s +	2827	Forward	4,022,042	3,142,045	3,029,424	10,716,657
Prunus serotina (host) 26s -	2827	Reverse	7,360	3,855	1,528	1,657
Prunus dulcis (host) 28s +	3172	Forward	2,671,432	1,947,632	1,773,477	7,200,428
Prunus dulcis (host) 28s -	3172	Reverse	7,374	4,066	1,857	2,199
Chloroplasts (contribute to 4.5s, 5s, 16s and 23s rRNA) +	157087	Forward	6,519,321	12,001,735	3,890,395	13,346,351
Chloroplasts (contribute to 4.5s, 5s, 16s and 23s rRNA) -	157087	Reverse	10,569,787	6,616,948	3,744,331	5,323,496
Total			**28,246,699**	**27,351,027**	**15,670,835**	**48,468,144**

Figure 10.4: Distribution of mapped small RNA reads onto reference sequences including a virus (Palv), rRNAs (5S, 5.8S, 26S and 28S) and chloroplast genome.

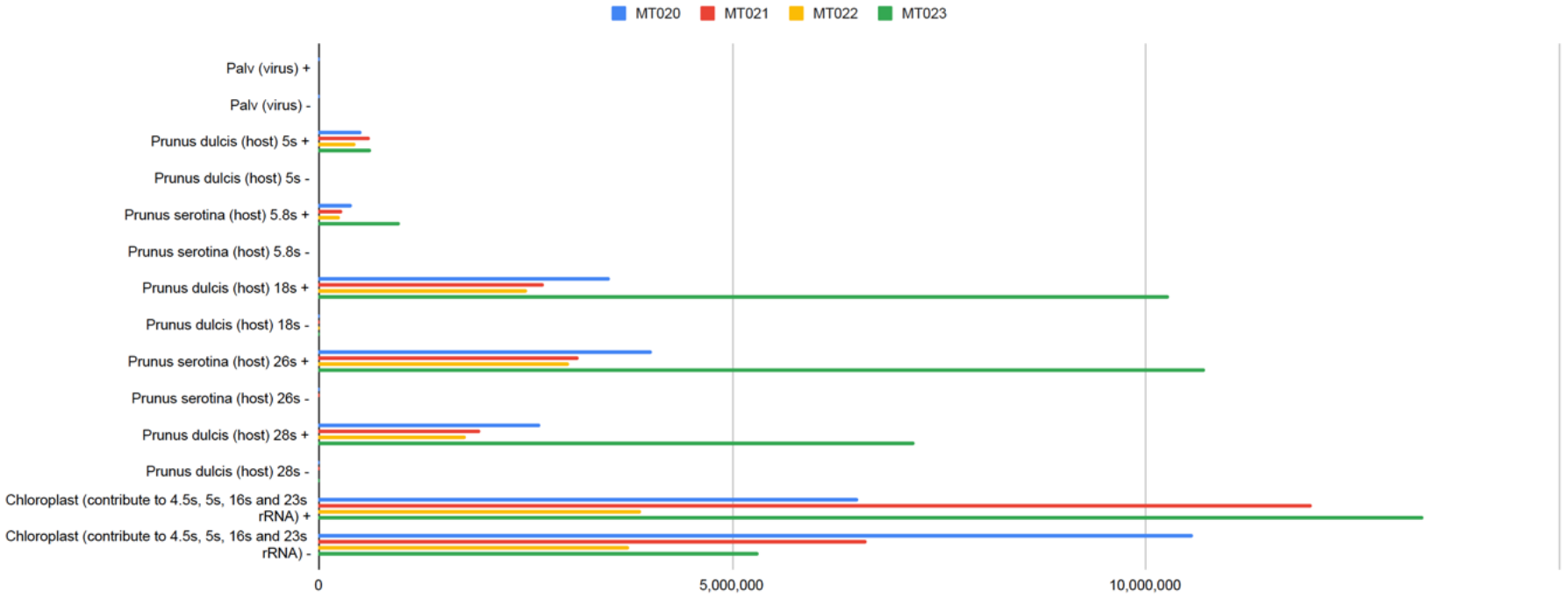


Figure 10.5: Mapping of reads onto Prunus 5.8S reference sequence. Reads are mapped along the 159 bp reference 5.8S rRNA sequence.

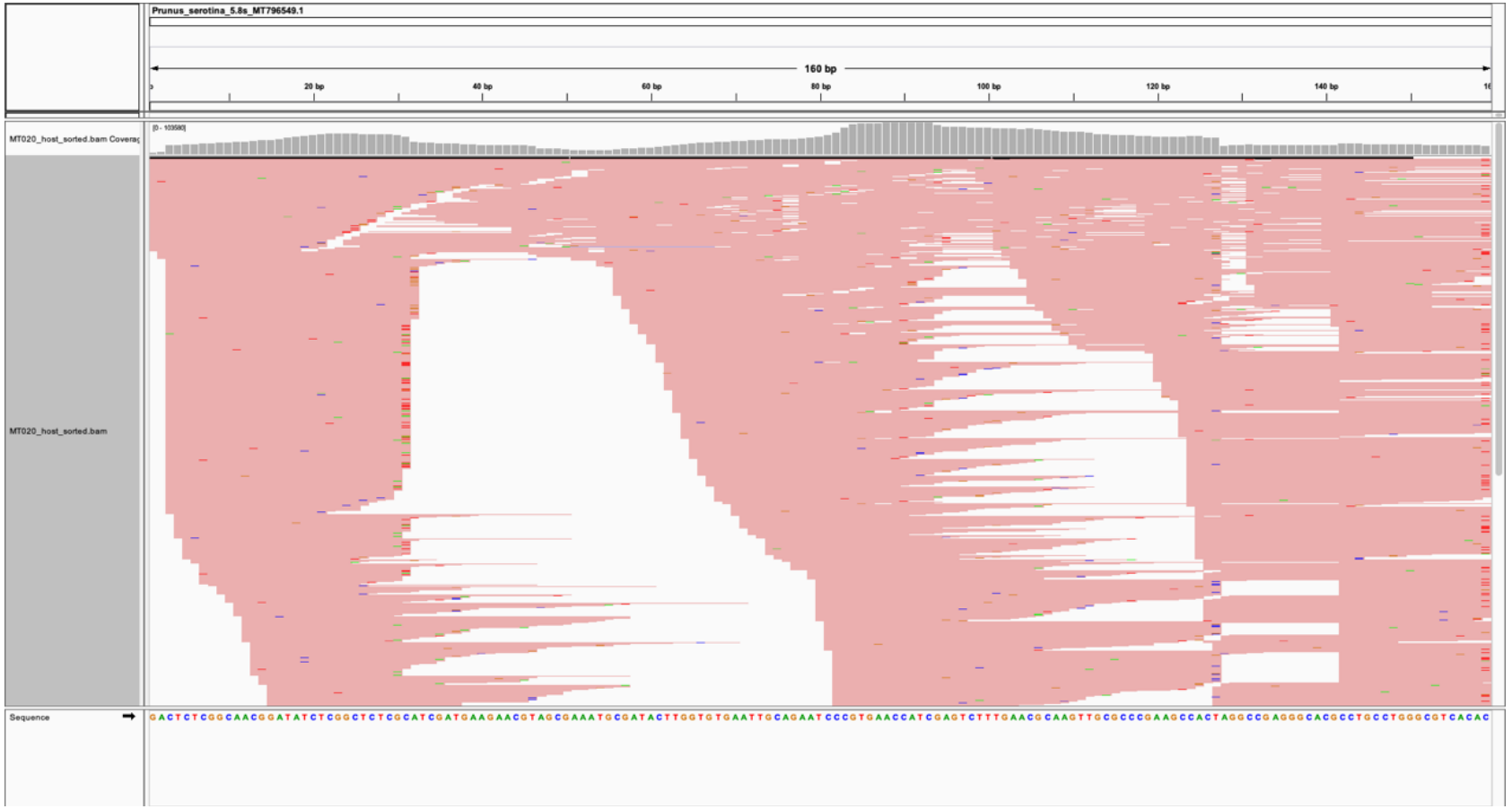


Figure 10.6: Mapping of reads onto Prunus 5.8S reference sequence. Reads are mapped along the 2,827 bp reference 26S rRNA sequence. Red = forward mapped reads; blue= reverse strand mapped reads.

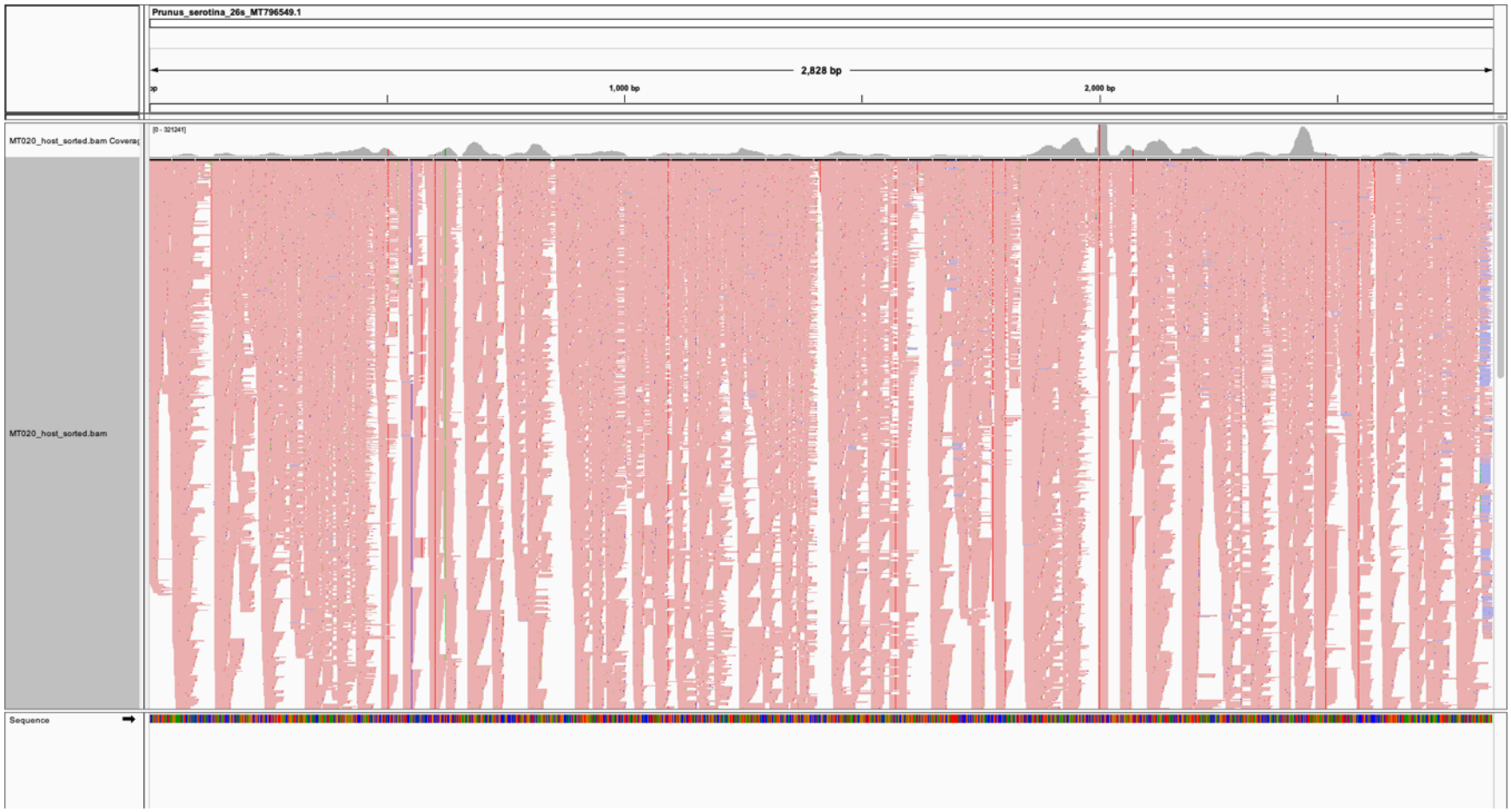


Figure 10.7: Mapping of reads onto a reference virus Palv (accession KY635989). Red = forward mapped reads; blue= reverse strand mapped reads.



Part 2: Extension of Quality Control assessment – implementing a pipeline for RNA source distribution and application to a trouble shooting experiment.

RNA (including small RNAs; sRNAs) within plant cells can come from a broad variety of sources including the chloroplast and mitochondria genomes, ribosomal RNAs (rRNAs), plant microRNA and contaminants if present (i.e., PhiX and vector sequences)

Aims:
- Estimate amount of HTS data derived from various RNA sources
- Determine if enough HTS data is available for virus diagnosis
- Identify potential issues if present

Method and Results

Adaptor clipped reads are mapped using bowtie (http://bowtie-bio.sourceforge.net/index.shtml) onto various RNA sources allowing one mismatch (equivalent to 4-5% sequence divergence). Reads are initially mapped onto ribosomal RNAs, unmapped reads are then mapped onto mitochondrial genes, and so on (**Figure 10.8** Left Panel). We used previously collected NGS data for controls plants from BGI and AGRF to establish and assess the outputs of this pipeline (**Figure 10.8** Right Panel).

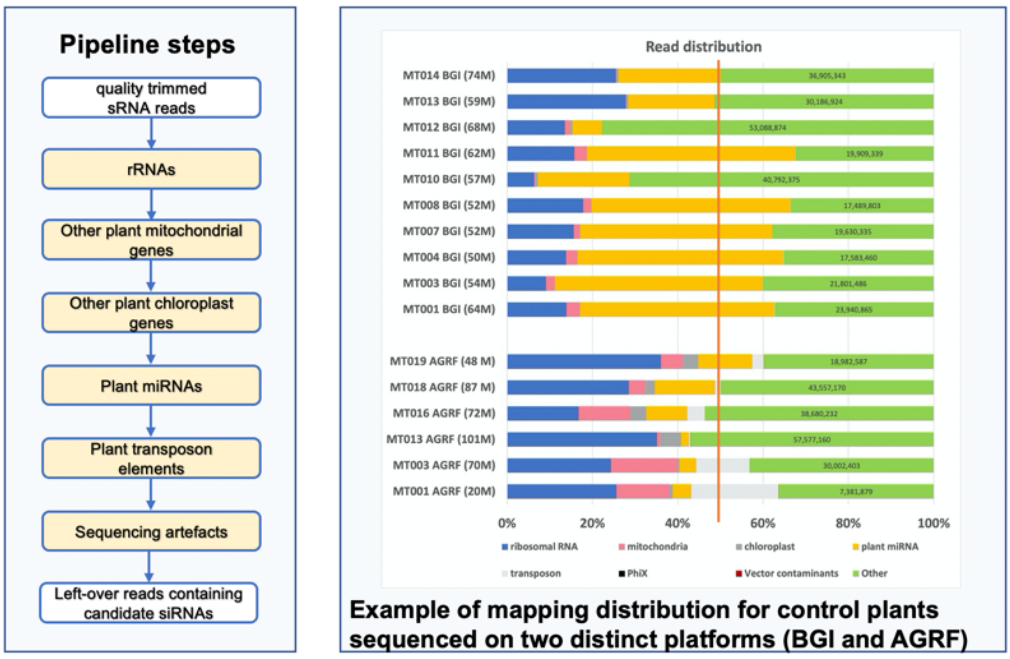


Figure 10.8: Implementation of an RNA source assignment pipeline. Left panel: individual mapping steps of quality trimmed reads onto various reference sequences – left over sequences or unmapped reads onto the various reference sequences are used for virus diagnosis. Right Panel: example of RNA source assignment of quality trimmed reads.

Trouble shooting experiment to identify technical challenges impacting NGS data
In collaboration with AGRF and NEB (New England Biolabs), 2 control (MT004 and MT018) and 2 Batch1 (MT021 and MT100) samples using the original NEB library prep kit and a new NEB library prep kit. All selected samples show high quality total RNA profiles (**Figure 10.9**).

Aims:
- Re-assess RNA quality
- Establish whether the shift in RNA profile caused by a poor sequencing reagent lot.

Evaluation of RNA integrity

Total RNA samples were assessed for their RNA integrity using Agilent's bioanalyzer. Control and batch 1 samples showed RIN scores above 7.0 (**Figure 10.9**). All samples passed QC check prior preparation of NGS library using NEBNext Small RNA Library Prep Set for Illumina.

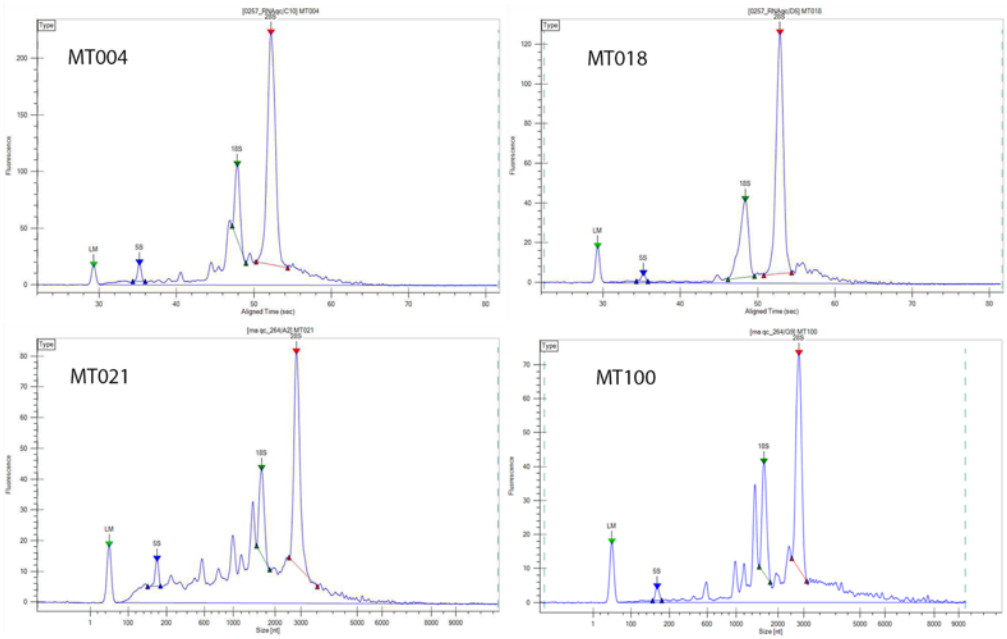


Figure 10.9: RNA integrity evaluation of total RNA samples for control (MT004, MT018) and batch 1 (MT021, MT100) samples. All samples passed QC assessment.

Reproducible RNA content profiles across samples

We observed strong reproducible results for control and batch 1 samples as well as a human brain control. Using an 'original' lot kit used to prepare batch 1 samples or a 'new' lot of reagents, both sets of library prep reagents resulted in reproducible RNA content profiles for all samples tested (**Figure 10.10**). These results strongly suggest that a defective reagent lot was not the cause of the increase amount of ribosomal RNAs. Furthermore, it suggests that ribosomal RNA degradation may have occurred during the library prep of batch 1 samples. We continue to work with the service provider to resolve the technical issue experienced with the first large scale batch of samples.

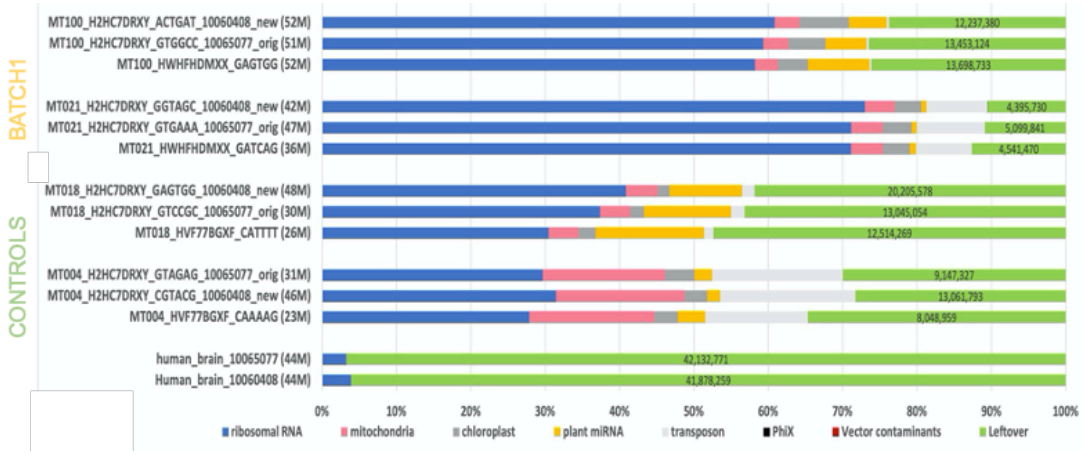


Figure 10.10: RNA content profile for trouble shooting samples.

Batch 1 samples show a ribosomal RNA degradation profile

Mapping of reads onto reference ribosomal RNA (rRNAs) sequences showed that batch 1 samples displayed an increased amount of rRNAs-derived reads of 18 to 40 nt in length. Such profile was observed among batch 1 samples contrasting with control plants that showed discrete peaks of rRNAs (**Figure 10.11**). Furthermore, we noticed an increased background noise of reads from 18 to 40 nt in lengths that were not mapped onto reference sequences (rRNAs, chloroplast, mitochondria, etc) and are available for the diagnosis of viruses and viroids. Typical plant small RNA profiles show peaks of small RNAs at 21nt, 22nt and 24nt lengths. Although, batch 1 samples show such expected peaks

it also shows that there was a large accumulation of sequences between 32nt -37nt in length, this secondary group of sequences does not map onto tested reference sequences and further investigation is necessary to determine its origin.

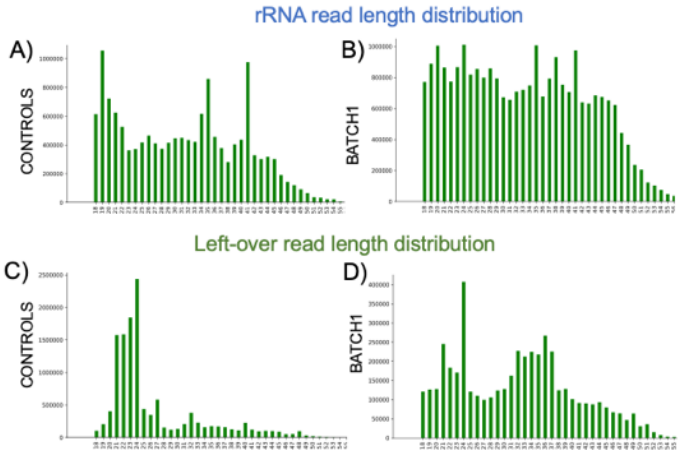


Figure 10.11: Size distribution of reads mapped onto rRNAs (A and B) and 'left-over' unmapped reads available for virus/viroid diagnosis (C and D) for control (A and C) and batch 1 (B and D) samples.

Part 3: Side-by-side comparison of PEQ and NGS methods for batch 1 samples

PEQ protocols identified 8 out 90 (8.8%) plants infected with quarantinable viruses, while preliminary small RNA-Seq NGS results identified the same plants except for sample MT098 (**Table 10.2**), where it missed to detect a Cherry Virus A (CVA) in a Prunus sample. For another two rosa plants it detected the presence of Prunus necrotic ringspot virus (PNRSV) but missed to detect a Prune Dwarf Virus (PDV) in these samples (MT100 and MT102) (**Table 10.4**).

Additional non-quarantinable viruses were also found in 12 batch 1 plants, representing six distinct viruses and viroids (**Table 10.5**). Analysis of Peach Latent Mosaic Virus (PLMVd) found in samples MT085, MT091 and MT102 showed that this viroid is highly expressed in sample MT091 and that its presence in the other two samples may be due to an 'index-hoping' (sample cross contamination) event. Strikingly, we found that the same single nucleotide polymorphisms (SNPs) are shared by the PLMVd sequences found in all three plants (**Figure 10.12**). Further, PCR validation is required to determine if the presence of PLMVd in samples MT085 and MT102 are genuine or a case for index hopping.

Overall, despite the limited amount of NGS data available for virus/viroid diagnosis for batch 1 samples, NGS was able to detect the majority of identified quarantinable viruses by PEQ protocols and flagged a handful of plants for the presence of additional viruses and viroids.

Table 10.4: Diagnosis of viruses and viroids in batch 1 samples using PEQ and small RNA-seq NGS methods.

		PEQ detected	sRNA HTS results				
Sample	**Commodity**	**Virus acronym**		**Accession**	**Virus genome length (nt)**	**Coverage (%)**	**Identity (%)**
MT039	Prunus	PDV	PDV RNA1	KY883327	3339	10.3	97.97
			PDV RNA2	AF277662	2593	10.4	96.68
			PDV RNA3	KY883331	2051	34	98.3
MT040	Prunus	PNRSV	PNRSV RNA1	MH727233	3332	3.7	100
MT093	Prunus	CVA	CVA	KY510884	7438	2.2	99.39
MT094	Prunus	CVF	CVF	MG925373	3857	6.6	98.82
MT098	Prunus	CVA	CVA				
MT100	Rosa	PNRSV	PNRSV RNA1	AF278534	3332	17.2	98.26
			PNRSV RNA2	MH727232	2591	1.9	97.92
			PNRSV RNA3	JN416776	1951	29.9	98.63
		PDV	PDV				
MT101	Rosa	PNRSV	PNRSV RNA1	AF278534	3332	38.5	98.52
			PNRSV RNA2	MH727232	2591	22.6	97.27
			PNRSV RNA3	JN416776	1951	52.9	98.35
MT102	Rosa	PNRSV	PNRSV RNA3	FJ610344	1957	20.8	98.55
		PDV	PDV				

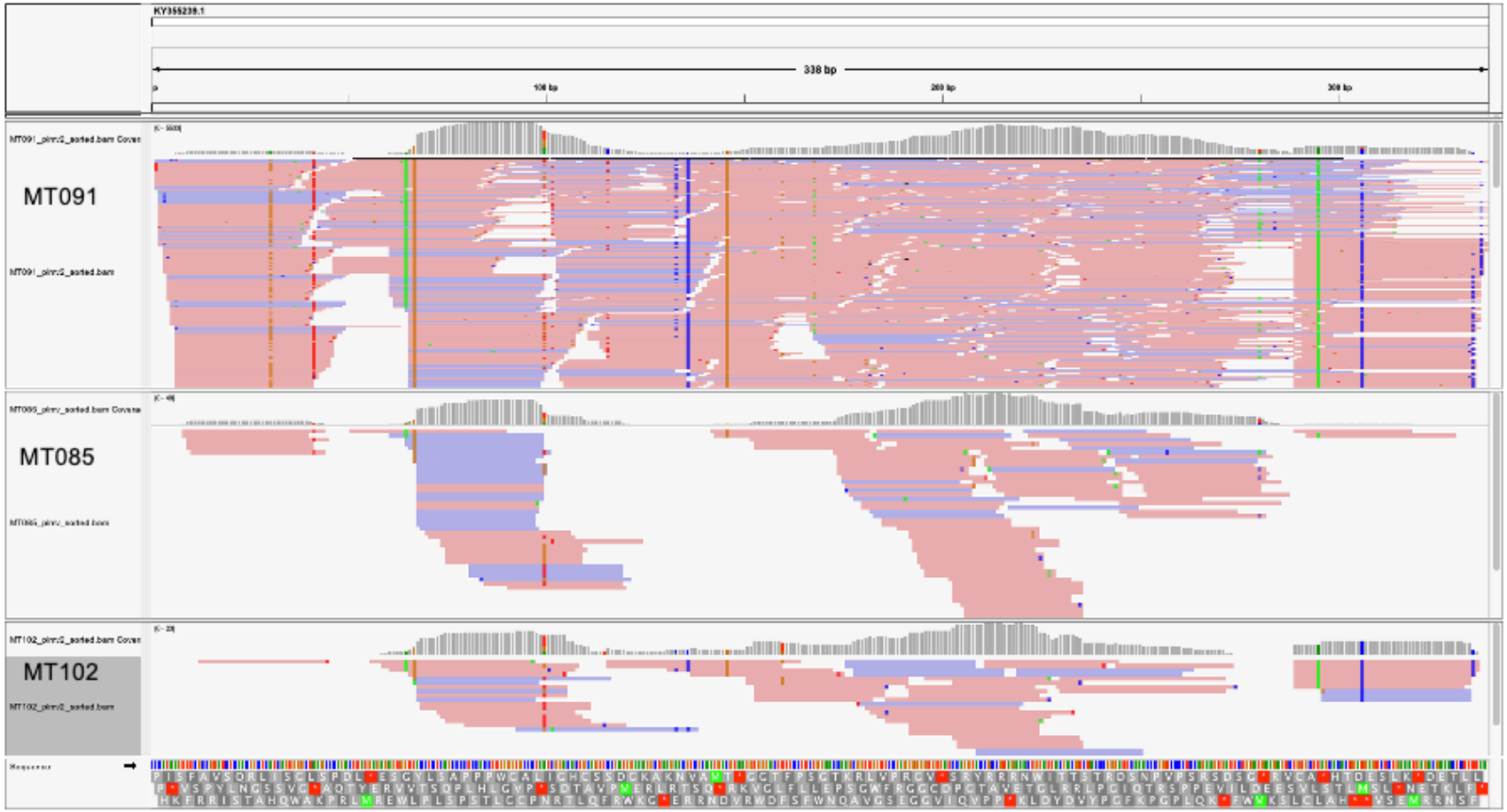
CVA Cherry virus A
CVF Cherry Virus F
PDV Prune dwarf virus
PNRSV Prunus necrotic ringspot virus

Table 10.5: Diagnosis of additional non- quarantinable viruses and viroids in batch 1 samples by small RNA-seq NGS method. Orange background denotes candidate index hopping events.

Sample	Commodity	Virus	Accession	Virus genome length (nt)	Coverage (%)	Identity (%)
MT020	Prunus	PaLV	MH229943	1169	24	97.51
MT029	Prunus	PrGVA	MF504040	3174	4.4	100
MT031	Prunus	PrGVA	MF504038	3174	6.9	100
MT032	Prunus	PrGVA	MF504040	3174	4.4	99.29
MT039	Prunus	VWDaV	KX196166	3320	4.5	83.33
MT055	Prunus	NSPaV	MK361454	4578	3.8	98.84
MT063	Prunus	NSPaV	KT273410	4991	7	98.57
MT085	Prunus	PLMVd	DQ680759	338	55	100
MT091	Prunus	PLMVd	JN680785	302	57.9	100
MT092	Prunus	PaLV	KY635989	5814	1.7	98.97
MT092	Prunus	ChALV	KY686298	5815	1.7	95.83
MT102	Rosa	PLMVd	KY355239	337	19	100

ChALV = Cherry associated luteovirus
NSPaV = Nectarine stem pitting-associated virus
PaLV = Peach associated luteovirus
PLMVd = Peach latent mosaic viroid
PrGVA = Prunus geminivirus A
VWDaV = Viola white distortion associated virus

Figure 10.12: Mapping of NGS reads (21-22nt) onto reference PLMVd genome. IGV genome browser shows SNPs across MT091, MT085 and MT102 samples. Common SNPs (green, red, blue and dark orange bars) suggests a common genetic origin.



Appendix 11: Cultured bacteria high throughput sequencing pipeline validation

Introduction

Culturing pathogenic bacteria for whole genome sequencing is a means to isolate pathogen DNA from host and environmental DNA. In the previous milestone, we described a high throughput sequencing (HTS) pipeline to identify plant pathogenic bacteria from acquisition of bacterial cultures to bioinformatics data analysis as outlined in Figure 11.1. This includes protocols for DNA extraction, library preparation, sequencing and data analysis. Furthermore, we showed that shallow short read sequencing (~300 MB) using the Illumina platform or long read sequencing using the Oxford Nanopore Technology (ONT) MinION sequencing of genomic DNA extracted from cultured bacteria is often sufficient for identification of many pathogenic species down to the subspecies/pathovar level.

In this report, we further validate and optimize this pipeline by applying it to bacterial isolates from post entry quarantine (PEQ) samples supplied by the Department of Agriculture, Water and Environment (DAWE) Science and Surveillance Group (SSG). To do so, PEQ sourced genomic DNA of 37 pure cultured bacteria isolated from various infected plants were run through the HTS pipeline. Progress is outlined in this report.

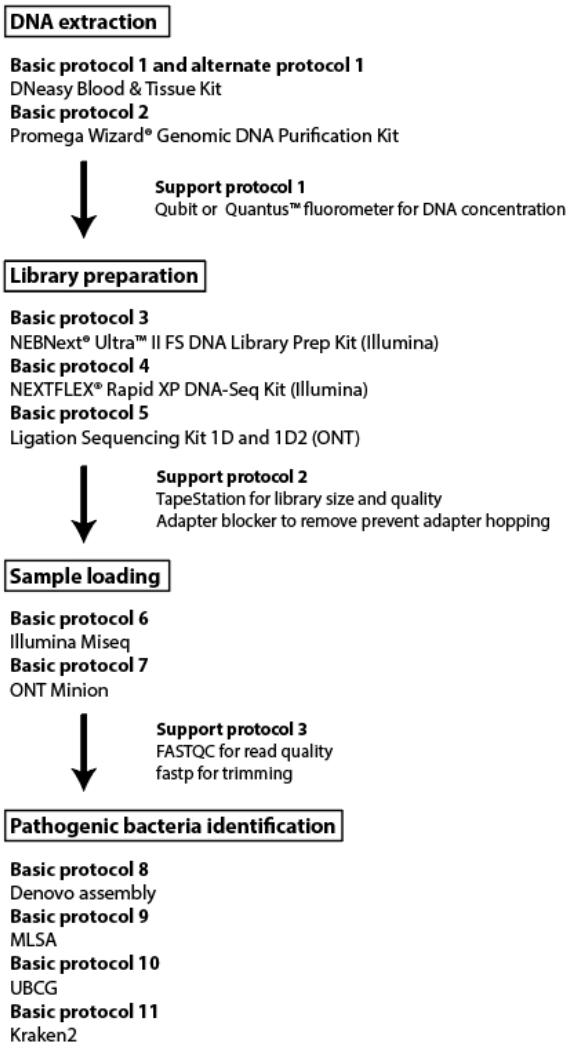


Figure 11.1. The high throughput sequencing and bioinformatics pipeline used to identify culturable plant pathogenic bacteria

Material and Methods

PEQ samples

Aliquots of 10 µL genomic DNA of 37 cultured, pre-identified bacterial isolates (Table 11.1) were supplied by PEQ and the SSG community for pipeline validation.

Table 11.1 Post entry quarantine (PEQ) supplied bacterial gDNA samples for pipeline validation including host and PEQ identification (ID)

Sample ID	Host	PEQ Molecular ID
PEQ -1	*Sansevieria*	*Pantoea stewartii*
PEQ -2	*Solanum*	*Sphingomonas* sp.
PEQ -3	*Actinidia*	*Pseudomonas citronellolis* or *Hyphomicrobium facile*
PEQ -4	*Prunus*	*Paenibacillus* sp.
PEQ -5	*Actinidia*	*Staphylococcus epidermidis*
PEQ -6	*Prunus*	*Bacillus pumilus*, *invictae* or *safensis*
PEQ -7	*Solanum nigrum*	*Pseudomonas viridiflava*
PEQ -8	*Apium graveolens*	*Pectobacterium carotovorum*
PEQ -9	*Fragaria*	*Dyella* sp.
PEQ -11	*Juglans*	*Pantoea agglomerans* / *vagans*
PEQ -12	*Solanum*	*Paenibacillus* sp.
PEQ -13	*Solanum*	*Microbacterium paludicola*
PEQ -14	*Solanum*	*Chryseobacterium taeanense*
PEQ -15	*Solanum*	*Rhizobium radiobacter*
PEQ -17	*Solanum*	*Alcaligenes faecalis*
PEQ -18	*Epipremnum*	*Stenotrophomonas maltophila*
PEQ -20	*Solanum melongena*	*Bacillus pumilus* or *B. safenis*
PEQ -21	*Solanum melongena*	*Bacillus* sp, *B. Licheniformis*, *B. subtilis* or *B. tequilensis*
PEQ -22	*Solanum melongena*	*Bacillus pumilus* or *B. safenis*, *B. australimari* or *B. invictae*
PEQ -23	*Brassica oleracea*	*Bacillus subtilis*
PEQ -24	*Ficus*	*Acinetobacter* sp.
PEQ -25	*Lilium*	*Rahnella aquatilis*
PEQ -26	*Phalaenopsis*	*Dickeya* sp.
PEQ -27	*Cynodon*	*Stenotrophomonas maltophilia*
PEQ -28	*Phalaenopsis*	*Dickeya fangzhongdai*
PEQ -29	*Fragaria*	*Luteibacter* sp. or *Dyella* sp.
PEQ -30	*Fragaria*	*Luteibacter yeojuensis*
PEQ -31	Sugar snap peas	*Pseudomonas syringae*
PEQ -32	Sugar snap peas	*Pseudomonas syringae*
PEQ -33	*Phalaenopsis*	*Dickeya fangzhongdai*
PEQ -34	*Phalaenopsis*	*Dickeya fangzhongdai*
PEQ -35	*Phalaenopsis*	*Dickeya fangzhongdai*
PEQ -36	King oyster mushrooms (multiple)	*Ewingella americana*
PEQ -37	King oyster mushrooms (multiple)	*Ewingella americana*
PEQ -38	King oyster mushrooms (multiple)	*Ewingella americana*
PEQ -39	Chrysenthemum	*Pseudomonas* sp.

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, concentration of DNA was measured and ~8 ng DNA was used for each library preparation. For samples with low concentrations such as PEQ-12, PEQ-15 and PEQ-30 (0.226-0.456 ng/µL) and no detection of DNA for PEQ-16, 8 µL of sample was used. Libraries were generated using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A no template control library was also generated. Libraries were then paired end sequenced (2 × 250 bp) using the Illumina NovaSeq.

Bioinformatic analysis

The bioinformatic section of the cultured bacteria HTS identification pipeline is being further validated through analysis of the data generated from the 37 PEQ samples. Important considerations including index hopping, bioinformatic tools for differentiation of closely related bacteria and access to appropriate genomic databases or a creation of an in-house genomic database will also be assessed.

Results and Discussion

Sequencing output

Information of trimmed reads and contigs for each sample are listed in Table 11.2. A small number of reads were observed in the negative control (Blank) which could be contamination from surrounding samples, index hopping or low-quality reads that will be removed during the trimming stages. Further analysis is required to confirm the source of reads. Sequencing depth for each sample ranged between 275 MB to 1.2 GB except for PEQ-12 and PEQ-16 which had 176 MB and 21 MB respectively. The low sequence output for PEQ-12 and PEQ-16 is not unexpected due to low concentrations of DNA available for input. Sequence output of two other samples with low DNA input, PEQ-15 and PEQ-30, was comparable to output of other samples due increased DNA input.

Table 11.2 Sequence output for PEQ samples.

Sample ID	Sequence (MB)	Total reads	Sample ID	Sequence (MB)	Total reads
Blank	0.241	270	PEQ-20	398	3581744
PEQ-1	1200	10756220	PEQ-21	428	3785320
PEQ-2	762	7006942	PEQ-22	540	4799350
PEQ-3	607	5421992	PEQ-23	275	2410648
PEQ-4	577	5391712	PEQ-24	504	4514220
PEQ-5	389	3394528	PEQ-25	363	3142136
PEQ-6	510	4465548	PEQ-26	526	4725898
PEQ-7	586	5446496	PEQ-27	289	2830906
PEQ-8	551	4996890	PEQ-28	406	3820630
PEQ-9	587	5364106	PEQ-29	526	4787732
PEQ-10	543	4761860	PEQ-30	508	4600396
PEQ-11	618	5706106	PEQ-31	535	4852388
PEQ-12	176	1490704	PEQ-32	791	6871406
PEQ-13	710	6471434	PEQ-33	340	3063316
PEQ-14	602	5284776	PEQ-34	354	3170166
PEQ-15	422	3763484	PEQ-35	290	2719566
PEQ-16	21	189342	PEQ-36	399	3533330
PEQ-17	626	5804108	PEQ-37	559	5292286
PEQ-18	576	5050458	PEQ-38	856	7565444
PEQ-19	743	6904354	PEQ-39	861	7696370

Appendix 12: High throughput sequencing protocols for key fastidious and unculturable plant pathogenic bacteria

Introduction

Fastidious and unculturable bacteria pose a problem for post entry quarantine (PEQ) testing as confirmation through isolation is not an option, pathogens can exist in low titre and numerous specific molecular tests are relied upon for detection of a range of exotic pathogens in the one host/sample. Streamlining of this testing into one simplified test to reduce resources and turnaround time is highly desirable and multiple high throughput sequencing (HTS) approaches hold promise to aid in achieving this goal.

In the following sections we describe a HTS pipeline to identify fastidious and unculturable plant pathogenic bacteria from sampling to bioinformatic data analysis as outlined in Figure 12.1. This includes protocols for DNA extraction, library preparation (Illumina), target enrichment, sequencing and bioinformatic analysis.

Using the wet lab section of this pipeline, we have prepared and sequenced the following fastidious bacterial pathogens as case studies;

- three phytoplasmas *in planta*, *Candidatus* Liberibacter africanus, *Ca*. L. solanacaereum and *Ca*. L. brunswickensis *in planta* and in their respective vectors;
- the following pathogens spiked into host DNA at varying concentrations;
    - *Xanthomonas citri* subsp. *citri* in lemon leaf DNA,
    - *Xylella fastidiosa* in olive leaf DNA; and
    - *Erwinia amylovora* in apple leaf DNA

Sequencing of these case studies are presented in Appendices 13 and 14.

Protocols

Starting samples

Detection of phytopathogenic bacteria *in planta* can be aided by understanding the life cycle of bacterial pathogens. Firstly, the tissue sampled. Composition of bacterial communities can vary among different parts of the plant, and different phytopathogenic bacteria have varied distribution within their hosts. For instance, phytoplasma and *Liberibacter* species are restricted to the sieve tubes in the phloem and often follow a 'source to sink' pattern of distribution within their host plant across the growing season (Constable et al., 2003; Tran-Nguyen et al., 2007). For this reason, the appropriate tissue for sampling many hosts for multiple pathogens will vary from planting to maturity. *Clavibacter michiganensis* subsp. *michiganensis* are known to be localized in the leaflets of tomatoes in the early stages of infection (Gleason et al., 1993). *X. fastidiosa* and *Xanthomonas albilineans*, bacterial pathogens of multiple important crop species are located solely within the xylem of a plant (Pieretti et al., 2009). The distribution and concentration of *X. fastidiosa* within plants can be variable and is dependent upon plant species type, seasonal and environmental factors (Sicard et al., 2018). To maximize the likelihood of detection, it is recommended that sampling should be performed during the period of active growth of the plants (Hopkins, 1981), usually from late spring to autumn in temperate regions but potentially all year round for tropical hosts such as coffee (EPPO, 2018). Hence, a range of plant tissues would need to be sampled for testing of multiple pathogens in any given host. Furthermore, some parts of the plant, such as Betula leaves, grape leaves, mature strawberry leaves and rhizomes of turmeric and ginger, can contain high concentrations of polysaccharide and polyphenols which increase the difficulty of extraction of high quality nucleic acid (Porebski et al., 1997; Rezadoost et al., 2016; Syamkumar et al., 2003). Hence, if possible, it should be avoided, and an alternate sample site chosen instead.

Secondly, the timing of collection. Different environmental conditions influence bacterial communities, this includes the growth stage of the host (Constable et al., 2003). Understanding the life cycle of target pathogens can help ascertain the optimal sampling time (or conditions) when pathogenic bacteria are at their highest density *in planta* (e.g during active growth of the plant). Manipulating growth conditions in the glasshouse could increase the chances of detecting specific pathogens.

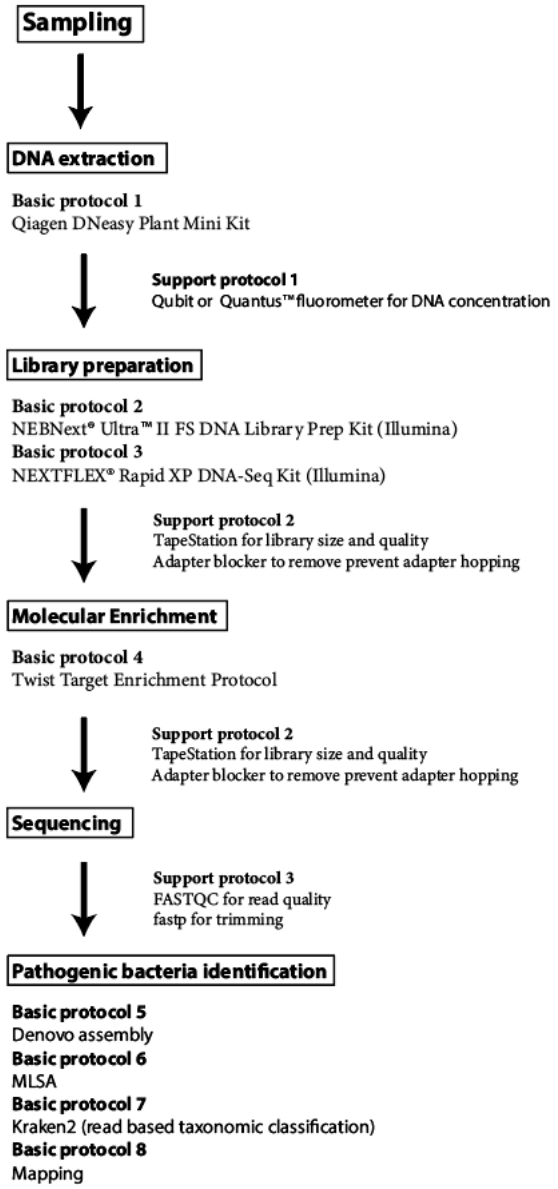


Figure 12.1 The high throughput sequencing and bioinformatics pipeline used to identify fastidious and unculturable plant pathogenic bacteria

DNA extractions

Many commercial DNA extraction methods (e.g. Qiagen DNeasy Plant Mini Kit and Promega Wizard® Genomic DNA Purification Kit etc) have been used to extract DNA for library preparation in the literature (Zheng et al., 2016). CTAB methods or CTAB in combination with commercial kits have also been used to accommodate contaminants such as polyphenols and polysaccharides from a variety of plant species (Green et al., 1999; Li et al., 2008; Porebski et al., 1997; Rezadoost et al., 2016; Selvaraj et al., 2018).

The Qiagen DNeasy Plant Mini Kit was used to extract DNA for Illumina library preparation in this work as it is the validated bacterial extraction kit in use at Agriculture Victoria Research (AVR) (Green et al., 1999). Other DNA extraction kits (including specific HMW extraction kits) can also be used but need to be effectively validated for DNA extraction quality suitable for each sequencing technology.

Quality checking of input DNA is important for library preparation as input quality can greatly impact the resulting library. For example, quantity of DNA is important as the amount of DNA can modify library preparation protocols. For this, the Quantus™ or Qubit fluorometer is suggested to measure the concentration of the extracted gDNA. Gel-electrophoresis or a

TapeStation can be used to assess the DNA size distribution (e.g., shearing) which is important for long read sequencing such as Oxford Nanopore Technology (ONT) which works optimally with HMW DNA. Where possible, TapeStation should be the preferred method to assess genomic DNA size distribution.

Controls to be used during this step is a negative control (blank) to assess cross contamination.

Library preparation

There are multiple commercial kits that can be used to prepare libraries for Illumina sequencing. Essentially, the overall process begins with fragmentation of extracted DNA to appropriate sizes; between 150 – 300 bp for HiSeq and NovaSeq platforms and 300 – 600 bp for the MiSeq platform if using the MiSeq Reagent Kit v3. Most library preparation kits provide a sheering option which can include heat and enzymatic fragmentation. Alternate options include the Covaris g-TUBE; a single-use device that shears genomic DNA into selected fragments sizes ranging from 6 to 20 kb using a centrifuge. For some kits, an additional end repair step is added to accommodate fragmentation to ensure that DNA fragments have blunt ends for sequence adaptor ligation. Fragment size of libraries is an important consideration as insert sizes influence outputs of downstream bioinformatic analysis. For example, larger insert sizes can result in substantial assembly improvements (Huptas et al., 2016). Following sequence adapter ligation individual indexes are added using PCR in a library amplification step for multiplexing multiple samples in the one sequencing run. The workflow of most library preparations is similar; however, some kits have minor modifications or patented technology in their workflow. For example, enzymes in the Nextera® DNA Library Preparation enable fragmentation and ligation of adapters in the same step (called tagmentation) and the NEXTFLEX Rapid XP DNA-Seq Kit which provides the adapter and indexes already together and only requires one ligation step.

Choice of library preparation kit for Illumina sequencing can be influenced by multiple factors. Firstly, the quantity and quality of the DNA can influence kit choice. Library preparation kits such as the NEBNext® Ultra™ II FS DNA Library Preparation Kit can accommodate down to 0.1 ng of DNA input and can be lenient with the quality of DNA. The cost effectiveness of library preparation kits can also influence their use in routine diagnostics. The PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit is known to be one of the cheaper commercial kits available and has been used at half-reaction as shown in this report. Lastly, the sequencing platform being used also influences library kit choice. Due to the smaller throughput of the MiSeq platform (~15 GB), most library preparation kits, including those with limited numbers of indexes (e.g. NEBNext® Ultra™ II FS Library Preparation Kit, Nextera® DNA Library Preparation Kit, NEXTFLEX Rapid XP DNA-Seq Kit, etc.) can be used with this platform. However, as the throughput of the Illumina platforms increases, (e.g. up to 850GB per lane on the NovaSeq), the library preparation kits need to be sourced from companies that have or can adapt to more than 192 compatible indexes for multiplexing of samples. Commercially available indexes such as the PerkinElmer NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) can accommodate the sample throughout of the Novaseq (up to 300 samples/lane). PerkinElmer indexes rely on the use of an A-tailing and ligation step that is similar to Truseq and NEBNext® kits making them compatible with a variety of library preparation kits. As such, NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) have been recommended for sequencing with the NovaSeq platform.

Indexes are important for differentiating sequence reads of individual samples when multiplexing samples on any of the sequencing platforms. If indexes are not compatible, the demultiplexing procedure will fail and sequencing reads will not be assigned to the corresponding biological samples (Varet et al., 2018). When many samples are multiplexed on larger sequencers, such as HiSeq and NovaSeq platforms, unique dual indexing (UDI) is highly recommended to reduce the phenomenon known as index hopping which is the incorrect assignment of indexes to a sample. When different indexing kits are being used, websites such as https://checkmyindex.pasteur.fr/ can be used to check index compatibility (Varet et al., 2018). Even with the same indexing kit, some companies provide pooling guides to ensure sequence read associated to specific samples can be differentiated. For instance, Figure 2 shows the index adapters pooling guide of IDT for Illumina–TruSeq UDI for a four-plex samples which changes for different numbers of samples.

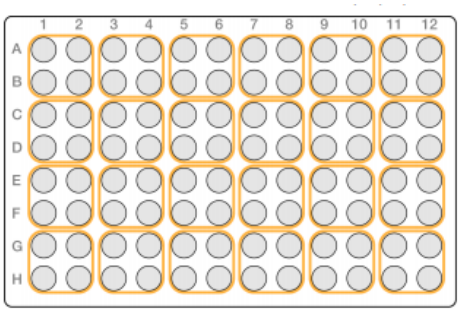


Figure 12.2 Pooling guide for four-plex dual indexing of IDT for Illumina–TruSeq UDI.

There is a need to adapt multiple library preparation kits to one indexing kit due to the different uses of each library preparation kit (e.g. RNA vs DNA library preparation). Most library preparation kits and indexes come with their own adapter sequences and chemistries rendering them incompatible. The PerkinElmer NEXTFLEX® Unique Dual Index Barcodes (384 and 1,536 UDI barcodes) are compatible with library preparation kits that use A-tailing ligation, of which there are many. However, this means that it is not compatible with the commonly used Nextera® DNA Library Preparation Kits, hence, it is not advised to use the Nextera® library preparation kits for sequencing on the NovaSeq.

A comparison analysis was done in the previous milestone on the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes); NEBNext® Ultra™ II FS DNA Library Preparation Kit with NEBNext® Multiplex Oligos for Illumina® (96 UDI Primer Pairs); and NEBNext® Ultra™ II FS DNA Library Preparation Kit with PerkinElmer NEXTFLEX® UDI Barcodes (384 barcodes). The kits were tested using a sample of Fig (*Ficus carica*) DNA spiked with 1% and 0.1% total amount of DNA of *Ralstonia solanacearum* and *Pantoea ananatis* respectively. Results showed that the NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) gave better assemblies than the others in the mock metagenomic analysis, hence are in use at AVR.

Controls

Controls to be used during library preparation include the negative control used in the previous step, no template control to detect cross contamination and a positive control of a known quality and quantity DNA. These controls do not need to be sequenced in depth (200-300 MB worth of data is sufficient), however, their use will ensure that contamination can be detected (negative controls) and the library preparation process was successful.

The 2200 TapeStation system (Agilent technologies) should be used to check the quality of libraries. Sheering of libraries, adapter/primer dimers and size of libraries all can be detected on the TapeStation which can affect sequencing quality. For Illumina sequencing, adapter blockers should be used to prevent index hopping.

Enrichment

Pathogenic bacteria cell enrichment increases the ratio of bacterial cells in comparison to plant cells to improve sensitivity for pathogen detection. Issues in obtaining sufficient quantities of uncultivable plant pathogenic bacteria DNA for detection and characterisation (e.g. phytoplasmas) is common. Using centrifugation protocols, such as ultracentrifugation and caesium chloride buoyant gradient centrifugation, organisms like phytoplasmas can be physically enriched for downstream processes (Frampton et al., 2018; Tran-Nguyen et al., 2007). It has also been reported that coupling 0.9% NaCl and 0.063% sodium dodecyl sulfate (SDS) during DNA extraction can be used for microbial enrichment in plants (Ou et al., 2019; H. X. Wang et al., 2008). These type of mechanical steps aids in sequencing full genomes for characterisation, however, these processes can be tedious and may not be practical in routine or high-throughput diagnostics.

Molecular target enrichment captures or amplifies, selected genomic regions for further analysis. These forms of enrichment aid in detection of low titre pathogens and are often time and cost-effective, easing the bioinformatic component of the pipeline.

A common target enrichment technique is amplicon sequencing. This includes 16S sequencing and metabarcoding that can identify individual species from environmental samples. Essentially, target regions are isolated and amplified using PCR

primers on a mixed DNA sample from any origin which will subsequently be sequenced by HTS to identify the species composition of the sample.

For bacterial species, genes such as gyrase B (*gyrB*), 60 kDa chaperonin (*cpn60*), β-subunit of DNA polymerase (*rpoB*), *recN* and *dnaX* have been used to enhance resolution of bacterial taxa (Barret et al., 2015; Links et al., 2014; Tian et al., 2016; Vos et al., 2012; Zeigler, 2003). However, identifying conserved primer binding regions in these genes across diverse taxa can be difficult and can create biases for a subset of bacteria (Bulman et al., 2018). Subsequently, this would create an extra burden when identifying multiple pathogenic bacteria within a host sample as multiple primers would need to be used which can cause failed and/or nonspecific amplification due to interactions between primers (Cho et al., 1999; D. G. Wang et al., 1998). With new strains of bacterial pathogens being discovered, new primer sets would need to be continually designed, optimized and validated which can be time and labour intensive. If Illumina sequencing were to be used, the size limit of 200 bp for optimal amplicon sequencing can also be prohibitive in discriminating between closely related species. Though not included in this report, amplicon sequencing for MinION is being explored.

An alternate to amplicon sequencing is the design and utilisation of hybridization probes to capture specific DNA or RNA targets within a sample. Hybridization probes are biotinylated DNA or RNA fragments less than 500 bp long that are designed to bind to your targeted regions (Karamitros et al., 2015; Samorodnitsky et al., 2015). Once added to a sample and hybridisation enabled, these probes can be captured separating the target DNA from the rest of the sample. For bacterial diagnostic purposes, these probes could either target specific bacterial pathogens for direct identification (Gasc et al., 2017) or target a barcoding region (Gasc et al., 2018; Metsky et al., 2019). Unlike amplicon sequencing, where size limit of amplicons is dependent on the sequencer used, there is no limit to target size when using hybridization probes (Cai et al., 2019; D. G. Wang et al., 1998).

There are a number of potential drawbacks with the use of hybridisation probes in diagnostic testing. One is that hybridisation probes are less specific than amplicon sequencing as off-target binding occurs at a higher percentage leading to the capture of non-target sequences (Liu et al., 2007; Samorodnitsky et al., 2015). However, the capture of these sequences should not lead to false positive results and could provide assurance that false negative results are not occurring (e.g. genetic variants are not being missed with the technology). The impact of non-target binding will be monitored and assessed in validation experiments in this project. Additionally, if a sample did not contain a target pathogen, no DNA would be captured for sequencing leading to no output. This could indicate that the samples were negative or that the test had failed. To avoid this issue a number of controls are being validated to ensure that a negative test can be differentiated from a failed test. These include housekeeping probes to show that probing of individual library preparations has been successful, positive controls to show the probes are working and negative controls to identify any potential contamination points throughout the pipeline.

The use of hybridization probes as part of this work is reported in further detail in Appendix IV.

Bioinformatic analysis: Bacterial detection from HTS data

Multiple bioinformatic tools will be trialled for identification of fastidious and unculturable bacterial pathogens using both enriched and unenriched Illumina sequencing data. This includes tools the use trimmed reads such as the k-mer-based analysis tool, Kraken2 (Wood et al., 2019) and the mapping tool, Bowtie2 (Langmead et al., 2012); and tools that compare *de novo* assembled contigs to sequence databases such as GenBank using BLAST (Altschul et al., 1990).

For some of these bioinformatic tools, Conda, a source package management system and environment management system will be used. Each bioinformatic tool is dependent on multiple software and programming languages, such as Java and Python, and occasionally these software can clash with one another rendering the bioinformatic tool inoperative. In this project, Conda has successfully been used to create individual environments for Kraken2 where they were installed with their dependencies (Milestone 104). Within these environments, bioinformatic tools can be easily run and be updated. Hence, as a prerequisite, Conda would need to be installed.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Barret, M., Briand, M., Bonneau, S., Preveaux, A., Valiere, S., Bouchez, O., . . . Jacquesa, M. A. (2015). Emergence shapes the structure of the seed microbiota. *Appl Environ Microbiol, 81*(4), 1257-1266. doi:10.1128/aem.03722-14

Bulman, S. R., McDougal, R. L., Hill, K., & Lear, G. (2018). Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. *Australasian Plant Pathology, 47*(5), 467-474. doi:10.1007/s13313-018-0579-3

Cai, W., Nunziata, S., Rascoe, J., & Stulberg, M. J. (2019). SureSelect targeted enrichment, a new cost effective method for the whole genome sequencing of Candidatus Liberibacter asiaticus. *Scientific Reports, 9*(1), 18962. doi:10.1038/s41598-019-55144-4

Cho, R. J., Mindrinos, M., Richards, D. R., Sapolsky, R. J., Anderson, M., Drenkard, E., . . . Oefner, P. J. (1999). Genome-wide mapping with biallelic markers in Arabidopsis thaliana. *Nature Genetics, 23*(2), 203-207. doi:10.1038/13833

Constable, F. E., Gibb, K. S., & Symons, R. H. (2003). Seasonal distribution of phytoplasmas in Australian grapevines. *Plant Pathology, 52*(3), 267-276. doi:https://doi.org/10.1046/j.1365-3059.2003.00849.x

EPPO, E. a. M. P. P. O. (2018). Xylella fastidiosa. *PM 7/24(3). EPPO Bulletin, 40*, 175–218.

Frampton, R. A., Thompson, S. M., Kalamorz, F., David, C., Addison, S. M., & Smith, G. R. (2018). Draft Genome Sequence of a "Candidatus Liberibacter europaeus" Strain Assembled from Broom Psyllids (Arytainilla spartiophila) from New Zealand. *Genome Announcements, 6*(20), e00430-00418. doi:10.1128/genomeA.00430-18

Gasc, C., Constantin, A., Jaziri, F., & Peyret, P. (2017). OCaPPI-Db: an oligonucleotide probe database for pathogen identification through hybridization capture. *Database, 2017*. doi:10.1093/database/baw172

Gasc, C., & Peyret, P. (2018). Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome, 6*(1), 61-61. doi:10.1186/s40168-018-0442-3

Gleason, M. L., Gitaitis, R. D., & Ricker, M. D. (1993). Recent progress in understanding and controlling bacterial canker of tomato in eastern North America. *Plant Disease, 77*(11), 1069-1076.

Green, M. J., Thompson, D. A., & MacKenzie, D. J. (1999). Easy and efficient DNA extraction from woody plants for the detection of phytoplasmas by polymerase chain reaction. *Plant Disease, 83*(5), 482-485.

Hopkins, D. L. (1981). Seasonal concentration of the Pierce's disease bacterium in grapevine stems, petioles, and leaf veins. *Phytopathology, 71*(4), 415-418. doi:10.1094/phyto-71-415

Huptas, C., Scherer, S., & Wenning, M. (2016). Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC Research Notes, 9*(1), 269. doi:10.1186/s13104-016-2072-9

Karamitros, T., & Magiorkinis, G. (2015). A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic acids research, 43*(22), e152-e152. doi:10.1093/nar/gkv773

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Li, R., Mock, R., Huang, Q., Abad, J., Hartung, J., & Kinard, G. (2008). A reliable and inexpensive method of nucleic acid extraction for the PCR-based detection of diverse plant pathogens. *Journal of Virological Methods, 154*(1), 48-55. doi:https://doi.org/10.1016/j.jviromet.2008.09.008

Links, M. G., Demeke, T., Grafenhan, T., Hill, J. E., Hemmingsen, S. M., & Dumonceaux, T. J. (2014). Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. *New Phytol, 202*(2), 542-553. doi:10.1111/nph.12693

Liu, W.-T., Guo, H., & Wu, J.-H. (2007). Effects of Target Length on the Hybridization Efficiency and Specificity of rRNA-Based Oligonucleotide Microarrays. *Applied and Environmental Microbiology, 73*(1), 73. doi:10.1128/AEM.01468-06

Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., . . . Viral Hemorrhagic Fever, C. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology, 37*(2), 160-168. doi:10.1038/s41587-018-0006-x

Ou, T., Xu, W.-f., Wang, F., Strobel, G., Zhou, Z.-y., Xiang, Z.-h., . . . Xie, J. (2019). A Microbiome Study Reveals Seasonal Variation in Endophytic Bacteria Among different Mulberry Cultivars. *Computational and Structural Biotechnology Journal, 17*, 1091-1100. doi:https://doi.org/10.1016/j.csbj.2019.07.018

Pieretti, I., Royer, M., Barbe, V., Carrere, S., Koebnik, R., Cociancich, S., . . . Rott, P. (2009). The complete genome sequence of Xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. *BMC Genomics, 10*(1), 616. doi:10.1186/1471-2164-10-616

Porebski, S., Bailey, L. G., & Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter, 15*(1), 8-15. doi:10.1007/BF02772108

Rezadoost, M. H., Kordrostami, M., & Kumleh, H. H. (2016). An efficient protocol for isolation of inhibitor-free nucleic acids even from recalcitrant plants. *3 Biotech, 6*(1), 61-61. doi:10.1007/s13205-016-0375-0

Samorodnitsky, E., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Lyon, E., . . . Roychowdhury, S. (2015). Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human mutation, 36*(9), 903-914. doi:10.1002/humu.22825

Selvaraj, V., Maheshwari, Y., Hajeri, S., Chen, J., McCollum, T. G., & Yokomi, R. (2018). Development of a duplex droplet digital PCR assay for absolute quantitative detection of "Candidatus Liberibacter asiaticus". *PLOS ONE, 13*(5), e0197184. doi:10.1371/journal.pone.0197184

Sicard, A., Zeilinger, A. R., Vanhove, M., Schartel, T. E., Beal, D. J., Daugherty, M. P., & Almeida, R. P. P. (2018). Xylella fastidiosa: Insights into an Emerging Plant Pathogen. *Annual Review of Phytopathology, 56*(1), 181-202. doi:10.1146/annurev-phyto-080417-045849

Syamkumar, S., Lowarence, B., & Sasikumar, B. (2003). Isolation and amplification of DNA from rhizomes of turmeric and ginger. *Plant Molecular Biology Reporter, 21*(2), 171-171. doi:10.1007/BF02774243

Tian, Q., Zhao, W., Lu, S., Zhu, S., & Li, S. (2016). DNA Barcoding for Efficient Species- and Pathovar-Level Identification of the Quarantine Plant Pathogen Xanthomonas. *PLOS ONE, 11*(11), e0165995. doi:10.1371/journal.pone.0165995

Tran-Nguyen, L. T., & Gibb, K. S. (2007). Optimizing Phytoplasma DNA purification for genome analysis. *J Biomol Tech, 18*(2), 104-112.

Varet, H., & Coppée, J.-Y. (2018). checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes. *Bioinformatics, 35*(5), 901-902. doi:10.1093/bioinformatics/bty706

Vos, M., Quince, C., Pijl, A. S., de Hollander, M., & Kowalchuk, G. A. (2012). A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLOS ONE, 7*(2), e30600. doi:10.1371/journal.pone.0030600

Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., . . . Lander, E. S. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science, 280*(5366), 1077. doi:10.1126/science.280.5366.1077

Wang, H. X., Geng, Z. L., Zeng, Y., & Shen, Y. M. (2008). Enriching plant microbiota for a metagenomic library construction. *Environ Microbiol, 10*(10), 2684-2691. doi:10.1111/j.1462-2920.2008.01689.x

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Zeigler, D. R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International Journal of Systematic and Evolutionary Microbiology, 53*(6), 1893-1900. doi:https://doi.org/10.1099/ijs.0.02713-0

Zheng, Z., Xu, M., Bao, M., Wu, F., Chen, J., & Deng, X. (2016). Unusual Five Copies and Dual Forms of nrdB in “Candidatus Liberibacter asiaticus”: Biological Implications and PCR Detection Application. *Scientific Reports, 6*(1), 39020. doi:10.1038/srep39020

Appendix 13: High throughput sequencing metagenomics sensitivity

Introduction

High throughput sequencing (HTS) has been an essential tool for metagenomic studies, the study of genetic material from environmental samples (Ghosh et al., 2019). As sequencing technology improved and more microbial genomes were published, shotgun metagenomics became an applicable option for population studies and diagnostics (Escobar-Zepeda et al., 2015). Metagenomic sequencing has the capability to perform unbiased identification of known and unknown plant pathogens as total DNA samples are sequenced. This permits the identification of essential genes that may further clarify genus and species simultaneously. Long read sequencing has also contributed greatly to metagenomics as it enables the assembly full genomes, or at least the creation of large contigs, from complex samples thereby linking informative genes and reducing the complexity of downstream data analysis (in comparison to short read sequencing) (Bronzato Badial et al., 2018; Chalupowicz et al., 2019; Hu et al., 2019). Furthermore, due to the lack of amplification, metagenomics has the potential to quantify pathogen populations within a plant.

One of the main limitations of metagenomic sequencing for pathogen detection is that the titre of a target pathogens DNA can be very low compared to that of host genome. This means that to sequence pathogen reads very deep sequencing may be required or in some cases the pathogen may not be detected. Consequently, for many pathogens, metagenomics might not be as sensitive as amplification-based molecular tests such as polymerase chain reaction or as targeted HTS.

In this report, the detection limit of metagenomic sequencing of spiked samples at three sequencing depths (2 GB, 5 GB and 10 GB) is compared to current diagnostic approaches including qPCR and LAMP with the aim of getting the baseline for using HTS sequencing without enrichment. Additionally, the level of taxonomic identification (e.g., species, subspecies and sequencing type (ST)) that can be determined from the metagenomic data was also assessed.

Material and Methods

Preparation of DNA for test sensitivity comparison

In order to compare the sensitivity of HTS diagnostic techniques to currently used molecular tests like qPRC and LAMP, Olive DNA with various concentrations of spiked *Xylella fastidiosa* (Xf) ST 2 DNA were prepared. Nucleic acids were extracted from olive leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and were shown to be Xf negative using species-specific PCR tests. DNA of Xf isolate ICMP15971, at a concentration of 4.75 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of $4.75x10^{-8}$ ng/µL. A volume of 50 µL of each *Xylella* dilution was then spiked into 950 µL of 25 ng/µL olive DNA to reach a final concentration range of 0.2375 ng/µL to $2.375x10^{-8}$.

qPCR

Each qPCR reaction was performed in triplicate, in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL qPCR primer mixture (primers and final concentrations listed in Table 13.1), 7 µL molecular biology grade water and 2 µL of template DNA. Each test included a negative template control (water) and a positive control (Xf ICMP15971 DNA). The cycling conditions were 50 °C for 2 minutes, 95 °C for 2 minutes, followed by 40 cycles of 95 °C for 10 seconds and 62 °C for 30 seconds on a QuantStudio 3 (Thermo Fisher). Where multiple replicates were run, the mean Ct (threshold cycle) was calculated and recorded.

LAMP

Each LAMP reaction was performed in duplicate, in a final volume of 25 µL. The reaction mixture contained 15 µL Isothermal Master Mix (ISO-004), 2 µL LAMP primer mixture (primers and final concentrations listed in Table 13.1), 6 µL molecular biology grade water and 2 µL of template DNA. Each test included a negative template control (water) and a positive control (Xf ICMP15971 DNA). The reactions were run for 65 °C 30 mins, anneal 98 °C to 80 °C (rate 0.05 °C/sec) on either a Genie III (OptiGene) or a QuantStudio 3 (Thermo Fisher). If target DNA was present, dsDNA was amplified and an increase in

fluorescence from the dsDNA intercalating dye was visualised on the Genie III and the time taken recorded (amplification time). Where multiple replicates were run, the mean amplification time was recorded.

Table 13.1 List of primers and final reaction concentrations (Conc.) used in *Xylella fastidiosa* (Xf) LAMP and qPCR tests taken from Harper et al. (2010)

Assay, Primer	Primer sequence (5'-3')	Conc.
qPCR		
XF-F	CACGGCTGGTAACGGAAGA	0.3 µM
XF-R	GGGTTGCGTGGTGAAATCAAG	0.3 µM
XF-P	TCGCATCCCGTGGCTCAGTCC	0.1 µM
LAMP		
XF-F3	CCGTTGGAAAACAGATGGGA	0.2 µM
XF-B3	GAGACTGGCAAGCGTTTGA	0.2 µM
XF-FIP	ACCCCGACGAGTATTACTGGGTTTTTCGCTACCGAGAACCACAC	1.6 µM
XF-BIP	GCGCTGCGTGGCACATAGATTTTTGCAACCTTTCCTGGCATCAA	1.6 µM
XF-LF	TGCAAGTACACACCCTTGAAG	0.8 µM
XF-LB	TTCCGTACCACAGATCGCT	0.8 µM

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, concentration of DNA was measured and ~50 ng DNA was used for each library preparation. Libraries for the spiked Xf serial dilutions were generated in duplicate using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A plant only library and a no template control were also generated. Libraries were then paired end sequenced (2 × 150 bp) using the Illumina NovaSeq at 2 GB, 5 GB and 10 GB sequencing depth.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen et al., 2018). Trimmed reads were analysed using Kraken 2 (Wood et al., 2019) and assembled using SPAdes v3.6.0 (Bankevich et al., 2012). A nucleotide BLAST search (BLASTn) (Altschul et al., 1990) was conducted on the assembled contigs. Bracken analysis was conducted on Kraken2 results to obtain fraction of reads assigned as Xf. Genome coverage was calculated by determining the total length of assembled contigs and dividing it by the average Xf genome size. A k-mer based mapping tool, StringMLST v. 0.6.3 (Gupta et al., 2017), was used to determine if the sequence type (ST) could be assigned for each Xf dilution at the varying sequence depths.

Results

Sequencing output

Reads for each data set were quality trimmed and the average data output from replicates is shown in Table 13.2. In depth analysis revealed contamination in one of the replicates for *Xylella* concentration 0.24 ng/µL sequenced at 10 GB, hence it was removed from further analysis. Final sequence output for samples was between 11-62% higher than estimated with an average of 30% higher data output for all samples.

Sensitivity test

The LAMP and qPCR assays detected Xf at femtogram levels (Table 13.3), 2.4 x 10^{-05} and 2.4 x 10^{-06} ng/µL, respectively. Using Kraken2 software to analyse individual reads from the metagenomic sequencing results, Xf was detected up to 2.4 x 10^{-04} ng/µL of DNA for all sequencing depths assessed, 10 X and 100 X less sensitive than LAMP and qPCR respectively.

Using StringMLST on trimmed reads, the sequence type (ST) was able to be identified in all sequencing depths of the first dilution, 0.24 ng/µL, however, ST was only identified at concentration 2.4 x 10^{-02} ng/µL when sequenced at 10 GB depth. StringMLST was also able to identify some MLST genes for 2 and 5 GB sequencing depth at concentration 2.4 x 10^{-02} ng/µL and 10 GB at concentration 2.4 x 10^{-03} ng/µL, however, it was insufficient to allocate a ST.

As expected, total *Xylella* reads increased as sequencing depth increased and decreased as *Xylella* concentration decreased. Results for largest *Xylella* contig and coverage were linear with total *Xylella* reads with sequencing depth of 10 GB at the highest *Xylella* concentration giving almost full genome coverage at 99.67%.

Table 13.2 Sequencing output average for olive DNA spiked with a serial dilution of *Xylella fastidiosa* DNA at 2 GB, 5 GB and 10 GB sequencing depth done in duplicates.

	Total reads			Sequencing size (GB)		
Concentration of *Xylella* DNA (ng/µL)	**2 GB**	**5 GB**	**10 GB**	**2 GB**	**5 GB**	**10 GB**
0.24	16,432,731	30,743,618	71,689,668[1]	2.97	5.55	12.87[1]
2.4 x 10^{-02}	16,290,745	35,641,047	74,609,748	2.95	6.45	13.53
2.4 x 10^{-03}	12,963,860	35,228,220	76,802,908	2.36	6.39	13.91
2.4 x 10^{-04}	16,600,916	35,627,430	63,908,493	3.01	6.42	11.50
2.4 x 10^{-05}	15,136,800	31,586,701	64,101,188	2.75	5.74	11.64
2.4 x 10^{-06}	17,920,463	32,046,680	72,343,394	3.26	5.83	13.10
2.4 x 10^{-07}	15,443,806	33,350,697	67,454,921	2.80	6.05	12.09
2.4 x 10^{-08}	17,662,231	32,886,638	62,315,591	3.17	5.97	11.25
Olive DNA	13,493,506	-	-	2.44	-	-
BLANK	1,649	-	-	323 KB	-	-

[1]Replicate was removed due to contamination.

Table 13.3 LAMP, qPCR and metagenomic results for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to calculate fraction of reads while coverage was calculated using total length of combined contigs and total genome size. ST2 = sequence type of *Xylella* isolate used in this experiment, UD = Undetermined, Ct = threshold cycle of amplification in qPCR.

Concentration of *Xylella* DNA (ng/µL)	qPCR (Ct)	LAMP (amp time)	StringMLST			Total *Xylella* reads			Total *Xylella* contig			Largest *Xylella* contig			Genome coverage (%)		
			2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB	2 GB	5 GB	10 GB
0.24	21.09	9.9	ST2	ST2	ST2[1]	38162	66031	147981[1]	2561	1486	318[1]	12785	33334	129197[1]	77.5	93.62	99.67[1]
2.4×10^{-02}	24.52	11.2	XF	XF	ST2	2089	4686	9542	133	488	1233	690	1570	3106	1.84	7.23	20.59
2.4×10^{-03}	27.87	12.9	UD	UD	XF	190	501	1155	3	11	45	523	439	575	0.05	0.142	0.619
2.4×10^{-04}	31.17	14.7	UD	UD	UD	13	57	96	1	1	1	317	281	333	0.012	0.011	0.013
2.4×10^{-05}	34.21	20.4	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-06}	37.45	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-07}	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD
2.4×10^{-08}	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD	UD

[1] One replicate was contaminated and was removed from analysis.

Discussion

In this report, we showed that metagenomic HTS is not as sensitive for detection of low titre bacterial pathogens as qPCR and LAMP that are currently used in laboratory testing. Results from the sensitivity test using LAMP, qPCR and metagenomics, showed that qPCR was 100 times more sensitive than metagenomic sequencing (using 10 GB of sequence data). However, it should be noted than the qPCR and LAMP used in this analysis could only identify Xf to the species level and metagenomic sequencing, though less sensitive, could provide species, subspecies and even ST level identification at some concentrations. The knowledge that a sample with a qPCR Ct of <24.5 sequenced to a depth of 10 GB could provide enough sequence data to identify an Xf isolate to the ST level is very useful baseline information for use in a biosecurity situation, although for PEQ it may be enough to know Xf is present. Being able to differentiate sequence types of other bacteria could be useful when some types are considered quarantine pathogens, but other types are present.

As expected, an increase in total sequencing depth generated more Xf reads, increased the Xf genome coverage and improved the resolution of Xf taxonomic identification. This experiment did not include sequencing of a single sample any deeper than 10 GB per sample. It is of interest to the project team to explore whether the sensitivity of qPCR detection of Xf can be achieved with a greater sequencing depth of metagenomic data, however, based on current sequencing costs this exercise would likely be an academic one. Even with the cheapest sequencing on the NovaSeq being approx. $22 per GB, an increased depth of 20 GB would add up to $440 per sample. It is unlikely that this would be a viable option for PEQ diagnostics. For this reason, more cost-effective techniques that facilitate the selective enrichment and sequencing of multiple target pathogens from complex samples are being explored in this project. As an example, the hybridisation probe assay designed (Appendix 13) is expected to add approximately $50 to the cost of processing each sample but will drastically reduce the amount of sequencing required. Further cost/benefit analysis will be conducted as these protocols are evaluated.

Xylella dilutions created in this experiment have been created in bulk to provide a set of standards to benchmark the sensitivity of new HTS tests designed and evaluated in this project.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Bronzato Badial, A., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., & King, J. (2018). Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. *Plant Dis, 102*(8), 1648-1652. doi:10.1094/pdis-04-17-0488-re

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., . . . Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology, 68*(2), 229-238. doi:10.1111/ppa.12957

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics, 6*(348). doi:10.3389/fgene.2015.00348

Ghosh, A., Mehta, A., & Khan, A. M. (2019). Metagenomic Analysis and its Applications. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 184-193). Oxford: Academic Press.

Gupta, A., Jordan, I. K., & Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics, 33*(1), 119-121. doi:10.1093/bioinformatics/btw586

Harper, S. J., Ward, L. I., & Clover, G. R. G. (2010). Development of LAMP and Real-Time PCR Methods for the Rapid Detection of Xylella fastidiosa for Quarantine and Field Applications. *Phytopathology®, 100*(12), 1282-1288. doi:10.1094/PHYTO-06-10-0168

Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., & Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. *Phytobiomes Journal, 3*(2), 92-101. doi:10.1094/pbiomes-01-19-0004-r

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Appendix 14: Hybridization Probe Design for Targeted Sequencing

Introduction

Hybridization probes for targeted sequencing

The simplest and least biased option for pathogen detection using high throughput sequencing (HTS) sequencing is to sequence the plant samples directly, inclusive of host DNA and the entire microbiome (including the pathogen). This is called a metagenomic sequencing approach. However, in cases where the pathogen is present in low titre within the host, metagenomic sequencing is often not sensitive enough to detect the pathogen. An initial trial comparing metagenomics, qPCR and LAMP, discussed in this report (Appendix III), indicated that LAMP and qPCR were more sensitive at detecting *Xylella fastidiosa* (Xf), than direct sequencing of infected plants when up to 10 GB of data was sequenced per sample. This indicates a need for enrichment of the pathogen or pathogen DNA if HTS sequencing is going to be used for detection of low titre fastidious pathogens at the same sensitivity as qPCR.

The use of hybridisation probes prior to HTS enables capture of target nucleic acids from a broader pool, essentially enriching the sample for a specific target and reducing the amount of sequencing required. Hybridisation probes can be used for a range of applications including diversity studies, molecular diagnostics and breed/cultivar improvement (Donaldson et al., 2017; Mascher et al., 2013; Muraya et al., 2015; Pankin et al., 2014; Whiston et al., 2017). For instance, exome capture, or whole exome sequencing, commonly uses hybridization capture to probe and sequence protein coding regions in the genome to detect variants (Oono et al., 2020; Whiston et al., 2017). This is particularly useful in eukaryotic organisms where exons make up a small proportion of the genome yet contain a lot of meaningful information. For example, exomes constitute ~ 1% of the human genome but harbour 85% of the known genetic disease-causing mutations (Gasc et al., 2018a; Z. Wang et al., 2013) therefore offering a cost-effective approach for detection of variation in coding regions than whole genome sequencing. In plant science, exome sequencing for agriculturally important crops has also been frequently used as a low cost, less complex alternative to whole genome sequencing for detecting genetic variation affecting production (e.g. biomass and early maturation) (Bustos-Korts et al., 2019; Muraya et al., 2015; Pankin et al., 2014). Like in human genomes, a low proportion of higher plant genomes are made up of exons, with exome capture assays designed for barley and rice capturing ~1.2% and 15% of the total genome, respectively (with the latter also covering miRNAs) (Mascher et al., 2013; Oono et al., 2020). In contrast, exomes of prokaryotic organisms make up a much larger portion of their genomes. In fact, the majority of bacterial and archaeal genomes contain only 6–14% non-coding DNA (Rogozin et al., 2002).

Hybridization probes have also been used to capture ancient DNA which are known to be hard to sequence as they are often highly degraded and diluted in environmental DNA (Barta et al., 2014). Hybridization probes for ancient DNA have been designed that use distantly related species as a genome reference and has been shown to enrichment ancient DNA of mammals and plants up to 159-fold (Ávila-Arcos et al., 2011; Carpenter et al., 2013).

Hybridization probes have been used to reduce the bias in 16S rRNA amplicon sequencing of microbial communities (Cariou et al., 2018; Rassoulian Barrett et al., 2020). It has been noted that in highly diverse samples, hybridization probes are more accurate than amplicon sequencing due to primer bias, however in less diverse samples, both enrichment techniques generate similar results (Cariou et al., 2018; Gasc et al., 2018b). Custom hybridization probes have also been made for enrichment of DNA for specific microbial taxa. For example, probe sets designed for enrichment of 356 viral species reported to infect humans were shown to enrich unique viral content 18 fold on average, enabling assembly of genomes that could not be assembled from raw metagenomic data (Metsky et al., 2019). Another example of the use of hybridisation probes in bacterial research is the targeted detection of antibiotic resistance determinants in metagenomic samples (Guitor et al., 2019). Genes encoding antibiotic resistance represent less than 0.1% of the metagenome and capture of these regions using hybridisation probes provided an accurate and sensitive method to detect rare and common resistance genes in surveillance of microbial communities (Guitor et al., 2019).

Hybridization probe process

Hybridization probes are biotinylated DNA or RNA fragments, either less than 120 bp for Illumina libraries or less than 500 bp for long read sequencing, that are designed to bind to your targeted regions (Karamitros et al., 2015; Samorodnitsky et al., 2015). Once added to a sample and targets are hybridised, these probes can be captured using streptavidin-coated

magnetic beads, effectively capturing target DNA while background DNA is washed away. Alternatively, the process can be subtractive, and the targeted probes can remove unwanted regions of DNA (e.g. host or rRNA).

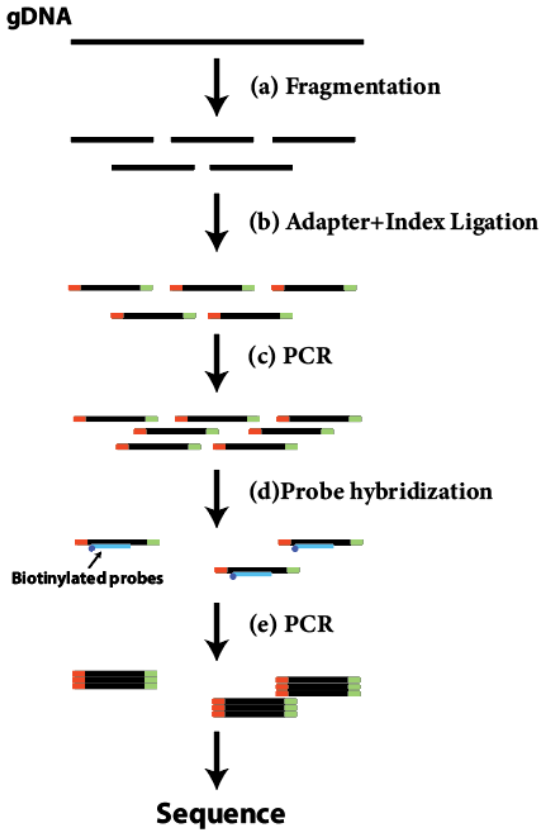


Figure 14.1 Hybridization capture flowchart for targeted high throughput sequencing. The hybridization sequencing workflow begins with DNA isolation from a sample of interest (potential pathogen host plant or vector). (a) DNA is fragmented through heat, sonication, or enzymes. (b, c) PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes are ligated and amplified. (d) Target sequences are captured through hybridization with biotinylated probes and streptavidin-coated magnetic beads. (e) Captured libraries are amplified through PCR and submitted for sequencing.

In this work, hybridization probes with custom targets from Twist Bioscience have been utilised to assess this technology. Twist hybridization probes are 120 bp dsDNA designed to target selected genomic regions of high priority bacterial pathogens. Essentially, genomic DNA are sheered to 150-200 bp fragments and subjected to library preparation. These libraries can then be pooled in eights to a maximum of 1.5 µg of DNA libraries, to increase efficiency (time and cost) before starting the hybridization process which includes target binding by adding the hybridization probe panel, probe capture using streptavidin-coated magnetic beads, PCR amplification to increase concentration of targeted DNA and finally a bead clean up and elution (Figure 13.1).

Regions to enrich

Imported plants for plant primary industries are at risk of bringing in exotic bacterial plant pathogens. Detection of these pathogens in Post Entry Quarantine (PEQ) is essential to prevent their introduction into Australia. Subsequently, detection of these pathogens needs to be specific and accurate. Conversely, extended surveillance for low titre, difficult to detect pathogens could prevent new genetic plant varieties or cultivars from being released, hence, impacting industry. Within this report, we present the design, synthesis and validation strategy for a hybridisation probe capture assay for the detection of pathogenic bacteria within the National Priority Plant Pests list (2019), many of which can be present in low titre in their plant hosts and be difficult to detect with routine diagnostics. Multiple genomic targets are included for each bacterial pathogen with the design strategy varying depending on genome sequence availability. Target identification strategies include pan-genome analysis with closely related bacterial species to identify unique regions or targeting MLST genes due to a lack of complete genome sequences and/or importance to identify genetic diversity (e.g. identification of sequence types

of Xf. Other target regions that could also be incorporated into the probe set are specific genes involved in pathogen-host interactions. These include pathogenicity, virulence and effector genes (Urban et al., 2015).

Xylella fastidiosa

Xylella fastidiosa (Xf) is a complex of several subspecies and genetic variants that causes diseases in a wide array of plant species. This includes *X. fastidiosa* subsp. *fastidiosa* (*Xff*), which causes Pierce's disease in grapevines; *X. fastidiosa* subsp. *pauca* (*Xfp*), which causes citrus variegated chlorosis; and a genetically diverse subspecies, *X. fastidiosa* subsp. *multiplex* (*Xfm*), which causes leaf scorch diseases in a large number of tree species. Management of Xf relies on early detection, especially in PEQ before being introduced into a Xf free country like Australia. However, early detection can be difficult as many hosts are asymptomatic and some strains can be difficult to detect in planta (Bragard et al., 2019; Scortichini et al., 2019). For this purpose, PCR-based detection methods, such as qPCRs, digital PCRs, nested PCRs, LAMP and multilocus sequence typing (MLST) have been implemented (Baldi et al., 2017).

Detailed identification of Xf is important due to its genetic diversity. A frequently used method for bacterial identification is MLST, where differences in nucleotide sequences of seven housekeeping genes are used to identify bacterial strains (Bergsma-Vlami et al., 2017; Scally et al., 2005; Xiaoli Yuan et al., 2010). MLST relies on an assigned number given to an allele of each gene of each characterized strain and enables classification of sequence types (ST) for tracing activities and host prediction. Subsequently, targeting hybridization probes on MLST genes would permit fast, reproducible and easy interpretation of sequencing data.

Erwinia amylovora

Erwinia amylovora (Ea) is the causal agent of fire blight disease, a devastating necrotic disease affecting plant species in the *Rosaceae* family. In the USA, the pathogen causes significant economic losses in horticultural crops such as apples and pears. Ea is native to North America and has now spread to many countries including Europe, Mediterranean countries, central Asia and Asia such as Korea (Ait Bahadou et al., 2018; Bonn et al., 2000; Dardouri et al., 2017; Park et al., 2017). It remains a major biosecurity threat to many other production areas, such as Australia and Japan, that do not have the pathogen.

Many commonly used conventional PCR diagnostics for the detection of Ea are broad and not specific for the pathogen (Powney et al.,2011). More specific PCR tests to detect Ea have targeted the pEA29 plasmid (Bereswill et al., 1992; Llop et al., 2000; McManus & Jones, 1995; Salm and Geider 2004; De Bellis et al. 2007). However, it has been found that not all Ea strains harbour the plasmid, leading some strains to be undetected (Llop et al., 2006). Chromosomally targeted PCR tests, qPCR, LAMP (Shin et al., 2018) and digital PCR (dPCR) (Dreo et al., 2014; Santander et al., 2019), have shown promise in terms of specificity and sensitivity.

The limited target regions used in diagnostic test design indicate a need to identify other regions within the genome. A possible target are virulence factors and their effectors. Due to a high degree of specialization to a narrow ecological niche, Ea genomes have low genetic diversity, however, there are significant differences in the level of virulence (Cabrefiga et al., 2005; Mann et al., 2013; Sholberg et al., 2001). Like many other Gram-negative plant-pathogenic bacteria, Ea also contains an *hrp*- type III secretion system (T3SS) that delivers effector proteins into host plants. Proteins involved include; effector proteins such as DspA/E, Eop1, Eop3, Eop4 and HopPtoC; hairpins, such as HrpN and HrpW; and other proteins such as Eop2, HrpJ, HrpK and an HrpL regulon (McNally et al., 2012; Xiaochen Yuan et al., 2020). Another pathogenicity factor of *E. amylovora* is the amylovoran biosynthesis. Amylovoran biosynthetic genes are located within a 12-gene amylovoran biosynthetic (ams) operon, designated as amsA to amsL, amyR (Piqué et al., 2015). Regulators for the amylovoran biosynthesis are the Rcs phosphorelay system and AmyR, which is also required for virulence (D. Wang et al., 2012a; D. Wang et al., 2012b).

Potential targeted regions for hybridization probes could encompass regions from chromosomally targeted primers, virulence factors and their effectors, and also unique regions from pan-genome analysis conducted by (Mann et al., 2013) and Mann (2013b). Within Mann (2013b), the pan-genome analysis provided Ea-specific regions with associated features for target analysis so regions including those likely to be located on mobile genetic elements or IS elements could be avoided wherever possible for diagnostic targeting.

Xanthomonas citri subsp. *citri*

Xanthomonas citri subsp. *citri* (Xcc) is the causal agent of the citrus bacterial canker type A which is the most severe and widely spread of the citrus bacterial canker. Within the type A there are pathotypes, A, A*, and Aw (Webster et al., 2020). There is no curative treatment for plants infected leading to significant economic losses due to intensive eradication (Behlau, 2020). Preventative measures have been applied, such as choosing less susceptible cultivars whenever possible and spraying copper-based bactericides in high risk areas, however, for countries without Xcc, prevention of entry is vital (Behlau, 2020). Many primers, conventional, LAMP and qPCR, used for Xcc has been aimed to encompass all Xanthomonas strains that causes citrus bacterial canker such as *Xanthomonas citri* pv. *aurantifolii* that causes citrus bacterial canker type B and C (Cubero et al., 2002; Fonseca et al., 2019; Rigano et al., 2010). There are also primers that have been used to identify down to the subtype levels, though some proved promising, others observed amplification of other *Xanthomonas* species (Coletta-Filho et al., 2006; Cubero et al., 2002; Koupaei et al., 2014; Yu et al., 2017). Primers were also designed for qPCR and ddPCR (Golmohammadi et al., 2007; Yun Zhao et al., 2016). Even with primers specifically targeting Xcc, genomic regions for hybridization capture remains limited.

Candidatus Liberibacter species

Liberibacter species are unculturable phloem-limited α-proteobacteria. Among them, '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso) are economically important phytopathogens, causing diseases such as zebra chip and citrus greening, and are all exotic plant pathogens to Australia.

Further classifications have been given to CLaf and CLso. For CLso, five haplotypes have been identified so far (LsoA, LsoB, LsoC, LsoD and LsoE). As for CLaf, subspecies such as CLaf subsp. clausena (CLafCl), CLaf subsp. vepridis (CLafV), and CLaf subsp. tecleae (CLafT) CLaf subsp. capensis (CLafC) and CLaf subsp. zanthoxyli (CLafZ) have been identified (Roberts et al., 2017). Primers have been designed for qPCRs to detect CLas, Clam, CLaf and CLso (Li et al., 2009; Li et al., 2006). However, due to the inability to culture *Ca.* Liberibacter species, phylogenetic analysis on either 16S rRNA, 50S rRNA or housekeeping genes (**Table 14.1**) have been predominantly used for molecular characterisation of CLas, CLaf, CLam and CLso differentiating between species and haplotype/subspecies (Morris et al., 2017; Roberts et al., 2017; Ruiz-Padilla et al., 2020; Teresani et al., 2014). Hence, potential targets for hybridization capture should include these regions.

Table 14.1 List of target genes used for molecular characterization of 'Candidatus Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso) through amplicon sequencing and phylogeny.

Target genes for characterization	Source
adk, *gyrB*, *dnaG*, *fumC, grpE*, *icdA*, *metG*, *mutS*, *purA,* and *recA*	(Ruiz-Padilla et al., 2020)
adk, fumC, dnaG, grpE, icdA, kdsA, metG, mutS, purA and *recA*	(Glynn et al., 2012)
acnA, atpD, ftsZ, glnA, glyA, gnd and *groEL*	(Hajri et al., 2019)
omp	(Roberts et al., 2017)
dnaG, *gyrB*, *mutS*, *nusG*, *rplA*, *rpoB* and *tufB*	(Morris et al., 2017)
tufB, secE, nusG, rplKAJL, rpoB	(Teixeira et al., 2008)

Phytoplasma 16SrI

Phytoplasmas are associated with diseases in more than 1,000 plant species including economically important crops and ornamental plants. They are unculturable, wall-less bacteria and have some of the smallest bacteria genomes. Phytoplasma transmit from plant to plant through phloem feeding insects of the order Hemiptera such as leafhoppers, planthoppers and psyllids (Weintraub et al., 2006). Phytoplasma 16SrI (*Candidatus* phytoplasma asteris or aster yellows) strains affect a wide range of plant hosts causing a great impact on agriculture worldwide (Lee et al., 2004). The phytoplasma 16SrI group is further diverged into subgroups with the most detected being 16SrI-A and 16SrI-B (Yan Zhao et al., 2009). As control of phytoplasma can get time and resource intensive due to screening, vector control, symbiont control and systemic acquired resistance, prevention of entry into the country is important (Weintraub, 2007).

Due to the lack of complete phytoplasma genomes caused by the inability to culture them in vitro, identifying unique genes within the 16SrI group would be risky as the genetic diversity of the majority of phytoplasmas are unknown (Maejima et al., 2014). Hence, targeting common genes used for phytoplasma identification through amplicon sequencing may be the only option to clearly differentiate between 16S groups and subgroups. Furthermore, commonly sequenced genes would provide a database for phylogenetic analysis for phytoplasma identification.

From literature, common targeted region for molecular characterization of phytoplasmas through amplicon sequencing is the 16S rRNA and 50S rRNA regions such as *rps19*, *rpl22* and *rps3* (Fernández et al., 2020; Lee et al., 2004; Pereira et al., 2017; Sichani et al., 2014; Taloh et al., 2020). The *secY* gene has also been used for phylogeny analysis and been reported to be more variable among phylogenetic markers to aid in differentiating phytoplasmas (Lee et al., 2010; Roberto Pierro et al., 2020; Shao et al., 2006). Additional sequencing of amplicons include *tufB*, *stamp*, *map*, *vmp* genes, *secA*, *rbfA*, *groEL* and *inmp* genes (Al-Subhi et al., 2020; Roberto Pierro et al., 2020; R. Pierro et al., 2017; Quaglino et al., 2015; Shao et al., 2006). A common molecular marker for all bacteria, the gene encoding the 60 kDa chaperonin protein (*cpn*60) has also been used for phytoplasma identification, with software like CpnClassiPhyR used to aid in classification (Muirhead et al., 2019).

Materials and Methods

Hybridization Capture Targets

Target identification

Genes used as targets for hybridisation probe design were identified two ways: by pan-genome analysis where multiple genomes were available, to identify core genes amongst strains of the same species that are not present in closely related species, and by review of literature and sequence databases to identify genes for species where limited sequence data was available (e.g., phytoplasma species).

Genomes were annotated using Prokka v1.14.5 (Seemann, 2014) to identify potential target enrichment regions and to identify MLSA and MLST genes. Annotation files from Prokka were submitted to Roary pan-genome pipeline v3.13.0 (Page et al., 2015) to perform a pan-genome analysis for genomes listed in Tables 13.3, 14.3, 14.5 and 14.7. Pan-genome analysis of Xf and closely related genus (Table 13.3) was done to identify the core genes using a 99% cut-off, with a BLASTP identity cut-off of at 70%. Similarly, pan-genome analysis of Xcc and closely related species (Table 13.3) was done to identify the core genes using a 99% cut-off, with a BLASTP identity cut-off of at 70%. Pan-genome analysis using all available CLas, CLaf, Clam and CLso (Table 14.5), where whole genomes were available, to identify the core genes using a 99% cut-off, with a BLASTP identity cut-off of at 95% to search for MLSA genes with high identity within the species. Similarly, pan-genome analysis using all available phytoplasma 16SrI species, where whole genomes were available, to identify core genes between CLas, CLaf, Clam and CLso (Table 14.7) using a 99% cut-off, with a BLASTP identity cut-off of at 95% to search for MLSA genes with high identity within the species. Target genes for hybridization capture of Ea relied on a previous pan-genome analysis by Mann et al. (2013). Annotation was taken from NCBI for the strain *Erwinia amylovora* CFBP1430 (NC_013961.1) and the BLASTn algorithm was used to assess genes and primers chosen from Mann et al. (2013) and Mann (2013b).

Unique genes from the pan-genome analysis of each bacterial species that were larger than 1kbp and sequences of genes chosen from literature were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990) to identify any potential cross reaction with non-target species.

Custom probe design

Target regions identified for probe design were supplied to Twist Bioscience for custom capture library tiling design and synthesis. Probes were designed removing regions of low complex with 4 X tiling to increase capture rate of low titre pathogens. Each probe consists of a biotinylated 120 mer dsDNA so once bound to targeted regions of sample DNA, it can be captured using Streptavidin bound to magnetic beads.

Validation of hybridization probes

DNA extraction

Nucleic acids were extracted from olive, lemon and apple leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Olive, lemon and apple trees are known to be plant hosts of Xf, Xcc and Ea respectively, hence, used as background DNA for positive samples used in validation experiments.

Positive samples for hybridization probe capture validation

Positive samples use for hybridisation probe capture validation are listed in Table 11.2 Genomic DNA for each species was tested using PCR and/or qPCR to confirm the identify, except for the phytoplasma DNA, which was validated in another AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria). Identification/confirmation of Xf DNA was reported in Appendix I. The concentration of each sample was measured using the Quantus™ fluorometer following manufacturer's instructions.

Table 14.2 Bacterial pathogen isolates used for validation of targeted high throughput sequencing using hybridization capture.

Sample	Species	Host background	Origin
ICMP 15971	*Xylella fastidiosa* subsp. *fastidiosa*	Pure culture	USA
CLbr A1	*Candidatus* Liberibacter brunswickensis	*Solanum melongena*	Australia
5 Vic #7 (CLbr)	*Candidatus* Liberibacter brunswickensis	*Accizia solanicola*	Australia
1 Lso N (CLso)	*Candidatus* Liberibacter solanacearum	*Solanum tuberosum*	New Zealand
TPP6 (CLso)	*Candidatus* Liberibacter solanacearum	*Bactericera cockerelli*	New Zealand
BRIP 70839 (CLaf)	*Candidatus* Liberibacter africanus	*Citrus sinensis* (L.) Osbeck	South Africa
Xcc 83	*Xanthomonas citri* subsp *citri*	Pure culture	Australia
Ea322A	*Erwinia amylovora*	Pure culture	Attenuated strain generated and supplied by Cornell University (USA). Original Ea322 strain from France
StLL (16SrXXXVII)	*Ca.* phytoplasma stylosanthis	*Solanum tuberosum*	Australia
AGY (16SrXII)	*Ca.* phytoplasma australiense	*Vitis vinifera*	Australia
WF-B (16SrII)	*Ca.* phytoplasma australasiae	*Solanum tuberosum*	Australia

qPCR

Successful plant DNA extraction was assessed using qPCR targeting the *cox* gene. Plant DNA was further diluted 1:10 as high concentration of contaminants could inhibit PCR reactions. Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL cox primer mixture (COX primers and final concentrations listed in Table 13.1), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water) and a positive plant control. The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

The identity of *Ca.* liberibacter DNA samples were confirmed using qPCR. Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL HLB primer mixture (HLB primers and final concentrations listed in Table 13.1), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water). The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 10 seconds and 58°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

Table 14.3 List of primers and final reaction concentrations (Conc.) used in qPCR tests to confirm presence of plant DNA, Candidatus Liberibacter asiaticus' (CLas), 'Ca. Liberibacter americanus' (CLam), 'Ca. Liberibacter africanus' (CLaf) and 'Ca. Liberibacter solanacearum' (CLso) DNA.

Primer	Gene Target	Species	Primer sequence	Conc. (µM)	Reference
COX-F			CGTCGCATTCCAGATTATCCA	0.25	
COX-R	*cox*	Plants	CAACTACGGATATATAAGAGCCAAAACTG	0.25	(Weller et al., 2000)
COX-P			TGCTTACGCTGGATGGAATGCCCT	0.1	
HLBaf	16S rRNA	CLaf	TCGAGCGCGTATGCAATACG	0.25	(Li et al., 2006)
HLBso	16S rRNA	CLso	GTCGAGCGCTTATTTTTAATAGGA		(Li et al., 2009)
HLBr	16S rRNA	CLaf, CLam, CLas, CLso	GCGTTATCCCGTAGAAAAAGGTAG	0.25	(Li et al., 2009; Martoni et al., 2021)
HLBp	16S rRNA	CLaf, CLam, CLas, CLso	AGACGGGTGAGTAACGCG	0.1	(Li et al., 2009)

PCR

DNA extracted from apple leaf tissue, neat and 1:10 dilution, was tested for the presence of Ea using PCR. Each Ea PCR reaction was performed in a final volume of 25 µL using primer sets Ea1-F&R and Ea2-F&R (Table 13.2). The reaction mixture contained 1 x PCR buffer (Invitrogen, Netherlands), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each forward and reverse primer, 1 U of Platinum Taq polymerase (Invitrogen, Netherlands) and topped up to 25µL using molecular biology grade water. The PCR conditions for both primer sets were as follows: denaturation at 95°C for 2 mins, 35 cycles of 95°C for 30 s, 64°C for 30 s and 72°C for 30 s, and a final hold step at 72°C for 7 mins. Each test included a negative template control (water) and a positive control (Ea322A DNA).

DNA extracted from lemon leaf tissue, neat and 1:10 dilution, was tested for the presence of Xcc using PCR. Each Xcc PCR reaction was performed in a final volume of 25 µL. The reaction mixture contained 1 x PCR buffer (Invitrogen, Netherlands), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each R22 forward and reverse primer, 0.8 µM of each R8-PP forward and reverse primer (Table 13.2), 1 U of Platinum Taq polymerase (Invitrogen, Netherlands) and topped up to 25µL using molecular biology grade water. The PCR conditions for both primer sets were as follows: denaturation at 95°C for 2 mins, 30 cycles of 95°C for 30 s, 58°C for 30 s and 72°C for 30 s, and a final hold step at 72°C for 5 mins. Each test included a negative template control (water) and a positive control (Xcc 83 DNA).

PCR products for Ea PCR reactions were visualised by agarose gel electrophoresis (1% agarose gel) with SYBR safe DNA gel stain (Invitrogen, Netherlands). PCR products for Xcc PCR reactions were visualised by agarose gel electrophoresis (2.5% agarose gel) with SYBR safe DNA gel stain (Invitrogen, Netherlands).

Table 14.4 List of primers used in PCR tests to confirm presence of Erwinia amylovora and Xanthomonas citri subsp. citri.

Target	Primer	Primer sequence	Expected Product size
E. amylovora	Ea1-F	GACACCGGCTATCTGCATCA	235 bp
	Ea1-R	AAATCCTTTGATCGGCGGGT	
	Ea2-F	AAC GCA TCG TCC GGG CAA ATG T	388 bp
	Ea2-R	AGC GCT GTC TGG AAA CCC ATG C	
X. citri subsp. *citri*	R8- PP1F	GACACCGGCTATCTGCATCA	200 bp
	R8- PP1R	AAATCCTTTGATCGGCGGGT	
	R22- 1F	ACCACCACTAACATCCGGAG	400 bp
	R22- 1R	TTCTCAAAAGAGGCTACAATGAAGA	

Hybridization probe sensitivity test and metagenomic comparison

DNA of Ea isolate 322A, at a concentration of 1.94 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of 1.94 x 10^{-8} ng/µL. A volume of 10 µL of each Ea dilution was then spiked into 37.5 µL of 51.8 ng/µL apple DNA to reach a final concentration range of 0.409 ng/µL to 4.09 x 10^{-8} ng/µL which was equivalent to 1% to 1 x 10^{-7}% of total DNA.

DNA of Xcc isolate 83, at a concentration of 1.4 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of 1.4 x 10^{-8} ng/µL. A volume of 10 µL of each Xcc dilution was then spiked into 21.3 µL of 65.8 ng/µL olive DNA to reach a final concentration range of 0.448 ng/µL to 4.48 x 10^{-8} which was equivalent to 1% to 1 x 10^{-7}% of total DNA.

DNA samples extracted from host samples for the *Ca.* Liberibacter and phytoplasma species were sourced from other projects. CLaf was provided by Biosecurity Queensland (Department of Agriculture and Fisheries) which was sourced from CRI facilities, Nelspruit, South Africa. Phytoplasmas were provided by AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria). CLso samples were provided by Plant and Food Research, New Zealand. *Ca.* Liberibacter brunswickensis samples were provided by an AVR PhD project. Bacterial spiked plant DNA (8 dilutions of Ea, Xf and Xcc spiked in plant DNA), 3 phytoplasma samples and five *Ca.* Liberibacter samples (CLaf, 2 CLso and 2 *Ca.* Liberibacter brunswickensis) were used for probe and no-probe comparisons in duplicates.

Target enrichment: library preparation, hybridization, enrichment and sequencing

Using the Quantus™ fluorometer, DNA concentration was measured for all samples. Approximately 50 ng of Xf, Ea and Xcc DNA serial diluted and spiked in plant DNA at different concentrations and ~3 ng of plant or insect samples naturally infected with phytoplasma and *Ca.* Liberibacter species was used for library preparation with the exception of the *Ca.* Liberibacter solanacearum infected tomato potato psyllid sample TPP6, where 5 µL containing ~ 0.8 ng of DNA was used as the sample volumes was limited. Libraries were generated with ~50 ng starting DNA using half reaction volumes and ~3 ng starting DNA using full reaction volumes of the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit and PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). Four replicates of each library were prepared along with a plant only library and a no template control library.

Using the Quantus™ fluorometer, DNA concentration of each library was quantified. Four libraries per sample were created. Two libraries were further used for probing while the remaining two were used as metagenomic controls. Around ~220 ng DNA of Xf, Ea and Xcc serial dilution duplicates were pooled eight samples per hybridization capture, which was equivalent to 6 hybridization capture reactions. The concentration of *Ca.* Liberibacter and phytoplasma libraries were low (between 1-

3.4 ng/µL) with one of the duplicates for TPP6 being 0.27 ng/µL, so full volume libraries were pooled in eight samples per hybridization capture, which was equivalent to two hybridization capture reactions. Pooled libraries were subjected to hybridization capture using the Twist Target Enrichment Protocol (Twist Bioscience, USA) according to manufacturer's instructions.

Libraries for the non-captured TPP6 duplicates gave 0.164 ng/µL and 0.14 ng/µL, lower than all other libraries. The 2200 TapeStation system (Agilent technologies) was used to assess library quality and size distribution of libraries. Libraries for the non-captured TPP6 duplicates were not detected on the TapeStation, however, they were still added to the sequencing run using 15 µL. Hybridization captured and non-captured libraries were paired end sequenced (2 x 250 bp) using the Illumina NovaSeq on a SP flow cell.

Results

Hybridization capture targets

Xylella fastidiosa

Pan-genome analysis of various Xf strains representing a diversity of subspecies, *Xylella taiwanensis* and two *Xanthomonas* species was generated using Roary with a reduced cut-off of 70% BLASTp identity to stringently search for unique genes (Table 13.3). Results showed both *Xylella* species and 2 *Xanthomonas* species shared 972 core genes while the 6 *Xylella* strains share 1,440 core genes. Prokka annotated between 2322 to 2797 genes for *Xylella fastidiosa* strains and 2817 genes for *Xylella taiwanensis*.

Table 14.5 Genomes of *Xanthomonas* and *Xylella* isolates used for annotation and pan-genome analysis. Annotation results using Prokka gave total genes annotated within the genome while pan-genome analysis using Roary gave unique genes for each genome.

Species/subspecies/pathovar	Strain ID	NCBI assembly no.	Total genes	Unique genes
Xanthomonas arbicola	17	GCA_000972745.1	4420	688
Xanthomonas citri pv *glycines*	2098	GCA_007724205.1	4114	381
Xylella fastidiosa	TemeculaL	GCA_006370155.1	2405	120
Xylella fastidiosa subsp fastidiosa	Bakersfield-1	GCA_009664125.2	2468	78
Xylella fastidiosa subsp multiplex	Griffin	GCA_000466025.1	2322	249
Xylella fastidiosa subsp pauca	De Donno	GCA_002117875.1	2441	318
Xylella fastidiosa subsp sandyi	Ann-1	GCA_000698805.1	2797	478
Xylella taiwanensis	PLS235	GCA_003352785.1	2817	1176

Unique genes for probe design were chosen based on length, region in the chromosome and low to no BLASTn results to genomes other than *Xylella*. Although Roary was able to identify unique genes to each sub-species, comparison of these genes to the non-redundant database at NCBI using BLASTn identified that there was high sequence identity to genes found in other bacterial species and were discounted. MLST genes were also chosen as capture regions targeting *Xylella* to aid in identifying sequence type (Table 14.1). Chosen hybridization target genes were further compared with *Xylella* genomes from Table 14.10 using the BLASTn algorithm and a representative strain was chosen within each subspecies and *X. taiwanensis* based on low identity with other strains to cover genetic diversity of genes. Genes chosen as target regions for hybridization capture are listed in Table 14.1.

Table 14.6 Chosen genes for development of targeted hybridization capture high throughput sequencing for detection of *Xylella fastidiosa* and *X. taiwanensis* and their presence in representative strains.

		Presence/absence, gene size (base pair)				
Prokka annotation		**Xf subsp. sandyi Ann-1**	**XF Bakersfield**	**Xf De dono**	**Xf Filmore**	**Xf taiwanensis PLS229**
group_806	hypothetical protein	2195	2195	2195	2198	2501
group_1256	hypothetical protein	1379	17159	1388	1316	
mshA	D-inositol-3-phosphate glycosyltransferase	2582	2582	2582	2579	2517
group_7658	hypothetical protein			2093		
copB	Copper resistance protein B	1004	944	914	869	
group 1350	hypothetical protein	1388	1370	1370	1379	1382
dacC	D-alanyl-D-alanine carboxypeptidase DacC	1391	1443	1497	1452	1449
group_852	hypothetical protein	2219	2219		2212	
gltT	C4-dicarboxylate transport protein 2	1313	1313	1311	1313	1313
lacF	Lactose transport system permease protein LacF	881	881	881	881	881
nuoL	NADH-quinone oxidoreductase subunit L	2141	2141	2141	2141	2147
holC	DNA polymerase III subunit chi	425	425	425	425	425
leuA	2-isopropylmalate synthase	1559	1559	1559	1559	1559
cysG	Siroheme synthase	1435	1436	1435	1435	1405

Erwinia amylovora

Target genes for hybridization capture of Ea was taken from Mann et al. (2013) and Mann (2013b) with gene sequences taken from the gene annotation of the strain *Erwinia amylovora* CFBP1430 from NCBI (NC_013961.1). Genes with the least cross reaction with non-target species were chosen based on a BLASTn search using the NCBI database excluding Ea strains. Genes and BLASTn results are listed in Table 14.2.

Table 14.7 Chosen genes for development of targeted hybridization capture high throughput sequencing for detection of *Erwinia amylovora* and Blastn results to the NCBI database excluding *E. amylovora* strains.

Genes	Size (bp)	BLASTn results
eop2	1991	Less than 45% query cover and less than 70% identity to strains of other species
WP_004154715.1_7-hypothetical protein	6255	Less than 83% identity to 5 strains of other species
WP_004155245.1_424-hypothetical protein	3285	Less than 75% identity to 3 strains of other species. Remaining hits had query cover less than 10% with less than 86% identity.
amsE	801	75-83% identity to 5 strains of other species. Remaining hits had query cover less than 12% with less than 76% identity.
amsD	1056	67-85% identity to 7 strains of other species. Remaining hits had query cover less than 18% with less than 87.5% identity.

Xanthomonas citri* subsp. *citri

Pan-genome analysis of various *Xanthomonas* species (Table 14.3) were generated using Roary with a reduced cut-off of 70% BLASTp identity to stringently search for unique genes. Results showed all 11 *Xanthomonas* strains shared 1816 core genes. Prokka annotated 4613 genes in Xcc. Among them, 405 genes were unique with 97 unique genes being above 1 Kbp.

Table 14.8 Genomes of Xanthomonas isolates used for annotation and pan-genome analysis. Annotation results using Prokka gave total genes annotated within the genome.

Xanthomonas species	NCBI assembly no.	Total genes
Xanthomonas axonopodis pv *commiphoreae* LMG26789	GCA_003698225.1	3989
Xanthomonas axonopodis pv *dieffenbachiae*	GCA_001564415.1	4223
Xanthomonas campestris pv *musacearum*	GCA_000277895.2	4156
Xanthomonas citri pv *glycines* strain K2	GCA_007724385.1	4479
Xanthomonas citri pv *phaseoli varfuscans*	GCA_002759395.1	4283
Xanthomonas citri pv *vignicola* strain CFBP7113	GCA_002218285.1	4321
Xanthomonas citri pv *viticola*	GCA_000723725.1	4252
Xanthomonas citri pv *punicae* LMG859	GCA_006228545.1	4428
Xanthomonas oryzae BB151-3	GCA_004319485.1	4550
Xanthomonas translucens pv *translucens*	GCA_900094325.1	3957
Xanthomonas citri subsp *citri* Aw12879	GCA_000349225.1	4613

Unique genes for probing were chosen based on length, region in the chromosome and low to no BLASTn results to genomes other than Xcc (Table 9). Though Roary was able to identify unique genes to Xcc comparison of these genes to the non-redundant database at NCBI using BLASTn results identified that most of these had high sequence identity to genes found in other bacterial species and were discounted.

Table 14.9 Chosen genes for development of targeted hybridization capture high throughput sequencing for detection of *Xanthomonas citri* subsp. *citri*.

Prokka annotation		Size (bp)
gtaB	AMDFNCPH_00526	1518
hypothetical protein	AMDFNCPH_00529	1185
ndhJ	AMDFNCPH_01080	2025
hypothetical protein	AMDFNCPH_01855	1185
paoB	AMDFNCPH_01857	1380
btsT	AMDFNCPH_01859	2988
macB	AMDFNCPH_02036	2109
hypothetical protein	AMDFNCPH_02047	2118
hypothetical protein	AMDFNCPH_02358	10173
hypothetical protein	AMDFNCPH_02724	1911
hypothetical protein	AMDFNCPH_03192	1218

Candidatus Liberibacter species

Annotation of CLaf, CLas, Clam and CLso (Table 14.5) were generated using Prokka to aid in identifing MLSA genes. Pan-genome analysis was conducted using Roary with a higher cut-off of 95% BLASTp identity to stringently search for high identity among MLSA genes. The high cut-off Roary analysis did not detect any core genes among the *Ca* Liberibacter strains, even the MLSA genes. Prokka annotated between 983 to 1128 genes for the *Ca* Liberibacter strains.

Multiple MLSA genes from literature were identified in the *Ca.* Liberibacter strains and compared to the non-redundant database at NCBI using the BLASTn algorithm to identify any potential cross reaction with non-target species. BLASTn results showed the genes *rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG* and *glnA* had the least identity to non-*Ca.* Liberibacter species Table 14.6.

Table 14.10 Genomes of *Candidatus* Liberibacter strains used for annotation and pan-genome analysis. Annotation results using Prokka gave total genes annotated within the genome.

Ca. Liberibacter strains	NCBI assembly no.	Total genes
Ca. Liberibacter africanus PTAPSY	GCA_001021085.1	1066
Ca. Liberibacter americanus Sao Paulo	GCA_000496595.1	983
Ca. Liberibacter asiaticus PA19	GCA_013309695.1	1074
Ca. Liberibacter asiaticus gxpsy	GCA_000346595.1	1128
Ca. Liberibacter solanacearum ZC1	GCA_000183665.1	1126

Table 14.11 Chosen genes for development of targeted hybridization capture high throughput sequencing for detection of '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso) and their presence in representative strains.

	***Candidatus* Liberibacter asiaticus PA19**		***Candidatus* Liberibacter asiaticus gxpsy**		***Candidatus* Liberibacter africanus**		***Candidatus* Liberibacter solanacearum**		***Candidatus* Liberibacter americanus**	
Genes	**Prokka annotation**	**Size (bp)**	**Prokka annotation**	**Size (bp)**	**Prokka annotation**	**Size (bp)**	**Prokka annotation**	**Size (bp)**	**Prokka annotation**	**Size (bp)**
rplL	JAMHBJMG_00015	381	EMJBICAA_00009	381	ADEABLJM_00866	378	DJHFHIGO_00986	375	CBFBGEJB_00456	396
rplJ	JAMHBJMG_00016	519	EMJBICAA_00010	519	ADEABLJM_00865	519	DJHFHIGO_00987	522	CBFBGEJB_00455	519
metG	JAMHBJMG_00058	1377	EMJBICAA_00083	630	ADEABLJM_00095	1539	DJHFHIGO_00587	1539	CBFBGEJB_00710	1536
adk	JAMHBJMG_00111	1566	EMJBICAA_00107	606	ADEABLJM_00178	603	DJHFHIGO_01039	606	CBFBGEJB_00914	606
mutS	JAMHBJMG_00452	606	EMJBICAA_00451	2763	ADEABLJM_00422	2724	DJHFHIGO_00158	2739	CBFBGEJB_00207	2703
grpE	JAMHBJMG_00499	648	EMJBICAA_00497	648	ADEABLJM_00378	606	DJHFHIGO_00114	639	CBFBGEJB_00250	636
dnaG	JAMHBJMG_00714	1947	EMJBICAA_00711	1947	ADEABLJM_00162	1947	DJHFHIGO_01027	1953	CBFBGEJB_00546	1935
glnA	JAMHBJMG_00058	1376	EMJBICAA_00052	1376	ADEABLJM_00060	1377	DJHFHIGO_01026	1389	CBFBGEJB_00901	1404

16SrI group phytoplasmas

Annotation of phytoplasma 16SrI species and strains (Table 14.7) were generated using Prokka to aid identify MLSA genes. Pan-genome analysis was conducted using Roary with a higher cut-off of 95% BLASTp identity to stringently search for high identity among MLSA genes. The high cut-off Roary analysis detected 111 core genes among the phytoplasma strains. Prokka annotated between 529 to 954 genes in each phytoplasma species.

Table 14.12 Genomes of 16SrI phytoplasma species used for annotation and pan-genome analysis. Annotation results using Prokka gave total genes annotated within the genome.

16SrI group phytoplasma species or strain	NCBI assembly no.	Total genes
Aster yellows witches-broom phytoplasma	GCA_000012225.1	687
Brassica napus phytoplasma TW1	GCA_003181115.1	732
Catharanthus roseus aster yellows phytoplasma DeVilla chromosome	GCA_004214875.1	561
Chrysanthemum coronarium phytoplasma OY-V	GCA_000744065.1	770
Chrysanthemum yellows phytoplasma CYP	GCA_000803325.1	600
Maize bushy stunt phytoplasma M3	GCA_001712875.1	529
New Jersey aster yellows phytoplasma	GCA_002554195.1	730
Onion yellows phytoplasma OY-M	GCA_000009845.1	954
Periwinkle leaf yellowing phytoplasma DY2014	GCA_005093185.1	901
Rapeseed phyllody phytoplasma RP166	GCA_013372025.1	856
Wheat blue dwarf phytoplasma	GCA_000495255.1	549

Multiple MLSA genes identified from the literature were compared to the non-redundant database at NCBI using the BLASTn algorithm to find any potential cross reaction with non-target species. BLASTn results showed the genes *rbfA*, *nusA*, *rpsC-rpsS*, *rplF*, *secY*, *map* and *secA* had the least identity to non-phytoplasma species and are core genes in the pan-genome analysis (Table 14.8). Among the 11 16Sr I phytoplasma species or strains, Aster yellows witches-broom phytoplasma had the least identity with Chrysanthemum coronarium phytoplasma OY-V, with the lowest identity being 94.6% for the gene *secY* gene.

Table 14.13 Chosen genes for development of targeted hybridization capture high throughput sequencing for detection of the phytoplasma 16SrI group and their presence in representative strains.

	Chrysanthemum coronarium phytoplasma OY-V		Aster yellows witches-broom phytoplasma	
Genes	**Prokka annotation**	**Size (bp)**	**Prokka annotation**	**Size (bp)**
rbfA	KOJGJCGG_00276	342	FMMPGCON_00602	342
nusA	KOJGJCGG_00279	1074	FMMPGCON_00599	1074
rpsC-rpsS	KOJGJCGG_00359	1429	FMMPGCON_00545	1434
rplF	KOJGJCGG_00370	555	FMMPGCON_00536	555
secY	KOJGJCGG_00375	1242	FMMPGCON_00531	1242
map	KOJGJCGG_00377	744	FMMPGCON_00529	744
secA	KOJGJCGG_00481	414	FMMPGCON_00327	2508

Probe design

A total of 5,901 dsDNA hybridization probes were designed for *Xylella fastidiosa*, *X. taiwanensis, Xanthomonas citri* subsp. *citri*, *Erwinia amylovora*, '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso), and phytoplasma 16SrI. Low complex regions with 4 X tiling were removed to increase capture rate of low titre pathogens. Each probe consists of a 120 mer dsDNA with a biotin molecule at one end to capture targeted genes using Streptavidin bound to magnetic beads.

Genomic DNA validation

Table 14.9 shows that the *cox* qPCR assay on olive, apple and lemon DNA detected the plant *cox* gene of olive DNA and both neat and diluted apple and lemon DNA (1:10). PCR results also showed that apple, olive and lemon DNA were negative for the exotic pathogens Ea, Xf and Xcc, respectively (Table 14.9). Identity of positive genomic DNA of Ea322A, Xcc 83 and Xf ICMP 15971 was also confirmed as the correct species using species-specific PCRs and qPCR.

Species identity of *Ca* Liberibacter in naturally infected plant and insect DNA to be used for probe and no-probe comparisons were confirmed using species-specific qPCR. Positive qPCR results are indicated by CT below 40 cycles. Phytoplasma DNA was validated in another AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria).

Sequencing

A total of 128 probed and non-probed libraries, plant only libraries (olive, apple and lemon DNA) and negative control libraries were prepared in duplicate and loaded onto an Illumina NovaSeq for sequencing. Sequencing results and analysis will be reported in milestone 106.

Discussion

Hybridization targets

Pan-genome analysis on representative subspecies of Xf (*fastidiosa*, *multiplex* and *pauca*) and two *Xanthomonas* species aided in identifying common regions within Xf that were unique from other genera. These genes were further filtered by comparison to sequence databases using BLASTn and unique genes within different chromosomal regions were identified for Xf diagnostic probe design. MLST regions were also selected for Xf probe design as inclusion would enable identification of ST.

Pan-genome analysis was done on Xcc and 10 other *Xanthomonas* strains with five of them being the same species (*Xanthomonas citri*). These additional genomes aided in identifying unique genes within Xcc. However, similar to Xf, not all unique genes from the pan-genome analysis were unique to Xcc when compared to the non-redundant database at NCBI and further filtering of genes ensued. Genes core to Xcc with the least identity to genes of closely related species located within different chromosomal regions were selected for probe design.

As *Ca.* Liberibacter species (CLam, CLaf, CLas and CLso) and phyotoplasma species were unculturable with limited full genome sequences available for analysis, MLSA genes of *Ca.* Liberibacter species were chosen for hybridization capture. As multiple MLSA genes were proposed and used in literature (Glynn et al., 2012; Morris et al., 2017; Teixeira et al., 2008), each MLSA gene was compared to the non-redundant database at NCBI using the BLASTn algorithm. MLSA genes with the least identity and coverage to non-*Ca.* Liberibacter species were chosen for hybridization capture. Stringent pan-genome analysis showed that the *Ca.* Liberibacter species were different enough from each other that no core genes were identified among them. This is in accordance with literature as identity among housekeeping genes of Clam, CLaf and CLas can be as low as 76% (Morris et al., 2017). Hence, each MLSA gene was represented by all analysed *Ca.* Liberibacter species.

Pan-genome analysis of phytoplasmas identified 111 core genes and among them were the MLSA genes. As multiple MLSA genes were proposed and used in literature, MLSA genes of phytoplasma strains were compared to the non-redundant database at NCBI using the BLASTn algorithm. MLSA genes with the least identity and coverage to non-phytoplasma species were chosen for hybridization capture. BLASTn results did show minor diversity for MLSA genes among the phytoplasma 16SrI species, hence, Aster yellows witches-broom phytoplasma and Chrysanthemum coronarium phytoplasma OY-V, which had the least identity across MLSA genes, were used as representative genomes to accommodate diversity within the phytoplasma 16SrI species.

The total length of targeted regions was 167,634 bp across 15 different species targeting almost 130 genes. A variable size of genes was also targeted ranging from 342 bp to 10,173 bp. Additionally, we targeted both unique regions for specific species and MLSA regions that have similarities to genes from other species. The diversity and number of genes for each target organism included in the synthesised hybridisation probes should provide redundancy in their use for pathogen detection (e.g. if one provides ambiguous results others should clarify pathogen detection) and will aid in future optimisation of regions utilized for diagnostic hybridization capture. As mentioned previously, one of the benefits of hybridization probe capture is that new probes can be continually added and probes for use in detection of some species used as cases studies in this work (e.g. phytoplasmas) are expected to be included once the technology is established.

DNA QA/QC

Results from qPCR targeting the *cox* gene (Table 14.9) showed successful amplification of extracted olive DNA and both neat and diluted DNA of apple and lemon leaves with no signs of inhibitory contaminants. The lack of Ea in apple DNA, Xcc in lemon DNA and Xf in olive DNA was necessary to ensure that it will not affect the probe and no-probe comparison tests as DNA of those pathogens were added in at very low concentrations.

All strains of pathogens used in this report were sequenced with different plant and/or insect background DNA replicating a range of host DNA. This is to confirm whether host DNA can affect the quality of hybridization capture such as off-target capture or harbouring inhibitory molecules interfering with hybridization capture.

With these types of probes, if a sample did not contain a target pathogen, no DNA would be captured for sequencing leading to no output. This could indicate that the samples were negative or that the test had failed. To get around this issue several controls are being validated to ensure that a negative test can be differentiated from a failed test. These include housekeeping probes on hosts such as plants and insects to show that probing of individual library preparations has been successful, positive controls to show the probes are working and negative controls to identify any potential contamination points throughout the pipeline.

As there is a high possibility that DNA will be extracted from asymptomatic plants within PEQ, titre of targeted bacteria would either be non-existent or very low titre. Hence, there is low probability of oversaturation of probes indicating that there is a possibility to trial with more than 8 samples. This can reduce cost of probes per sample as current costs without optimization is ~$50 for the probes not including library preparation and sequencing costs. The protocol for Twist does have limitations in total DNA, as only 1.5 µg of total DNA can be pooled per hybridization reaction. There have been report of other research laboratories increasing that amount without saturating the hybridization probes, hence it is a possible avenue to trial.

References

Ait Bahadou, S., Ouijja, A., Karfach, A., Tahiri, A., & Lahlali, R. (2018). New potential bacterial antagonists for the biocontrol of fire blight disease (Erwinia amylovora) in Morocco. *Microbial Pathogenesis, 117*, 7-15. doi:https://doi.org/10.1016/j.micpath.2018.02.011

Al-Subhi, A. M., Hogenhout, S., Al-Yahyai, R. A., Al-Ghaithi, A. G., & Al-Sadi, A. M. (2020). Evaluating the effect of tuf and secA gene sequence length for discrimination of phytoplasmas. *Canadian Journal of Plant Pathology*, 1-7. doi:10.1080/07060661.2020.1792995

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Ávila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., . . . Gilbert, M. T. P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports, 1*(1), 74. doi:10.1038/srep00074

Baldi, P., & La Porta, N. (2017). Xylella fastidiosa: Host Range and Advance in Molecular Identification Techniques. *Frontiers in Plant Science, 8*(944). doi:10.3389/fpls.2017.00944

Barta, J. L., Monroe, C., Teisberg, J. E., Winters, M., Flanigan, K., & Kemp, B. M. (2014). One of the key characteristics of ancient DNA, low copy number, may be a product of its extraction. *Journal of Archaeological Science, 46*, 281-289. doi:https://doi.org/10.1016/j.jas.2014.03.030

Behlau, F. (2020). An overview of citrus canker in Brazil. *Tropical Plant Pathology*. doi:10.1007/s40858-020-00377-2

Bergsma-Vlami, M., van de Bilt, J. L. J., Tjou-Tam-Sin, N. N. A., Helderman, C. M., Gorkink-Smits, P. P. M. A., Landman, N. M., . . . Westenberg, M. (2017). Assessment of the genetic diversity of Xylella fastidiosa in imported ornamental Coffea arabica plants. *Plant Pathology, 66*(7), 1065-1074. doi:https://doi.org/10.1111/ppa.12696

Bonn, W. G., & van der Zwet, T. (2000). Distribution and economic importance of fire blight. *Fire blight: the disease and its causative agent, Erwinia amylovora, 37*, 53.

Bragard, C., Dehnen-Schmutz, K., Di Serio, F., Gonthier, P., Jacques, M. A., Miret, J. A. J., . . . Milonas, P. (2019). Update of the Scientific Opinion on the risks to plant health posed by Xylella fastidiosa in the EU territory. *Efsa Journal, 17*(5).

Bustos-Korts, D., Dawson, I. K., Russell, J., Tondelli, A., Guerra, D., Ferrandi, C., . . . van Eeuwijk, F. A. (2019). Exome sequences and multi-environment field trials elucidate the genetic basis of adaptation in barley. *The Plant Journal, 99*(6), 1172-1191. doi:https://doi.org/10.1111/tpj.14414

Cabrefiga, J., & Montesinos, E. (2005). Analysis of Aggressiveness of Erwinia amylovora Using Disease-Dose and Time Relationships. *Phytopathology®, 95*(12), 1430-1437. doi:10.1094/PHYTO-95-1430

Cariou, M., Ribière, C., Morlière, S., Gauthier, J.-P., Simon, J.-C., Peyret, P., & Charlat, S. (2018). Comparing 16S rDNA amplicon sequencing and hybridization capture for pea aphid microbiota diversity analysis. *BMC research notes, 11*(1), 461-461. doi:10.1186/s13104-018-3559-3

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., . . . Bustamante, C. D. (2013). Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American journal of human genetics, 93*(5), 852-864. doi:10.1016/j.ajhg.2013.10.002

Coletta-Filho, H. D., Takita, M. A., de Souza, A. A., Neto, J. R., Destéfano, S. A. L., Hartung, J. S., & Machado, M. A. (2006). Primers based on the rpf gene region provide improved detection of Xanthomonas axonopodis pv. citri in naturally and artificially infected citrus plants. *Journal of Applied Microbiology, 100*(2), 279-285. doi:https://doi.org/10.1111/j.1365-2672.2005.02787.x

Cubero, J., & Graham, J. H. (2002). Genetic Relationship among Worldwide Strains of *Xanthomonas* Causing Canker in Citrus Species and Design of New Primers for Their Identification by PCR. *Applied and Environmental Microbiology, 68*(3), 1257. doi:10.1128/AEM.68.3.1257-1264.2002

Dardouri, S., Chehimi, S., Murillo Martínez, J., & Hajlaoui, M. R. (2017). Molecular characterization of Tunisian strains of Erwinia amylovora. *Journal of Plant Pathology, vol. 99, n. 2 (2017), p. 331-337*.

Donaldson, M. E., Rico, Y., Hueffer, K., Rando, H. M., Kukekova, A. V., & Kyle, C. J. (2017). Development of a genotype-by-sequencing immunogenetic assay as exemplified by screening for variation in red fox with and without endemic rabies exposure. *Ecology and evolution, 8*(1), 572-583. doi:10.1002/ece3.3583

Dreo, T., Pirc, M., Ramšak, Ž., Pavšič, J., Milavec, M., Žel, J., & Gruden, K. (2014). Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: a case study of fire blight and potato brown rot. *Analytical and Bioanalytical Chemistry, 406*(26), 6513-6528. doi:10.1007/s00216-014-8084-1

Fernández, F. D., Galdeano, E., & Conci, L. R. (2020). Phytoplasmas diversity and identification of new aster yellows subgroup (16SrI) associated with weed species in Argentina. *International Journal of Systematic and Evolutionary Microbiology, 70*(1), 35-43. doi:https://doi.org/10.1099/ijsem.0.003704

Fonseca, N. P., Patané, J. S. L., Varani, A. M., Felestrino, É. B., Caneschi, W. L., Sanchez, A. B., . . . Setubal, J. C. (2019). Analyses of Seven New Genomes of Xanthomonas citri pv. aurantifolii Strains, Causative Agents of Citrus Canker B and C, Show a Reduced Repertoire of Pathogenicity-Related Genes. *Frontiers in Microbiology, 10*(2361). doi:10.3389/fmicb.2019.02361

Gasc, C., & Peyret, P. (2018a). Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome, 6*(1), 61-61. doi:10.1186/s40168-018-0442-3

Gasc, C., & Peyret, P. (2018b). Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome, 6*(1), 61. doi:10.1186/s40168-018-0442-3

Glynn, J. M., Islam, M. S., Bai, Y., Lan, S., Wen, A., Gudmestad, N. C., . . . Lin, H. (2012). MULTILOCUS SEQUENCE TYPING OF 'CANDIDATUS LIBERIBACTER SOLANACEARUM' ISOLATES FROM NORTH AMERICA AND NEW ZEALAND. *Journal of Plant Pathology, 94*(1), 223-228. Retrieved from http://www.jstor.org.ez.library.latrobe.edu.au/stable/45156028

Golmohammadi, M., Cubero, J., Peñalver, J., Quesada, J. M., López, M. M., & Llop, P. (2007). Diagnosis of Xanthomonas axonopodis pv. citri, causal agent of citrus canker, in commercial fruits by isolation and PCR-based methods. *Journal of Applied Microbiology, 103*(6), 2309-2315. doi:https://doi.org/10.1111/j.1365-2672.2007.03484.x

Guitor, A. K., Raphenya, A. R., Klunk, J., Kuch, M., Alcock, B., Surette, M. G., . . . Wright, G. D. (2019). Capturing the Resistome: a Targeted Capture Method To Reveal Antibiotic Resistance Determinants in Metagenomes. *Antimicrobial Agents and Chemotherapy, 64*(1), e01324-01319. doi:10.1128/AAC.01324-19

Hajri, A., Cousseau-Suhard, P., Gentit, P., & Loiseau, M. (2019). New Insights into the Genetic Diversity of the Bacterial Plant Pathogen 'Candidatus Liberibacter solanacearum' as Revealed by a New Multilocus Sequence Analysis Scheme. *bioRxiv*, 623405. doi:10.1101/623405

Karamitros, T., & Magiorkinis, G. (2015). A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic acids research, 43*(22), e152-e152. doi:10.1093/nar/gkv773

Koupaei, F. Y., Taghavi, S. M., Shiotani, H., & Tsuyumu, S. (2014). A PCR-based assay for differentiating A- and A*-type strains of Xanthomonas citri subsp. citri, the causal agent of Asiatic citrus canker. *Journal of General Plant Pathology, 80*(1), 85-89. doi:10.1007/s10327-013-0494-y

Lee, I.-M., Bottner-Parker, K. D., Zhao, Y., Davis, R. E., & Harrison, N. A. (2010). Phylogenetic analysis and delineation of phytoplasmas based on secY gene sequences. *International Journal of Systematic and Evolutionary Microbiology, 60*(12), 2887-2897. doi:https://doi.org/10.1099/ijs.0.019695-0

Lee, I.-M., Gundersen-Rindal, D. E., Davis, R. E., Bottner, K. D., Marcone, C., & Seemüller, E. (2004). 'Candidatus Phytoplasma asteris', a novel phytoplasma taxon associated with aster yellows and related diseases. *International Journal of Systematic and Evolutionary Microbiology, 54*(4), 1037-1048. doi:https://doi.org/10.1099/ijs.0.02843-0

Li, W., Abad, J. A., French-Monar, R. D., Rascoe, J., Wen, A., Gudmestad, N. C., . . . Levy, L. (2009). Multiplex real-time PCR for detection, identification and quantification of 'Candidatus Liberibacter solanacearum' in potato plants with zebra chip. *Journal of Microbiological Methods, 78*(1), 59-65. doi:https://doi.org/10.1016/j.mimet.2009.04.009

Li, W., Hartung, J. S., & Levy, L. (2006). Quantitative real-time PCR for detection and identification of Candidatus Liberibacter species associated with citrus huanglongbing. *Journal of Microbiological Methods, 66*(1), 104-115. doi:https://doi.org/10.1016/j.mimet.2005.10.018

Maejima, K., Oshima, K., & Namba, S. (2014). Exploring the phytoplasmas, plant pathogenic bacteria. *Journal of General Plant Pathology, 80*(3), 210-221. doi:10.1007/s10327-014-0512-8

Mann, R. (2013). (Unpublished doctoral dissertation thesis). *La Trobe Univeristy, Melbourne*.

Mann, R. A., Smits, T. H. M., Bühlmann, A., Blom, J., Goesmann, A., Frey, J. E., . . . Rodoni, B. (2013). Comparative genomics of 12 strains of Erwinia amylovora identifies a pan-genome with a large conserved core. *PLOS ONE, 8*(2), e55644-e55644. doi:10.1371/journal.pone.0055644

Martoni, F., Nogarotto, E., Piper, A. M., Mann, R., Valenzuela, I., Eow, L., . . . Blacket, M. J. (2021). Propylene Glycol and Non-Destructive DNA Extractions Enable Preservation and Isolation of Insect and Hosted Bacterial DNA. *Agriculture, 11*(1), 77. Retrieved from https://www.mdpi.com/2077-0472/11/1/77

Mascher, M., Richmond, T. A., Gerhardt, D. J., Himmelbach, A., Clissold, L., Sampath, D., . . . Stein, N. (2013). Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *The Plant journal : for cell and molecular biology, 76*(3), 494-505. doi:10.1111/tpj.12294

McNally, R. R., Toth, I. K., Cock, P. J. A., Pritchard, L., Hedley, P. E., Morris, J. A., . . . Sundin, G. W. (2012). Genetic characterization of the HrpL regulon of the fire blight pathogen Erwinia amylovora reveals novel virulence factors. *Molecular Plant Pathology, 13*(2), 160-173. doi:https://doi.org/10.1111/j.1364-3703.2011.00738.x

Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., . . . Viral Hemorrhagic Fever, C. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology, 37*(2), 160-168. doi:10.1038/s41587-018-0006-x

Morris, J., Shiller, J., Mann, R., Smith, G., Yen, A., & Rodoni, B. (2017). Novel 'Candidatus Liberibacter' species identified in the Australian eggplant psyllid, Acizzia solanicola. *Microbial Biotechnology, 10*(4), 833-844. doi:https://doi.org/10.1111/1751-7915.12707

Muirhead, K., Pérez-López, E., Bahder, B. W., Hill, J. E., & Dumonceaux, T. (2019). The CpnClassiPhyR Is a Resource for cpn60 Universal Target-Based Classification of Phytoplasmas. *Plant Disease, 103*(10), 2494-2497. doi:10.1094/PDIS-03-19-0454-A

Muraya, M. M., Schmutzer, T., Ulpinnis, C., Scholz, U., & Altmann, T. (2015). Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition. *PLOS ONE, 10*(7), e0132120. doi:10.1371/journal.pone.0132120

Oono, Y., Ichida, H., Morita, R., Nozawa, S., Satoh, K., Shimizu, A., . . . Hase, Y. (2020). Genome sequencing of ion-beam-induced mutants facilitates detection of candidate genes responsible for phenotypes of mutants in rice. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 821*, 111691. doi:https://doi.org/10.1016/j.mrfmmm.2020.111691

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., . . . Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics, 31*(22), 3691-3693.

Pankin, A., Campoli, C., Dong, X., Kilian, B., Sharma, R., Himmelbach, A., . . . von Korff, M. (2014). Mapping-by-Sequencing Identifies HvPHYTOCHROME C as a Candidate Gene for the early maturity 5 Locus Modulating the Circadian Clock and Photoperiodic Flowering in Barley. *Genetics, 198*(1), 383. doi:10.1534/genetics.114.165613

Park, D., Lee, Y.-G., Kim, J.-S., Cha, J.-S., & Oh, C.-S. (2017). Current status of fire blight caused by Erwinia amylovora and action for its management in Korea. *Journal of Plant Pathology, 99*, 59-63.

Pereira, T. B. C., & Bedendo, I. P. (2017). A 'Candidatus phytoplama asteris' (I6SrI group) phytoplasma associated with delayed maturity in soybean plants in Brazil. *Australasian Plant Disease Notes, 12*(1), 50. doi:10.1007/s13314-017-0273-2

Pierro, R., Panattoni, A., Passera, A., Materazzi, A., Luvisi, A., Loni, A., . . . Quaglino, F. (2020). Proposal of A New Bois Noir Epidemiological Pattern Related to 'Candidatus Phytoplasma Solani' Strains Characterized by A Possible Moderate Virulence in Tuscany. *Pathogens (Basel, Switzerland), 9*(4), 268. doi:http://dx.doi.org/10.3390/pathogens9040268

Pierro, R., Passera, A., Panattoni, A., Casati, P., Luvisi, A., Rizzo, D., . . . Materazzi, A. (2017). Molecular Typing of Bois Noir Phytoplasma Strains in the Chianti Classico Area (Tuscany, Central Italy) and Their Association with Symptom Severity in Vitis vinifera 'Sangiovese'. *Phytopathology®, 108*(3), 362-373. doi:10.1094/PHYTO-06-17-0215-R

Piqué, N., Miñana-Galbis, D., Merino, S., & Tomás, J. M. (2015). Virulence factors of Erwinia amylovora: a review. *International journal of molecular sciences, 16*(6), 12836-12854.

Quaglino, F., Kube, M., Jawhari, M., Abou-Jawdah, Y., Siewert, C., Choueiri, E., . . . Bianco, P. A. (2015). 'Candidatus Phytoplasma phoenicium' associated with almond witches'-broom disease: from draft genome to genetic diversity among strain populations. *BMC Microbiology, 15*(1), 148. doi:10.1186/s12866-015-0487-4

Rassoulian Barrett, S., Hoffman, N. G., Rosenthal, C., Bryan, A., Marshall, D. A., Lieberman, J., . . . Salipante, S. J. (2020). Sensitive Identification of Bacterial DNA in Clinical Specimens by Broad-Range 16S rRNA Gene Enrichment. *Journal of Clinical Microbiology, 58*(12), e01605-01620. doi:10.1128/JCM.01605-20

Rigano, L. A., Marano, M. R., Castagnaro, A. P., Do Amaral, A. M., & Vojnov, A. A. (2010). Rapid and sensitive detection of Citrus Bacterial Canker by loop-mediated isothermal amplification combined with simple visual evaluation methods. *BMC Microbiology, 10*(1), 176. doi:10.1186/1471-2180-10-176

Roberts, R., & Pietersen, G. (2017). A novel subspecies of 'Candidatus Liberibacter africanus' found on native Teclea gerrardii (Family: Rutaceae) from South Africa. *Antonie van Leeuwenhoek, 110*(3), 437-444. doi:10.1007/s10482-016-0799-x

Rogozin, I. B., Makarova, K. S., Natale, D. A., Spiridonov, A. N., Tatusov, R. L., Wolf, Y. I., . . . Koonin, E. V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic acids research, 30*(19), 4264-4271. doi:10.1093/nar/gkf549

Ruiz-Padilla, A., Redondo, C., Asensio, A., Garita-Cambronero, J., Martínez, C., Pérez-Padilla, V., . . . Cubero, J. (2020). Assessment of Multilocus Sequence Analysis (MLSA) for Identification of Candidatus Liberibacter Solanacearum from Different Host Plants in Spain. *Microorganisms, 8*(9). doi:10.3390/microorganisms8091446

Samorodnitsky, E., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Lyon, E., . . . Roychowdhury, S. (2015). Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human mutation, 36*(9), 903-914. doi:10.1002/humu.22825

Santander, R. D., Meredith, C. L., & Aćimović, S. G. (2019). Development of a viability digital PCR protocol for the selective detection and quantification of live Erwinia amylovora cells in cankers. *Scientific Reports, 9*(1), 11530. doi:10.1038/s41598-019-47976-x

Scally, M., Schuenzel, E. L., Stouthamer, R., & Nunney, L. (2005). Multilocus sequence type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. *Applied and Environmental Microbiology, 71*(12), 8491. doi:10.1128/AEM.71.12.8491-8499.2005

Scortichini, M., & Cesari, G. (2019). An Evaluation of Monitoring Surveys of the Quarantine Bacterium Xylella Fastidiosa Performed in Containment and Buffer Areas of Apulia, Southern Italy. *Applied Biosafety, 24*(2), 96-99. doi:10.1177/1535676019845738

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Shao, J., Jomantiene, R., Dally, E. L., Zhao, Y., Lee, I. M., Nuss, D. L., & Davis, R. E. (2006). PHYLOGENY AND CHARACTERIZATION OF PHYTOPLASMAL NusA AND USE OF THE nusA GENE IN DETECTION OF GROUP 16SrI STRAINS. *Journal of Plant Pathology, 88*(2), 193-201. Retrieved from http://www.jstor.org/stable/41998311

Shin, D.-S., Heo, G.-I., Son, S.-H., Oh, C.-S., Lee, Y.-K., & Cha, J.-S. (2018). Development of an Improved Loop-Mediated Isothermal Amplification Assay for On-Site Diagnosis of Fire Blight in Apple and Pear. *The plant pathology journal, 34*(3), 191-198. doi:10.5423/PPJ.FT.03.2018.0055

Sholberg, P. L., Bedford, K. E., Haag, P., & Randall, P. (2001). Survey of Erwinia amylovora isolates from British Columbia for resistance to bactericides and virulence on apple. *Canadian Journal of Plant Pathology, 23*(1), 60-67. doi:10.1080/07060660109506910

Sichani, F. V., Bahar, M., & Zirak, L. (2014). Characterization of phytoplasmas related to aster yellows group infecting annual plants in Iran, based on the studies of 16S rRNA and rp genes. *Journal of Plant Protection Research, 54*(1).

Taloh, A., Raju, D. V. S., Banyal, N., Kumar, G., Panda, P., Manimekalai, R., . . . Rao, G. P. (2020). Genetic diversity of phytoplasma strains infecting chrysanthemum varieties in India and their possible natural reservoirs. *3 Biotech, 10*(9), 411. doi:10.1007/s13205-020-02407-x

Teixeira, D. C., Eveillard, S., Sirand-Pugnet, P., Wulff, A., Saillard, C., Ayres, A. J., & Bové, J. M. (2008). The tufB–secE–nusG–rplKAJL–rpoB gene cluster of the liberibacters: sequence comparisons, phylogeny and speciation. *International Journal of Systematic and Evolutionary Microbiology, 58*(6), 1414-1421. doi:https://doi.org/10.1099/ijs.0.65641-0

Teresani, G. R., Bertolini, E., Alfaro-Fernández, A., Martínez, C., Tanaka, F. A. O., Kitajima, E. W., . . . Font, M. I. (2014). Association of 'Candidatus Liberibacter solanacearum' with a Vegetative Disorder of Celery in Spain and Development of a Real-Time PCR Method for Its Detection. *Phytopathology®, 104*(8), 804-811. doi:10.1094/PHYTO-07-13-0182-R

Urban, M., Irvine, A., Cuzick, A., & Hammond-Kosack, K. (2015). Using the pathogen-host interactions database (PHI-base) to investigate plant pathogen genomes and genes implicated in virulence. *Frontiers in Plant Science, 6*(605). doi:10.3389/fpls.2015.00605

Wang, D., Korban, S. S., Pusey, P. L., & Zhao, Y. (2012a). AmyR Is a Novel Negative Regulator of Amylovoran Production in Erwinia amylovora. *PLOS ONE, 7*(9), e45038. doi:10.1371/journal.pone.0045038

Wang, D., Qi, M., Calla, B., Korban, S. S., Clough, S. J., Cock, P. J., . . . Zhao, Y. (2012b). Genome-wide identification of genes regulated by the Rcs phosphorelay system in Erwinia amylovora. *Molecular plant-microbe interactions, 25*(1), 6-17.

Wang, Z., Liu, X., Yang, B.-Z., & Gelernter, J. (2013). The Role and Challenges of Exome Sequencing in Studies of Human Diseases. *Frontiers in Genetics, 4*(160). doi:10.3389/fgene.2013.00160

Webster, J., Bogema, D., & Chapman, T. A. (2020). Comparative Genomics of Xanthomonas citri pv. citri A* Pathotype Reveals Three Distinct Clades with Varying Plasmid Distribution. *Microorganisms, 8*(12), 1947.

Weintraub, P. G. (2007). Insect vectors of phytoplasmas and their control-an update. *Bulletin of Insectology, 60*(2), 169.

Weintraub, P. G., & Beanland, L. (2006). Insect vectors of phytoplasmas. *Annu. Rev. Entomol., 51*, 91-111.

Weller, S. A., Elphinstone, J. G., Smith, N. C., Boonham, N., & Stead, D. E. (2000). Detection of Ralstonia solanacearum strains with a quantitative, multiplex, real-time, fluorogenic PCR (TaqMan) assay. *Applied and Environmental Microbiology, 66*(7), 2853-2858. doi:10.1128/aem.66.7.2853-2858.2000

Whiston, R., Finlay, E. K., McCabe, M. S., Cormican, P., Flynn, P., Cromie, A., . . . Meade, K. G. (2017). A dual targeted β-defensin and exome sequencing approach to identify, validate and functionally characterise genes associated with bull fertility. *Scientific Reports, 7*(1), 12287. doi:10.1038/s41598-017-12498-x

Yu, S. M., Ramkumar, G., & Lee, Y. H. (2017). DETECTION OF XANTHOMONAS CITRI subsp. CITRI A*, A ^W AND X. FUSCANS subsp. AURANTIFOLII B, C USING PCR AND REAL-TIME PCR. *Journal of Plant Pathology, 99*(2), 461-467. Retrieved from http://www.jstor.org.ez.library.latrobe.edu.au/stable/44686792

Yuan, X., Hulin, M. T., & Sundin, G. W. (2020). Effectors, chaperones, and harpins of the Type III secretion system in the fire blight pathogen Erwinia amylovora: a review. *Journal of Plant Pathology*. doi:10.1007/s42161-020-00623-1

Yuan, X., Morano, L., Bromley, R., Spring-Pearson, S., Stouthamer, R., & Nunney, L. (2010). Multilocus Sequence Typing of Xylella fastidiosa Causing Pierce's Disease and Oleander Leaf Scorch in the United States. *Phytopathology®, 100*(6), 601-611. doi:10.1094/PHYTO-100-6-0601

Zhao, Y., Wei, W., Lee, I.-M., Shao, J., Suo, X., & Davis, R. E. (2009). Construction of an interactive online phytoplasma classification tool, iPhyClassifier, and its application in analysis of the peach X-disease phytoplasma group (16SrIII). *International Journal of Systematic and Evolutionary Microbiology, 59*(Pt 10), 2582-2593. doi:10.1099/ijs.0.010249-0

Zhao, Y., Xia, Q., Yin, Y., & Wang, Z. (2016). Comparison of Droplet Digital PCR and Quantitative PCR Assays for Quantitative Detection of Xanthomonas citri Subsp. citri. *PLOS ONE, 11*(7), e0159004. doi:10.1371/journal.pone.0159004

Appendix 15 – Manuscript on side-by-side comparison of PEQ and NGS methods

Type of the Paper (Article)

Side-by-side comparison of post-entry quarantine and high throughput methods for the diagnosis of plant viruses and viroids

Marie-Emilie Gauthier[1], Ruvini Lelwala[1,2], Candace Elliott[2], Joanne Mackie[2], Craig Windell[1], Douglas Beattie[1], Sonia Fiorito[2], Adrian Dinsdale[3], Mark Whattam[3], Julie Pattemore[2] and Roberto A. Barrero[1*]

[1] *eResearch Office, Queensland University of Technology, Brisbane QLD 4001, Australia*
[2] *Science and Surveillance Group, Post Entry Quarantine, Department of Agriculture, Water and the Environment, Mickleham VIC 3064, Australia*
[3] *Plant Innovation Centre @ Post Entry Quarantine, Department of Agriculture, Water and the Environment, Mickleham VIC 3064, Australia*
* Correspondence: roberto.barrero@qut.edu.au

Abstract: Rapid and safe access to new plant genetic stocks is crucial for plant primary industries to remain profitable, sustainable, and internationally competitive. Imported plant species may spend several years in Post Entry Quarantine (PEQ) facilities undergoing pathogen testing, which can impact the ability of plant industries to quickly adapt to new global market opportunities and access new varieties. Advances in high throughput sequencing (HTS) technologies provide new opportunities for a broad range of fields including phytosanitary diagnosis. In this study, we compare the performance of two HTS methods (RNA-Seq and sRNA-Seq) with that of existing PEQ molecular assays in detecting and identifying viruses and viroids from various plant commodities. To analyse the data, we tested several bioinformatics tools which rely on different approaches: direct-read, *de novo* and reference-guided assembly. We included a new portable, scalable, and reproducible end-to-end workflow of Virus Surveillance and Diagnosis (VSD) implemented in a nextflow pipeline framework. We bring awareness to the need to routinely evaluate for cross contamination and select methods that mitigate index hopping when analysing HTS data. Overall, our results suggest that sRNA analysed using nextflow-VSD provides opportunities to accelerate quarantine testing at PEQ by detecting regulated exotic viruses from imported plants in a single assay.

Keywords: *high throughput sequencing; plant virus and viroid detection; phytosanitary diagnostic assay, plant siRNA, post-entry quarantine facility*

1. Introduction

Pathogenic viruses and viroids infecting plants can result in major economic and ecological losses. Preventing the introduction of such pests is therefore a fundamental concern to national biosecurity and border control agencies [1-3]. Virus and viroids are conventionally considered the most difficult plant pathogens to detect. Traditional diagnostic tools such as serological or molecular assays (e.g., PCR, ELISA) are specific to the pathogens being screened for but require *a priori* knowledge of their identity to test for their presence. While non-specific tests such as inoculation of test plants, electron microscopy, woody indexing, or visual inspection do not require this prior knowledge, they generally provide a lower level of detection resolution which can be limited to pest features or higher-level taxonomy (e.g., genus). The traditional tests are also unable to detect multiple pathogens in one assay. Additionally, plants being tested often require extended periods of time in post-entry quarantine (PEQ) which significantly adds costs to importers.

The use of high-throughput sequencing (HTS) technologies for detecting plant viruses started in 2009 [4-6]. Since then, these methods have become an integral part of research and diagnostics on plant viruses and viroids, as they provide unbiased high-resolution detection of highly diverse viruses without needing prior characterisation [7-10]. HTS also enables the identification of viruses that can cause non-specific plant disease symptoms and are thus problematic for traditional diagnostics tools [11]. With HTS increased affordability and portability, international groups and agencies are looking into its viability as a method for the rapid screening of viral pathogens for plants going through importation certification schemes in quarantine facilities, aiming to improve the rate at which propagation material is made available to producers. Nevertheless, the validation of HTS-based virus detection schemas in such diagnostics settings is still in its infancy. The number of different approaches in nucleic acid preparation, sequencing, and sequence data analysis highlight the challenges that the development of standard operating procedures may present, with several studies reporting result discrepancies between different HTS or analytical methods [12-14]. And challenges specific to HTS (e.g. adapter dimer formation, barcode misassignment, chimeric contig assembly) need to be carefully assessed [15].

While some studies have evaluated the sensitivity of specific HTS approaches compared to particular traditional tests in a given commodity [16, 17], HTS technologies still need to pass rigorous tests and systematically prove equal or better than the

current standards in sensitivity, specificity and reproducibility before being approved for routine use [8, 18]. For this reason, they are not yet currently considered to be stand-alone replacements for traditional methods, with academic and governmental research both acknowledging there is still a need for extant methods to independently confirm the biological significance of detected viruses via HTS. However international efforts are being made to critically catalogue methods available [15, 19, 20] as well as provide researchers and diagnostics labs with open-source reference datasets to compare and validate their detection pipelines [21, 22]. There is also an increase in comparative studies reporting the performance of different approaches [13, 14, 22-24]. Together, this will help improve and accelerate harmonisation, comparability, and reproducibility of HTS methods.

Different types of nucleic acids can be used as inputs for the detection of plant viruses and viroids via HTS, with total RNA and sRNA being the two most widely used approaches for HTS analysis of plant tissues infected with viruses [15]. In comparison to the total RNA approach, the sRNA method specifically targets small-interfering RNA (siRNA) for sequencing, which has the advantage of biasing the method of detection toward viruses and viroids that have driven an immune response in the plant. The plant response drives the cleavage by Dicer proteins of double-stranded forms of viral RNA into sRNAs that are 21 nucleotide (nt), 22 nt and 24 nt long [20]. HTS techniques such as RNA sequencing (RNA-Seq) and sRNA sequencing (sRNA-Seq) yield short sequences of viral RNA or sRNA entity extracted from plants. These can be detected and identified using bioinformatics tools that can either perform a direct search on the sequenced reads or assemble these first to generate longer sequences prior to annotation.

This study performs a comparison between two HTS methods (RNA-Seq and sRNA-Seq) analysed using different bioinformatics tools, and PEQ traditional assays for the detection of regulated viruses and viroids to assess the feasibility of adopting HTS testing as a diagnostics tool in quarantine facilities.

2. Materials and Methods

Plant material preparation

To fulfill a side-by-side comparison between HTS methods and PEQ assays, 15 plants were selected that were diagnosed by PEQ as carriers one or more viral or viroid infections. These plants comprised a diverse range of commodities, namely citrus (6), prunus (1), grass (1), sweet potato (1), raspberry (2), strawberry (3) and iris (1) (**Table 1**). Sampling was based on current PEQ protocols and carried out in September 2019. Collection was completed when plants were actively growing. Up to three leaves of different age were selected from different parts of the plant and combined in an aggregate for extraction. If

symptoms were observed, samples were collected from diseased parts of the plant. Tissue was collected either with a single-use sterile blade or by hand, using single-use gloves or equivalent. Samples were immediately placed in an insulated cool container until they were further processed in the laboratory. Approx. 50 - 100 mg of leaf/midrib was then subsampled for extraction.

Table 1. This is a table. Tables should be placed in the main text near to the first time they are cited.

Sample ID	Commodity	Positive detection in PEQ
MT001	*Citrus Troyer x Frost-Lisbon*	CEVd
MT002	*Prunus persica Nectarine*	PNRSV
MT003	*Citrus aurantiifolia (Christm.) Swingle*	CTV, CVEV, CVd-III, HSVd
MT004	*Citrus medica L. (Etrog)*	CEVd, CTV, HSVd
MT005	*Rubus idaeus*	RBDV, endoRYNV
MT006	*Rubus idaeus*	endoRYNV
MT007	*Citrus sp.*	CVd-III, HSVd
MT008	*Citrus Sweet Orange*	CVd-III, HSVd
MT010	*Miscanthus sinensis 'Morning light'*	Novel potyvirus
MT011	*Citrus medica L.*	CTV, CVd-VI, HSVd
MT012	*Iris Crimson collosus 'Deep Crimson'*	ISMV, TRSV
MT013	*Fragaria Alpine (Fragaria vesca)*	SMoV
MT014	*Fragaria UC4 (Fragaria vesca hybrid)*	SMoV
MT015	*Fragaria sp.*	SMoV
MT016	*Ipomoea batatas GRF0334*	SPFMV

CEV= Citrus exocortis viroid, CTV = Citrus tristeza virus, CVEV = Citrus vein enation virus, CVd-III = Citrus viroid III/Citrus dwarfing viroid, CVd-VI = Citrus viroid VI, HSVd = Citrus viroid II/ Hop stunt viroid, ISMV = Iris severe mosaic virus, PNRSV = Prunus necrotic ringspot virus, RBDV = Raspberry bushy dwarf virus, endoRYNV = Endogenous Rubus yellow net virus, SMoV = Strawberry mottle virus, TRSV = Tobacco ringspot virus.

Total RNA extraction was performed using an automated approach on the Maxwell® Rapid Sample Concentrator Instrument with a SimplyRNA Tissue kit (AS1340, Promega). 100 mg of sample was placed with a 5 mm stainless steel bead in a 2 ml round bottom tube and stored at -80°C for 30 min. The tube was then placed into a TissueLyser (Qiagen) for 2 min at 50 Hz. 600 µL of homogenisation solution was added to the tube, followed by 300 µl of lysis solution. The tubes were subsequently vortexed and spun for 5 min at 20,000 g. The quality of the RNA extracted was assessed by QuantiFluor using SimplyRNA Tissue protocol (AS 1340) (**Supplementary Table 1**). The same RNA

extraction from each plant was used for the preparation of both small RNA (sRNA)-Seq and RNA-Seq libraries by two separate sequencing providers (SPs) which we will refer to as SP1 and SP2 onwards. RNA aliquots were transported to the sequencing providers at -80°C. All fifteen samples were sent to SP1 but only ten samples were provided to SP2 (**Supplementary Table 1**).

Conventional assays

The Rubus yellow net viruses detected in sample MT005 and MT006 are predicted to be endogenous [25]. The virus detected in sample MT010 is a novel potyvirus which shows highest homology to both Sorghum mosaic virus and Maize dwarf mosaic virus (**Table 1**).

Nucleic acid extraction, library construction and sequencing

RNA integrity and quantification were assessed with a Bioanalyser 2100 (Agilent) by both sequencing providers. For the RNA-Seq library preparation, both sequencing providers processed the samples using TruSeq Stranded Total RNA. Ribosomal RNA (rRNA) was removed with the Ribo-Zero Plant kit. Dual index adapters were used for the SP2 samples. Different methods were used by the two sequencing providers for the sRNA-Seq library preparation. Samples were processed by SP2 with an NEBNext Multiplex Small RNA Library kit and the final size selection was performed on the Pippin prep system (using 3% cassettes with the P marker and size selection parameters of 125-175 bp). SP1 used a proprietary sRNA library preparation protocol. This included two PAGE electrophoresis steps: a first one to separate RNA which was 18-30 bp long from total RNA before library preparation and a second one to separate final PCR products in the range of 100-120 bp. Unique molecular identifiers (UMIs) were also incorporated after the 3′ ligation step.

The libraries obtained were sequenced on a SEQ-500 (using a rolling cycle sequencing approach) at the SP1 and on an llumina NextSeq 500 for the sRNA Seq libraries and on a NovaSeq for the RNASeq libraries at the SP2. Paired-end 2x150 bp reads were generated for all the RNA-Seq samples. For sRNA-Seq samples, 50 bp and 75 bp single-end reads were generated at the SP1 and the SP2 respectively. All raw fastq files have been deposited at SRA submission SUB10125984, project number PRJNA752836 (reviewer's link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA752836?reviewer=bct21ptue1tfv4r51bt1fe7rj0).

Bioinformatics analyses

Post-sequencing, reads generated by RNA-Seq were filtered for low quality regions and sequencing adapters using Trim Galore [26]. The adapters were Illumina TruSeq adapter (AGATCGGAAGAGC) for the SP2 samples and the specific forward (AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA) and reverse (AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGA GTTG) adapters for the SP1 samples. The sRNA reads provided

by SP2 were filtered for Illumina adapters and low quality regions using Fastp [27]. The sRNA data provided by SP1 was already filtered for low quality sequences and adapters using SOAPnuke [28]. We performed an additional filtering step for low quality regions using Fastp. FastQC was used to check the quality of reads prior to and after filtering for both RNA-Seq and sRNA-Seq samples [29]. To measure the effectiveness of ribodepletion of the samples prior to sequencing, the amount of rRNA was assessed by aligning the reads to Bowtie 1 using the Silva database.

Four bioinformatics approaches were tested on the RNA-Seq samples to detect viruses and viroids. We tested two approaches that use direct annotation on quality checked reads without any assembly: the plant virus detection pipeline (PVDP) [30] and Kodoja [31]. These workflows were run on default parameters using the pre-computed database kodojaDB_v1.0 for Kodoja and the PlantVirusesDB_0420v4_masked.fa for PVDP. We also tested two *de novo* transcriptome assemblers: the *rnaviral* mode *of SPAdes 3.15 [32] and Trinity-v2.11.0 using singularity [33]. For Trinity, the quality-trimmed* reads were filtered using a blacklist of ribosomal RNAs, representative plant chloroplasts, mitochondria, and miRNAs, and sequence artefacts before performing *de novo* assembly, following best practice recommendations for this tool (https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html). *SPAdes 3.15 was run using the rnaviral mode on default parameters, using k-mer lengths of 33 and 49.* The contigs obtained from the *de novo* assembly approaches were *collapsed* into scaffolds using CAP3 [34]. Nucleotide alignment against the NCBI nt nucleotide database was then performed on these scaffolds using megablast [35].

The presence of virus and viroids in the sRNA-Seq reads was assessed using two methods: VirusDetect 1.7 [36] and a new implementation of the Virus Surveillance and Diagnostics (VSD) tool [37] in a Nextflow framework. Before running VirusDetect, reads were first filtered for ribosomal RNA and host representative sequences (mitochondria and chloroplast) as recommended by the developer. The software was run using the cut-off depth parameter of 5 (default). VSD can be run using two different assemblers: SPAdes or Velvet. We found that for this study, the optimal k-mer range for SPAdes (v3.14) which maximised recovery of viruses and viroids was 9 to 21 increasing by increments of 2; and a k-mer size of 15 for Velvet (1.2.10). For both VirusDetect and VSD, we restricted the assembly to 21-22 nt long reads and 24 nt reads. The SPAdes contigs obtained by VSD were further collapsed and blasted against NCBI as per the RNA-Seq reads. A custom python script was applied to derive the top virus or viroid hit and use it as a reference to map the filtered 21-22 nt and 24 nt long reads with Bowtie. PCR duplicates were identified using UMICollapse [38] for the libraries which incorporated UMIs for the sRNA samples from

the SP1. From the obtained alignments, we derived the total number of aligned reads, deduplicated reads (for samples from SP1) and normalised our counts using reads per million (RPM): total number of aligned reads x 1,000,000 / total number of quality-filtered reads. Coverage and average depth were calculated using Picard tools. The scripts used in this study are available at [eResearch github repository]

We implemented a contamination flag (CF) which was applied on Kodoja, PVDP, VSD and VirusDetect. Candidate contamination events were flagged based on the reads per million (RPM) value reported for each target detected in the final virus table output. We either used reads per million values derived by the bioinformatics method (eg PVDP, VirusDetect) or derived it independently (Kodoja, VSD). The CF assumptions are that: 1) virus or viroid present in high load in a given sample are likely to be the source of contamination 2) the abundance of reads matching to this viral target in the samples where contamination occurred is in comparison significantly much lower. We first calculate the maximum RPM value recorded for each virus and viroid identified on a run. If for a given virus, the RPM value reported for a sample represented less than 0.1% (PVDP) and 1% (Kodoja, VirusDetect and VSD) of this maximum RPM value, it was then flagged as a contamination event.

3. Results

3.1 General sequencing statistics

RNA sequencing generated on average 37.8 million reads for the SP1 samples and 46.3 million reads for the SP2 samples. The number for reads recovered per sample was relatively consistent across sample for a given service provider. Quality filtering removed on average 5% and 6% of the reads for the SP1 and the SP2 samples respectively. Successful removal of rRNA of the samples prior to sequencing was variable with an average of 14 to 30% of the quality filtered reads in the SP1 samples, compared to 7% to 33% in the SP2 samples, that aligned to the SILVA database and thus were deemed residual rRNAs. There was no correlation between the commodity the sample originated from and the level of residual rRNA present *(**Figure 1, Supplementary Table 2**).*

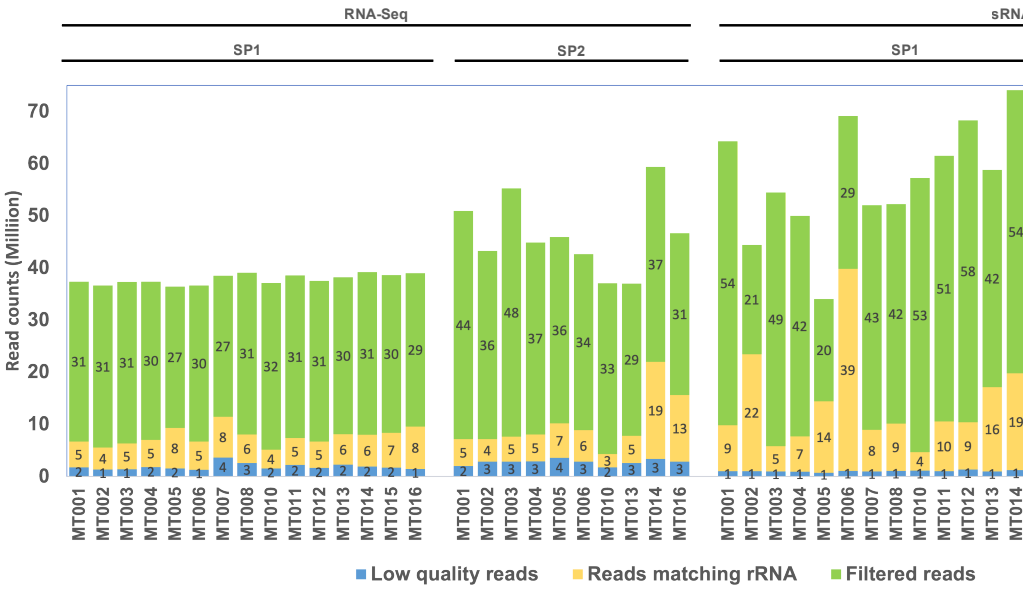


Figure 1. Total number of quality-filtered reads left-over after removing low-quality reads and reads matching ribosomal RNA.

sRNA sequencing generated on average 56.4 million reads for the SP1 samples and 27.4 million reads for the SP2 samples. The number of reads recovered per sample was quite variable for each service provider. Quality filtering resulted in the removal of 17% reads on average for the SP2 samples and 2% for the SP1 samples, which was expected since these latter were provided as quality-filtered by the sequencing provider. The content of rRNA was highly variable in the sRNA samples, with 7 to 58% of the quality filtered reads in the SP1 samples, compared to 15 to 62% in the SP2 samples, that aligned to the SILVA database. There was a correlation between the commodity and the percentage of rRNA detected in the samples, with the grass, sweet potato and citrus samples being the least affected while the strawberry, peach and raspberry were the most affected across both the SP1 and the SP2 samples (***Figure 1, Supplementary Table 2***). These rRNAs mostly originated from plants, but we also recovered matches to insects such as scale insects and aphids (data not shown).

3.2 Assembly statistics

For the RNA-Seq samples, the *SPAdes workflow recovered on average 80,591 contigs for the SP1 samples and 75,900 contigs for the SP2 samples* (post CAP3 scaffolding). *In comparison, the Trinity assemblies contained approximately 25% more contigs, with an average of 107,291 contigs for the SP1 samples and 101,692 contigs for the SP2 samples. (**Supplementary Table 3A**).* For the sRNA-Seq samples, the size of the assembly was highly dependent on the sequencing technology as well as

the method and assembler used. *The assemblies contained significantly more contigs in the samples from the SP1 than with SP2; and both Virus Detect and VSD-Velvet generated more contigs than VSD-SPAdes (**Supplementary Table 3A**). After performing homology searches, we only retained contigs that mapped to viral sequences. We recovered twice as many viral contigs in the SP1 samples compared to the SP2 samples across sequencing platforms and bioinformatics tools tested. For the RNA-Seq samples, the number of viral contigs recovered by RNA-Seq was on par between SPAdes and Trinity but for the sRNA-Seq samples, the assemblies derived with VirusDetect and the ones obtained using VSD-Velvet contained a high number of contigs than the ones obtained using VSD-SPAdes (**Figure 2, Supplementary Table 3B)**.*

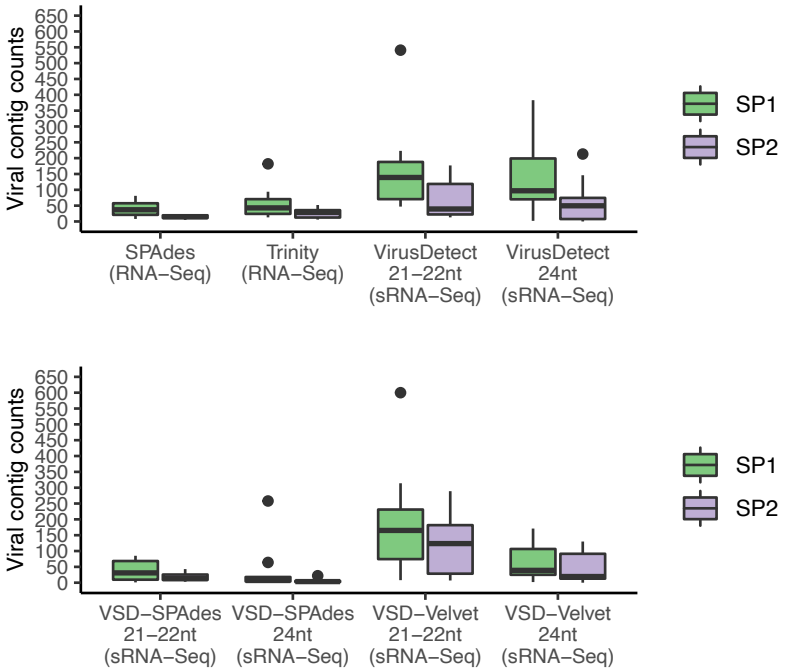


Figure 2. Total number of viral contigs recovered for each program tested that use an assembly-based approach.

3.3 Detection of virus and viroids

*Looking across both read-based and assembly-based software approaches, the final number of unique viruses and viroids recovered per sample varied significantly (**Figure 3A**). The total number of viral candidates detected in the SP1 samples was on average higher than in the SP2 samples for both sequencing technologies (**Figure 3B)**.*

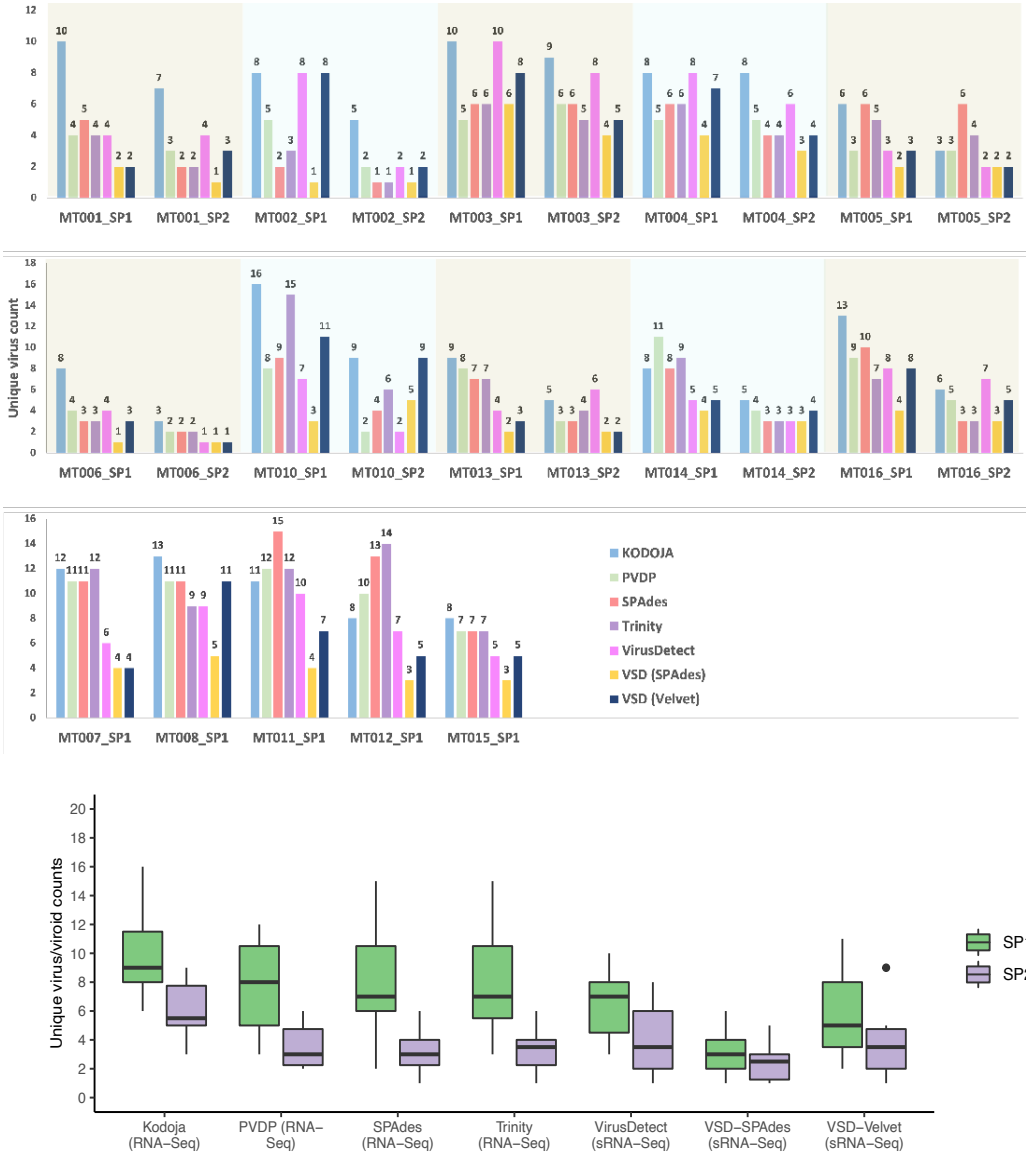


Figure 3. Side-by-side comparison of the total number of unique viruses and/or viroids detected by RNA-Seq and sRNA-Seq depending on software and sequencing provider used. **A.** Separate viral counts detected for each sample. **B.** Average number of unique viruses and/or viroids recovered across samples by each software. The counts are split between sequencing providers.

To measure the sensitivity of each method and HTS technology combination, we checked their ability to detect a total of 42 viruses and viroids across both sequencing platforms (ie 26 across 15 plant samples for the SP1 and 16 across 10 samples for the SP2). For the RNA-Seq samples, only Kodoja was able to detect all the viruses and viroids in the SP1 samples while PVDP, *SPAdes* and Trinity missed a Hop stunt viroid present in sample MT003. We recovered very few reads matching this viroid

which might explain why several softwares failed to detect it. PVDP,

Kodoja and *SPAdes* were able to detect all the viruses and viroids that were identified by PEQ in the SP2 samples while Trinity missed the Hop stunt viroid present in sample MT003. Thus, across the sequencing platforms, Trinity achieved overall the lowest sensitivity (40/42, 95.2%), followed by *SPAdes* and PVDP (41/42, 97.6%), and finally Kodoja (42/42, 100%) (**Table 2**, **Supplementary Table 4**).
For the sRNA-Seq data, VSD detected all the viruses and viroids from the data generated from both sequencing providers when using Velvet as the assembler (sensitivity of 100%, 42/42). It also recovered all targets for the SP1 when using *SPAdes* but missed three targets in one sample for the SP2 (sensitivity of 92.9%, 39/42*); this sample suffered from low sequencing depth and when additional sequencing was performed, VSD-SPAdes was able to recover the missing viral targets (data not shown).* VirusDetect detected all the viruses and viroids from the data generated from both sequencing providers and thus achieved an overall sensitivity of 100% (42/42) (**Table 2**, **Supplementary Table 4**).

Table 2. Average detection sensitivity for the different methods and technology tested in this study.

Sequencing provider	Sequencing technology	Software	Sensitivity (%)
SP1	RNA-Seq	Kodoja	100 (26/26)
		PVDP	96.1 (25/26)
		SPAdes	96.1 (25/26)
		Trinity	96.1 (25/26)
	sRNA-Seq	VirusDetect	100 (26/26)
		VSD-*SPAdes*	100 (26/26)
		VSD-*Velvet*	100 (26/26)
SP2	RNA-Seq	Kodoja	100 (16/16)
		PVDP	100 (16/16)
		SPAdes	100 (16/16)
		Trinity	93.8 (15/16)
	sRNA-Seq	VirusDetect	100 (16/16)
		VSD-*SPAdes*	81.3 (13/16)
		VSD-*Velvet*	100 (16/16)

When we checked the viral contigs assembled for the sRNA data, we found that the contigs generated by VirusDetect (which relies on Velvet for its *de novo* assembly step) and VSD-Velvet were shorter than the ones generated by VSD-*SPAdes. However, we observed a significant amount of misassembled contigs in the assemblies derived with VSD-SPAdes. We could identify that these were caused by shared nucleotide stretches of up to 20 nucleotides present in the plant and virus sequences from our libraries, which caused the formation of*

chimeric plant-virus contigs. When we checked each SPAdes log after assembly, we found that the software output warnings stating it failed to determine erroneous k-mer threshold and valley value was estimated improperly, suggesting the lower bound of the k-mer range implemented was likely yielding suboptimal results. The SPAdes software developers also stated that their tool had not been tested on short reads and recommended not to use k-mers below 15 to reduce the risk of misassembly.

Additional viruses and viroids were detected via HTS that were not targeted and thus not tested for by PEQ. These include Citrus virus A in MT007 and MT008, Cherry rasp leaf virus in sample MT012, Strawberry vein banding virus in sample MT013, MT014 and MT015, Beet pseudoyellows virus in MT014, Sweet potato symptomless mastrevirus 1, Sweet potato badnavirus and Sweet potato virus C in MT016. As these were not confirmed through an independent method, they will not be further discussed in this paper.

3.4 Detection of contamination

Some of the viruses and viroids that were identified by PEQ in **Table 1** were also detected in other samples by both RNA-Seq and sRNA-Seq; they were either present in a commodity that is not known to be a host to the virus/viroid or in a known host, but PEQ did not detect their presence through traditional testing. These events were treated as candidate contamination events (**Figure 4**). For the RNA-Seq samples processed by the SP1, each tool detected events across all 15 samples from the SP1; Kodoja reported a total of 62, PVDP 51, *SPAdes* 40 and Trinity 40 events respectively. For those processed by the SP2, Kodoja identified 22 candidate events across all 10 samples, PVDP 7 events across 6 samples, *SPAdes* 1 event and Trinity 2 events across 2 samples (**Figure 3**).

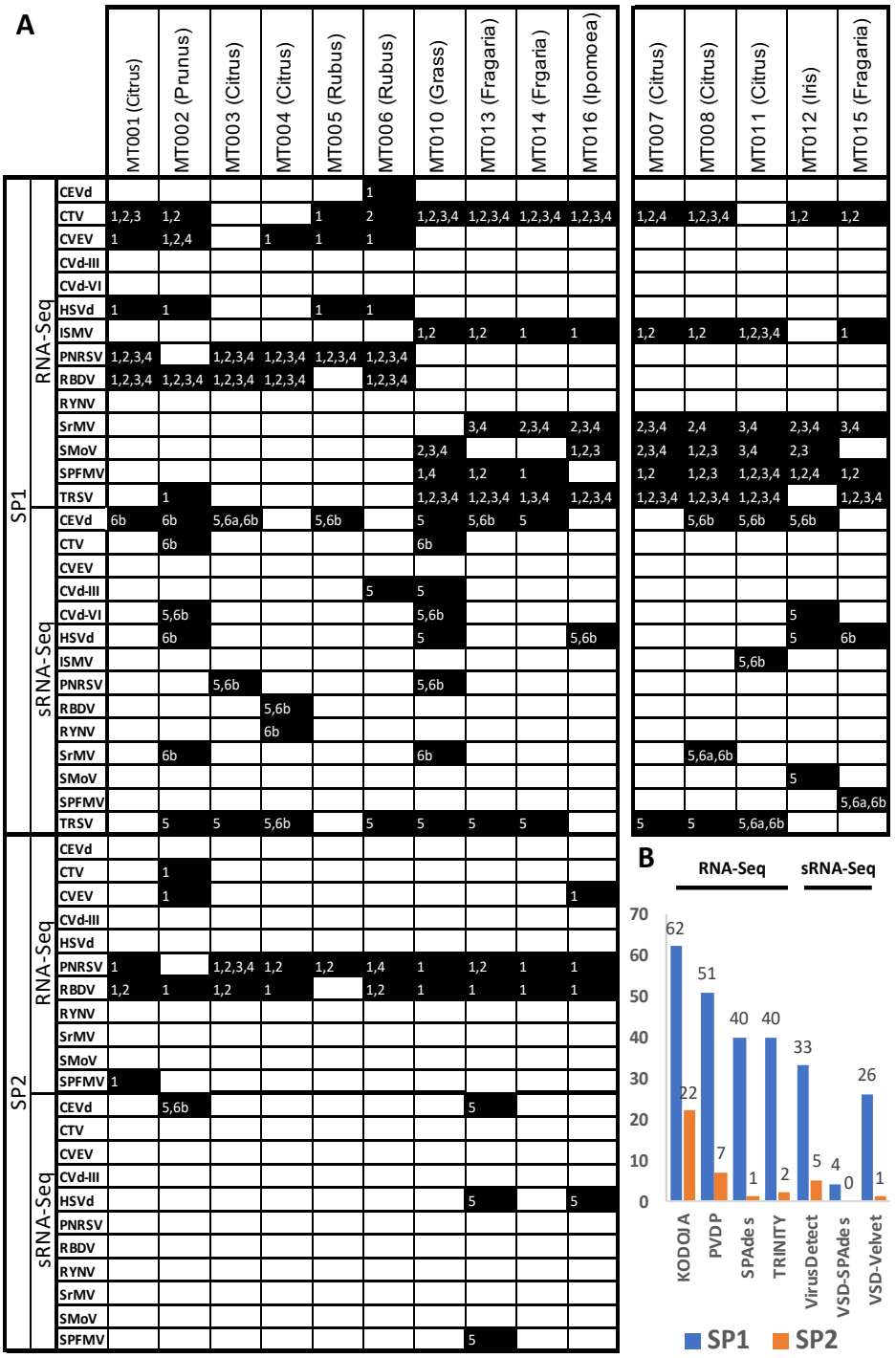


Figure 4. Viruses and viroids detected across samples suspected to be the result of contamination. **A.** Details of all predicted viral contaminants detected in each sample sequenced by SP1 and SP2. Each number in the table corresponds to one of the softwares tested: numbers 1 to 4 refer to methods applied to RNA-Seq data; numbers 5 to 6 refer to methods applied to sRNA-Seq data. 1 = Kodoja, 2 = PVDP, 3 = *SPAdes, 4 = Trinity, 5 = VirusDetect, 6a* = VSD-*SPAdes, 6b* = VSD-*Velvet*. ***B.*** *Total number of events detected per method and sequencing technology tested.*

For the sRNA-Seq samples processed by the SP1, VirusDetect identified 33 contamination events across 13 samples, VSD-Velvet 26 events across 12 samples and VSD-*SPAdes* 4 events across 4 samples. For the SP2 samples. VirusDetect detected 4 candidate events across 3 samples while VSD-Velvet detected 1 and VSD-*SPAdes* none.

We implemented a Contamination Flag (CF) for data derived by Kodoja, PVDP, VirusDetect and VSD. For the RNA-Seq data, two out of the 62 events were missed using this flag for the SP1 data derived by Kodoja (**Supplementary Table 5 and 6**), while all events were detected for the SP2 derived data. We were otherwise able to detect all events for data derived by PVDP (**Supplementary Table 7 and 8**). Similarly, for the sRNA data, the CF identified all contamination events detected by VSD-*SPAdes* (**Supplementary Table 9**), VSD-Velvet (**Supplementary Table 10 and 11**) and VirusDetect (**Supplementary Table 12 and 13**).

4. Discussion

This study contributes to the growing body of research aiming to accelerate the validation and adoption of HTS-based diagnostic tools for the detection and identification of viruses and viroids in quarantine facilities. We had access to different plant commodities whose disease profile for regulated viruses and viroids had been established through traditional methods. We were interested to assess the performance of two HTS methods (RNA-Seq and sRNA-Seq) as well as different bioinformatics analyses tools in independently detecting and identifying these pathogens. Several studies have previously reported result discrepancies between different HTS or analytical methods [12-14]. This study also highlighted the need to carefully check the plant virus and viroid sequences detected using a given HTS technology, chemistry, and analytical method. We found that for the RNA-Seq samples, recovery of all true positives was dependent on the sequencing chemistry and bioinformatics tool combination used, with some tools struggling to detect a viroid target that was present at low titer in sample MT003 in one sequencing provider. The direct annotation method Kodoja was the only one among the four tools tested that was able to detect it in both chemistry tested. In comparison, all true positives were detected irrespective of the chemistry and analytic tool used (ie VirusDetect and VSD-Velvet) in the sRNA-Seq samples.

The total number of unique viruses and viroids recovered within a sample was also highly variable depending on the sequencing technology, the sequencing approach and the analyses tool used which pushed us to investigate what might have driven these discrepancies. We found that the chemistry used by SP1 was particularly prone to contamination, which inflated the total number of viruses and viroids detected in a

given sample in both RNA-Seq and sRNA-Seq samples. While the use of unique molecular identifiers (UMIs) by the SP2 for sRNA data permitted to collapse the number of duplicates in highly abundant viruses and viroids in some samples, it did not help identify lowly abundant pathogens that occurred due to contamination, as these did not display high level of duplication.

Cross-contamination can occur at different steps of the laboratory process: nucleic acid extraction, library preparation and sequencing. We hypothesise that the contamination we observe in our study is due to index-hopping, which is a phenomenon that can occur due to low levels of free index primers present in the sample multiplexed pool during sequencing. This problem has been exacerbated by recent Illumina technology that uses patterned flow cells such as the NovaSeq, with the extent of exclusion amplification cross-contamination being up to 10%, resulting in appreciable false-positive detections for sensitive applications [39, 40]. Our library preparations include samples with high loads of a pathogen and our sequencing runs have been sequenced at high depth and both of these factors are known to make sequenced samples particularly prone to bleeding [41]. Therefore, it is essential that the interpretation of results for pathogen detection (here viruses and viroids) include an evaluation for possible cross-contamination.

We detected much lower contamination from the data derived by the SP2. Interestingly, dual index barcoding, in which unique index barcodes are placed on either side of the DNA insert, was able to mitigate but not eliminate contamination from index-hopping in RNA-Seq samples for the SP2. Similar observations have been reported by others [41]. On the other hand, index hopping events were detected at much lower levels in the matching samples sequenced using sRNA-Seq on a NextSeq 500, by this same sequencing provider with VirusDetect and VSD-Velvet detecting some and VSD-SPAdes no events across samples. Therefore overall, we found that the best performing combination for the set of samples tested in this study was the use of sRNA sequencing technology from the SP2 sequenced on a NextSeq and analysed using either VSD or VirusDetect. We were able to achieve highest accuracy, by detecting all expected viruses and viroids on one hand and minimise index switching events on another.

The occurrence of index switching events in our data drove us to evaluate how best to flag these from true positives. PVDP reports useful parameters such as evenness, read dispersion and certainty score to help identify false positives [30]. However, because the reads resulting from index hopping events were distributed evenly across the virus or viroid genome they matched to, these measurements did not help flag these. We implemented a Contamination Flag (CF) and tested it on the direct read annotation approaches (Kodoja and PVDP) for the RNA-Seq data and on VSD and VirusDetect for the sRNA-Seq

data. The CF flag enabled us to detect all index hopping events in the RNA-Seq data for SP2 but not SP1, as the difference in abundance of the false positives was much more variable in this latter for viroids, preventing a clear separation of false positives from true positives. All index hopping events detected by VSD and VirusDetect were flagged in the sRNA-Seq data for SP1. While the CF is a useful tool to make sense of the data generated, index hopping predictions would still need to be independently confirmed by a secondary technique (e.g. PCR or MinION screening) as false positives. For this reason, we would strongly recommend the use of chemistries and technologies which minimise contamination, and incorporate dual-indexing or unique molecular identifiers, to mitigate the risks of contamination which really jeopardises both the sensitivity and the potential rapid turn-around times of HTS technology in diagnostics.

For the RNA-Seq dataset, the direct read annotation approaches were more sensitive at detecting pathogens present at low titer than the *de novo* assembly ones, as these latter require reads to be present in sufficient amount to assemble into contigs for validation of viral presence. Another advantage of these tools over the *de novo* assembly ones is the ease of installation and the speed at which they run. They can therefore be a great first pass tool to detect quickly whether there is virus or viroid signal in some samples and thus guide downstream bioinformatics procedures such as reference-guided assembly, as pointed out by Guttierez et al (2021). However, their sensitivity also makes direct read annotation methods more prone to detect index hopping events. Therefore, it is important to carefully check and apply downstream filtering steps to the table of candidate viruses and viroids generated by this approach.

One important consideration for total RNA extraction, is that the high abundance of nucleic acids from the host plant co-extracted with viral nucleic acids can greatly limit sequencing sensitivity and to mitigate this, the relative proportion of viral sequences in the extracted RNA can be increased by performing a plant ribosomal RNA (rRNA) depletion step. This method was applied on both RNA-Seq and sRNA-Seq samples however its success was variable between sequencing technology and across the commodity tested. rRNA depletion was more successful in the RNA-Seq samples with residual rRNA representing a maximum of 30% of the filtered reads compared to 60% in sRNA-Seq samples. There was a correlation between the commodity the sample was derived from and the amount of residual rRNA remaining in the sRNA-derived data, with the peach, raspberry and strawberry being the least affected. The ribosomal RNA content is known to be highly variable across plants, with differences observed in both their rRNA gene repertoire and copy number [42]. This could explain the difference we observe between commodities. Unfortunately, the variability in the success of rRNA depletion from samples

originating from different commodities can bring unpredictability on the final number of reads that will be available to perform diagnostics, even if the starting number of sequencing reads might seem adequate. It is therefore important to assess the residual rRNA content of samples when performing quality control of a sequencing run for virus and viroid detection.

5. Conclusions

Side by side comparison of HTS methodologies highlights that small RNA-seq is less prone to index hopping events when a de novo assembly method is used as a primary screening approach. Ribosomal depleted RNA-seq data showed varying degrees of index hoping events and caution should be exercised when utilizing such approach for phytosanitary diagnosis of viruses and viroids.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title.

Author Contributions: Conceptualization, R.A.B. and M.W.; methodology, M.G., R.A.B.; software, C.W., M.G.; validation, M.G., R.L., C.E., S.F. and J.M.; formal analysis, M.G., R.L.; resources, J.P., C.E.; data curation, M.G., D.B.; writing—original draft preparation, M.G.; writing—review and editing, R.A.B.; supervision, R.A.B.; funding acquisition, J.P., R.A.B., M.W., A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by Hort Innovation, grant MT18005, using the Hort Innovation Citrus, Nursery, Rubus, Table Grape, Fresh Potato and Processed Potato industry research and development levy, and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generated in this study is available at the Short Read Archive database under BioProject PRJNA752836.

Acknowledgments: We thank members of the Hort Innovation MT18005 program and Industry Representatives for their input and feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rybicki, E. P.; A Top Ten list for economically important plant viruses. *Archives of Virology* **2015,** 160, (1), 17-20.

2. Loebenstein, G., Plant Virus Diseases: Economic Aspects. In *Encyclopedia of Virology (Third Edition)*, Mahy, B. W. J.; Van Regenmortel, M. H. V., Eds. Academic Press: Oxford, 2008; pp 197-201.

3. FAO; Preparing to use high-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes. Commission on Phytosanitary Measures Recommendation No. 8. Rome. Published by FAO on behalf of the Secretariat of the International Plant Protection Convention (IPPC). 6 pp. Licence: CC BY-NC-SA 3.0 IGO. In 2019.

4. Kreuze, J. F.; Perez, A.; Untiveros, M.; Quispe, D.; Fuentes, S.; Barker, I.; Simon, R.; 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* **2009,** 388, (1), 1-7.

5. Al Rwahnih, M.; Daubert, S.; Golino, D.; Rowhani, A.; Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* **2009,** 387, (2), 395-401.

6. Adams, I. P.; Glover, R. H.; Monger, W. A.; Mumford, R.; Jackeviciene, E.; Navalinskiene, M.; Samuitiene, M.; Boonham, N.; Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol Plant Pathol* **2009,** 10, (4), 537-545.

7. Jones, S.; Baizan-Edge, A.; MacFarlane, S.; Torrance, L.; Viral diagnostics in plants using next generation sequencing: computational analysis in practice. *Frontiers in plant science* **2017,** 8, 1770.

8. Villamor, D.; Ho, T.; Al Rwahnih, M.; Martin, R.; Tzanetakis, I.; High throughput sequencing for plant virus detection and discovery. *Phytopathology* **2019,** 109, (5), 716-725.

9. Visser, M.; Bester, R.; Burger, J. T.; Maree, H. J.; Next-generation sequencing for virus detection: covering all the bases. *Virology journal* **2016,** 13, (1), 85.

10. Mehetre, G. T.; Leo, V. V.; Singh, G.; Sorokan, A.; Maksimov, I.; Yadav, M. K.; Upadhyaya, K.; Hashem, A.; Alsaleh, A. N.; Dawoud, T. M.; Almaary, K. S.; Singh, B. P.; Current developments and challenges in plant viral diagnostics: A systematic review. *Viruses* **2021,** 13, (3).

11. Matsumura, E. E.; Coletta-Filho, H. D.; Nouri, S.; Falk, B. W.; Nerva, L.; Oliveira, T. S.; Dorta, S. O.; Machado, M. A.; Deep sequencing analysis of RNAs from citrus plants grown in a citrus sudden death-affected area reveals diverse known and putative novel viruses. *Viruses* **2017,** 9, (4).

12. Massart, S.; Chiumenti, M.; De Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Komínek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L.; Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. *Phytopathology* **2019,** 109, (3), 488-497.

13. Brinkmann, A.; Andrusch, A.; Belka, A.; Wylezich, C.; Höper, D.; Pohlmann, A.; Nordahl Petersen, T.; Lucas, P.; Blanchard, Y.; Papa, A.; Melidou, A.; Oude Munnink, 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.; Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets. *J Clin Microbiol* **2019,** 57, (8).

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

15. 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.; Mavrič Pleško, I.; McGreig, S.; Miozzi, L.; Remenant, B.; Reynard, J. S.; Rollin, J.; Rott, M.; Schumpp, O.; Massart, S.; Haegeman, A.; A primer on the analysis of high-throughput sequencing data for detection of plant viruses. *Microorganisms* **2021,** 9, (4).

16. Al Rwahnih, M.; Daubert, S.; Golino, D.; Islas, C.; Rowhani, A.; Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. *Phytopathology* **2015,** 105, (6), 758-63.

17. Santala, J.; Valkonen, J. P. T.; Sensitivity of Small RNA-Based Detection of Plant Viruses. *Frontiers in Microbiology* **2018,** 9, (939).

18. Martin, R. R.; Constable, F.; Tzanetakis, I. E.; Quarantine regulations and the impact of modern detection methods. *Annual Review of Phytopathology* **2016,** 54, (1), 189-205.

19. Höper, D.; Mettenleiter, T. C.; Beer, M.; Metagenomic approaches to identifying infectious agents. *Rev Sci Tech* **2016,** 35, (1), 83-93.

20. Vivek, A. T.; Zahra, S.; Kumar, S.; From current knowledge to best practice: A primer on viral diagnostics using deep sequencing of virus-derived small interfering RNAs (vsiRNAs) in infected plants. *Methods* **2020,** 183, 30-37.

21. Tamisier, L.; Haegeman, A.; Foucart, Y.; Fouillien, N.; Rwahnih, M. A.; Buzkan, N.; Candresse, T.; Chiumenti, M.; De Jonghe, K.; Lefebvre, M.; Margaria, P.; Reynard, J. S.; Stevens, K.; Kutnjak, D.; Massart, S.; Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. *Zenodo* **2021,** 4273791.

22. 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.; Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. *Phytopathology* **2019,** 109, (3), 488-497.

23. Cao, M.; Zhang, S.; Li, M.; Liu, Y.; Dong, P.; Li, S.; Kuang, M.; Li, R.; Zhou, Y.; Discovery of four novel viruses associated with flower yellowing disease of green sichuan pepper (Zanthoxylum Armatum) by virome analysis. *Viruses* **2019,** 11, (8), 696.

24. Maachi, A.; Torre, C.; Sempere, R. N.; Hernando, Y.; Aranda, M. A.; Donaire, L.; Use of high-throughput sequencing and two RNA input methods to identify viruses infecting tomato crops. *Microorganisms* **2021,** 9, (5), 1043.

25. Ho, T.; Broome, J. C.; Buhler, J. P.; O'Donovan, W.; Tian, T.; Diaz-Lara, A.; Martin, R. R.; Tzanetakis, I. E.; Characterization of endogenous Rubus yellow net virus in raspberries. *bioRxiv* **2021**, 2021.06.17.448838.

26. Krueger, F.; Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* **2015,** 516, 517.

27. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J.; fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018,** 34, (17), i884-i890.

28. Chen, Y.; Chen, Y.; Shi, C.; Huang, Z.; Zhang, Y.; Li, S.; Li, Y.; Ye, J.; Yu, C.; Li, Z.; Zhang, X.; Wang, J.; Yang, H.; Fang, L.; Chen, Q.; SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **2018,** 7, (1), 1-6.

29. Andrews, S.; FastQC: a quality control tool for high throughput sequence data. In Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom: 2010.

30. Gutiérrez, P.; Rivillas, A.; Tejada, D.; Giraldo, S.; Restrepo, A.; Ospina, M.; Cadavid, S.; Gallo, Y.; Marín, M.; PVDP: A portable open source pipeline for detection of plant viruses in RNAseq data. A case study on potato viruses in Antioquia (Colombia). *Physiological and Molecular Plant Pathology* **2021,** 113, 101604.

31. Baizan-Edge, A.; Cock, P.; MacFarlane, S.; McGavin, W.; Torrance, L.; Jones, S.; Kodoja: A workflow for virus detection in plants using k-mer analysis of RNA-sequencing data. *J Gen Virol* **2019,** 100, (3), 533-542.

32. Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A.; Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **2020,** 70, (1), e102.

33. Haas, B. J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P. D.; Bowden, J.; Couger, M. B.; Eccles, D.; Li, B.; Lieber, M.; MacManes, M. D.; Ott, M.; Orvis, J.; Pochet, N.; Strozzi, F.; Weeks, N.; Westerman, R.; William, T.; Dewey, C. N.; Henschel, R.; LeDuc, R. D.; Friedman, N.; Regev, A.; De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **2013,** 8, (8), 1494-1512.

34. Huang, X.; Madan, A.; CAP3: A DNA sequence assembly program. *Genome research* **1999,** 9, (9), 868-877.

35. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L.; BLAST+: architecture and applications. *BMC bioinformatics* **2009,** 10, (1), 421.

36. Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Ling, K.-S.; Kreuze, J.; Fei, Z.; VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **2017,** 500, 130-138.

37. Barrero, R. A.; Napier, K. R.; Cunnington, J.; Liefting, L.; Keenan, S.; Frampton, R. A.; Szabo, T.; Bulman, S.; Hunter, A.; Ward, L.; Whattam, M.; Bellgard, M. I.; An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. *BMC Bioinformatics* **2017,** 18, (1), 26.

38. Liu, D.; Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *PeerJ* **2019,** 7, e8275-e8275.

39. 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.; Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv* **2017**, 125724.

40. 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.; Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **2018,** 19, (1), 332.

41. Gu, W.; Miller, S.; Chiu, C. Y.; Clinical metagenomic next-generation sequencing for pathogen detection. *Annu Rev Pathol* **2019,** 14, 319-338.

42. Rogers, S. O.; Bendich, A. J.; Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Molecular Biology* **1987,** 9, (5), 509-520.

Appendix 16: Custom Plant Viruses and Viroids Database (PVirDB) for diagnosis of quarantinable and non-quarantinable viruses

Conception and rationale

The non-redundant NCBI nucleotide (NT) database is widely used for performing sequence similarity searches for a broad range of studies including the detection of viruses and viroids using HTS methods. One of the major challenges of public databases is the lack of a uniform and consistent description of the data deposited that can pose a challenge to virus diagnostics, especially if the same virus sequence has been described and assigned distinct descriptors making it a challenge to discriminate between genetically related sequences. Another significant challenge in using non-reductant databases is the amount of public genetic data that continues to grow more rapidly over the last five years. Also, diagnostics against large public databases require appropriate IT infrastructure available to users including data storage and larger compute resources to scrutinise user data against millions of public genetic records (i.e., GenBank). Thus, we recognise the need for a curated plant virus and viroids database that addresses the above issues and captures virus taxonomic information to facilitate pathogen identification and reporting.

We have created a curated **Plant Viruses and viroids Database** (**PVirDB**) with harmonised sequence descriptors and by selecting representative viruses/viroid genomes for closely related genetic strains enabling accurate and unambiguous pest detection. Furthermore, PVirDB is portable and can be used in a range of computer resources.

PVirDB is a custom curated nucleotide database consisting of sequences of quarantinable viruses and viroids of PEQ and non-quarantinable viruses of interest to domestic plant industries. Descriptors include taxonomic information (Genus, Family) as captured by GenBank records. Future improvements include verification of taxonomic information as per guidelines from the International Committee on Taxonomy of Viruses (ICTV) as well as include PEQ assay information for each quarantinable pest. We continue to liaise with the project team and policy regulators in related activities to improve and optimise the scope and design of PVirDB.

This database has been tested on the original pilot control dataset to validate that is recovered all expected targets. It is now being run in parallel with NR for megablast homology searches in the VSD nextflow latest beta version.

Implementation and improvements

PVirDBv.1

- Nucleotide sequences of quarantinable and non-quarantinable viruses of interest were retrieved from NCBI nucleotide (nt) database as depicted in Figure 16.1 in fasta format.
- CD-Hit tool (Li & Godzik, 2006) was used for collapsing the sequences with 98% sequence similarity into clusters.
- The first version of PVirDB included a total of 7,412 nucleotide sequences of 244 pathogens (228 viruses and 16 viroids) that belong to 18 families and 51 genera after removing the duplicates.

PVirDBv.20210331

- Sequences of viruses among top 40 exotic pathogens of Australia was added to the database.
- Authors of PVDP tool (Gutiérrez et al., 2021) shared the script they used to generate a custom virus database of sequences with a title containing the accession code, species name, segment, genus, family, isolate name, host, and PubMed identifier.
- The input sequences for this script need to be in GenBank format and therefore sequences were retrieved in GenBank format for all the pathogens.

- The unique combined quarantinable and non-quarantinable sequences were first passed through PVDP script and then the resulting renamed sequences were collapsed into clusters with 99% sequence similarity threshold using CD-Hit.
- False positive non-targeted viruses were recovered during the validation of the database with control plants because of the bias introduced by not having host plant sequences in the database. We noted that sequences with similarity to a non- quarantinable virus had higher sequence similarity to plant chloroplast sequences. To address this issue, we added 54 representative chloroplast genomes to PVirDB from a broad range of plant commodities. The addition of host plant chloroplast genomes prevented the reporting of false positives.
- The NCBI's LandPlant Refseq virus sequences were also added to prevent false positives resulting from the potential presence of non- quarantinable virus in HTS data.
- This version of PVirDB consisted of a total of 19, 751 sequences from 246 quarantinable viruses/viroids and 11 non-quarantinable viruses/viroids requested by PEQ, 13 additional viruses from top 40 exotic pathogens in Australia, Land plant virus reference sequences and 54 representative chloroplast genomes from host plant commodities.

PVirDBv.20210506

- Sequences with more than 10% of their bases corresponding to Ns were removed from the database to avoid getting top similarity to unreliable sequences. 59 quarantinable sequences and 4 non- quarantinable sequences were removed accordingly.
- Tag descriptors 'Type:quarantinable' or 'Type:non-quarantinable' or 'Type:LandPlantVirus' or 'Type:ChloroplastGenome' was added to the fasta headers of the sequences respectively.
- The filters and tagged sequences were combined to generate the database consisting of 16,704 quarantinable, 386 non-quarantinable, 2544 LandPlantVirus and 54 chloroplast genomes.

PVirDBv.20210517

- The quarantinable and non- quarantinable viruses that share accessions with the LandPlantVirus set were removed in this version to avoid missing potential top hits to viruses of interest due to their potential best match to a sequence in LandPlantVirus set.

PVirDBv.20210611

- Quarantinable and non-quarantinable viruses with top similarity to Landplant viruses were being filtered from summary outputs due to being tagged as 'LandPlantVirus'.
- BlastN comparison were conducted between individual sets of LandPlant Viruses vs quarantinable /non-quarantinable, and the outputs were filtered with coverage >=95% and identity >=95%.
- A total of 99 LandPlantVirus sequences with >=98% sequence similarity to 'quarantinable' viruses and 1 LandPlantVirus sequence with strong similarity to 'non-quarantinable' viruses were identified and were re-assigned as 'Type:quarantinable and 'Type:non-quarantinable, respectively.

Incorporation of ICTV species demarcation criteria to outputs

International Committee of Taxonomy of Viruses (ICTV) has defined species demarcation criteria for virus genera based on the percent sequence identity of two nucleotide/ amino acid sequences. This information can be used to further inform the decision of determining the quality of detections reported by GA-blastn hits. Table 16.1 below denotes the sequence identity thresholds for nucleotide and protein sequences for specific viral genus of interest encompassed among quarantinable and non-quarantinable viruses of interest. **Figure 16.1** depicts the workflow for generating and validating PVirDB.

Table 16.1. Species demarcation criteria of genera of quarantinable and non-quarantinable viruses included in the PVirDB

Family	Genus	Nucleotide cutoff -%	Amino acid cutoff - %
Geminiviridae	Begomovirus	91	
	Mastrevirus	78	
	Curtovirus	77	
Potyviridae	Ipomovirus	76	82
	Tritimivoris	76	82
	Potyvirus	76	82
Bromoviridae	Alfamovirus	Not applicable-single species in the genus	
	Anulavirus	53	34-50.4
	Bromovirus	50-80	
	Cucumovirus	65	
	Ilarvirus	Not defined	
Betaflexiviridae	Oleavirus	Not applicable-single species in the genus	
	Capilovirus	72	80
	Carlavirus	72	80
	Citrivirus	Not applicable-single species in the genus	
	Foveavirus	72	80
	Tepovirus	Not applicable-single species in the genus	
	Trichovirus	72	80
	Vitivirus	72	80
Tymoviridae	Tymovirus	80	90
	Marafivirus	80	90
	Maculavirus	80	90
Secoviridae	Comovirus		75 - 80
	Nepovirus		75 - 80
	Cheravirus		75 - 80
	Sadwavirus		75 - 80
	Waikavirus		75 - 80
Virgaviridae	Hordeivirus	Not established	
	Pecluvirus	Not established	
	Pomovirus	80	90
Pospiviroidae	Apsaviroid	90	
	Cocaviroid	90	
	Hostuviroid	90	
	Pospiviroid	90	
Avsunviroidae	Avsunviroid	90	
	Pelamoviroid	90	
Alphaflexiviridae	Allexivirus	72	80
	Mandarivirus	72	80
	Potexvirus	72	80
Aspiviridae	Ophiovirus	85	
Luteoviridae	Luteovirus		90
	Enamovirus	Not established	
Caulimoviridae	Badnavirus	80	
	Caulimovirus	80	
	Cavemovirus	80	

	Soymovirus	80	
	Tungrovirus	Not applicable-single species in the genus	
Solemoviridae	Solemoviridae	75	
Closteroviridae	Closteroviridae		75
	Ampelovirus		75
	Closterovirus		75
	Crinivirus		75
	Velarivirus		75
Tomusviridae	Tombusvirus		85
	Aureusvirus		40-80
	Necrovirus		87-93
	Panicovirus	Not established	
	Machlomovirus	Not established	
Reoviridae	Fijivirus	55	40
	Phytoreovirus	90	56
Kitaviridae	Bunervirus	Not established	
	Cilevirus	Not established	
Rhabdoviridae	Alphanucleorhabdovirus	75	
	Cytorhabdovirus	75	
Totiviridae			50
Mayoviridae		Not applicable-single species in the genus	

The above information was incorporated for each species of interest in the database in the ICTV_taxonomy_MinIdentity_Species*.tsv file.

Reporting bash script (run_VSD_report*.sh)

A custom bash script was developed for summarising GA-blastn outputs.

This script takes one or more tabular outputs from blastn as the input and requires to be in the same folder as the blastn outputs, BlastTool.jar and ICTV_taxonomy_MinIdentity_Species*.tsv files.

This script processes the blastn outputs in the following order:

01. Fetch top hit for each contig
02. Reorder the columns in the blastn output in a way compatible with the BlastTools.jar tool.
03. Use the blastTools.jar tool to summarise the blastn outputs to give all the contig IDs that matches to a certain subject accession and report average number of hits to subject and length, percent identity, coverage of the total alignment.
04. Search for ‘quarantinable’, ‘non-quarantinable’ and ‘LandPlantVirus’ in the summary outputs
05. Fetch unique names from Species column
06. Fetch for best hi for each virus/viroid by selecting the longest alignment and highest coverage.
07. Fetch for ICTV information in the ICTV_taxonomy_MinIdentity_Species*.tsv for each virus/viroid and append the information to the header

The reporting bash script summarise and filter the blast outputs and generate 3 summary*.txt files for each of blast outputs.

1. ***NON_QUARANTINABLE_viruses_viroids_ICTV.txt** - hits to non- quarantinable pests of interest to the local plant industries.
2. ***QUARANTINABLE_viruses_ICTV.txt output** - hits to pests of biosecurity concern
3. ***.tabular.txt** output will be hits to quarantinable pests, non-quarantinable pest of interest, host chloroplast genomes and other pests that are not defined as non-quarantinable pests of interest.

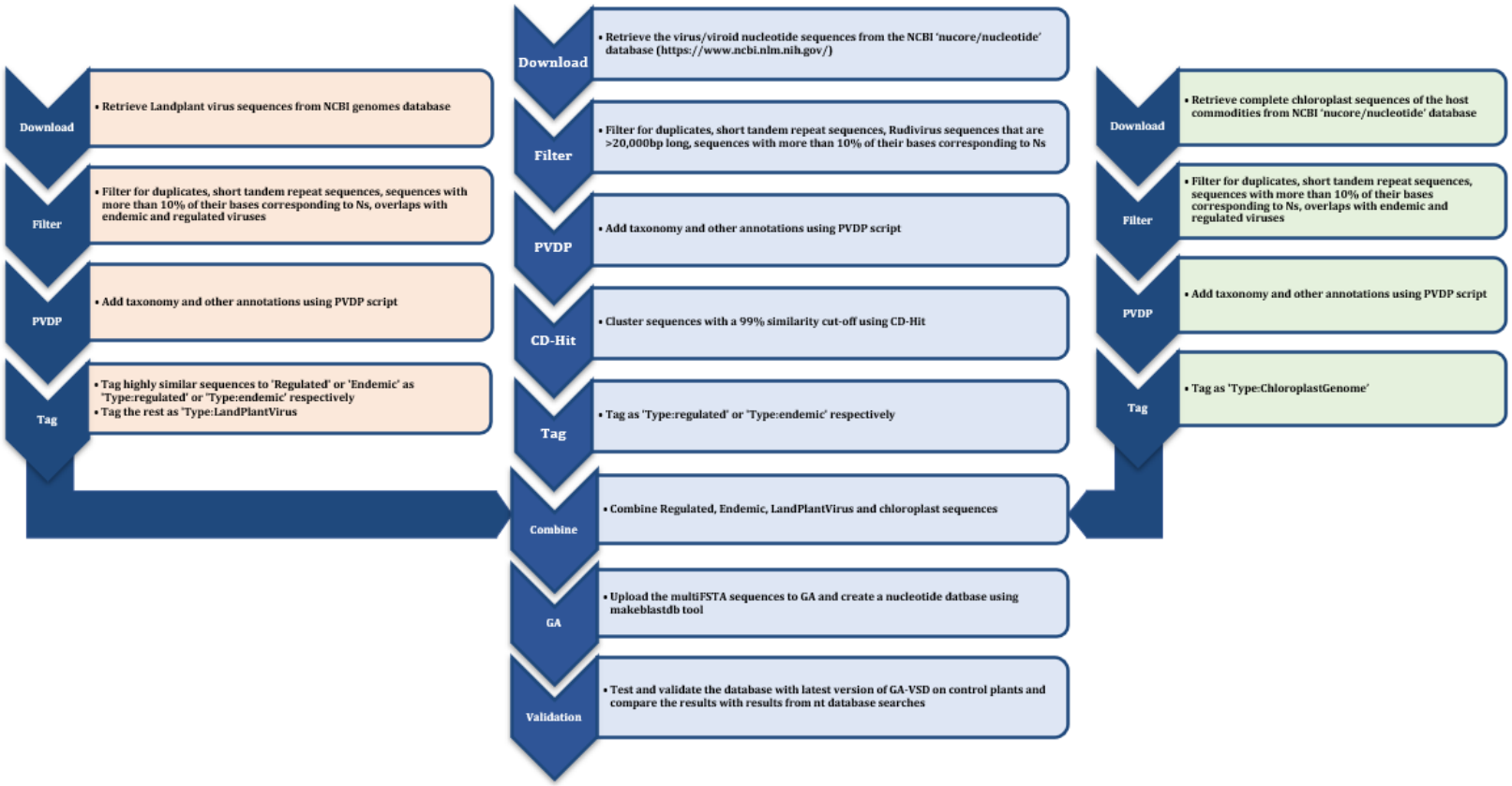


Figure 16.1. Workflow for generating and validating PVirDB

Appendix 17: Improvements and optimisation of the Virus Surveillance and Diagnosis pipeline

Quality trimming

Bioinformatics tools can detect the automatically adaptor sequences ligated to HTS data, or the user provide sequence information of the adaptor for removal. We found that fastp tool previously used in our implementation was not consistent and required multiple QC rounds to clean raw reads. We switched to Cutadapt (http://dx.doi.org/10.14806/ej.17.1.200) as it provided more consistent results. In the updated VSD pipeline we use Cutadapt to clip adaptors, filter poor quality reads/bases, and remove untrimmed reads. High quality reads passing all filters with a length between 18 to 25 nt are retained for further processing.

RNA source origin

High quality reads are then mapped onto various RNA databases (see below list) to determine the proportion of informative sequences generated for each individual sample. The mapping of quality filtered reads is done using an in-house script that uses the aligner Bowtie allowing up to two sequence polymorphisms (equivalent to ~8% sequence divergence) when mapping reads to reference sequences (see script below).

Categories of inspected RNA sources:

- rRNA: 5S rRNAs, SSU (16S/18S) and LSU (23S/28S) nuclear rRNAs, mitochondrial rRNAs and plastid rRNAs
- Plant mitochondrial genes (other than rRNAs)
- Plant plastid genes (other than rRNA)
- Plant miRNAs and other miRNAs (miRBase)
- tRNAs
- Plant nc genes (snRNAs, snoRNAs, tasiRNAs)
- Plant transposons
- Sequencing artefacts (PhiX and vector contaminants)
- Plant viruses and viroids
- Unclassified

Examples of output are reported in **Appendix 18**.

QC check script:

```
#!/bin/bash -l
#PBS -N qc_check
#PBS -l select=1:ncpus=4:mem=60gb
#PBS -l walltime=4:00:00
#PBS -j oe

#usage
# qsub -v FQ=MT027_HWHFHDMXX_GATCAG_L002_R1_trimmed.fastq QC_check.sh
#

cd $PBS_O_WORKDIR
#module load java

#SAMPLE=$(echo "$FQ" | cut -f 1-3 -d '_')
#SAMPLE=`basename -s '_trimmed.fastq' $FQ`
#echo $SAMPLE

#####
#USAGE:
#BGI control data:
#for i in MT002_S1_R1_001 MT003_S2_R1_001 MT005_S3_R1_001 MT006_S4_R1_001
MT007_S5_R1_001 MT008_S6_R1_001 MT010_S7_R1_001 MT012_S8_R1_001; do qsub -v
FQ=${i},kit=nextflex -q testvm
/work/hia_mt18005/diagnostics/scripts/virussentinel/qc_check.sh; done

SCRIPTSDIR=/work/hia_mt18005/diagnostics/scripts/virussentinel

if [ ! -d qc_check ]
   then
      mkdir qc_check
fi

conda activate cutadapt

if [[ ${kit} == "nebnextnovaseq" ||  ${kit} == "bgi" ]];
```

```
    then
       SAMPLE=$(echo "$FQ")
elif [[ ${kit} == "nebnextnovaseq2" ||  ${kit} == "bgi" ]];
    then
       SAMPLE=$(echo "$FQ" | cut -f 1 -d '_');
elif [[ ${kit} == "qiaseq" ]];
    then
       SAMPLE=$(echo "$FQ" | sed 's/_S.*//')
elif [[ ${kit} == "takara" || ${kit} == "nextflex" ]];
    then
       SAMPLE=$(echo "$FQ" | cut -f 1 -d '_')
fi

if [[ ${kit} == "nebnextnovaseq" ]] || ${kit} == "nebnextnovaseq2" ]];
   then
       cutadapt -j 8 \
       --untrimmed-output qc_check/${SAMPLE}_untrimmed.fastq \
       --no-indels -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
       -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG  \
       --trim-n --max-n 1 -m 5 --nextseq-trim=25 \
       -o qc_check/${SAMPLE}_trimmed.fastq \
       ${SAMPLE}.fastq.gz

elif [[ ${kit} == "bgi" ]];
#reads are already trimmed for adapter so just perform quality trim
   then
       cutadapt -j 8 --trim-n \
       --max-n 1 -m 5 -q 25 \
       -o qc_check/${SAMPLE}_trimmed.fastq \
       ${SAMPLE}.fastq

elif [[ ${kit} == "takara" ]];
   then
       cutadapt -j 8 --trim-n \
       --untrimmed-output qc_check/${SAMPLE}_untrimmed.fastq \
       --max-n 1 -m 5 -q 25 \
       -u 3 -a AAAAAAAAAA \
       -o qc_check/${SAMPLE}_trimmed.fastq \
       ${FQ}.fastq.gz

elif [[ ${kit} == "qiaseq" ]];
    then
       cat ${FQ}_L001_R1_001_trimmed.fastq.gz ${FQ}_L002_R1_001_trimmed.fastq.gz >
qc_check/${SAMPLE}_umi_cleaned.fastq.gz;
       cutadapt -j 8 --trim-n \
       --max-n 1 -m 10 -q 25 \
       -o qc_check/${SAMPLE}_trimmed.fastq \
       qc_check/${SAMPLE}_umi_cleaned.fastq.gz

elif [[ ${kit} == "nextflex" ]];
    then
       cutadapt -j 8 --trim-n \
       --max-n 1 -m 5 -q 25 \
       -o qc_check/${SAMPLE}_trimmed.fastq \
       $PBS_O_WORKDIR/cutadapt/${FQ}_trimmed3.fastq
fi

cd qc_check

#derive distribution for quality filtered reads > 5 bp long
perl ${SCRIPTSDIR}/fastq2fasta.pl ${SAMPLE}_trimmed.fastq > ${SAMPLE}_trimmed.fasta
conda activate py27
python ${SCRIPTSDIR}/read_length_dist.py --input ${SAMPLE}_trimmed.fasta

DB_DIR=/work/hia_mt18005/databases/read_distribution
module load bowtie

read_distribution () {
        echo -e ${SAMPLE} > ${SAMPLE}_file.log;
        echo -e '\n5S rRNA genes alignment:' >> ${i}_bowtie.log;

        bowtie -q -v 2 -k 1 -p 4 \
               --al ${i}_5S_rRNA_sRNA.fq \
               --un ${i}_5S_rRNA_cleaned_sRNA.fq \
               ${DB_DIR}/5S_rRNA \
               ${i}.fastq \
```

```
        ${i}_5S_rRNA_match 2>>${i}_bowtie.log;

    echo -e 'nc SSU and LSU rRNA genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_rRNA_sRNA.fq \
        --un ${i}_nc_rRNA_cleaned_sRNA.fq \
        ${DB_DIR}/rRNA \
        ${i}_5S_rRNA_cleaned_sRNA.fq \
        ${i}_rRNA_match 2>>${i}_bowtie.log;
    echo -e 'mt rRNA genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_mt_rRNA.fq \
        --un ${i}_mt_rRNA_cleaned_sRNA.fq \
        ${DB_DIR}/plant_mt_rRNA_genes \
        ${i}_nc_rRNA_cleaned_sRNA.fq \
        ${i}_mt_rRNA_match 2>>${i}_bowtie.log;

    echo -e 'pt rRNA genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_pt_rRNA.fq \
        --un ${i}_pt_rRNA_cleaned_sRNA.fq \
        ${DB_DIR}/plant_pt_rRNA_genes \
        ${i}_mt_rRNA_cleaned_sRNA.fq \
        ${i}_pt_rRNA_match 2>>${i}_bowtie.log;

    echo -e 'mt other genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_mt_other.fq \
        --un ${i}_mt_other_cleaned_sRNA.fq \
        ${DB_DIR}/plant_mt_other_genes \
        ${i}_pt_rRNA_cleaned_sRNA.fq \
        ${i}_mt_other_match 2>>${i}_bowtie.log;

    echo -e 'pt other genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_pt_other.fq \
        --un ${i}_pt_other_cleaned_sRNA.fq \
        ${DB_DIR}/plant_pt_other_genes \
        ${i}_mt_other_cleaned_sRNA.fq \
        ${i}_pt_other_match 2>>${i}_bowtie.log;

    echo -e 'plant miRNA genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_plant_miRNA.fq \
        --un ${i}_plant_miRNA_cleaned_sRNA.fq \
        ${DB_DIR}/plant_miRNA \
        ${i}_pt_other_cleaned_sRNA.fq \
        ${i}_plant_miRNA_match 2>>${i}_bowtie.log;

    echo -e 'other miRNA (miRBase) genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_other_miRNA.fq \
        --un ${i}_other_miRNA_cleaned_sRNA.fq \
        ${DB_DIR}/miRBase \
        ${i}_plant_miRNA_cleaned_sRNA.fq \
        ${i}_other_miRNA_match 2>>${i}_bowtie.log;

    echo -e 'tRNA genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_plant_tRNA.fq \
        --un ${i}_plant_tRNA_cleaned_sRNA.fq \
        ${DB_DIR}/plant_tRNA \
        ${i}_other_miRNA_cleaned_sRNA.fq \
        ${i}_tRNA_match 2>>${i}_bowtie.log;

    echo -e 'plant nc genes alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_plant_nc.fq \
        --un ${i}_plant_nc_cleaned_sRNA.fq \
        ${DB_DIR}/plant_nc_genes \
        ${i}_plant_tRNA_cleaned_sRNA.fq \
        ${i}_plant_nc_match 2>>${i}_bowtie.log;

    echo -e 'plant transposons alignment:' >> ${i}_bowtie.log;
    bowtie -q -v 2 -k 1 -p 4 \
        --al ${i}_transposon.fq \
```

```
              --un ${i}_transposon_cleaned_sRNA.fq \
              ${DB_DIR}/plant_transposons \
              ${i}_plant_nc_cleaned_sRNA.fq \
              ${i}_transposon_match 2>>${i}_bowtie.log;

       echo -e 'PhiX alignment:' >> ${i}_bowtie.log;
       bowtie -q -v 2 -k 1 -p 4 \
              --al ${i}_PhiX.fq \
              --un ${i}_PhiX_cleaned_sRNA.fq \
              ${DB_DIR}/phiX174 \
              ${i}_transposon_cleaned_sRNA.fq \
              ${i}_PhiX_sRNA_match 2>>${i}_bowtie.log;

       echo -e 'Vector alignment:' >> ${i}_bowtie.log;
       bowtie -q -v 2 -k 1 -p 4 \
              --al ${i}_UniVec.fq \
              --un ${i}_UniVec_cleaned_sRNA.fq \
              ${DB_DIR}/UniVec \
              ${i}_PhiX_cleaned_sRNA.fq \
              ${i}_UniVec_match 2>>${i}_bowtie.log;

       echo -e 'Virus and viroid alignment:' >> ${i}_bowtie.log;
       bowtie -q -v 2 -k 1 -p 4 \
              --al ${i}_viral.fq \
              --un ${i}_final_unaligned_sRNA.fq \
              ${DB_DIR}/virus \
              ${i}_UniVec_cleaned_sRNA.fq \
              ${i}_viral_match 2>>${i}_bowtie.log;

       rm ${i}_*_cleaned_sRNA.fq
       grep 'one reported alignment' ${i}_bowtie.log > ${i}_read_counts.txt;
       grep -m1 processed ${i}_read_counts.txt >> ${i}_read_counts.txt;
       sed -i 's/# reads .*: //' ${i}_read_counts.txt;
       sed -i 's/ .*$//' ${i}_read_counts.txt;
}

#derive read category distribution for trimmed and untrimmed reads
if [[ ${kit} == "nebnextnovaseq" || ${kit} == "nebnextnovaseq2" || ${kit} == "takara"
]];
   then
       for i in ${SAMPLE}_untrimmed ${SAMPLE}_trimmed;
           do read_distribution;
           done

elif [[ ${kit} == "bgi" || ${kit} == "qiaseq"  || ${kit} == "nextflex" ]];
    then
       for i in ${SAMPLE}_trimmed;
           do read_distribution;
           done
fi
```

De novo assembly

We have also updated the *de novo* assembler in the VSD pipeline to Velvet as we found that it outperformed the formerly used software (i.e., SPAdes 3.14). Comparison of the performance of the two assemblers was tested using positive control pilot samples used in the HTS service provider 'Ring' study.

Contamination flag

We also developed a python script that filters the final list of detected viruses and viroids and provides specific coverage statistics. Processed data is then parsed by second script which derives a summary at the run level and flags potential index-hopping events.

Script 1 – contamination flag

```
#!/usr/bin/env python
import argparse
import pandas as pd
import numpy as np
import os
import subprocess
from functools import reduce
import glob

def main():
    ##############################################################################
    parser = argparse.ArgumentParser(description='Load VSD pipeline results')
```

```
    # All the required arguments #
    parser.add_argument("--results", type=str)
    parser.add_argument("--fastq", type=str)
    parser.add_argument("--rawfastq", type=str)
    parser.add_argument("--sample", type=str)
    parser.add_argument("--cov", help="derive coverage stats",
                        action="store_true")
    parser.add_argument("--read_size", type=str)
    parser.add_argument("--targets", type=str)

    args = parser.parse_args()
    results_path = args.results
    sample = args.sample
    fastq_file = args.fastq
    coverage_cal = args.cov
    raw_fastq_file = args.rawfastq
    read_size = args.read_size
    targets = args.targets

    print(raw_fastq_file)

    raw_data = pd.read_csv(results_path, header=0, sep='\t',index_col=None)

    #load list of target viruses and viroids and matching official ICTV name
    targets_df = pd.read_csv(targets, header=0, sep='\t')
    print (raw_data)
    if len(raw_data) == 0:

        print('DataFrame is empty!')
        csv_file1 = open(sample + "_" + read_size + '_all_targets_with_scores.txt',
"w")
        csv_file1.write("sacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tRNA_type\tTargetted_sp_generic_name\tnaccs_score\tlength_sco
re\tavpid_score\tcov_score\tgenome_score\tcompleteness_score\ttotal_score")
        csv_file1.close()
        csv_file2 = open(sample + "_" + read_size + '_top_scoring_targets.txt', "w")
        csv_file2.write("sacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tTargetted_sp_generic_name\tRNA_type\tTargetted_sp_generic_na
me\tnaccs_score\tlength_score\tavpid_score\tcov_score\tgenome_score\tcompleteness_sco
re\ttotal_score")
        csv_file2.close()
        csv_file3 = open(sample + "_" + read_size +
'_top_scoring_targets_with_cov_stats.txt', "w")
        csv_file3.write("Sample\tsacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tTargetted_sp_generic_name\tnaccs_score\tlength_score\tavpid_
score\tcov_score\tgenome_score\tcompleteness_score\ttotal_score\tMean coverage\tRead
count\tRead count norm\tSelected read count norm\tPCT_1X\tPCT_10X\tPCT_20X")
        csv_file3.close()
        exit ()

    print(raw_data)
    #print(raw_data).head(10)

    print("Cleaning up the data")
    print("Remove double spacing")
    raw_data = raw_data.replace('\s+', ' ', regex=True)

    print('Remove hyphens')
    raw_data["stitle"] = raw_data["stitle"].str.replace('-', ' ')

    print('Remove underscores')
    raw_data['stitle'] = raw_data['stitle'].str.replace('_',' ')

    print('Remove commas')
    raw_data['stitle'] = raw_data['stitle'].str.replace(',',' ')

    print('Remove problematic text in virus description names')
    #this fixes virus names like
'Prunus_necrotic_ringspot_virus_Acot_genomic_RNA,_segment_RNA2,_complete_sequence'
    raw_data['stitle'] = raw_data['stitle'].str.replace(' genomic RNA
segment','segment')

    raw_data['stitle'] = raw_data['stitle'].str.replace('{complete viroid
sequence}','', regex=False)
    #This is a complete genome
```

```
    raw_data['stitle'] =  raw_data['stitle'].str.replace('Rubus yellow net virus
isolate Canadian 2 hypothetical protein genes  partial cds; hypothetical proteins
polyprotein  ORF 6  and hypothetical protein genes  complete cds; and hypothetical
protein genes  partial cds','Rubus yellow net virus isolate Canadian 2 complete cds')
    #remove resistance genes from list of results
    raw_data = raw_data[~raw_data['stitle'].str.contains("resistance gene")]

    raw_data = raw_data.sort_values('stitle')
    print(raw_data.stitle)

    print("Assigning standardised species name for viruses of interest")

    def get_cat(x):
        for Species_Long, Species_Short in
targets_df[['Species_Long','Species_Short']].values:
            if Species_Long.lower() in x.lower():
                return Species_Short

    raw_data['Targetted_sp_generic_name'] = raw_data['stitle'].apply(get_cat)
    print("Remove hits that do not target viruses of interest")
    raw_data['Targetted_sp_generic_name'].isnull().values.any()
    raw_data = raw_data[raw_data['Targetted_sp_generic_name'].notna()]

    print("If present in original nomenclature, add RNA type information to virus
standardised species name")
    RNA1=("RNA1", "RNA 1")
    RNA2=("RNA2", "RNA 2")
    RNA3=("RNA3", "RNA 3")
    raw_data['RNA_type'] = np.where(raw_data.stitle.str.contains('RNA1|RNA 1|segment
1'), "RNA1",
                            np.where(raw_data.stitle.str.contains('RNA2|RNA 2|segment
2'), "RNA2",
                            np.where(raw_data.stitle.str.contains('RNA3|RNA 3|segment
3'), "RNA3", "NaN")))

    raw_data['Targetted_sp_generic_name_updated'] =
raw_data[['Targetted_sp_generic_name', 'RNA_type']].agg(' '.join, axis=1)
    raw_data['Targetted_sp_generic_name_updated'] =
raw_data['Targetted_sp_generic_name_updated'].astype(str).str.strip("NaN")
    raw_data['Targetted_sp_generic_name_updated'] =
raw_data['Targetted_sp_generic_name_updated'].astype(str).str.rstrip( )

    raw_data = raw_data.reset_index(drop=True)
    print (len(raw_data.Targetted_sp_generic_name.value_counts()))
    if len(raw_data.Targetted_sp_generic_name.value_counts()) == 0:
        print ('Dataframe has no targetted viruses or viroids')
        csv_file1 = open(sample + "_" + read_size + '_all_targets_with_scores.txt',
"w")
        csv_file1.write("sacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tRNA_type\targetted_sp_generic_name\tnaccs_score\tlength_scor
e\tavpid_score\tcov_score\tgenome_score\tcompleteness_score\ttotal_score")
        csv_file1.close()
        csv_file2 = open(sample + "_" + read_size + '_top_scoring_targets.txt', "w")
        csv_file2.write("sacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tTargetted_sp_generic_name\tRNA_type\tTargetted_sp_generic_na
me\tnaccs_score\tlength_score\tavpid_score\tcov_score\tgenome_score\tcompleteness_sco
re\ttotal_score")
        csv_file2.close()
        csv_file3 = open(sample + "_" + read_size +
'_top_scoring_targets_with_cov_stats.txt', "w")
        csv_file3.write("Sample\tsacc\tnaccs\tlength\tslen\tcov\tav-
pident\tstitle\tqseqids\tTargetted_sp_generic_name\tRNA_type\tTargetted_sp_generic_na
me\tnaccs_score\tlength_score\tavpid_score\tcov_score\tgenome_score\tcompleteness_sco
re\ttotal_score\tMean coverage\tRead count\tRead count norm\tSelected read count
norm\tPCT_1X\tPCT_10X\tPCT_20X")
        csv_file3.close()
        exit ()

    print("Applying scoring to blast results to select best hit")
    raw_data['naccs'] = raw_data['naccs'].astype(int)
    raw_data['naccs_score'] =
raw_data.groupby('Targetted_sp_generic_name_updated').apply(max_naccs)

    raw_data['length'] = raw_data['length'].astype(int)
    raw_data['length_score'] =
raw_data.groupby('Targetted_sp_generic_name_updated').apply(max_length)
```

```
    raw_data['av-pident'] = raw_data['av-pident'].astype(float)
    raw_data['avpid_score'] =
raw_data.groupby('Targetted_sp_generic_name_updated').apply(max_avpid)

    raw_data['cov'] = raw_data['cov'].astype(float)
    raw_data['cov_score'] =
raw_data.groupby('Targetted_sp_generic_name_updated').apply(max_cov)

    raw_data['genome_score'] = raw_data['stitle'].apply(genome_score)
    raw_data['completeness_score'] = raw_data['stitle'].apply(completeness_score)

    raw_data['total_score'] = raw_data['length_score'] + raw_data['naccs_score'] +
raw_data['avpid_score'] + raw_data['cov_score'] + raw_data['genome_score'] +
raw_data['completeness_score'].astype(int)

    print("Output all hits that match species of interest")
    raw_data.to_csv(sample + "_" + read_size + '_all_targets_with_scores.txt',
index=None, sep="\t" )

    print("Remove seconday hits based on contig name")
    unique_contigs = list(set([i.strip() for i in
','.join(raw_data['qseqids']).split(',')]))

    #print("Assigning standardised species name for viruses of interest")
    filtered_data = pd.DataFrame()
    for contig in unique_contigs:

        selected = pd.DataFrame()
        selected = raw_data[raw_data["qseqids"].str.contains(contig)]

        if len(selected) == 1:
            filtered_data = filtered_data.append(selected)
        elif len(selected)>1:
            #extract list of spp for a given contig
            Targetted_sp_generic_name_updated_list =
selected['Targetted_sp_generic_name_updated'].tolist()
            if len(Targetted_sp_generic_name_updated_list) == 1:
                filtered_data = filtered_data.append(selected)
            #if there are several contigs, retain the top hit, remove 2ary hits
            elif len(Targetted_sp_generic_name_updated_list) > 1:
                topmatch = selected['naccs'].max()
                selected = selected[selected['naccs'] == topmatch]
                #check if there was a tie when selecting by max naccs:
                Targetted_sp_generic_name_updated_list =
selected['Targetted_sp_generic_name_updated'].tolist()
                if len(Targetted_sp_generic_name_updated_list) == 1:
                    filtered_data = filtered_data.append(selected)
                #if there was a tie, selet next based on av-pidentity
                else:
                    topmatch = selected['av-pident'].max()
                    selected = selected[selected['av-pident'] == topmatch]
                    filtered_data = filtered_data.append(selected)

    filtered_data = filtered_data.drop_duplicates()

    print("Only retain the top hits")
    idx =
filtered_data.groupby(['Targetted_sp_generic_name_updated'])['total_score'].transform
(max) == filtered_data['total_score']
    filtered_data = filtered_data[idx]
    print(filtered_data.dtypes)

    #select one random hit if tie for top hits:
    print("If there is a tie, select a random sequence out of the top scoring hit")
    final_data =
filtered_data.drop_duplicates(subset="Targetted_sp_generic_name_updated",
keep="first")

    #By setting keep on False, all duplicates are True
    #If there are duplicates in species name (ie RNA types present), then it will
drop NaN
    final_data =
final_data[~((final_data["Targetted_sp_generic_name"].duplicated(keep=False))&(final_
data["RNA_type"].str.contains("NaN")))]
    final_data = final_data.drop(['Targetted_sp_generic_name'], axis=1)
```

```
    final_data = final_data.rename(columns={"Targetted_sp_generic_name_updated":
"Targetted_sp_generic_name"})
    final_data.to_csv(sample + "_" + read_size + '_top_scoring_targets.txt',
index=None, sep="\t")

    target_dict = {}
    target_dict =
pd.Series(final_data.Targetted_sp_generic_name.values,index=final_data.sacc).to_dict(
)
    print (target_dict)

    print("Align reads and derive coverage and depth for best hit")
    if coverage_cal:
        raw_read_counts = (len(open(raw_fastq_file).readlines(  ))/4)
        selected_read_counts = (len(open(fastq_file).readlines(  ))/4)
        print(raw_read_counts)
        print(selected_read_counts)
        cov_dict = {}
        PCT_1X_dict = {}
        PCT_10X_dict = {}
        PCT_20X_dict = {}
        read_counts_dict = {}
        read_counts_normalised_dict = {}
        selected_read_counts_normalised_dict = {}
        for refid, refspname in target_dict.items():
            print (refid)
            print (refspname)
            combinedid = str(refid + ' ' + refspname).replace(' ','_')

            blastdbpath = "/work/hia_mt18005/blastDB/nt"

            fastafile = (sample + '_' + read_size + "_" + combinedid +
'.fa').replace(' ','_')
            single_fasta_entry = open(fastafile, "w")
            command_line = ['blastdbcmd','-db', blastdbpath,'-entry', refid, \
                        '-outfmt','"%f"']
            subprocess.call(command_line, stdout=single_fasta_entry)
            single_fasta_entry.close()

            print("Building a bowtie index")
            index=(sample + '_' + read_size + '_' + combinedid).replace(' ','_')
            buildindex = ['bowtie-build','-f', fastafile, index]
            subprocess.call(buildindex)

            samoutput = str(index + ".sam")
            print("Aligning original reads")
            bowtie_output = str(index + "_bowtie_log.txt")
            aligning = ['bowtie', '-q', '-v', '2', '-k', '1', '-p', '4', '-x', index,
fastq_file, '-S', samoutput]
            proc = subprocess.call(aligning, stderr=open(bowtie_output,'w'))

            print("Derive a bam file")
            bamoutput = str(index + ".bam")
            derivebam = ['samtools', 'view', '-@', '4', '-bS', samoutput]
            subprocess.call(derivebam, stdout=open(bamoutput,'w'))

            print("Sorting bam file")
            sortedbamoutput = str(index + ".sorted.bam")
            sorting = ['samtools', 'sort', '-@', '4', bamoutput, '-o',
sortedbamoutput]
            subprocess.call(sorting)

            sortedbamoutput = str(index + ".sorted.bam")
            bamindex = str(index + ".sorted.bam.bai")
            print("Indexing bam file")
            indexing = ['samtools', 'index', sortedbamoutput]
            subprocess.call(indexing, stdout=open(bamindex,'w'))

            print("Running picard")
            picard_output = (index + "_picard_metrics.txt")
            picard = ['picard', 'CollectWgsMetrics', '-I', str(sortedbamoutput), '-
O', str(picard_output), '-R', str(fastafile), '-READ_LENGTH','22', "-COUNT_UNPAIRED",
"true"]

            subprocess.call(picard)
```

```
            subprocess.call(['rm','-r', samoutput])
            subprocess.call(['rm','-r', bamoutput])
            for fl in glob.glob(index + "*ebwt"):
                os.remove(fl)

            cov = ()
            PCT_1X = ()
            PCT_10X = ()
            PCT_20X = ()
            with open(picard_output) as f:
                a = ' '
                while(a):
                    a = f.readline()
                    l = a.find('MEAN_COVERAGE') #Gives a non-negative value when
there is a match
                    if ( l >= 0 ):
                        line = f.readline()
                        elements = line.split("\t")
                        cov, PCT_1X, PCT_10X, PCT_20X = elements[1],
elements[13],elements[15],elements[17]
                        print(cov)
            f.close()
            cov_dict[refspname] = cov
            PCT_1X_dict[refspname]  = PCT_1X
            PCT_10X_dict[refspname]  = PCT_10X
            PCT_20X_dict[refspname]  = PCT_20X

            read_counts = ()
            read_counts_normalised = ()
            selected_read_counts_normalised = ()
            with open(bowtie_output) as bo:
                a = ' '
                while(a):
                    a = bo.readline()
                    l = a.find('# reads with at least one alignment:') #Gives a non-
negative value when there is a match
                    if ( l >= 0 ):
                        print(a)
                        read_counts = a.split(" ")[7]
                        print(read_counts)
                        read_counts_normalised =
round(int(read_counts)*1000000/int(raw_read_counts))
                        print(read_counts_normalised)
                        selected_read_counts_normalised =
round(int(read_counts)*1000000/int(selected_read_counts))
                        print(selected_read_counts_normalised)

            read_counts_dict[refspname] = read_counts
            read_counts_normalised_dict[refspname] = read_counts_normalised
            selected_read_counts_normalised_dict[refspname] =
selected_read_counts_normalised

        cov_df = pd.DataFrame(cov_dict.items(),columns=['Targetted_sp_generic_name',
'Mean coverage'])
        read_counts_df =
pd.DataFrame(read_counts_dict.items(),columns=['Targetted_sp_generic_name', 'Read
count'])
        read_counts_norm_df =
pd.DataFrame(read_counts_normalised_dict.items(),columns=['Targetted_sp_generic_name'
, 'Read count norm'])
        selected_read_counts_norm_df =
pd.DataFrame(selected_read_counts_normalised_dict.items(),columns=['Targetted_sp_gene
ric_name', 'Selected read count norm'])
        PCT_1X_df =
pd.DataFrame(PCT_1X_dict.items(),columns=['Targetted_sp_generic_name', 'PCT_1X'])
        PCT_10X_df =
pd.DataFrame(PCT_10X_dict.items(),columns=['Targetted_sp_generic_name', 'PCT_10X'])
        PCT_20X_df =
pd.DataFrame(PCT_20X_dict.items(),columns=['Targetted_sp_generic_name', 'PCT_20X'])

        dfs = [final_data, cov_df, read_counts_df, read_counts_norm_df,
selected_read_counts_norm_df, PCT_1X_df, PCT_10X_df, PCT_20X_df]
        full_table = reduce(lambda left,right:
pd.merge(left,right,on='Targetted_sp_generic_name'), dfs)
        full_table['Mean coverage'] = full_table['Mean coverage'].astype(float)
        full_table['PCT_1X'] = full_table['PCT_1X'].astype(float)
```

```
        full_table['PCT_10X'] = full_table['PCT_10X'].astype(float)
        full_table['PCT_20X'] = full_table['PCT_20X'].astype(float)

        #full_table[["Mean coverage", "PCT_1X", "PCT_10X", "PCT_20X"]] =
full_table[["Mean coverage", "PCT_1X", "PCT_10X", "PCT_20X"]].apply(pd.to_numeric)
        full_table.insert(0, 'Sample', sample)
        print(full_table)
        full_table.to_csv(sample + '_' + read_size +
'_top_scoring_targets_with_cov_stats.txt', index=None, sep="\t",float_format="%.2f")

def max_avpid(df):
    max_row = df['av-pident'].max()
    labels = np.where((df['av-pident'] == max_row),
                      '1',
                      '0')
    return pd.DataFrame(labels, index=df.index).astype(int)

def max_length(df):
    max_row = df['length'].max()
    labels = np.where((df['length'] == max_row),
                        '2',
                        '0')
    return pd.DataFrame(labels, index=df.index).astype(int)

def max_naccs(df):
    max_row = df['naccs'].max()
    labels = np.where((df['naccs'] == max_row),
                        '1',
                        '0')
    return pd.DataFrame(labels, index=df.index).astype(int)

def max_cov(df):
    max_row = df['cov'].max()
    labels = np.where((df['cov'] == max_row),
                        '1',
                        '0')
    return pd.DataFrame(labels, index=df.index).astype(int)

def genome_score(x):
    if 'nearly complete sequence' in str(x):
        return -3
    elif 'complete sequence' in str(x):
        return 3
    elif 'complete genome' in str(x):
        return 3
    elif 'polyprotein gene complete cds' in str(x):
        return 3
    else:
        return 0

def completeness_score(x):
    if 'partial'in str(x):
        return -2
    elif 'polymerase protein' in str(x):
        return -2
    else:
        return 0

if __name__ == '__main__':
    main()
```

Script 2 – contamination flag

```
#!/usr/bin/env python
import argparse
import pandas as pd
import numpy as np
import os
import subprocess
from functools import reduce
import glob

run_data = pd.DataFrame()

for fl in glob.glob("*21-22_top_scoring_targets_with_cov_stats.txt"):
    sample_data = pd.read_csv(fl, header=0, sep='\t',index_col=None)
    run_data = run_data.append(sample_data)
```

```
run_data = run_data[['Sample','sacc','naccs','length','slen','cov','av-
pident','qseqids','Targetted_sp_generic_name','Mean coverage','Read count','Read
count norm','PCT_1X', 'PCT_10X', 'PCT_20X']]
run_data['read size'] = '21-22 bp'
run_data['count_max'] = run_data.groupby(['Targetted_sp_generic_name'])['Read count
norm'].transform(max)

run_data['threshold_value']=run_data['count_max']*0.1
#run_data['contamination_flag'] = ['true' if x < run_data['threshold_value'] else
'false' for x in run_data['Read count norm']]
run_data['contamination_flag'] = np.where(run_data["Read count norm"] <=
run_data["threshold_value"], True, False)
#df["equal"] = comparison_column
#compare `"col1"` and `"col2"`

run_data2 = pd.DataFrame()

for fl in glob.glob("*24-24_top_scoring_targets_with_cov_stats.txt"):
    sample_data = pd.read_csv(fl, header=0, sep='\t',index_col=None)
    run_data2 = run_data2.append(sample_data)

run_data2 = run_data2[['Sample','sacc','naccs','length','slen','cov','av-
pident','qseqids','Targetted_sp_generic_name','Mean coverage','Read count','Read
count norm','PCT_1X', 'PCT_10X', 'PCT_20X']]
run_data2['read size'] = '24 bp'

run_data2['count_max'] = run_data2.groupby(['Targetted_sp_generic_name'])['Read count
norm'].transform(max)

run_data2['threshold_value']=run_data2['count_max']*0.1
#run_data['contamination_flag'] = ['true' if x < run_data['threshold_value'] else
'false' for x in run_data['Read count norm']]
run_data2['contamination_flag'] = np.where(run_data2["Read count norm"] <=
run_data2["threshold_value"], True, False)
#df["equal"] = comparison_column
#compare `"col1"` and `"col2"`

merged_data = pd.DataFrame()
data_cont = [run_data, run_data2]
merged_data = pd.concat(data_cont)

merged_data = merged_data.sort_values(['read size', 'Sample'], ascending = (True,
True))
print(merged_data)

merged_data.to_csv('run_top_scoring_targets_with_cov_stats_with_cont_flag_0.1.txt',
index=None, sep="\t",float_format="%.2f")
```

An example of output is provided in **Table 17.1.**

Table 17.1 Output example showing identified viruses and viroids. A contamination flag highlighting a potential false positive (index hopping event) is shown in red fonts.

Sample	read size	Targetted_sp_generic_name	sacc	slen	cov	av-pident	Mean coverage	Read count	Read count norm	PCT_1X	PCT_10X	PCT_20X	count_max	threshold_value	Contamination flag
MT020	21-22 bp	Peach associated luteovirus	KY635989	5814	0.37	97.31	36.78	10488	1062	0.96	0.76	0.52	1062	106.2	FALSE
MT020	24 bp	Peach associated luteovirus	KY635989	5814	0.01	97.83	1.03	254	26	0.44	0.01	0	26	2.6	FALSE
MT029	21-22 bp	Peach latent mosaic viroid	MK766396	348	0.12	100	3.82	64	3	0.42	0.19	0.03	43634	4363.4	TRUE
MT029	21-22 bp	Prunus geminivirus A	MF504040	3174	0.55	99.67	69.8	12726	695	0.93	0.76	0.63	732	73.2	FALSE
MT029	24 bp	Prunus geminivirus A	MF504040	3174	0.18	100	6.65	888	48	0.76	0.21	0.06	172	17.2	FALSE
MT031	21-22 bp	Prunus geminivirus A	MF504038	3174	0.38	99.85	34.19	5123	444	0.94	0.63	0.49	732	73.2	FALSE
MT031	24 bp	Prunus geminivirus A	MF504038	3174	0.12	100	12.92	1928	167	0.82	0.35	0.19	172	17.2	FALSE
MT032	21-22 bp	Prunus geminivirus A	MF504038	3174	0.49	99.82	57.84	10010	732	0.95	0.73	0.61	732	73.2	FALSE
MT032	24 bp	Prunus geminivirus A	MF504038	3174	0.14	100	12.88	2357	172	0.84	0.31	0.17	172	17.2	FALSE
MT039	21-22 bp	Prune dwarf virus RNA1	KY883327	3339	0.43	97.6	90.11	19427	1487	0.98	0.88	0.79	1487	148.7	FALSE
MT039	24 bp	Prune dwarf virus RNA1	MK522387	3375	0.08	95.55	4.66	661	51	0.62	0.13	0.05	51	5.1	FALSE
MT039	21-22 bp	Prune dwarf virus RNA2	KY883329	2556	0.41	97.82	69.39	12896	987	0.95	0.81	0.64	987	98.7	FALSE
MT039	24 bp	Prune dwarf virus RNA2	AF277662	2593	0.07	98.49	4.95	539	41	0.65	0.16	0.06	41	4.1	FALSE
MT039	21-22 bp	Prune dwarf virus RNA3	KY883331	2051	0.27	97.33	204.36	79656	6098	0.99	0.97	0.95	6098	609.8	FALSE
MT039	24 bp	Prune dwarf virus RNA3	KY883331	2051	0.14	98.62	34.48	5307	406	0.95	0.72	0.55	406	40.6	FALSE
MT040	21-22 bp	Prunus necrotic ringspot virus RNA1	MH727237	3332	0.23	98.86	36.61	6407	535	0.96	0.75	0.53	1816	181.6	FALSE

Where: sacc = subject accession ID; slen = subject length; cov = coverage; av-pident= average percentage sequence identity; PCT_1X= proportion of genome covered by at least one mapped read at each position; PCT_10X and PCT_20X, similar to PCT_1X but requiring the mapping of at least 10 or 20 reads at each position, respectively.

Appendix 18: NGS service provider 'Ring' test identifies reliable commercial kit for data generation.

COVID19 has impacted and continues to impact the supply chain of reagents for HTS services. To enable HTS service redundancy and resilience to deliver to HTS services, we engaged various service providers to assess the generation of high quality HTS data and where applicable enable implementation/testing of alternative solutions.

Reliable diagnosis of quarantinable pest using HTS technologies relies on the generation of high-quality data. Recent, Illumina sequencing instruments (i.e., NovaSeq6000) is susceptible to generate cross sample contaminations, also known as index hoping event with a frequency of up to 2%. Therefore, commercial companies continue to develop new technical solutions to mitigate the formation of such non-specific products.

Here, in this HTS service provider 'ring' test we invited three providers to use two distinct commercial kits to generate small RNA-seq HTS data for selected positive control plants.

Aims

The main aims of this ring test is to assess the performance of various sRNA-seq library preparation protocols for the diagnosis of plant viruses and viroids.

Additional specific aims are:

- continue to optimise SOPs for sample collection and bioinformatics protocols
- increase specificity (reduce false positives/index hopping events) but maintain sensitivity
- reduce ribosomal RNA (rRNA) contamination
- increase flexibility/resilience of HTS services by tapping into the expertise of several service providers and/or enabling the implementation of new approaches

Four sRNA library kits: the QIAseq miRNA library kit (Qiagen), NEBNext Multiplex Small RNA Library (New England Biolabs), the SMARTer smRNA-Seq Kit (Takara Bio) and the NEXTFLEX Small RNA-Seq Kit v3 (PerkinElmer) were evaluated (**Table 18.1**).

The outcomes of this pilot study will determine which HTS methods are best suited to conduct follow up large scale side-by-side PEQ vs HTS comparisons.

Table 18.1 Details of the different sRNA library kit tested

Technology	Kit	Seq provider	Method	Seq platform
1) NEBNext Multiplex Small RNA Library	NEB	AGRF	Sequential ligation	NovaSeq6000
2) NEXTFLEX Small RNA-Seq Kit v3 + UMIs	PerkinElmer	AGRF	Sequential ligation	NovaSeq6000
3) NEXTFLEX Small RNA-Seq Kit v3 + UDIs	PerkinElmer	CARF	Sequential ligation	NovaSeq6000
4) SMARTer smRNA-Seq Ki Takara Bio includes Clontech product	Takara	CARF	Template switching	MiSeq

5) QIAseq FastSelect kit (rRNA removal) + QIAseq miRNA library kit + UMIs + UDIs	Qiagen	Ramaciotti	Sequential ligation	NovaSeq6000
6) QIAseq miRNA library kit + UMIs + UDIs	Qiagen	Ramaciotti	Sequential ligation	NovaSeq6000

Experimental design

For the ring study, we selected 10 PEQ control plants encompassing a diverse range of species including citrus, prunus, rubus, sweet potato, nursery, strawberry, and grapes. These control plants were known to be infected by one or more pathogen (**Table 18.2**). Nine of these overlapped with the samples that were selected for our original pilot study.

Table 18.2: Selected control plants for PEQ vs HTS comparisons

Sample ID	Commodity	Positive detection in PEQ
MT000	Rubus	Endogenous Rubus yellow net virus (endoRYNV)
		Raspberry bushy dwarf virus (RBDV)
MT001	Citrus	Citrus exocortis viroid (CEVd)
MT002	Prunus	Prunus necrotic ringspot virus (PNRSV)
MT003	Citrus	Citrus tristeza virus (CTV)
		Hop stunt viroid (HSVd)
		Citrus vein enation virus (CVEV)
		Citrus dwarfing viroid (CVd-III)
MT005	Rubus	Endogenous Rubus yellow net virus (endoRYNV)
		Raspberry bushy dwarf virus (RBDV)
MT007	Citrus	Citrus dwarfing viroid (CVd-III)
		Hop stunt viroid (HSVd)
MT008	Citrus	Citrus dwarfing viroid (CVd-III)
		Hop stunt viroid (HSVd)
MT010	Grass	Sorghum mosaic virus (SgMV)
MT012	Iris	Iris severe mosaic virus (ISMV)
		Tobacco ringspot virus (TRSV)
MT016	Ipomoea	Sweet potato feathery mottle virus (SPFMV)

Total RNA extraction was performed using an automated approach on the Maxwell® Rapid Sample Concentrator Instrument with a SimplyRNA Tissue kit (AS1340, Promega). The same RNA extraction from each plant was used for the preparation of sRNA-Seq libraries by all three participating sequencing providers (Ramaciotti Centre for Genomics at the UNSW in Sydney, the Central Analytical Research Facility at QUT in Brisbane and the Australian Genome Research Facility in Melbourne).

Different kind of biases and errors are introduced during sRNA library preparation and sequencing:

- The library construction starts with reverse transcription (RT) of sRNAs, producing complementary DNA (cDNA) for amplification and sequencing. As miRNA molecules are of short length, they are extended by ligation or polyadenylation, which introduce primer-binding sites for RT and subsequent amplification. The extension step, especially the ligation, is considered the most prominent source of bias in sRNA-Seq. Different affinity of adaptors to target molecules causes artificial changes in the true small RNA abundances, which can confound the analysis.
- The amplification step of cDNA molecules represents another source of bias, which originates from the differing efficiencies of the PCR in amplifying molecules of different lengths and secondary structures. This can lead to artificial shifts in the detected signals.
- Index hopping can occur during the sequencing step with a given read being assigned to the wrong library. Unique dual indexing might not be sufficient to mitigate false positives for the new Illumina sequencing machines with patterned flow cells (e.g., NovaSeq6000), especially for samples with heavy viral loads or that are sequenced at high depth.
- Finally, over-representation of ribosomal RNA in the libraries can also confound the analysis, preventing the recovery of low-titer virus signal.

The major differences between the sRNA library kits tested lie in the strategies they use to mitigate these biases:

- **Strategy 1**: NEBNext Multiplex Small RNA Library is the kit we are currently using to process the PEQ samples, and it uses the original two-adaptor ligation method with invariant adapters but adds polyethynol glycol as a strategy to reduce ligation bias. It uses single unique indexing for sequencing. We have found it performed well when used in combination with older generation Illumina sequencer such as NextSeq 550 sequencer (tested in the original 2020 pilot study), but AGRF is moving to the latest Illumina sequencing machines with patterned flow cells (i.e., NovaSeq6000) which are known to be more prone to index hopping. We tested this kit with the new sequencing machine to compare with our original pilot study data.
- **Strategy 2 & 3**: The NEXTFLEX Small RNA-Seq Kit v3 uses randomised adapters (adapters with four degenerate nucleotides at the ligation ends). It can incorporate random unique molecular identifiers (UMIs) to distinguish the reads that were generated by an identical molecule amplified by PCR. This permits to reduce over-representation of specific molecules and ensures the ones present at very low titer are still recovered. We tested this strategy with AGRF. It can also incorporate Unique dual indexing (UDIs) which permits to reduce the likelihood of index hopping, and we tested this strategy with CARF.
- **Strategy 4**: The SMARTer smRNA-Seq Kit is a ligation free method based on polyadenylation with template switching.
- **Strategy 5 and 6**: Finally, QIAseq mostly focuses on mitigating PCR bias and index hopping by incorporating UMIs and UDIs. We chose to test this kit with or without prior treatment with QIAseq FastSelect kit to remove rRNA content.

All libraries were sequenced on a NovaSeq except for the SMARTer smRNA-Seq ones that were sequenced on a MiSeq.

To evaluate the quality of our libraries, we categorised the reads by size and by RNA type after adapter removal and quality filtering. We then used three different bioinformatics approaches to detect and identify viruses and viroids: VirusDetect, VSD in Velvet mode (VSD-Velvet) and VSD in SPAdes mode (VSD-SPAdes).

Table 18.3. List of samples that were sequenced by each different sequencing provider.

Sample	BGI	NEBNext AGRF NextSeq	NEBNext AGRF NovaSeq	Nextflex-UMI AGRF NovaSeq	Nextflex-UDI CARF NovaSeq	Qiaseq rRNA + Ramaciotti NovaSeq	Qiaseq rRNA – Ramaciotti NovaSeq
MT000	✓	✓	✓	✓	✓	✓	✓
MT001	✓	✓	✓			✓	✓
MT002	✓	✓	✓		✓	✓	✓
MT003	✓	✓	✓	✓	✓	✓	✓
MT005	✓	✓	✓		✓	✓	✓
MT007	✓		✓		✓	✓	✓
MT008	✓		✓		✓	✓	✓
MT010	✓	✓	✓	✓	✓	✓	✓
MT012	✓		✓		✓	✓	✓
MT016	✓	✓	✓			✓	✓
Total	10	7	10	3	8	10	10

Major Findings

- The amount of raw sequencing output was highly variable depending on the kit considered. We recovered on average 56M reads per sample processed with BGI chemistry and 28M reads per sample processed with the NEBNext kit and sequenced with NextSeq in the pilot study. In comparison in the ring test, we obtained 102M reads per sample processed with the NEBNext kit and sequenced with NovaSeq, 49M reads per sample processed with the QIAseq kit, 59M reads per sample processed with the NEXTflex kit at AGRF, 109M per sample processed with the NEXTflex kit at CARF and 425k reads per sample processed with the SMARTer smRNA-Seq kit using a MiSeq. The number of sequences obtained with the latter was unfortunately too low to run our usual pipeline and so we were unable to proceed with virus and viroid detection with this specific dataset. Except for Qiaseq for which the raw sequencing output size was consistent across sample (between 40-60 million reads), the other 2 kits (NEBNext and NEXTflex) showed a lot more variability. For instance, some sample output for the NEXTflex kit from CARF were 2.5-fold that of others run on the same flowcell (**Figure 18.1**).
- Efficient analysis of sRNAs is dependent on the fraction of usable reads from our sequencing efforts. After performing adapter removal and quality filtering with Cutadapt, we categorised the initial raw reads into seven categories: 1) untrimmed reads which contained no adapters, 2) low quality inserts or inserts < 5 bp long, 3) inserts 5-17 bp long, 4) inserts 18-24 bp long (other than 21, 22 and 24 bp long), 5) inserts 21, 22 and 24 bp long (which contain siRNA reads we are interested in), 6) inserts 25-31 bp long, and 7) inserts > 31 bp long. We deemed that reads that were belonging to the fourth and fifth categories were informative as they contained reads that match the size of miRNA and siRNA which are targeted by the kits tested, while reads belonging to the other size category were classified as side products (**Figure 18.2 and 18.3**). We estimated that based on read size recovered, 85% of our sequencing effort is informative with the BGI kit, between 30-40% of our sequencing effort is informative for the NEB kit, the QIAseq kit and the NEXTflex kit (AGRF) and less than 20% for the Nextflex and the Takara kit (**Figure 18.4**). Quite a large proportion of reads was untrimmed in the samples processed with the NEXTflex kit at CARF and when we inspected these reads, we found that they contained a lot of chimeric sequences which consisted of a sRNA molecule at the 5'end and a longer sequence (either ribosomal or mitochondrial in

origin) at the 3'end. This suggests that something might have gone wrong during the ligation step for this library preparation.

- Another means of measuring the outcome of our sequencing effort is to check the RNA origin of the sequences we obtain after performing quality filtering. sRNAs can originate from a broad variety of sources (e.g. miRNAs, siRNAs) but also other less informative structural RNAs (e.g. tRNAs, rRNAs, snRNAs and snoRNAs). Therefore we used Bowtie to align any quality filtered reads longer than 5 bp to several sRNA source databases which match different RNA source categories: 1) rRNAs, 2) plant mitochondrial genes (other than rRNAs), 3) plant plastid genes (other than rRNAs), 4) plant miRNAs, 5) plant non-coding genes (snRNA, snoRNA, tasi), 6) plant transposons, 7) sequencing artefacts (PhiX and vector contaminants), 8) plant viruses and viroids, and 9) unclassified left-over. We further separated the rRNA category into 1) 5S rRNAs, 2) small subunit (16s/18S) and large subunit (23S/28S) nuclear rRNAs, 3) mitochondrial rRNAs and 4) chloroplast rRNAs. We found that there was a quite a lot of variability in the RNA content of each library between kit but also within kit (**Figure 18.5, 18.6 and 18.7**). Some kit-specific biases could be observed. For instance, the samples processed at BGI contained more miRNAs and virus derived siRNAs, while QIAseq derived libraries contained slightly more tRNAs and 5S RNAs. We did not observe a similar RNA source profile between samples originating from the same commodity (data not shown). The difference we observe between samples could thus reflect some potential issue with sample processing which drives degradation of larger RNA molecules and then their ultimate incorporation into our sRNA libraries. Ribosomal RNA was quite prevalent across kits trialled and highly variable across samples (**Figure 18.7, Table 18.3**). Ribosomal RNA removal with QIAseq FastSelect kit only removed ~4% of rRNA on average (**Table 18.3**) and so the improvement is probably too marginal to warrant the cost, however we have given this feedback to QIAGEN and they are keen to come up with a custom-built approach, with their research and development team, which might be more successful at removing the over-represented plant sequences that we are specifically detecting in our libraries.
- The Nextflex protocol carried at AGRF achieved lower sensitivity (87.5%) as it failed to detect the presence of Citrus vein enation virus (CVEV) in sample MT003. Otherwise, all other approaches achieved 100% sensitivity, with targets detected by at least one bioinformatics tool tested (**Table 18.4**). We were interested to understand why the Nextflex protocol carried at AGRF failed to recover this virus and we performed an alignment of the filtered reads against the matching CVEV reference genome and derived some coverage statistics. We found that although some reads were recovered for the virus, they were present in too low amount (<500 reads) to assemble when we performed *de novo* assembly. We recovered a mean coverage of less than 2 and only 35% of the genome was covered at 1X (**Table 18.5 and 18.6**).
- The shift from NextSeq500 to NovaSeq sequencing machine seems to have greatly increased the rate of index hopping we observe in the samples processed with the NEBNext kit by AGRF, when we compare the data originating from the pilot study and the ring study **(Table 18.7 and 18.8)**. We provided this feedback to AGRF and they commented that this would be expected as random clustering like the old NextSeq 500/550/HiSeq 2500/MiSeq are less prone to index hoping and that this issue was introduced with patterned flow cells (i.e. modern Illumina instruments such as NovaSeq/NextSeq 2000/HiSeq 4000). With these, index hopping is expected to occur at ~2% frequency (~1.2M reads at 60M depth). Furthermore, NovaSeq has also an increased data yield compared to the NextSeq as can be observed on **Figure 18.1** (we obtained five times more reads on average for the NEBNext samples sequenced with the NovaSeq than with the NextSeq 500). Therefore, we would expect the frequency of index hoping to increase even more, thus increasing the false positive rate as the number of "hopped" reads is sufficient to assemble the target viruses and viroids. Random clustering allowed for variations in data output so you could reduce the sequencing capacity as required if needed, whereas the NovaSeq flow-cells now have relatively fixed outputs. Furthermore, the NEBNext kit caps the number of samples that can be processed to 48 whereas other kits are more flexible. Unfortunately, this kit does not offer at present the option to incorporate UDIs which could mitigate the index hopping issues observed.
- Of all the kits tested, only the QIAseq approach showed no index hopping events. Some level of contamination was detected in all other kits **(Table 18.7 and 18.8)**. The QIAseq and the NEXTflex approach carried at CARF were the only 2 strategies trialled which used UDIs. It seems that for instruments which use patterned flow cell such as NovaSeq it is paramount to incorporate UDIs. It is likely that index hopping was still detected with the NEXTflex approach because of the much higher depth at which some of the samples were sequenced; on average twice as much data was generated in the NEXTflex approach than in the QIAseq, as shown on **Figure 18.1** (with an average of 49 M reads recovered per sample for the

samples generated with the QIAseq kit versus 109M for the ones generated with the NEXTflex kit at CARF).

Recommendations

- To prevent cross sample contamination (index hoping events), we recommend using QIAGEN's unique dual indexing (UDIs) for the library preparation of small RNA-seq data. Other commercial UDIs tested in the past did not efficiently prevent the presence of index hoping events.
- Use the QIAseq miRNA library preparation kit with both UMIs and UDIs to mitigate false positives and remove PCR artifacts.
- Partner with QIAGEN's Research and Development team to improve the QIAseq FastSelect product to remove rRNAs more efficiently and other over-represented non-informative sequences.

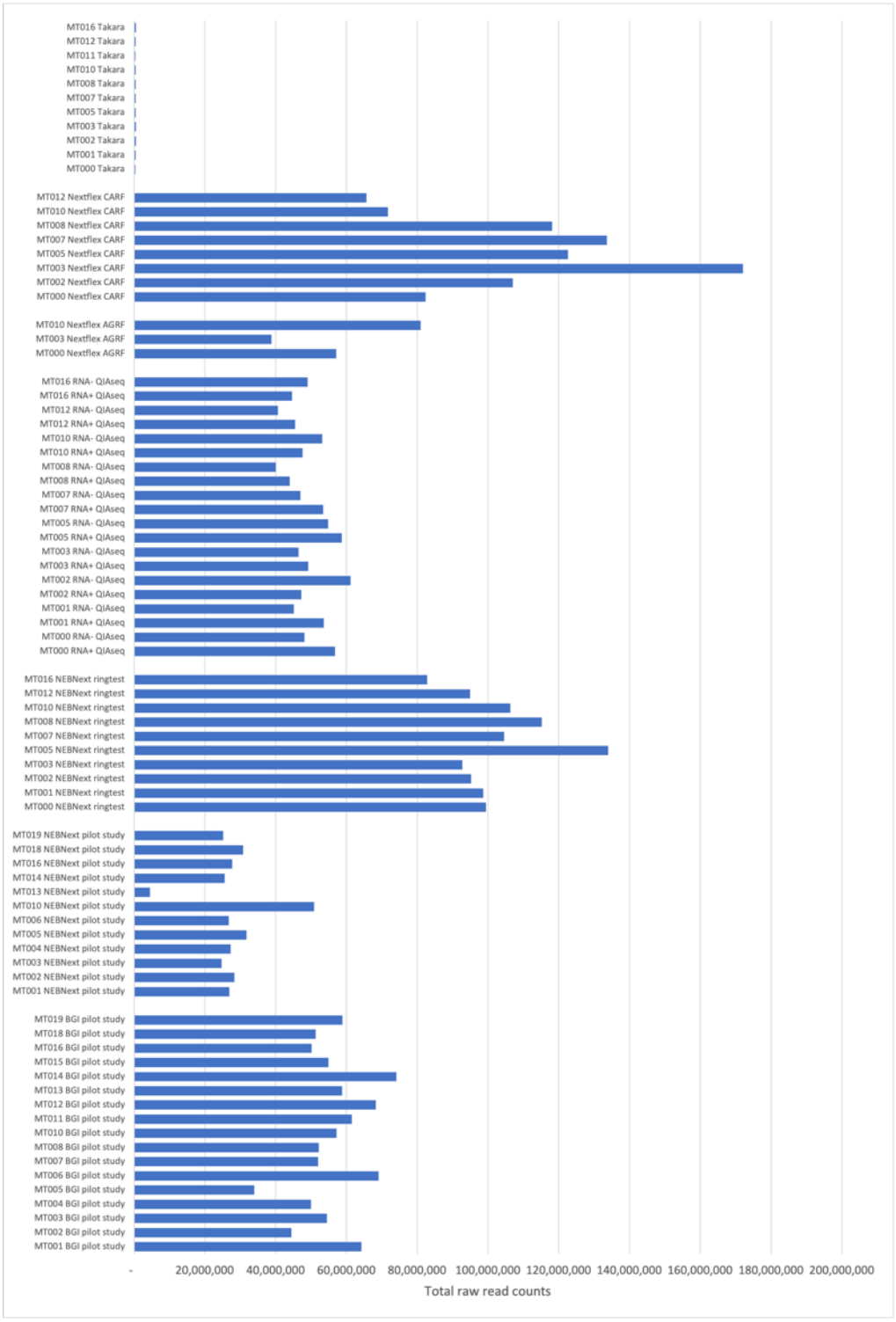


Figure 18.1. Total raw read counts recovered per sample for each sRNA library kit tested. The samples from the original study are included at the bottom of the figure for comparison.

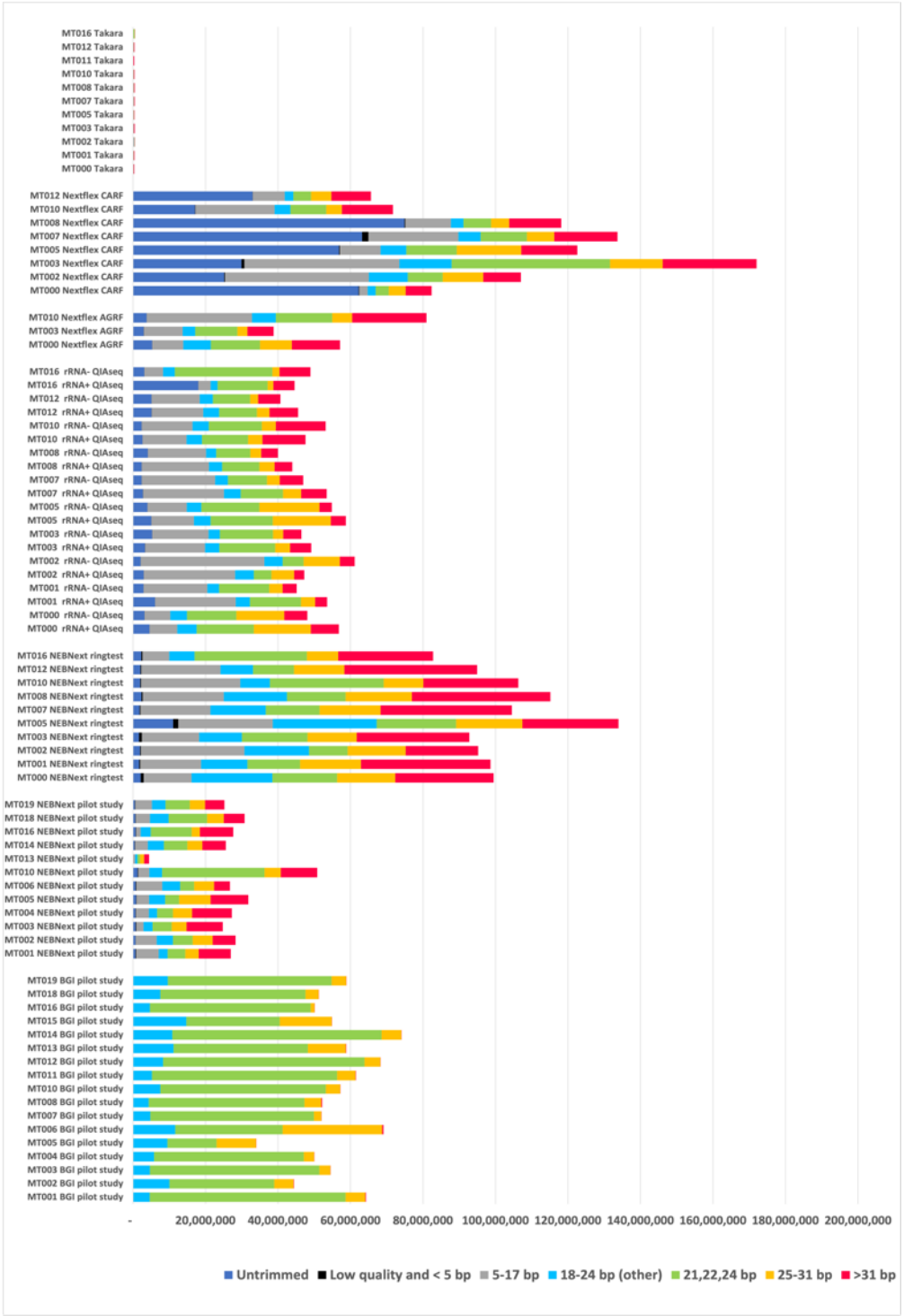


Figure 18.2. **Total number of informative reads and side-products in HTS samples**. Based on the results of adapter and quality filtering, the initial raw reads were subdivided into seven categories: (1) untrimmed reads which contained no adapters, (2) low quality inserts or inserts < 5 bp long, (3) inserts ≥ 5 bp and < 18 bp long, (4) inserts ≥18 and ≤ 24 bp long other than 21, 22 and 24 bp long, (5) inserts 21, 22 and 24 bp long, (6) inserts ≥ 25 and ≤ 31 bp and (7) inserts > 31 bp long. Given de size distribution of miRNAs and siRNAs, the fourth and fifth categories were considered to contain informative reads, while the others were deemed to contain side products.

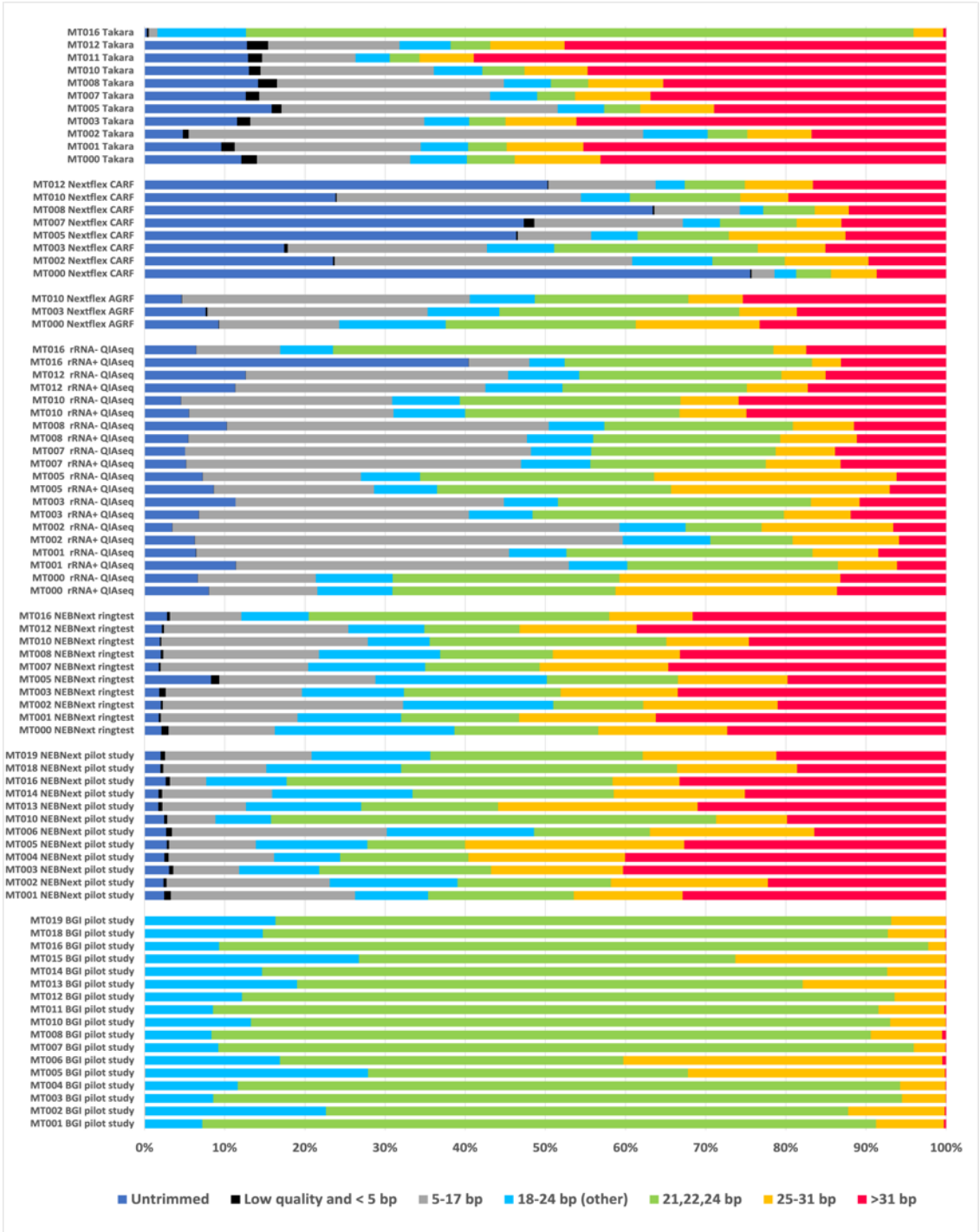


Figure 18.3. Percentage of informative reads and side-products in HTS samples. Based on the results of adapter and quality filtering, initial raw reads were subdivided into seven categories: (1) untrimmed reads which contained no adapters, (2) low quality inserts or inserts < 5 bp long, (3) inserts ≥ 5 bp and < 18 bp long, (4) inserts ≥18 and ≤ 24 bp long other than 21, 22 and 24 bp long, (5) inserts 21, 22 and 24 bp long, (6) inserts ≥ 25 and ≤ 31 bp and (7) inserts > 31 bp long. Given de size distribution of miRNAs and siRNAs, the fourth and fifth categories were considered to contain informative reads, while the others were deemed to contain side products.

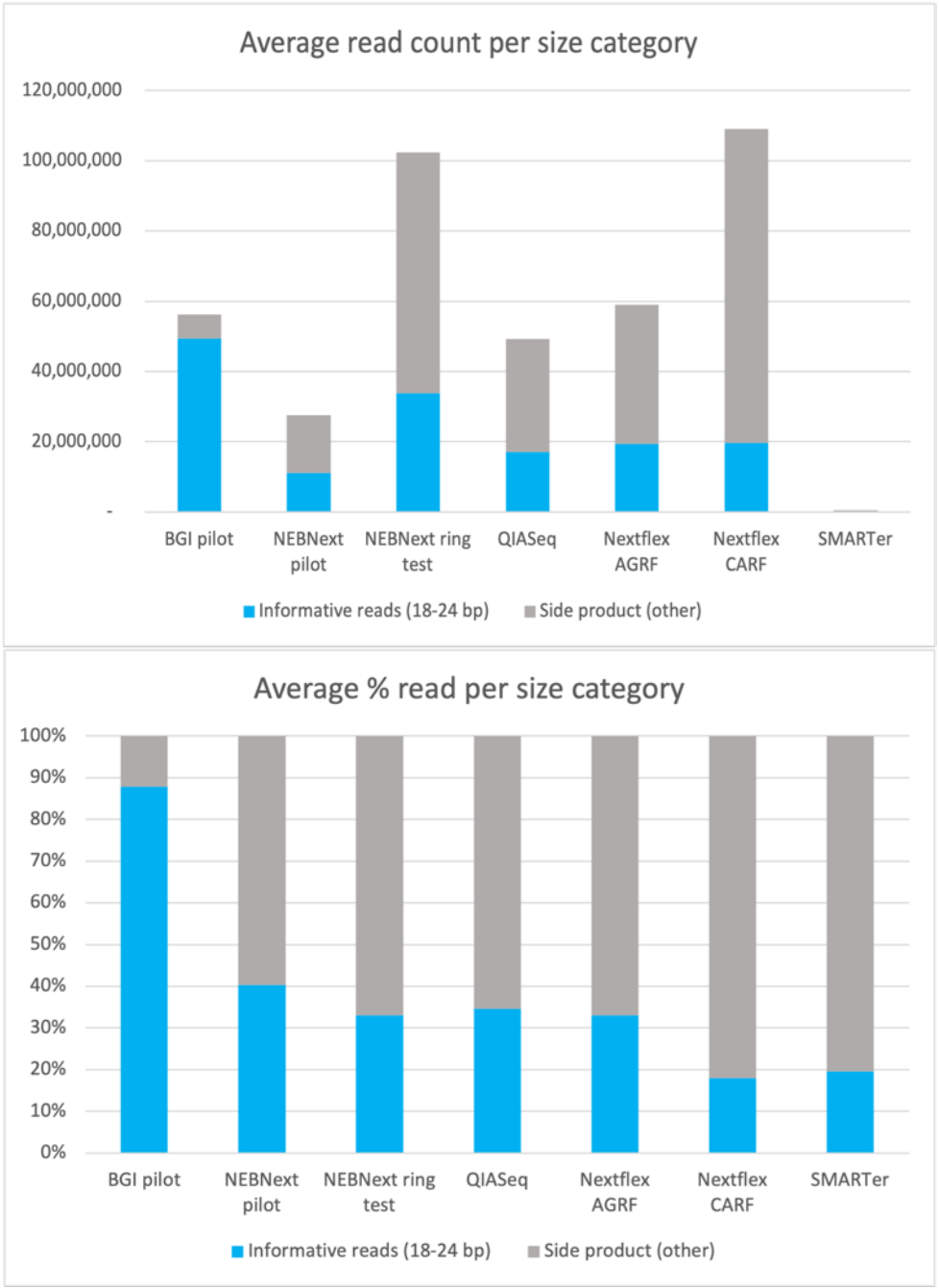


Figure 18.4 Distribution of small RNA-seq reads. (Top panel) Average informative read count recovered per commercial kit. (Bottom panel) Average % read recovered per commercial kit.

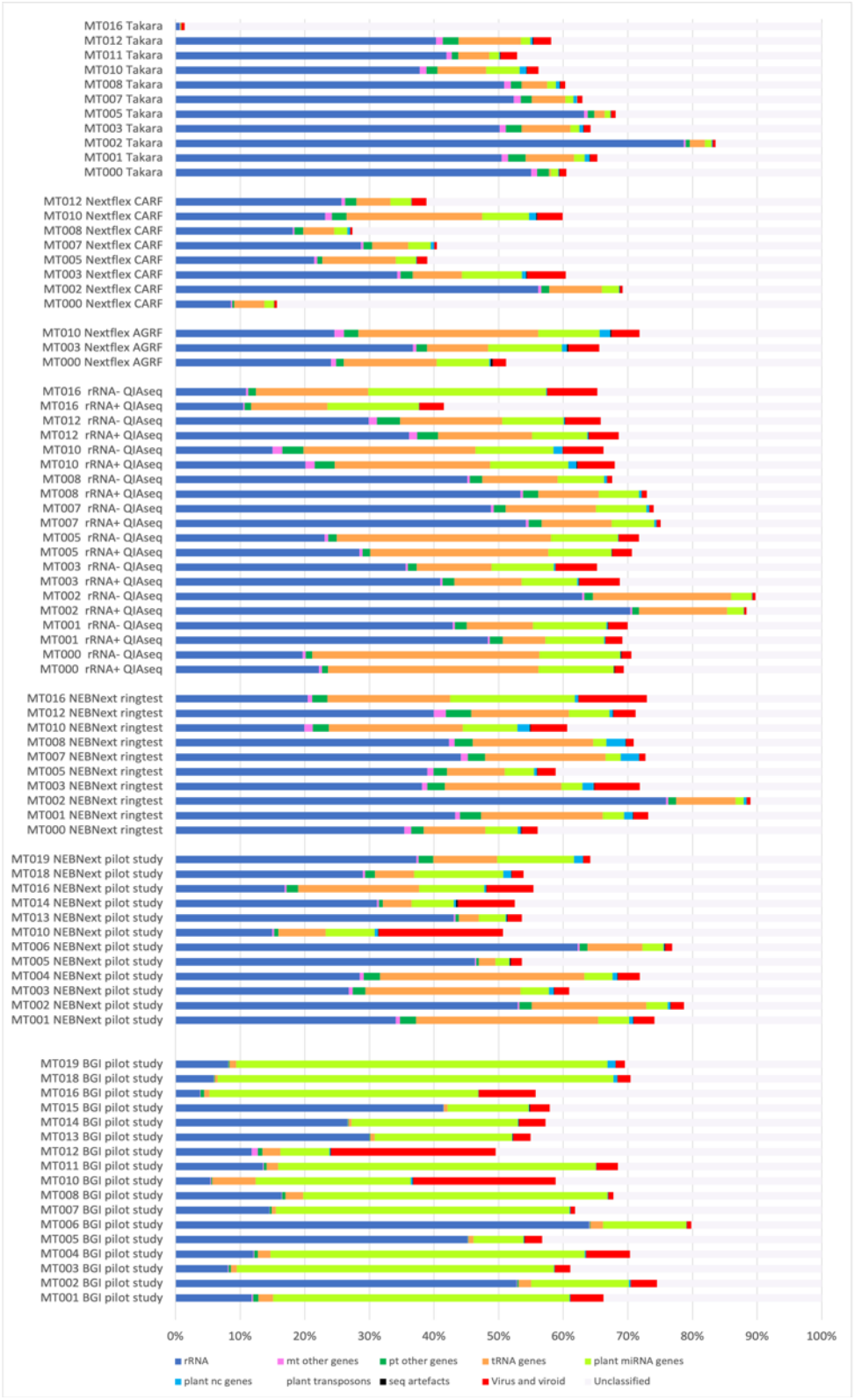


Figure 18.5. Percentage of quality filtered reads longer than 5 bp that match different RNA sources. RNA sources include: 1) rRNAs, 2) plant mitochondrial genes (other than rRNAs), 3) plant plastid genes (other than rRNAs), 4) plant miRNAs, 5) plant non-coding genes (snRNA, snoRNA, tasi), 6) plant transposons, 7) sequencing artefacts (PhiX and vector contaminants), 8) plant viruses and viroids, and 9) unclassified left-over.

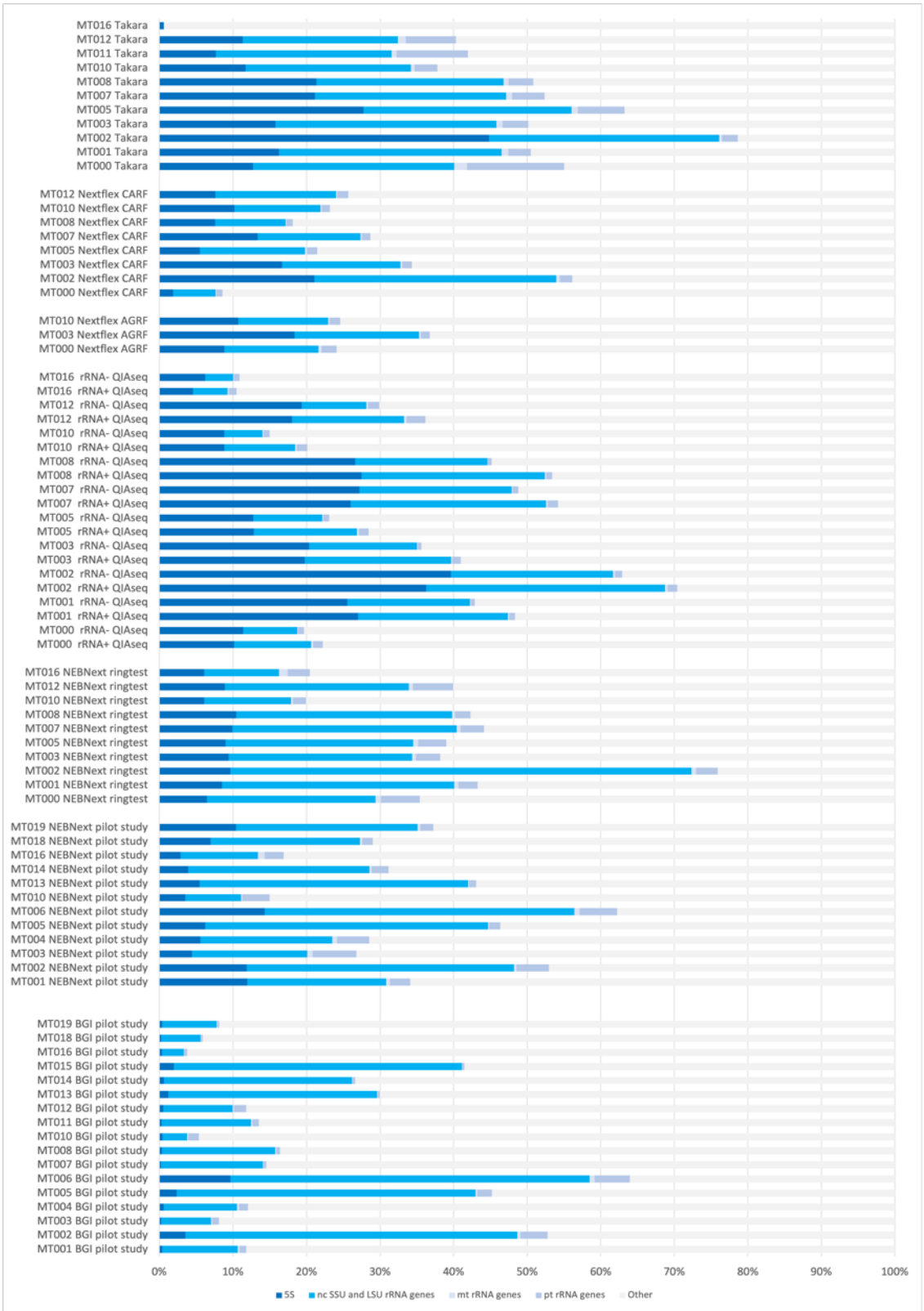


Figure 18.6. Percentage of quality filtered reads longer than 5 bp that match different ribosomal RNA source categories. We further separated the rRNA category from Figure 18.5 into: 1) 5S rRNA, 2) small subunit (16S/18S) and large subunit (23S/28S) nuclear rRNA, 3) mitochondrial rRNA and 4) chloroplast rRNA.

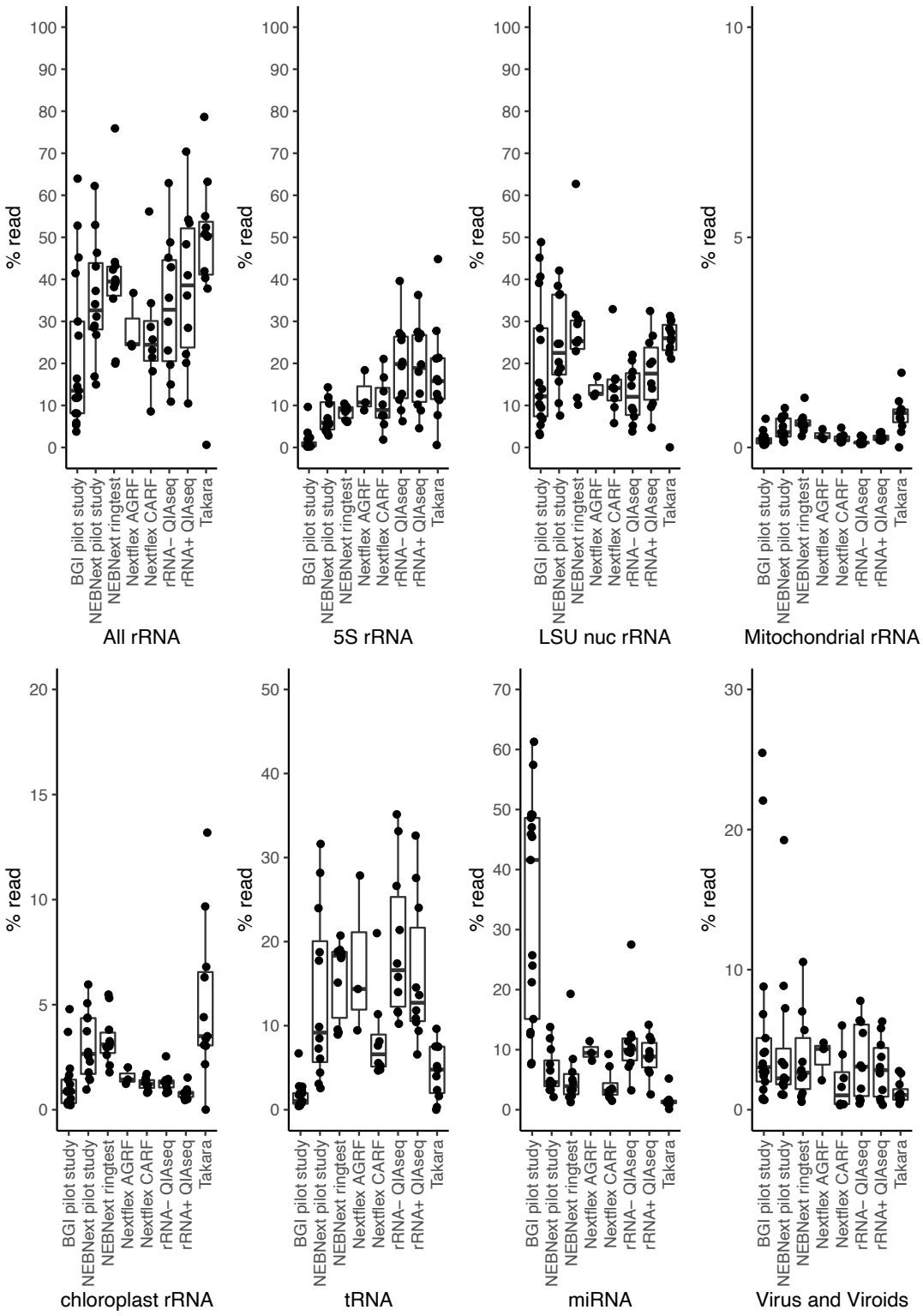


Figure 18.7. Distribution of RNA sources by service provider – library prep kit and approach

Table 18.3 Average percent ribosomal RNA content per kit trialled.

Kit/Seq provider	% rRNA read
BGI pilot study	21.8
NEBNext pilot study	35.3
NEBNext ring test	39.8
QIAseq RNA+	38.5
QIAseq RNA-	33.4
Nextflex AGRF	28.5
Nextflex CARF	27.0
Takara	47.4

Table 18.4. Sensitivity for each kit tested. The presence of a virus/viroid was considered as present in a kit if at least one bioinformatics tool detected it.

Sample	Virus/Viroid	BGI	NEBNext AGRF NextSeq	NEBNext AGRF NovaSeq	Nextflex-UMI AGRF NovaSeq	Nextflex-UDI CARF NovaSeq	Qiaseq rRNA + Ramaciotti NovaSeq	Qiaseq rRNA – Ramaciotti NovaSeq
MT000	endoRYNV	✓	✓	✓	✓	✓	✓	✓
	RBDV	NA	NA	✓	✓	✓	✓	✓
MT001	CEVd	✓	✓	✓	✓		✓	✓
MT002	PNRSV	✓	✓	✓	✓	✓	✓	✓
MT003	CTV	✓	✓	✓	✓	✓	✓	✓
	CVEV	✓	✓	✓	×	✓	✓	✓
	CVd-III	✓	✓	✓	✓	✓	✓	✓
	HSVd	✓	✓	✓	✓	✓	✓	✓
MT005	endoRYNV	✓	✓	✓	✓	✓	✓	✓
	RBDV	✓	✓	✓	✓	✓	✓	✓
MT007	CVd-III	✓		✓	✓	✓	✓	✓
	HSVd	✓		✓	✓	✓	✓	✓
MT008	CVd-III	✓		✓	✓	✓	✓	✓
	HSVd	✓		✓	✓	✓	✓	✓
MT010	SgMV	✓	✓	✓	✓	✓	✓	✓
MT012	ISMV	✓		✓	✓	✓	✓	✓
	TRSV	✓		✓	✓		✓	✓

MT016	SPFMV	✓	✓	✓	✓		✓	✓
Total number of targets detected/present		17/17	11/11	18/18	8/9	15/15	18/18	18/18
Sensitivity (%)		100	100	100	87.5	100	100	100

Table 18.5. Comparison of mean coverage and per cent base coverage at 1X for all viral targets across all kits

		NEBNext (AGRF)		Nextflex-UMI (AGRF)		Nextflex-UDI (CARF)		Qiaseq rRNA + (Ramaciotti)		Qiaseq rRNA - (Ramaciotti)	
Sample	**Virus/Viroid**	**Mean cov**	**% bases 1X**	**Mean cov**	**% bases 1X**	**Mean cov**	**% bases 1X**	**Mean cov**	**% bases 1X**	**Mean cov**	**% bases 1X**
MT000	endoRYNV	46.5	64.0	38.2	34.0	14.2	19.0	50.2	54.0	48.1	53.0
	RBDV-RNA1	97.0	93.0	75.6	74.0	45.9	67.0	109.4	94.0	99.4	94.0
	RBDV-RNA2	141.7	98.0	115.5	85.0	84.9	80.1	157.2	99.0	147.9	99.0
MT001	CEVd	238.8	100.0					239.5	96.0	236.7	100.0
MT002	PNRSV-RNA1	8.9	80.0			22.8	95.0	11	88.0	17.9	92.0
	PNRSV-RNA2	8.9	79.0			23.1	93.0	11.9	88.0	19.4	91.0
	PNRSV-RNA3	102.1	99.0			184.7	100.0	129	99.0	159.5	100.0
MT003	CTV	239.8	100.0	220.9	100.0	247.6	100.0	236.2	100.0	235.9	100.0
	CVEV	2.1	52.6	1.6	34.8	6.9	47.2	3.0	60.0	2.8	65.0
	CVd-III	224.0	97.0	167.7	81.0	183.2	81.0	232.3	98.0	233.2	97.0
	HSVd	171.9	99.0	182.5	99.0	242.3	100.0	205.3	99.0	204.7	100.0
MT005	endoRYNV	199.2	95.0			198.3	87.0	202.5	95.0	201	95.0
	RBDV-RNA1	195.9	96.0			205.5	93.0	195	97.0	194.3	97.0
	RBDV-RNA2	214.5	100.0			222.7	97.0	214.4	99.0	211.4	99.0
MT007	CVd-III	186.0	97.0			193.4	97.0	198.6	97.0	201.6	100.0
	HSVd	150.4	100.0			189.3	98.0	172.8	99.0	143.4	94.0
MT008	CVd-III	220.9	100.0			193.2	97.0	224.1	98.0	223.5	98.0
	HSVd	222.4	100.0			138.1	100.0	164.9	88.0	217.8	98.0
MT010	SgMV	177.5	79.0	46.7	31.0	42	28.0	40.6	30.0	41.4	29.0
MT012	ISMV	166.3	99.0			166.2	100	188.4	100.0	196.1	100.0
	TRSV-RNA1	187.7	96.0					187.5	94.0	199.7	93.0
	TRSV-RNA2	219.0	98.0					222.2	98.0	229.8	98.0
MT016	SPFMV	247.3	100.0					217.3	100.0	235.8	100.0

Table 18.6. Comparison total read counts and read counts per million (RPM) for all viral targets across all kits

		NEBNext (AGRF)		Nextflex-UMI (AGRF)		Nextflex-UDI (CARF)		Qiaseq rRNA + (Ramaciotti)		Qiaseq rRNA - (Ramaciotti)	
		Read #	RPM	Read #	RPM	Read #	RPM	Read #	RPM	Read #	RPM
MT000	endoRYNV	55,123	658	21,382	412	6804	391	62,261	1412	55,714	1487
	RBDV-RNA1	42,740	510	26,515	511	12,651	727	49,650	1126	43,371	1158
	RBDV-RNA2	44,223	528	18952	365	10,672	614	46,607	1057	35,207	940
MT001	CEVd	1,114,223	13,898					828,509	33,155	788,003	32,331
MT002	PNRSV-RNA1	1400	22			3592	87	1735	92	2832	115
	PNRSV-RNA2	1084	17			3064	74	1466	78	2369	96
	PNRSV-RNA3	11,670	180			30,660	740	15,378	815	25,206	1021
MT003	CTV	4,360,319	58,394	1,085,400	30294	6,497,914	66,679	2,133,217	73,524	2,042,601	80,410
	CVEV	608	8	464	19	1935	20	852	29	783	31
	CVd-III	10,622	142	6623	185	22,820	234	27,825	959	26,738	1053
	HSVd	3923	53	5349	149	21,205	218	12,219	421	11,745	462
MT005	endoRYNV	484,607	4695			371,632	6948	530,321	12,786	520,603	13,124
	RBDV-RNA1	348,809	3379			350,728	6557	374,193	9022	369,682	9319
	RBDV-RNA2	279,126	2704			196,094	3666	199,237	4804	165,009	4160
MT007	CVd-III	3332	40			5464	126	8219	293	7816	325
	HSVd	3043	36			4237	98	5720	204	5230	217
MT008	CVd-III	6324	70			4179	139	14,363	630	14,538	740
	HSVd	13,838	153			2365	79	19,328	848	23,622	1202
MT010	SgMV	72,233	938	383,656	4967	257,342	7963	340,614	10,498	363,836	10,006
MT012	ISMV	32,531	456			140,127	5958	278,755	10,756	302,576	13,766
	TRSV-RNA1	337,051	4721					404,754	15,618	288,713	12276
	TRSV-RNA2	323,879	4536					300,970	11,613	226,320	9623
MT016	SPFMV	6,520,025	88,434					1,087,848	47,417	2,558,661	63,589

Table 18.7. Details of index hopping events detected in each sample. Blue ticks = BGI pilot study, pink ticks = NEBNext NextSeq AGRF pilot study, black ticks = NEBNext NovaSeq (AGRF ring test), red ticks = Nextflex-UMI (AGRF), green ticks = Nextflex-UDI (CARF).

Virus/Viroid	MT000	MT001	MT002	MT003	MT005	MT007	MT008	MT010	MT012	MT016
CEVd	✓		✓✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓✓	✓
CTV	✓		✓		✓	✓		✓✓	✓	
CVd-I	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CVd-III	✓✓							✓✓		
HSVd	✓		✓	✓	✓			✓✓	✓	✓✓
PRNSV				✓				✓		
RBDV				✓						
SrMV	✓		✓	✓			✓	✓	✓	
SPFMV	✓	✓	✓	✓	✓	✓	✓	✓	✓	
TRSV	✓		✓	✓		✓	✓✓	✓✓		✓

Table 18.8. Specificity for each kit tested measured by the number of false positives detected by at least one bioinformatics tool. The grey cells show the samples which were not analysed for a given kit.

Sample	BGI	NEBNext AGRF NextSeq	NEBNext AGRF NovaSeq	Nextflex-UMI AGRF NovaSeq	Nextflex-UDI CARF NovaSeq	Qiaseq rRNA + Ramaciotti NovaSeq	Qiaseq rRNA – Ramaciotti NovaSeq
MT006	2	0	4	4	1	0	0
MT001	0	0	2			0	0
MT002	5	1	3		0	0	0
MT003	3	0	4	2	0	0	0
MT005	1	0	5		0	0	0
MT007	1	0	4		0	0	0
MT008	3	0	3		1	0	0
MT010	7	0	5	2	0	0	0
MT012	3	0	4		0	0	0
MT016	1	1	3			0	0
Total #	26	2	37	8	2	0	0

Appendix 19: Large scale side-by-side PEQ and HTS testing of diverse plant commodities.

Overall, a total of 199 plants representing eleven plant genera were tested using PEQ and HTS protocols. These included overall 82 Prunus, 38 Rubus, 53 Citrus, 9 Persea, 5 Malus, 4 Pyrus, 3 Rosa, 2 Solanum and as single plant for Olea, Lonicera and Ficus (**Table 19.1**).

Total RNA extracts for 199 quarantined plants were collected and submitted to two HTS service providers for sRNA-seq HTS data generation, namely:

- Batch 1 consisted of 90 plants and was sequenced by AGRF in Melbourne.
- Batch 2 consisted of 89 plants and was sequenced by BGI in Hong Kong.
- Only 20 plants from Batch 3 have been analysed at the time this report was drafted. These were sequenced at AGRF in Melbourne.

Data generation

- Batch 1 comprised of 90 samples was submitted to AGRF in November 2020 and HTS data was received on January 2021.
- Batch 2 comprising 89 samples was initially submitted to AGRF in December 2020, but due to a shortage in supplies caused by the COVID19 pandemic AGRF was not able to progress with service agreement. BGI Hong Kong was then selected as an alternative sequencing provider and batch 2 samples were sent to BGI in February 2020 and receiving data by the end of April 2021.
- Batch 3 samples: i) Citrus samples (n=20) collected at NSW DPI and data generated by AGRF by end of June; ii) PEQ samples (n=48) data received from AGRF by middle of August 2021.

Table 19.1. List of commodities that were sequenced in each sequencing batch.

Commodity	Batch 1	Batch 2	Batch 3	Total
Prunus	80	2		82
Rubus	1	37		38
Citrus	5	28	20	53
Malus	0	5		5
Persea	0	9		9
Pyrus	0	4		4
Rosa	3	0		3
Solanum	0	2		2
Lonicera	1	0		1

Ficus	0	1		1
Olea	0	1		1
Total	**90**	**89**	**20**	**199**

HTS service providers assessed the RNA integrity of submitted samples prior enrichment of the small RNA fraction, library preparation and HTS sequencing. Overall, a total of 10.3 billion HTS raw reads were generated for 199 samples. In average, 48.97, 44.1and 96.34 million raw reads were produced for individual samples in batch 1, 2 and 3, respectively. Raw reads for each sample were analysed using a newly implemented VSD pipeline in a Nextflow platform.

Major Findings

- Out of 199 imported plants tested, 31 plants (15.6% detection rate) were confirmed positive for the presence of 49 quarantinable viruses and viroids and 10 plants for 10 non-quarantinable viruses and viroids by PEQ or DPI protocols. The sRNA-Seq HTS method identified all of these pathogens in the same plants (**Tables 19.2 and 19.4**).
- We initially failed to detect five viroids present at low titer when we used the *de novo* assembler SPAdes 3.14 but these were all recovered when we upgraded our pipeline to use Velvet instead (**Table 19.4**).
- The HTS method detected CVEV in six citrus plants that were initially negative by PEQ testing. Follow up using MinION sequencing confirmed four of these pathogens were present in the samples while two samples still need to be validated (**Tables 19.3 and 19.4**).
- The HTS method detected 7 additional viruses which match Prunus latent virus and Nectarine stem pitting associated virus. These were not tested for detection by PEQ at the time of testing and PEQ is waiting to receive feedback from policy regulators as to whether they are a biosecurity concern PEQ has developed primers for Prunus latent virus and confirmed its presence in the samples in which they were detected via HTS (**Table 19.6**).
- The HTS method detected the presence of Rubus Yellow Net Virus (RYNV) in 18 plants (**Table 19.3 and 19.7**). Only two of these detections overlapped with PEQ by woody indexing. Assessment of sRNA-Seq HTS data from plants infected with RYNV showed an atypical small RNA mapping patterns displaying deep but patchy coverage of a reference RYNV genome suggesting that these may likely be host genome integrated sequences (**Figure 19.1**). Personal communication with Dr. Thien Ho (DRISCOLL) confirmed that a small subset of cultivar plants are progenitors of most commercial raspberries and are known to harbour up to six endogenous RYNV integrated sites in their genome. Endogenous RYNV do not cause plant disease, and this is supported by the negative woody indexing obtained by PEQ for most of these plants. Interestingly the only two plants that gave a positive woody indexing also harbour another virus, Raspberry bushy dwarf virus (RBDV), and thus it is likely to be the presence of this other virus and not RYNV, that might cause the symptoms detected by woody indexing in these plants. Secondary confirmation is currently performed on DNA and target conserved regions of the RYNV that are also found in endoRYNV. Thus, PEQ is looking at designing primers which could specifically identify the presence of endoRYNV versus epiRYNV.
- The HTS method detected 43 index hopping events in batches that were run at AGRF (23 were present in Batch 1 and 20 in Batch 3) (**Table 19.3 and 19.8**). These could clearly be identified based on read abundance with our index hopping flag method (see **Appendix 17).** These events originate from samples that have a particularly heavy load of a given pathogen (for example MT091 for the Peach latent mosaic viroid in Batch 1). The shift from NextSeq500 to NovaSeq6000 sequencing instrument seems to have greatly increased the rate of index hopping we observe in the samples processed with the NEBNext kit by AGRF **(Appendix 15)**. We provided this feedback to AGRF, and they commented that this would be expected as random clustering like the old NextSeq 500/550/HiSeq 2500/MiSeq are less prone to index hoping and that this issue was introduced with patterned flow cells (i.e., modern Illumina instruments such as NovaSeq/NextSeq 2000/HiSeq 4000). With these, index hopping is expected to occur at ~2% frequency (~1.2M reads at 60M depth). Furthermore, NovaSeq has also an increased data output which means the frequency of index hoping is expected to increase even more in a run.

Table 19.2: Comparison of HTS and PEQ/DPI detections of quarantinable and non-quarantinable viruses. Only plants confirmed by PEQ/DPI tests are shown.

Method	Status	Number of plants		Number of viruses/viroids	
		Quarantinable	Non-quarantinable	Quarantinable	Non-quarantinable
PEQ	positive	31	10	49	10
HTS	positive	31	10	49	10

Table 19.3: Additional detections by sRNA-Seq HTS method. *Denotes RYNV detections by HTS that were negative by woody indexing assays. 42 suspected index hopping events (sample cross contamination) from quarantinable or non-quarantinable viruses.

Method	Status	Number of plants			Number of viruses/viroids		
		Quarantinable	Non-quarantinable	Index hopping	Quarantinable	Non-quarantinable	Index hopping
PEQ	negative	0	0	0	0	0	0
HTS	positive	20 (18*)	0	41	20	0	42

Table 19.4. Overlap between HTS and PEQ/DPI. *These detections were only detected when our pipeline was upgraded to use Velvet instead of SPAdes 3.14 to perform *de novo* assembly; ^these CVEV detections were initially missed by PEQ but subsequently confirmed to be correct by MinION sequencing. Two of these CVEV detections still need to be validated (see **Table 3**).

Sample	Commodity	Pathogen name	Category
MT020	Prunus	Peach associated luteovirus	Quarantinable
MT091	Prunus	Peach latent mosaic viroid	Quarantinable
MT092	Prunus	Peach associated luteovirus	Quarantinable
MT094	Prunus	Cherry virus F	Quarantinable
MT098	Prunus	Cherry virus F*	Quarantinable
MT166	Citrus	Citrus tristeza virus	Quarantinable
MT167	Citrus	Citrus tristeza virus	Quarantinable
MT167	Citrus	Citrus viroid III*	Quarantinable
MT168	Citrus	Citrus tristeza virus	Quarantinable
MT168	Citrus	Citrus vein enation virus^	Quarantinable

MT168	Citrus	Hop stunt viroid	Quarantinable
MT169	Citrus	Citrus tristeza virus	Quarantinable
MT169	Citrus	Hop stunt viroid	Quarantinable
MT175	Citrus	Citrus tristeza virus	Quarantinable
MT175	Citrus	Citrus vein enation virus^	Quarantinable
MT176	Citrus	Citrus bent leaf viroid*	Quarantinable
MT176	Citrus	Citrus tristeza virus	Quarantinable
MT176	Citrus	Citrus viroid III*	Quarantinable
MT177	Citrus	Citrus tristeza virus	Quarantinable
MT177	Citrus	Citrus vein enation virus^	Quarantinable
MT177	Citrus	Citrus viroid III*	Quarantinable
MT184	Citrus	Citrus tristeza virus	Quarantinable
MT184	Citrus	Citrus vein enation virus^	Quarantinable
MT164	Malus	Citrus concave gum associated virus	Quarantinable
MT193	Malus	Citrus concave gum associated virus	Quarantinable
MT195	Malus	Citrus concave gum associated virus	Quarantinable
MT222	Citrus	Citrus tristeza virus	Quarantinable
MT223	Citrus	Citrus tristeza virus	Quarantinable
MT224	Citrus	Citrus tristeza virus	Quarantinable
MT225	Citrus	Citrus tristeza virus	Quarantinable
MT226	Citrus	Citrus tristeza virus	Quarantinable
MT227	Citrus	Citrus tristeza virus	Quarantinable
MT228	Citrus	Citrus tristeza virus	Quarantinable
MT229	Citrus	Hop stunt viroid	Quarantinable
MT229	Citrus	Citrus viroid III	Quarantinable
MT229	Citrus	Citrus tristeza virus	Quarantinable

MT230	Citrus	Citrus tristeza virus	Quarantinable
MT230	Citrus	Citrus vein enation virus	Quarantinable
MT230	Citrus	Citrus viroid VII	Quarantinable
MT231	Citrus	Citrus exocortis viroid	Quarantinable
MT232	Citrus	Citrus tristeza virus	Quarantinable
MT232	Citrus	Citrus viroid I LSS	Quarantinable
MT233	Citrus	Citrus viroid I LSS	Quarantinable
MT233	Citrus	Citrus tristeza virus	Quarantinable
MT234	Citrus	Citrus viroid II	Quarantinable
MT235	Citrus	Citrus viroid III	Quarantinable
MT235	Citrus	Citrus tatter leaf virus	Quarantinable
MT235	Citrus	Citrus tristeza virus	Quarantinable
MT236	Citrus	Citrus viroid V	Quarantinable
MT236	Citrus	Citrus viroid VI	Quarantinable
MT100	Rosa	Prunus necrotic ringspot virus	Non-quarantinable
MT101	Rosa	Prunus necrotic ringspot virus	Non-quarantinable
MT102	Rosa	Prunus necrotic ringspot virus	Non-quarantinable
MT039	Prunus	Prune dwarf virus	Non-quarantinable
MT040	Prunus	Prunus necrotic ringspot virus	Non-quarantinable
MT093	Prunus	Cherry virus A	Non-quarantinable
MT205	Rubus	Raspberry bushy dwarf virus	Non-quarantinable
MT230	Citrus	Citrus viroid VII	Non-quarantinable
MT237	Citrus	Citrus viroid VII	Non-quarantinable
MT238	Citrus	Citrus virus A	Non-quarantinable

Table 19.5 Additional pathogens found by HTS which were not detected by PEQ and who still need to be checked.

Sample	Commodity	Pathogen name	Category
MT169	Citrus	Citrus vein enation virus	Quarantinable
MT176	Citrus	Citrus vein enation virus	Quarantinable

Table 19.6 Additional pathogens found by HTS which were not targeted by PEQ at time of testing and whose status is now under review. PEQ subsequently developed PCR primers and confirmed the presence of Prunus latent virus A in MT029, MT031 and MT032.

Sample	Commodity	Pathogen name	Category
MT029	Prunus	Prunus latent virus (previously Prunus geminivirus A)	Likely of biosecurity concern
MT031	Prunus	Prunus latent virus (previously Prunus geminivirus A)	Likely of biosecurity concern
MT032	Prunus	Prunus latent virus (previously Prunus geminivirus A)	Likely of biosecurity concern
MT055	Prunus	Nectarine stem pitting associated virus	Likely non-quarantinable
MT063	Prunus	Nectarine stem pitting associated virus	Likely non-quarantinable
MT067	Prunus	Nectarine stem pitting associated virus	Likely non-quarantinable
MT088	Prunus	Nectarine stem pitting associated virus	Likely non-quarantinable

Table 19.7. Rubus yellow net virus detections detected by HTS which are predicted to be endogenous. The samples with an asterisk were also detected as positive for RYNV by woody indexing and PCR by PEQ

Sample	Commodity	Pathogen name	Category
MT116	Rubus	Rubus yellow net virus	Quarantinable
MT117	Rubus	Rubus yellow net virus	Quarantinable
MT118	Rubus	Rubus yellow net virus	Quarantinable
MT119	Rubus	Rubus yellow net virus	Quarantinable
MT120	Rubus	Rubus yellow net virus	Quarantinable
MT121	Rubus	Rubus yellow net virus	Quarantinable

MT122	Rubus	Rubus yellow net virus	Quarantinable
MT123	Rubus	Rubus yellow net virus	Quarantinable
MT139	Rubus	Rubus yellow net virus	Quarantinable
MT140	Rubus	Rubus yellow net virus	Quarantinable
MT143	Rubus	Rubus yellow net virus	Quarantinable
MT144	Rubus	Rubus yellow net virus	Quarantinable
MT145	Rubus	Rubus yellow net virus	Quarantinable
MT202	Rubus	Rubus yellow net virus	Quarantinable
MT204	Rubus	Rubus yellow net virus	Quarantinable
MT205*	Rubus	Rubus yellow net virus	Quarantinable
MT206*	Rubus	Rubus yellow net virus	Quarantinable
MT207	Rubus	Rubus yellow net virus	Quarantinable

Table 19.8. Index hopping detected by HTS

Sample	Commodity	Pathogen name	Category
MT029	Prunus	Peach latent mosaic viroid	Quarantinable
MT041	Prunus	Peach latent mosaic viroid	Quarantinable
MT042	Prunus	Peach latent mosaic viroid	Quarantinable
MT050	Prunus	Peach latent mosaic viroid	Quarantinable
MT051	Prunus	Peach latent mosaic viroid	Quarantinable
MT053	Prunus	Peach latent mosaic viroid	Quarantinable
MT062	Prunus	Peach latent mosaic viroid	Quarantinable
MT065	Prunus	Peach latent mosaic viroid	Quarantinable
MT067	Prunus	Peach latent mosaic viroid	Quarantinable
MT075	Prunus	Peach latent mosaic viroid	Quarantinable
MT076	Prunus	Peach latent mosaic viroid	Quarantinable
MT077	Prunus	Peach latent mosaic viroid	Quarantinable
MT078	Prunus	Peach latent mosaic viroid	Quarantinable
MT079	Prunus	Peach latent mosaic viroid	Quarantinable
MT085	Prunus	Peach latent mosaic viroid	Quarantinable
MT086	Prunus	Peach latent mosaic viroid	Quarantinable
MT088	Prunus	Peach latent mosaic viroid	Quarantinable

MT089	Prunus	Peach latent mosaic viroid	Quarantinable
MT090	Prunus	Peach latent mosaic viroid	Quarantinable
MT101	Rosa	Peach latent mosaic viroid	Quarantinable
MT102	Rosa	Peach latent mosaic viroid	Quarantinable
MT103	Citrus	Peach latent mosaic viroid	Quarantinable
MT222	Citrus	Citrus exocortis viroid	Quarantinable
MT223	Citrus	Citrus exocortis viroid	Quarantinable
MT224	Citrus	Citrus exocortis viroid	Quarantinable
MT225	Citrus	Citrus exocortis viroid	Quarantinable
MT226	Citrus	Citrus exocortis viroid	Quarantinable
MT227	Citrus	Citrus exocortis viroid	Quarantinable
MT228	Citrus	Citrus exocortis viroid	Quarantinable
MT229	Citrus	Citrus exocortis viroid	Quarantinable
MT230	Citrus	Citrus exocortis viroid	Quarantinable
MT231	Citrus	Citrus tristeza virus	Quarantinable
MT234	Citrus	Citrus viroid I-LSS	Quarantinable
MT235	Citrus	Citrus viroid I-LSS	Quarantinable
MT236	Citrus	Citrus viroid I-LSS	Quarantinable
MT237	Citrus	Citrus viroid I-LSS	Quarantinable
MT238	Citrus	Citrus viroid I-LSS	Quarantinable
MT239	Citrus	Citrus viroid I-LSS	Quarantinable
MT240	Citrus	Citrus viroid I-LSS	Quarantinable
MT241	Citrus	Citrus viroid I-LSS	Quarantinable
MT091	Prunus	Nectarine stem pitting associated virus	Likely non-quarantinable
MT232	Citrus	Citrus viroid VII	Non-quarantinable
MT241	Citrus	Citrus viroid VII	Non-quarantinable

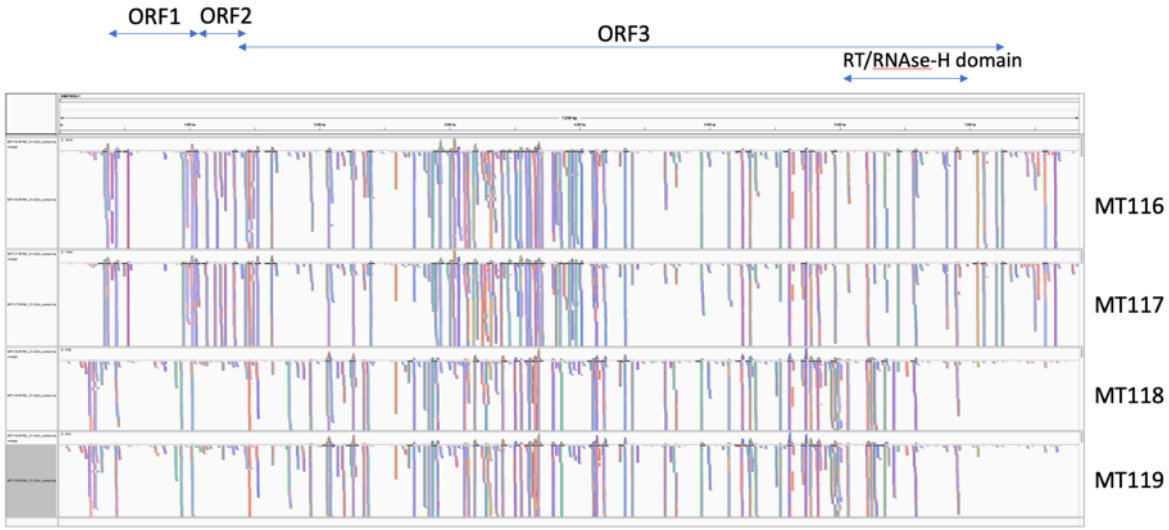


Figure 19.1. Mapping of small RNA reads onto the reference RYNV genome (Accession KM078034.1) shows scattered coverage despite the presence of large read peaks.

Appendix 20: A High Throughput Sequencing (HTS) framework, policies, and principles for plant diagnostics

DRAFT REPORT

[31 August 2021]

Prepared for QUT by

Lois Ransom PSM

LRM Consulting

M: +61466327114

E: loisransom@yahoo.com

A High-throughput Sequencing (HTS) framework, policies, and principles for plant diagnostics

Executive summary

This report guides the development of policies and processes for the adoption of HTS technology for routine phytosanitary screening of exotic viruses and viroids in plants in post entry quarantine as part a collaborative research project (MT18005) by the Queensland University of Technology, Agriculture Victoria and the Australian Government Department of Agriculture, Water and the Environment, with funding from Hort Innovation. Screening for regulated pests verifies the health status of the imported plant material and provides assurance that it achieves the appropriate level of biosecurity protection applied by Australia before being released for entry into the country.

This report considers best practice policy guidelines for the development, validation, and adoption of High-throughput Sequencing (HTS) diagnostic technologies across human, animal and plant health sectors and proposes a hierarchical policy framework and principles for the modular adoption of HTS technologies for plant pests of biosecurity concern. This framework provides a platform for standard operating procedures (SOPs), quality management and assurance, and guidelines for interpreting HTS diagnostic results, which provides confidence in the diagnostic results produced. Standards and guidelines will be informed by the research program and aligned with relevant best practices as they currently exist across the health sector.

The report is presented in three parts:

- **Part 1** summarises high level HTS policy recommendations for plants pest diagnostics and other relevant domains, with reference to HTS framework documents including the National Biosecurity Committee HTS report, Animal and Human HTS guidelines, and relevant overseas literature.
- **Part 2** presents a policy framework for the adoption and implementation of HTS technologies in plant pest diagnostics for PEQ testing and screening. It identifies key policy principles for integrating HTS technologies into existing operations that manage plant biosecurity risks. These principles underpin the equivalence of HTS technologies in general, as well as specifically sRNA-seq HTS, for detecting regulated viruses and viroids in high-risk plants for planting with existing PEQ diagnostic tools.
- **Part 3** guides the effective implementation of sRNA-seq HTS technology into PEQ testing protocols to detect viruses and viroids in selected high-risk hosts. It does this through a hierarchy of documents that outline the methodology and quality controls and assurances validated by research as equivalent to existing screening protocols. Screening for regulated pests in stonefruit (*Prunus* spp.) is used to illustrate the transition into operations of the HTS platform.

There is a high degree of consistency and synergy across health sectors for best practices in using HTS technologies. All sectors consider their use as a diagnostic tool should be subject to laboratory quality management processes that assure confidence in the test outcome, as they do for other test platforms. The key elements in best HTS practice have informed both the policy principles for HTS in plant PEQ and the policy framework.

The consistency of HTS platforms, application processes, drivers and controls highlight synergies and challenges, and offers opportunities for cross-sectoral collaboration. The use of bioinformatics differentiates HTS from other diagnostic processes. The selection, application and accurate interpretation of bioinformatics packages and outputs is critical to the diagnostic outcome achieved. Bioinformatics expertise will likely become a rate-limiting factor in the further evolution and application of HTS across all sectors that will need to be addressed. Similarly, data management demands are specific to the HTS platform. They demand super-computer analysis capability, and significant storage capacity. IT capacity and capability can be shared across sectors and functions, such as biosecurity, but will likely require further investment or this too will come a rate-limiting factor for HTS development and application.

Background

The Commission on Phytosanitary Measures (CPM) of the International Plant Protection Convention (IPPC) recognised in a policy statement [FAO, 2016] that accurate and rapid pest diagnosis underpins export certification, import inspections and the application of appropriate phytosanitary treatments, enables effective pest surveillance, and supports successful eradication programs. This is also recognised by the World Organisation for Animal Health (OIE).

The ongoing global development of diagnostics technologies presents significant opportunities to animal, plant, and human health practitioners, but there are also risks in adopting new technologies without considering and defining the policies, processes, and technical parameters for their appropriate and effective use.

Next generation sequencing (NGS) or as it is more recently known, high-throughput sequencing (HTS) is a diagnostic tool that has been evolving rapidly since the early 2000s. It relies on and has been enabled by super-computer technology and availability. The opportunities arising from the application of HTS vary across human, animal and plant health, but the methodology and controls around the use of the technology across these sectors are highly consistent. This is driven by the imperative for diagnostic outcomes to be accurate at the level of confidence required by the diagnostician.

This report was commissioned as part of a project being undertaken by the Queensland University of Technology with funding from Horticulture Innovation. The project is focussed on "Improving plant industry access to new genetics through faster and more accurate diagnostics using next generation sequencing" [Ref. https://www.horticulture.com.au/growers/help-your-business-grow/research-reports-publications-fact-sheets-and-more/mt18005/]. A key outcome is the adoption of HTS in the screening of imported horticultural plant material in post-entry quarantine facilities. Previous research enabled the adoption of HTS in the testing of imported ornamental grasses. The current project is providing the evidence and operational policies and protocols needed for the technology to be fully integrated into routine PEQ screening of viruses and viroids in imported plants.

This paper will inform the "revision of the end user guides and harmonisation with international best practices and standards documenting end-to-end SOPs, quality assurances and guidelines for interpretation of NGS diagnosis results for regulatory purposes completed and submitted to DAWE and NZ MPI to enable the adoption of NGS technology for routine phytosanitary screening of exotic viruses and viroids in PEQ." It is presented in three parts:

Part 1 Summarises high level HTS policy recommendations for plants pest diagnostics and other relevant domains, with reference to HTS framework documents including the National Biosecurity Committee HTS report, Animal and Human HTS guidelines, and relevant overseas literature.

Part 2 Presents a policy framework for the adoption and implementation of HTS technologies in plant pest diagnostics for PEQ testing and screening. It identifies key policy principles for integrating HTS technologies into existing operations that manage plant biosecurity risks. These principles underpin the equivalence of HTS technologies in general, as well as specifically sRNA-seq HTS, for detecting regulated viruses and viroids in high-risk plants for planting with existing PEQ diagnostic tools.

Part 3 Guides the effective implementation of sRNA-seq HTS technology into PEQ testing protocols to detect viruses and viroids in selected high-risk hosts. It does this through a hierarchy of documents that outline the methodology and quality controls and assurances validated by research as equivalent to existing screening protocols. Screening for regulated pests in stonefruit (*Prunus* spp.) is used to illustrate the transition into operations of the HTS platform.

Part 1: HTS best practice policy guidelines

International best practices and standards documenting end-to-end SOPs, quality assurances and guidelines for interpretation of NGS diagnosis results for regulatory purposes.

Background

HTS, bioinformatics and computational genomics is a powerful tool that enables the rapid and cost-effective simultaneous sequencing of potentially all organisms present in a sample. In short, DNA or RNA is extracted from the sample, cut into shorter lengths, and replicated. These fragmented sequences of nucleic acids are assembled and reassembled many times. The combinations are read, and statistical modelling maps the assembled sequences to reference libraries of known genomes. When used as a diagnostic tool, the HTS output and analysis enables the probability of origin of the extracted DNA or RNA to be determined to a level of statistical confidence.

The sample may be drawn from animal, including humans, and plant tissues, soil, water and microbial cultures. HTS can, without any *a priori* knowledge on the sample infection status, detect the nucleic acids of any organism, including variants and uncharacterised organisms present in a sample that might be potential threats to animal, human and/or plant health and biosecurity.

The purposes for which HTS is used are varied and provide greater or lesser opportunities across the human, animal and plant health and biosecurity sectors. They include but are not limited to:

- HTS as a tool that together with case and outbreak information assists the investigation of health issues
- Targeting detection and identification of organisms causing infectious diseases, as well as pests and pathogen vectors, from specimens *in vitro* and *in vivo*
- Confirmatory assays of organisms detected in some other primary assay
- Providing additional characterisation of organisms
- Detecting, identifying and characterising previously unidentified organisms
- Improving diagnosis of known organisms
- Improving diagnosis of emerging or re-emerging pests and diseases, including those that have a known, changed or unknown etiology and resultant epidemiology
- Developing a single 'universal' diagnostic assay, able to identify any potential pathogen (more difficult in plant pests given the number and range of organisms)
- Rapidly and simultaneously detecting multiple organisms in disease complexes
- Increasing capability to study evolutionary dynamics of pests and pathogens at farm, local, national and global levels. This has been applied to rapidly characterise and detect strains of the COVID-19 virus

- Better understanding the epidemiology of pests and diseases and the phylogenetic relationships of the same or similar organisms
- Enhancing the traceability of pests and diseases and modes of disease transmission
- Enabling more extensive characterisation of populations of known pathogens to facilitate effective control measures, such as the application and conservation of antimicrobials
- Better linking the genotype and phenotypes of organisms through full genome sequencing of multiple strains of a single species.

There are a range of HTS technologies being used, and more being developed. They all comprise a similar series of steps that are critical to the quality and accuracy of the diagnostic outcome and confidence with which the outcome can be used. Many of the steps require specific and specialised expertise, so training and knowledge transfer as well as publication of information, analyses, methodologies, and outcomes is critical to grow the use and rigour of this technology. Sharing knowledge and experience between practitioners and across the health sectors, will facilitate the effective evolution and uptake of this technology.

The bioinformatic analysis assembling the organism genome sequences from the raw data and the subsequent secondary analysis are the critical elements in HTS. Laboratories need to adopt policies and practices for generating analysing and managing genomic sequence data that are based on accurate information and rigorous interpretation. These will direct the generation, management and use of genomic sequence data within accepted practices of health and biosecurity investigations and within a laboratory's quality assurance system.

The three health sectors are continuing to develop and refine guidelines and standards for the application of HTS technologies. In general, the human health sector is more advanced and can offer guidance to other sectors from the lessons that have been learned. This experience has helped to establish a reasonably consistent platform for the development, validation, adoption and implementation of HTS technologies across the sectors. Some of these are described below, with a comparison of the key elements in using HTS across the health sectors presented in Table 1.

Animal health

The World Organization for Animal Health (OIE) considers HTS technologies as a diagnostic tool that assists animal health and food safety investigations, which should be used in accordance with standards for laboratory testing, like any other laboratory tool or procedure. In keeping with this consideration, the Chapter 1.1.7 of the Manual of Diagnostic Tests and Vaccines for Terrestrial Animals [OIE, 2021] was developed to assist laboratories by defining standards that will allow inclusion of this capability into a laboratory's scope of operations in a way in which users of the results can have confidence. The adoption of end-to-end standard operating procedures and best practices are key to reliably detect pests and diseases.

Human health

Several bodies have developed guidelines for integrating the use of HTS technologies in human health-driven diagnostics. These include the EuroGentest and European Society of Human Genetics Guidelines for NGS-based diagnostics (Matthijs et al, 2015), which guide the evaluation and validation of NGS applications for the diagnosis of genetic disorders, and the US Food and Drug Administration (FDA) Guidelines for NGS-based diagnostics (2018). The FDA regulatory vision for the regulatory oversight of this technology is that tests can be developed, validated, and offered for clinical use through a streamlined approach that leverages appropriate standards, quality systems controls and community assessment of clinical validity. The FDA guidelines were developed from a series of public workshops in 2015-16 that explored key aspects of the quality and performance of NGS technologies in practice. They provide a framework for designing, developing and validating these tests, that are consistent with those for plant and animal health laboratories, and make reference to the European guidelines.

An ISO standard has been developed to guide 'Quality control metrics for DNA sequencing, which provides a basis for sharing and for the application of "high quality" genomic data in the human health sector [ISO, 2020].

Plant health

Application of HTS technologies in the plant health sector is progressing but an international standard for the application and use of HTS is not yet in place. A CPM Recommendation [FAO, 2019] has been adopted and provides guidance to countries proposing to use HTS technologies and their results as the basis for appropriate phytosanitary regulations. As a contracting party to the IPPC, Australia is expected to follow this Recommendation.

A set of guidelines on HTS for plant health laboratories that parallel those in animal and human health and biosecurity sectors has been developed in Europe [Lebas and Massart, 2020]. They provide recommendations to guide plant health diagnostic laboratories on the selection, development, and optimisation of HTS tests, on their validation and verification, their routine application in plant pest diagnostics, including the use of internal and external quality checks, the interpretation and reporting of HTS test results. They are relevant to all HTS technologies for detection and identification of any plant pest from any source material and for any purpose.

There are critical steps for applying HTS in plant health that are common across the health sectors, and which provide opportunities for harmonisation of process, quality assurance, expertise and application. These are:

(i) Sampling
(ii) Nucleic acid extraction
(iii) Library preparation
(iv) Sequencing
(v) Analysis of raw reads
(vi) Identification of target(s)
(vii) Analysis of controls
(viii) Confirmation
(ix) Interpretation, and reporting
(x) Data management
(xi) Data sharing
(xii) Validations

Cross-sector harmonisation of HTS use in Australia

At the request of the National Biosecurity Committee, a national workshop was convened in Brisbane in February 2020 to (i) identify challenges using HTS and (ii) outline the minimum quality assurance standards required to use HTS in a biosecurity context. [NBC, 2020]. Subject matter experts represented the human health, animal and plant biosecurity sectors including HTS data service providers and policy developers, both state and federal.

The workshop participants identified synergies and challenges faced by each sector and emerging quality assurance themes drawing together a summary of collated experiences. The use of HTS across the sectors is at different levels of maturity, but its application is expanding rapidly.

The participants defined minimum quality standards for the application of HTS in the areas of sample preparation, laboratory processing, data analysis, storage and interpretation from a regulatory context, and proposed actions to implement a national biosecurity HTS framework enabling biosecurity diagnostics, surveillance and discovery.

The ongoing and rapid evolution of HTS technology means that prescribed operating standards quickly lose relevance or require ongoing updating. This impacts on laboratories seeking to maintain any national accreditations in the delivery of HTS-based diagnostic testing. The workshop outcomes recommend the development of a national HTS standard of practice rather than a single national standard of operation, that would be informed by the experiences of the human health sector. The HTS standards framework would be regularly reviewed and applied as the basis for harmonising use of the technology across the biosecurity sector. This would likely generate efficiency, underpin confidence in the performance and outcomes of HTS-based testing and provide a foundation for innovation and further harmonisation through shared experiences.

A HTS Standards framework should cover areas of common practice including, but potentially not limited to:

- Principles for the adoption and use of HTS to manage biosecurity risks
- Data generation
- Data storage, including a national repository or network of repositories
- Data analysis
- Data sharing
- Validation of new tests and platforms
- Terminology via a glossary of terms
- Quality assurance programs and potentially national accreditation standards
- Staff competencies and training.

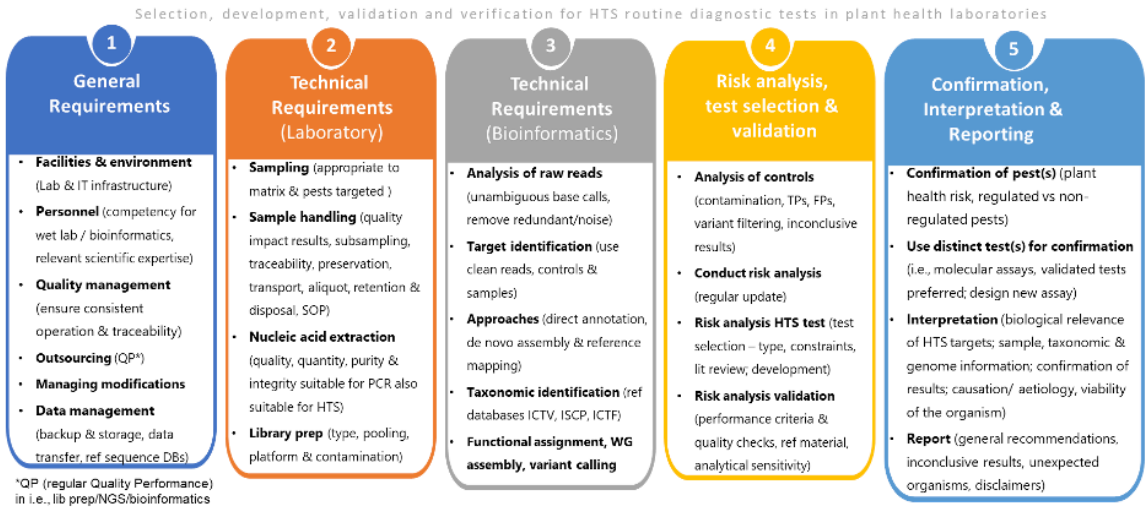
Recommendations for specific actions arising from the workshop included appropriate sampling for HTS; bioinformatic – supporting infrastructure and analysis; results interpretation; definition of a positive species identification; defining and re-analysis of unknown sequence data; staff skills, ISO and NATA standards and networking laboratories between and across health and biosecurity sectors.

The potential contribution to these actions by the plant biosecurity sector is outlined in Attachment 20.1, with reference to proposed standard operating procedures for integrating HTS into pathogen screening operations at the Australian Government post entry quarantine facility in Victoria.

High level policy elements

As noted above, there is a great deal of synergy and consistency in policy and operations for the use of HTS technologies in plant, animal and human health and biosecurity. The NBC workshop identified and explored many of these.

The following diagram summarises the key elements for selecting, developing, validating and verifying for routine HTS-diagnostic tests in plant health laboratories [Barrero *sensu* Lebas et al]. They should be reflected in the operating procedures for each HTS diagnostic test adopted as the basis for the confidence in the test result.



Policy elements for **choosing** to use an HTS-based diagnostic test are not inconsistent with choosing an appropriate diagnostic test method using other technology platforms including:

- The purpose and scope of the test or tests
- The efficacy, robustness, accuracy of the method – from validation against equivalent standard testing methods
- Test selection
- The target organism(s) eg. plant viruses and viroids diagnosis
- Their status as regulated or established organisms of interest
- The needs of the client eg. PEQ or industry
- The material being tested and its biological and epidemiological relationship to the target organism
- The level of diagnostic confidence expected and any additional confirmatory testing
- Any ethical or legal considerations that impact the choice or use of the test
- Disclosing the test results

Operationalising the use of HTS-based diagnostic tests require the same rigours of quality management, assurance and controls as other diagnostic platforms undertaken in laboratories. They include:

- The facilities necessary to undertake the testing and analysis of results – fit for purpose and applying best practice
- The right samples are taken in the right way, at the right time and handled in the right way for the best diagnostic outcome
- Staff undertaking the testing are competent, trained, have relevant scientific expertise, and are accredited as such
- Training and accreditation that is maintained and tested at appropriate intervals and frequencies
- End-to-end standard operational procedures are established, published, and followed
- Quality management is applied to all steps in the diagnostics processes from sample selection to biometric analysis and data management to reporting results
- Laboratory proficiency testing is undertaken to verify competency
- Any outsourcing of HTS-based diagnostic test functions is subject to regular quality performance assessment and audit
- A program of continuous improvement is applied to ensure changes to technology and methodology are assessed and integrated as appropriate
- Data is backed up, stored, curated and data systems are maintained to optimise analysis, security, and reporting
- ISO or NATA accreditation is gained and maintained for specific HTS tests, as appropriate

The **bioinformatics analysis and interpretation** are critical to the HTS platform. Demonstrating and assuring the efficacy and reliability of these analyses, their supporting pipeline tools, and technologies, as well as the experience and expertise of bioinformaticians underpins client confidence in the test results. This requires:

- Selection and effective application of the best pipeline for the test, as directed by the purpose
- The same care and rigour to selecting the genomic libraries and reference materials relevant to test purpose, method, and diagnostic outcome
- High level capability to read and interpret the outcomes of sequencing
- Ongoing training and mentoring of staff to ensure their skills are current and to foster innovation through peer and professional networks

Table 1: HTS framework elements across animal, plant and human health sectors.

Plant	Human	Animal #
Provides advice on use of HTS technologies to ensure any phytosanitary measures that may be applied are based on valid and accurate diagnostic results. As an international policy recommendation [IPPC Secretariat, 2019], protocols and outcomes are expected to be shared to promote harmonisation. Plant guidelines are in Lebas and Massart, 2020.	Guidelines have different structures but cover the same areas of use, performance, quality control and assurance, interpretation and reporting and validation. [US FDA, 2018] [Mattijs et al, 2015]	The OIE Chapter 1.1.7 of the Terrestrial Manual outlines [OIE, 2021] the use of HTS, bioinformatics and computational genomics as a laboratory tool or procedure that, is subject to standards for laboratory testing. Guidance defines standards to underpin confidence in the results obtained from this technology.
Standardise and apply best-practice operational guidelines for HTS including: · Interpretation of results · Consistent, harmonised implementation · Quality controls to ensure robust and accurate HTS data outputs that have biological significance in a phytosanitary context Plant guidelines cover: (i) Sampling (ii) Nucleic acid extraction (iii) Library preparation (iv) Sequencing (v) Analysis of raw reads (vi) Identification of target(s) (vii) Analysis of controls	Purpose, scope, specimen type Sample tracking and barcoding Performance needs Test elements and methods Standard operating procedures Detailed sample preparation Control samples Quality metrics: · Raw data base quality · Coverage (read depth and completeness) · Test run parameters · Performance thresholds · Test performance for distinct genetic samples Reporting: Relevant patient identification diagnosis, assay description, summary of results, major findings Policies for unsolicited data	· Sampling and reporting · Specimen and sample preparation · Generation of sequence data · Bioinformatics · Data management · Quality assurance · Interpretation of results

Plant	Human	Animal #
(vi) Confirmation, interpretation, and reporting		
Implement HTS training programmes, including online best laboratory practice courses		
Implement and maintain appropriate infrastructure and investments in Information Technology and bioinformatics, and education and training on bioinformatics, data storage and interpretation of test results	Bioinformatics detailed pipeline Analysis pipeline specific to gene panel and validation Diagnostic lab database for quality management of platform, assays and samples Platform-specific analysis pipeline [ISO, 2020]	Selection of technology platform – for the purpose it is being used
Validate reliability and accuracy of HTS by comparison trials against other diagnostic platforms used	Performance: · Accuracy · Precision (reproducibility and repeatability) · Limit of detection · Analytical specificity – interference, cross-reactivity, cross-contamination Validation studies, including revalidation after modifications Assess test limits Policy for variants	Validation of test systems for designated purposes to confirm sensitivity and specificity SOPs for all procedures: · Sampling · Sequencing · Bioinformatics · Reporting
Coordinate international proficiency testing to independently assess laboratory capability		
Communicate information on the interpretation of HTS results and resultant conclusions relevant to managing phytosanitary risks	Reporting - Relevant patient identification diagnosis, assay description, summary of results, major findings Policies for unsolicited data Variants database for each disease Diagnostics vs research: diagnostics to answer a clinical question of a single patient Research – hypothesis driven	
Publish protocols for HTS platforms Share guidelines and training material for transparency		
Publish information on the unexpected biological associations of quarantine organisms in plants and plant products that are revealed by HTS	Dissemination policy on unsolicited findings Variants reported to public databases	
Publish and apply guidelines to direct any phytosanitary actions if an unknown or non-viable organism is detected in plant material		

Operational implementation of NGS for animal diseases has been developed by the Laboratories for Emergency Animal Diseases Diagnosis and Response (LEADDR) in Australia https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/animal/ahl/next-generation-sequencing-guidelines.pdf .

Part 2: Policy framework and principles for the adoption of HTS technologies to manage plant biosecurity risks.

A policy framework for the adoption and implementation of HTS technologies in plant pest diagnostics for PEQ testing and screening.

Background

Phytosanitary measures are applied to imported plants and plant products, and the packaging and conveyances that carry them, to prevent the introduction of plant pests into areas where they do not occur and where they would cause significant economic and/or environmental damage if they were present.

The International Plant Protection Convention (IPPC) [FAO, 1997] is an international treaty that aims to protect cultivated and wild plants by preventing the introduction and international spread of plant pests. Its governing body, the Commission on Phytosanitary Measures (CPM), comprises all 180+ contracting parties and adopts International Standards for Phytosanitary Measures (ISPMs) and CPM Recommendations as a way of facilitating safe trade through harmonisation of the policies, processes and measures applied by trading partners to manage phytosanitary risks.

Post-entry quarantine is a phytosanitary measure that is applied by Australia to contain high risk plants for planting until their phytosanitary status is confirmed to be free of regulated pests. Australia is a signatory to the World Trade Organization Agreement on the Application of Sanitary and Phytosanitary Measures (the SPS Agreement) and the IPPC, so must meet rights and obligations of these treaties when applying phytosanitary measures. This includes the requirements that measures are risk based, technically justified and least trade restrictive to protect plant, animal and human health and life.

DAWE is responsible for ensuring that imported high risk plants for planting are released from PEQ only when they have been tested for regulated or quarantine pests of biosecurity concern, in this case predominantly diseases, and confirmed to be free.

The current suite of pest assays used during PEQ has evolved as new technologies have become available [Whattam et al, 2021]. In the earliest days of PEQ, plants were screened visually for disease symptoms. Biological assays were subsequently integrated into the PEQ screening tool kit, using virus indicators plants and woody indexing to elicit disease symptoms in expressive hosts. Electron microscopy is used to detect and characterise virus particles as indicators of pathogens and this has been complemented with molecular and serological assays using ELISA and, more recently, PCR. PEQ screening generally applies broad screening assays (visual, bioassays) to detect all potential pathogens as (knowns and unknowns) as well as pathogen specific assays to confirm the presence or absence of known regulated pests of the host plant.

Two thirds of all plant diagnostic tests conducted at PEQ are applied to detect over 250 regulated virus and viroids. Over 200 molecular tests are deployed annually to detect their presence in imported plant material, and more are being added as risk analyses are undertaken or reviewed. These molecular tests are effective but require significant operator-handling and are not conducive to improved efficiencies offered by automation and robotics. Importantly, these targeted molecular tests detect only those pathogens for which the tests have been designed for, so the risk of missing new and emerging biosecurity threats is real.

The development of high throughput sequencing (HTS) technologies offers another valuable tool for the rapid, accurate and cost-effective screening of plants for regulated pests during PEQ. This technology offers a transformational opportunity for plant PEQ operations by integrating a single small RNA-sequencing (sRNA-seq) HTS diagnostic test early in the PEQ period to detect and provisionally confirm the absence of regulated viruses and viroids in imported high-risk plants at the Australian Government Post Entry Quarantine (PEQ) facility at Mickleham in Victoria. The DAWE HTS development pathway is described in Attachment 20.2. Knowledge and experience gained along this pathway will inform the policy framework for general and specific application of HTS technologies. Many of the key requirements that should be captured in the framework are outlined in Attachment 20.1.

A policy framework is proposed for the adoption and implementation of HTS technologies in plant pest diagnostics, using the screening and detection of viruses and viroids in high-risk plants for planting at the PEQ plant facility at Mickleham, Victoria as a reference. The framework uses the documentation hierarchy structure in Figure 20.1. This hierarchy structure is used by the International Organization for Standardization (ISO) and has been applied in Australia for emergency response policy and operations.

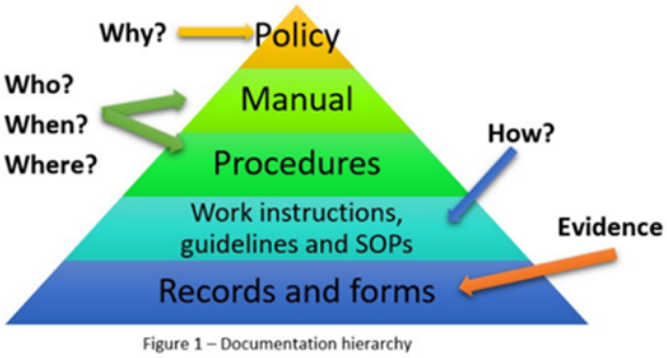


Figure 1 – Documentation hierarchy

Figure 20.1. Documentation hierarchy

The proposed hierarchy includes policy principles for integrating HTS technologies into existing plant PEQ operations consistent with IPPC policy on using HTS technologies to apply phytosanitary measures [FAO, 2019]. The principles are targeted for use at the PEQ facility for the routine use of HTS for virus and viroid detection, regardless of the type of HTS technology used. They do not yet apply to other organisms but will likely inform the validation and adoption of HTS platforms for regulated fungi, bacteria, nematodes, arthropods etc. as they are developed and mature.

The principles provide high level direction on the selection, development, and optimisation of HTS procedures, on their validation and verification, including the use of internal and external quality checks, data management, and interpretation and reporting of HTS test results. They consider the purpose of diagnostic testing and the biosecurity outcome that this enables and are informed by international standards that provide guidance on the characteristics of diagnostic protocols to ensure that there is confidence in the results, and on the analysis of equivalent systems and phytosanitary measures.

Underpinning the principles

The biosecurity outcome

As outlined above, the testing and screening of imported plants for regulated pests is undertaken during post entry quarantine to provide assurance of the health status of the imported material before it is released into the country. The biosecurity outcome sought, which applies Australia's ALOP for high-risk plants for planting, is *the verified absence of regulated pests at the level of detection achieved by current PEQ testing protocols*. These protocols are published in the document "Testing and screening of high-risk plant material in post-entry quarantine", which outlines the mandatory screening and testing requirements for imported high-risk nursery stock and seed lines undergoing post-entry quarantine (PEQ) and the target pathogens for which screening and testing is being conducted [DAWE, 2021]. The target pathogens have been determined to be regulated pests through pest risk analysis.

Efficacy and confidence in the diagnostic outcomes

The PEQ testing protocols are essentially diagnostic protocols as they describe the procedures and methods for the detection and identification for regulated pests. As described in ISPM 27: Diagnostic protocols for regulated pests [FAO, 2006], critical elements in the protocol ensure the sensitivity, specificity and reproducibility of the diagnostic method and protocols contain the methods and guidance necessary for the regulated pest(s) to be detected and positively identified by an expert or by competent staff who are specifically trained. The level of sensitivity, specificity and reproducibility of each method is known.

It may be necessary that some methods are validated before inclusion in the protocols. This validation may include, for example, the use of a set of known samples, including controls, prepared to verify sensitivity, specificity and reproducibility. These factors should ensure that the required confidence in the diagnosis is achieved every time. The risk of false negative or positive results can undermine confidence in the diagnostic protocol. The possible consequences of a false negative diagnostic result, which leads to the release of diseased plants, can be extreme.

Validation and equivalence

Validation may apply the concept of equivalence. The IPPC ISPM 24: Guidelines for the determination and recognition of equivalence of phytosanitary measures [FAO, 2017], describes the process of recognising equivalence as "the objective examination of alternative phytosanitary measures proposed to determine if they achieve the appropriate level of protection of an importing country as indicated by existing measures of that country."

The evaluation of equivalence is bound by the principles in the World Trade Organization Agreement on the Application of Sanitary and Phytosanitary Measures (the SPS Agreement) and the IPPC including, risk analysis; managed risk; transparency; harmonisation; non-discrimination and minimal impact.

As noted above, this report is part of a research project to operationalise HTS in plant PEQ. The results of validation studies will inform an equivalence assessment of the tested HTS technology against current PEQ screening and testing protocols. ISPM 24 offers the following framework for analysing equivalence, which may assist the presentation of validation outcomes:

(a) The description of the proposed alternative measures
(b) The effectiveness of the measures
(c) To the extent possible, the contribution of the proposed alternative measures in achieving the importing contracting party's appropriate level of protection
(d) Information regarding the method, action and operation of the proposed alternative phytosanitary measures in preventing or reducing the specified pest risk ie. Introduction into Australia with imported plants
(e) Information on how the measures were evaluated (e.g. laboratory testing, statistical analysis, practical operational experience), and the performance of the measures in practice
(f) A comparison between the proposed alternative measures and the importing contracting party's existing measures for same pest risk
(g) Information on technical and operational feasibility of the proposed alternative measures
(h) The operational and economic feasibility of adopting the proposed alternative phytosanitary measures.

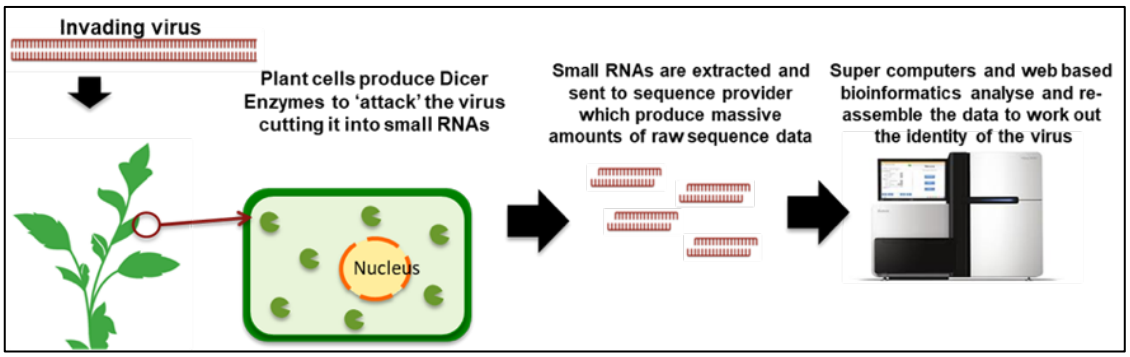
The hypothesis for using HTS in PEQ

The hypothesis being offered for integrating HTS as a testing and screening tool in plant PEQ is that "*HTS is an effective diagnostic tool, equivalent to or better than current molecular-based tools including PCR and ELISA, and better than bioassays and electron microscopy"*. Its application, as a primary screening tool for regulated viruses and viroids in PEQ screening protocols, initially as sRNA-seq HTS plus a confirmatory test where a regulated pest is detected, is equivalent to or better than current protocols based on systems comprising various combinations of electron microscopy, bioassays, ELISA and PCR.

The sRNA-seq HTS platform

As noted above, there are many HTS platforms in use across animal, human and plant health. The platform for initial use of HTS in PEQ using sRNA sequences is derived from the accumulation of virus- or viroid-derived small RNAs (sRNAs) in infected host cells from the host's anti-viral defence system. Plants recognise and respond to viral infections through dicer-like enzymes that cleave viral sequences into specific small RNA products. The sRNAs are predominantly 21, 22, and 24 nucleotides (nt) in length in the plant. These sequences are extracted and subsequently reassembled for mapping against sequence libraries to identify the virus or viroid that triggered the plant response. This process is described in Figure 20.2 [CRC presentation, 2016: https://www.slideshare.net/PlantBiosecurityCRC/rapid-post-entry-quarantine-disease-screening]

Figure 20.2. sRNA-seq HTS process



Developing the methodology

As summarised in Attachment 20.2, the department has been investigating the use of HTS in post entry quarantine since 2014. A research project by the Plant Biosecurity Cooperative Research Centre demonstrated an early proof of concept using sRNA-seq HTS for testing viruses and viroids in a broad range of imported plants in PEQ (Barrero et al, 2017). The project optimised the assembly of viral sequences by using specific small RNA subsets derived from plants held at quarantine glasshouse facilities in Australia and New Zealand. Early comparisons between conventional virus testing in PEQ and HTS are in Figure 20.3.

Sample*	Virus	Description	Genome length	Current PEQ protocols**	PBCRC2064 Toolkit***
	GRSPaV	Grapevine rupestris Stem pitting-associated virus	8,742	Yes	**76.07**
	GVB	Grapevine Virus B virus	7,601	Yes	**70.5**
	GFkV	Grapevine fleck virus	7,564	No	**92.62**
	GLRaV-3	Grapevine leafroll-associated virus 3	18,498	No	19.16
	HS viroid	Hop stunt viroid	264	No	**100**
	PNRV RNA1	Prunus necrotic ringspot virus	3,332	Yes	**82.65**
	PNRV RNA2		2,591	Yes	**54.65**
	PNRV RNA3		1,861	Yes	**100**
	RBDV RNA1	Raspberry bushy dwarf virus	5,449	Yes	**26.41**
	RBDV RNA2		2,231	Yes	**75.08**
	RYNV	Rubus yellow net virus	7,932	No	**66.17**
	SPSMV-1	Sweet potato symptomless virus 1	922	Yes	**98.81**
	SMoV RNA1	Strawberry mottle virus	7,036	Yes	**20.57**
	SMoV RNA2		5,919	Yes	**44.62**

* Samples known to be infected with viruses were selected in Module 1.
** Typical PCR based assays detect 2-4% of the viral genomes.
*** The percentage coverage of viral genomes are shown.
Red fonts = denote additional viral pathogens not found by current PEQ protocols

CRC PLANT biosecurity

biosecurity built on science

Figure 20.3. Plant Biosecurity Cooperative Research Centre project on Rapid Post Entry Quarantine disease screening. Update presented June 2016.

Subsequent research in a four year $1.5M project with Horticulture Innovation [Project MT18005], DAWE, Queensland University of Technology (QUT), and Agriculture Victoria, that commenced in 2019 aims to improve efficiency and effectiveness to allow industry to have more rapid, safe, and cost-efficient access to new plant genetics by:

- Optimising HTS methodology and comparing it with PEQ diagnostics
- Developing 'end to end' standard operating procedures
- Conducting side by side trials comparing HTS with PEQ diagnostics
- Developing a regulatory policy to adopt HTS in PEQ

This work is informing a complementary DAWE Biosecurity Innovation Project [BIP202048] for the 'transformation of exotic plant disease testing at PEQ, Mickleham by implementation of high throughput sequencing'. A significant focus of this project is the capture of standardised operational sRNA-HTS processes and procedures, and the training of scientific staff in sRNA-HTS technologies to enable its integration into PEQ operations from January 2023.

Validation

As outlined by Whattam et al, 2021, validating new technology in PEQ applications by comparing it with existing pest screening protocols before it is implemented is critical to ensure that it achieves an equivalent or better biosecurity outcome. Project MT18005 is confirming the rigour of sRNA HTS as a PEQ screening and testing tool and validating its equivalence to current PEQ diagnostics, while also defining the operating standards and critical controls that underpin the confidence in the diagnostic outcome that DAWE staff require when making biosecurity decisions. Recent results, including those presented in Figure 20.4 are confirming the efficacy of the HTS platform by direct comparison of pest status defined by HTS pipelines and test kits with pest status determined by PEQ-tested samples.

Figure 20.4. Side by side comparison of two HTS pipelines with PEQ (Batch 1). Project update June 2021.

Batch 1 comprised Prunus (80 plants), Rubus (1 plant), Citrus (5 plants), Rosa (3 plants), Lonicera (1 plant). Of the 90 plants tested, 15 were infected with pest, of which 11 endemic pests were detected and five regulated pests were found from four plants. There were two index-hopping events. As outlined in the Figure, two regulated viruses and one regulated viroid were detected. Five endemic viruses were detected. [Project MT18005: Improving plant industry access to new genetics through faster and more accurate diagnostics using next generation sequencing]

Sample	Commodity	Pathogen name	Category	PEQ	GA-VSD	HPC-VSD
MT020	Prunus	Peach associated luteovirus	Regulated	Y	Y	Y
MT091	Prunus	Peach latent mosaic viroid	Regulated	Y	Y	Y
MT092	Prunus	Peach associated lutovirus	Regulated	Y	Y	Y
MT094	Prunus	Cherry virus F	Regulated	Y	Y	Y
MT100	Rosa	Prunus necrotic ringspot virus	Endemic	Y	Y	Y
MT101	Rosa	Prunus necrotic ringspot virus	Endemic	Y	Y	Y
MT102	Rosa	Prunus necrotic ringspot virus	Endemic	Y	Y	Y
MT039	Prunus	Prune dwarf virus	Endemic	Y	Y	Y
MT040	Prunus	Prunus necrotic ringspot virus	Endemic	Y	Y	Y
MT093	Prunus	Cherry virus A	Endemic	Y	Y	Y
MT029	Prunus	Prunus geminivirus A	Land Plant	Y	Y	Y
MT031	Prunus	Prunus geminivirus A	Land Plant	Y	Y	Y
MT032	Prunus	Prunus geminivirus A	Land Plant	Y	Y	Y
MT085	Prunus	Peach latent mosaic viroid	Index hopping	N	Y	Y
MT102	Rosa	Peach latent mosaic viroid	Index hopping	N	Y	Y

Getting it right every time

As noted above and in Whattam et al, 2021 it is imperative that testing methods for regulated pests in plants undergoing PEQ are rigorous, always accurate and applied in full confidence that the result will be right - that is - clear outcomes with no ambiguity, false negatives, or false positives. New technologies must be assessed, tested, validated, and integrated into operations through processes that achieve the level of confidence and assurance required to safeguard Australia from the introduction of regulated pests on imported high-risk plants. HTS technologies require critical controls that provide this level of assurance. These are described in Part 1 of this report and expanded in Attachment 20.1.

The sRNA-seq HTS method, controls and quality assurances are essential outputs of the research program. The project has already identified issues that will need to be resolved prior to implementation. These include sampling time and the potential for seasonal variations in virus titre, the comparative efficacy of different pipelines and kits, identifying and defining controls for each step of the HTS process, variable results between PEQ and research comparisons and the impact of contaminants on the clarity of HTS results and subsequent confidence in the outcome. These issues, and likely others that will emerge over the next 12 months, will need to be resolved before the method is adopted.

To ensure consistency and repeatability, the HTS methodology will need to be sufficiently prescriptive, and implementing staff appropriately trained, to minimise variable and ambiguous test results. Opportunities for international harmonisation of best practice should also be pursued. Ongoing engagement with HTS pipeline service providers has the dual benefit of educating them on the specific needs of biosecurity regulators and collaborating on the evolution of the technology to ensure that bioinformatics software is appropriate to the test purpose and test platform.

Similarly, ongoing engagement with biosecurity agencies, such as the Plant Health Quadrilateral group (PH Quads - comprising the USA, Canada, NZ, Australia) and HTS users in Europe may facilitate the development of standardised controls that enables wider comparison of platforms, pipelines, and test results. One potential area for collaboration is in the identification and use of control sequences from alien species with a known non-related virus, or by using synthetic novel sRNA sequences to spike samples. International collaboration will also support the capture of test methodology in standardised protocols based on validation research, including the performance criteria of bioinformatics packages. A protocol for testing and validating new software packages in the suite of standard operating procedures, is advisable and

may be developed as a future ISPM to facilitate international harmonisation of HTS applications to manage phytosanitary risks.

Making regulatory decisions

The PH Quads Working Group:100, 2021 is drafting a position paper [A Dinsdale, pers.comm] that explores the challenges involved in making regulatory decisions based on HTS data; discusses considerations which should be accounted for when managing these regulatory issues; makes recommendations to inform regulatory decisions; and presents case studies which demonstrate the potential advantages of HTS in identifying various plant pests, and the associated regulatory implications. The Working Group defined three categories of HTS-related diagnostics from which regulatory actions are drawn. These were:

(i) Detecting specific pests and pathogens

(ii) Screening plants with symptoms but no known pests were detected using conventional methods, and

(iii) Screening plants that do not show obvious symptoms, and the intent of the diagnostics are investigational or regulatory in nature, such as demonstrating freedom of a regulated pest for market access.

An advanced draft advises regulatory authorities to exercise caution when applying the outputs of HTS to managing biosecurity risks. It advocates the development and use of guidelines and decision-making matrices to ensure appropriate and consistent diagnoses, to ensure consistent and confident decision making. An example of a decision-making matrix is shown in Figure 20.5.

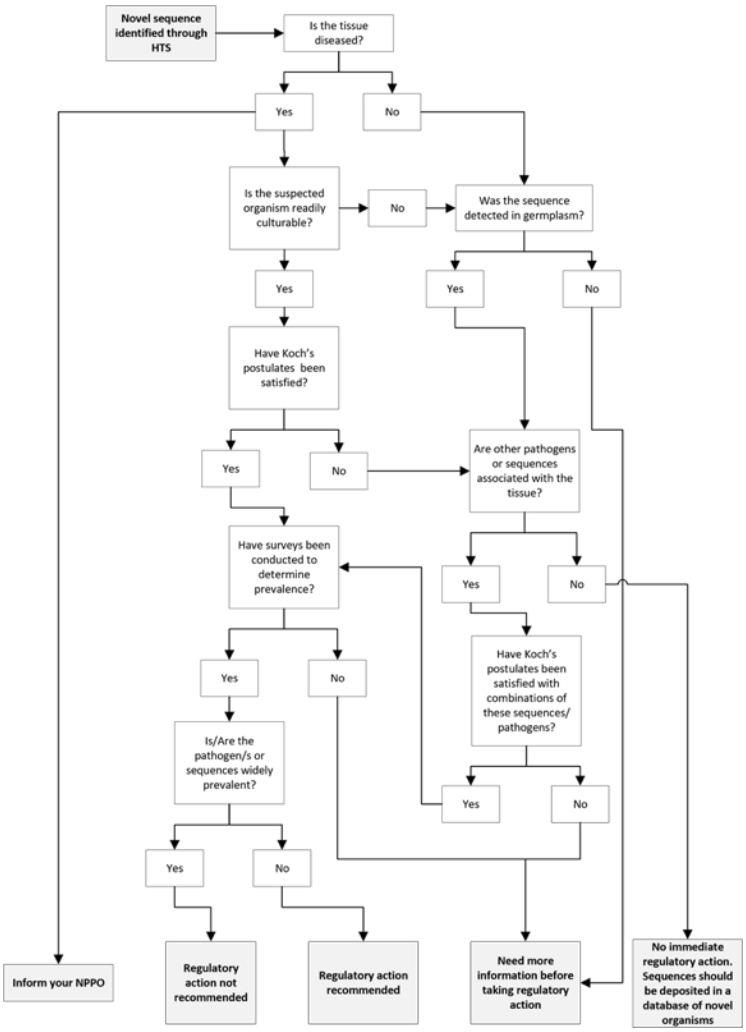


Figure 20.5. An example of a decision-making matrix from the draft PH Quads Working Group:100

The Working Group further recommends that:

1. The use of methods arising from new technologies in regulatory decision making should be discussed and mutually recognized by trading partners prior to their implementation.

2. New technologies require the implementation of appropriate quality assurance measures, and until such time they should not be used independently for regulatory decision-making purposes.

3. Regulatory action should be considered for those sequences with a clear association to disease or on the basis of biosecurity-risk likelihood and that National Plant Protection Organisations should also consider retaining a database of novel sequences or pests without an association to disease for future analysis.

4. International and regional organizations such as the International Plant Protection Convention (IPPC), European Plant Protection Organization (EPPO), and North American Plant Protection Organization (NAPPO) should engage in the development of standards and guidelines for the use and best practices of new technologies.

In summary, regulatory decisions cannot be made on the basis of novel sequences in the absence of biological studies, including fulfilling Koch's postulates to demonstrate pathogenicity, host range, transmission and geographic distribution. This also applies to non- or low-pathogenic viruses that may generate disease by acting synergistically in mixed infections.

Applying HTS – what research tells us

Research that compares HTS technology with conventional methods for detecting and identifying plant pests has largely determined that it offers opportunities that other methods can't – primarily for screening samples that may contain multiple viruses and viroids, including unknown and non-symptomatic agents. Studies have also identified constraints or limitations of the technology that are particularly relevant to managing regulated pests. As a result, the adoption of HTS as a diagnostic tool for regulated pests has progressed conservatively. Canada is likely the most advanced and have implemented only to the extent that diagnostic results from HTS/NGS are confirmed using a conventional method.

Three contemporary studies have been summarised as relevant to this report. They include observations on the utility and development of the technology as a reliable diagnostic tool.

1. Canada

Rott et al, 2017, compared convention bioassays with next generation sequencing (NGS) for detecting known and novel viruses and viroids in tree fruits. They concluded that NGS is a useful screening method for large numbers of viruses that may be present, and/or where this a need to identify the cause of virus-like symptoms but that it should be used in conjunction with a confirmatory method to verify the agent. This may include the need for biological studies to verify pathogenicity. They also concluded that bioinformatics must be effectively managed for efficient and accurate interpretation of results, and care taken to prevent cross contamination of samples with non-target sequences.

2. USA

Al Rwahnih et al, 2015 from the University of California, Davis. "NGS analysis was found to be superior to standard bioassay in detection of viruses of agronomic significant, including virus infections at low titers." Broader range of pathogens detected including strain variants and asymptomatic viruses. Bioassays are dependent on the success of bud grafting and the climatic conditions necessary for symptom expression. NGS includes no biological observations and any unknown detected would need further investigation, if required by the purpose of the testing. Some factors reducing confidence in NGS are low virus titres, the presence of closely related co-infecting viruses and contamination. Considered that NGS could replace bioassays where no pathogens were detected in the tested plants or where it was solely infected by well-known pathogens.

3. Europe

HTS of sRNA populations followed by bioinformatic analysis was first used in 2009 to identify and characterise viruses and viroids that infect plants. Massart et al, 2019, investigated the reliability of HTS-based virus detection protocols, with particular focus on their sensitivity and reliability for diagnostic purposes. In a ring test involving 21 plant virology laboratories and extracts from virus-infected apple, potato, and grapevine plants, they considered the impact of factors including sequence depth, various bioinformatic pipelines, and sRNA sequences on the diagnostic result achieved. They found that the rate of false positives was very low, and reproducibility was high. Bioinformatics strategies were critical for dealing with novel viruses, low virus titre, key parameters at assembly and annotation steps, the completeness of reference sequence databases and the level of scientific expertise in virus taxonomy and data analysis needed to interpret pipeline results. This study provides useful guidance on factors critical to the accuracy and efficacy of sRNA-HTS diagnostic tests and should inform the development and validation of this technology.

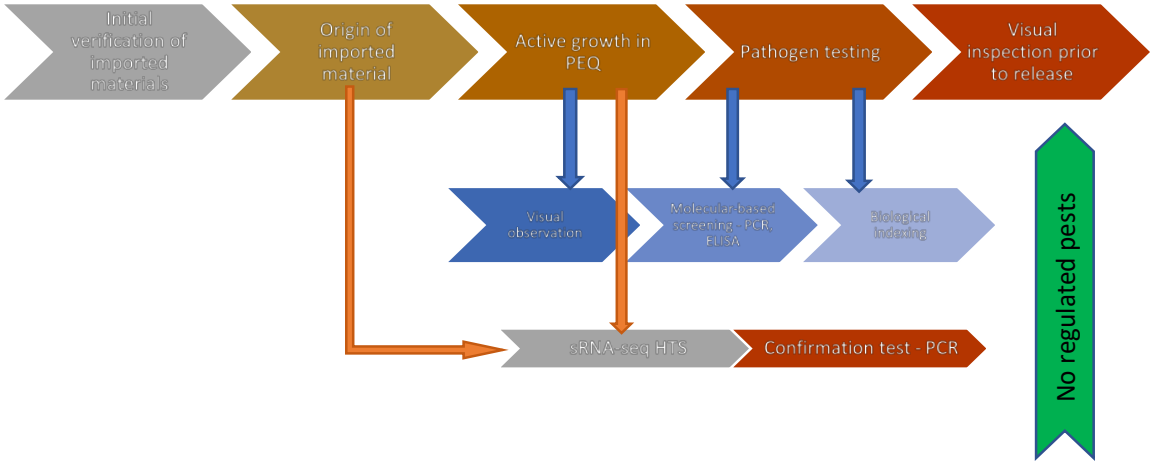
The Policy framework

The purpose of the policy framework is to direct the use of HTS broadly, and sRNA-seq specifically, as a testing and screening tool in plant PEQ to verify the absence of regulated pests at the level of detection achieved by current PEQ testing protocols.

The diagram in Figure 20.6 outlines the progress of imported high risk plants for planting through the PEQ process. All high-risk hosts undergo a period of active growth in PEQ and visual observation for symptoms of disease. Depending on the host species, imported plants may undergo biological indexing for specific pests, or as a general pathogen screen, and they may also be subjected to a range of targeted tests for specific pest genera or species. These generally use PCR or ELISA platforms.

HTS screening can be applied to plants during active growth, or as part of a targeted testing protocol at a specific time in the PEQ process. General screening may indicate the presence of regulated pests as well as unregulated and unknown organisms based on the presence and frequency of genomic sequences. Any viruses or viroids detected by HTS screening must be verified by a confirmation test. This is likely to be a PCR-based method, but it could also be ELISA or herbaceous indexing. Woody indexing should not be required to secure confirmation of the causal agent. The use of HTS in PEQ is intended to be limited to regulated pests. These should be published in BICON.

Figure 6. PEQ process for high-risk plants for planting



Policy Principles for using HTS technologies in plant Post Entry Quarantine

The development of high throughput sequencing (HTS) technologies offers another valuable tool for the rapid, accurate and cost-effective screening of plants for regulated pests during PEQ. The primary principle directing this policy framework is that HTS-based diagnostic tests are demonstrably able to achieve a diagnostic outcome at a level of confidence that is equivalent to or better than existing technologies or otherwise clearly defined.

These policy principles direct the integration and routine use of HTS technologies for the detection and identification of viruses and viroids from high-risk nursery stock (plants for planting) at the DAWE national Post Entry Quarantine (PEQ) facility at Mickleham, Victoria. They have been informed by best practices and other considerations outlined in Parts 1 and 2 of this report.

Associated manuals, operating procedures, guidelines, and work instructions for each HTS method applied during PEQ will enable the implementation of this policy and underpin confidence in the results of screening. These will be based on the research and resultant outcomes from the MT18005 project.

Policy principles

The HTS technology must be:

1. Fit for purpose
2. Efficient
3. Appropriate to the context in which it is used eg. Detecting potential plant pests in PEQ; routine diagnosis of established pests
4. Demonstrably able to achieve a diagnostic outcome at a level of confidence that is equivalent to or better than existing technologies or otherwise clearly defined
5. Reviewed and assessed as new methods or platforms develop or are refined as directed by agreed triggers

The selected HTS methodology must be:

6. Able to achieve the diagnostic sensitivity established for the specific method
7. Repeatable and reproducible
8. Documented and apply best practice
9. Validated
10. Quality and performance assured against specific controls and quality assurance parameters
11. Reviewed and revised to accommodate operational experience

The integration of the specific HTS method into existing systems must be

12. Validated as alternative and equivalent to current testing methods by comparative testing prior to integration
13. Operationally robust and stable, with decision-making criteria established and applied
14. Standardised and documented including any limits, boundaries, risks and caveats on the use of the method
15. Underpinned by robust quality controls and assurance
16. Enabled by access to sufficient biometric and bioinformatic capacity and capability including staff expertise, IT and data storage
17. Supported by confirmatory testing in the system, as appropriate
18. Approved as valid, equivalent, and technically to effectively manage biosecurity risks
19. Subject to periodic review to ensure ongoing best practice

The ongoing use of the specific HTS method is subject to

20. Policy and operational requirements including:
    a. Confidence in HTS as a robust diagnostic platform
    b. Operational policies
    c. Access to and performance of biometrics packages
    d. The adoption of other diagnostic tools that replace or promote additional HTS platforms
21. Training to maintain the competency of technical and operational staff to implement the method, analyse and interpret diagnostic outcomes and make decisions regarding the release or otherwise of the imported plants

The documentation hierarchy

The hierarchy in Figure 1 provides for:

- Overarching policy and context for the actions to be performed, including the policy principles, operational policies and relevant legislation
- Operational guidance and standards for consistency, transparency and efficacy. These may be in the form of manuals, procedures, work instructions and standard operational procedures (SOPs). They provide the basis for training of operating staff and integrate the quality controls and assurances for confidence in the operational outcome
- Records and forms to collect the evidence of operations and support assurance.

The content of the documents becomes more prescriptive towards the base of the triangle. Operational documents are version-controlled and adopted and maintained through defined approval and governance arrangements.

Manuals

Current testing and screening protocols are documented in the *Testing and screening of high-risk plant material in post-entry quarantine* "manual" [DAWE, 2021]. Each chapter is host-specific and outlines the regulated pests, the tests required and the testing period. Footnotes provide further advice relevant to the use of the manual. The PEQ protocol for stone fruit (*Prunus* spp.) is in Attachment 20.3. This has been modified to integrate sRNA-seq HTS for illustrative purposes. The final protocol will be developed on completion of validation research.

Work instructions

Staff undertaking PEQ operations are directed by work instructions that are published internally by DAWE. These include functional duties, managing the facility and decision-making. These will need to be reviewed and revised to integrate HTS technologies.

There are aspects of HTS that will require determinations by a decision-maker. These might include decisions to confirm a virus or viroid based on test results; decisions on actions arising from the detection of unknown or unregulated organisms that are detected by HTS platforms; decisions based on the application and/or outcomes of confirmatory testing using other diagnostic platforms.

Standard Operating Procedures (SOPs)

The SOPs describe the methodology for PEQ operations. The SOPs for using sRNA-seq HTS to detect viruses and viroids in plant tissue direct best practice so that the necessary level of confidence in the test result is achieved. The following SOPs are proposed but will be ultimately determined by the research that validates the technology and its application to the PEQ function and outcomes. They have been informed by the analysis in Part 1 of this report and actions outlined in Attachment 20.1.

SOP 1 Sampling

SOP 2 Nucleic acid extraction

SOP 3 Library preparation

SOP 4 Sequencing

SOP 5 Analysis of raw reads

SOP 6 Identification of target(s)

SOP 7 Analysis of controls

SOP 8 Confirmation

SOP 9 Interpretation and reporting

SOP 10 Data management

SOP 11 Data sharing

SOP 12 Validations

SOP 13 Training

Records and forms

Data entry standards will need to be established for HTS platforms. This may require modification of current databases or the establishment of systems appropriate to data capture, storage and reporting of HTS diagnostic outcomes. A database of novel sequences or pests without an association to disease for future analysis is being recommended to the PH Quads.

Reporting templates will likely require review to integrate the reporting and interpretation of HTS outcomes.

Transition to operations

A sequence of actions is required to integrate sRNA-seq HTS for the detection of regulated viruses and viroids into PEQ operations. These will ensure that the risks associated with the transition are managed and that the biosecurity outcome continues to be achieved. These actions include but may not be limited to:

- Policy review and any legal determinations that may be required to allow changes to import testing requirements for (some/all) plants for planting undergoing PEQ at Mickleham based on a minimum equivalence with existing screening protocols
- Operational standards based on verification research
- Changes to testing the manual
- Changes to instructional materials
- Staff training
- On-site proficiency verification
- Implementing IT and bioinformatics requirements
- Any outsourcing arrangements/contracts, procurement of machinery and consumables
- Any other administrative changes to import permits and to import conditions in BICON.

Lessons from other sectors will usefully guide planning and implementing these actions. The following experiences from Queensland Health (per A Dinsdale, 2021) are both practical and informative. It would be wise to apply them to implementing HTS into plant PEQ.

- *"Don't underestimate the importance of computing support – need to ensure organisation invests in solution architects (for cloud computing solutions), IT officers and bioinformaticians. This resourcing of staff is often not supported up front and means things are not set up properly and accreditable from the beginning and may increase costs through hardware purchasing and re-purchasing etc.*

- *Quality of the DNA matters, and quality controlling your DNA and libraries at every step will help to avoid headaches and failed sequencing and analyses.*
- *Start small to get comfortable with the technology before ramping up, rather trying to start with 100s per week. Utilise pilot studies/research projects to gently introduce the technology.*
- *Automation definitely helps when there are high numbers of samples. Robotics are a lot of work to set up but well worth it and sequencing demand in our experience takes off very quickly.*
- *Networking implementation, validation, standardisation, agreed acceptance criteria and methods for data sharing at a multi-lab, state or national level is invaluable at driving progress.*
- *With validation – try to validate end to end process – rather than trying to validate each piece of equipment in isolation."*

Considering the benefits of starting small, the following three-phase implementation plan might be considered. This will allow ongoing validation research to support the transition of HTS into full operation by January 2025. This includes the removal of bioassay requirements and sufficient time to secure and validate any new confirmation tests that may be required. It will also allow time for processes to be fully supported by documented procedures and training.

Phase	Date	PEQ Function	Host	Virus/viroid testing protocol
I	January 2023	Primary screen Virus/viroid	Two or three hosts chosen from validation research eg. *Prunus* spp. Ongoing validation to support transition of next hosts	Nucleic acid extraction for HTS, PCR to confirm regulated viruses/viroids Residual herbaceous indexing may be required
II	January 2023	Primary screen Virus/viroid	Next tranche of hosts Ongoing validation to support removal of all woody indexing Validating HTS for screening regulated bacteria and/or fungi	Nucleic acid extraction for HTS, PCR to confirm regulated viruses/viroids No herbaceous indexing
		Remove bioassay*	All woody indexing is replaced by HTS	HTS primary screen
III	January 2025	Primary screen Virus/viroid Bacteria Fungi	All virus/viroid screening by HTS Validated screening Bacterial/fungal by HTS operational	

* May require more than one HTS procedure

Discussion

Diagnosis of pests and pathogens in PEQ requires absolute confidence in the outcomes of the diagnostic test every time. The consequences of not detecting a regulated pest in plants for planting that are subsequently released can be extreme in terms of economic, environmental, and social impacts at local, regional and national levels. General guidance on the development of diagnostic protocols is provided by ISPM 27: Diagnostic protocols for regulated pests [FAO, 2006]. There are currently no HTS-based protocols adopted as annexes to ISPM 27, but it is almost certain that there will be in future.

HTS offers a technological leap forward for screening plants undergoing PEQ. However, the molecular science that it is based on is not new, having been applied through PCR and other molecular testing platforms for many years. The strengths, weaknesses and limitations of PCR and others are well understood and accounted for in PEQ screening processes.

Similarly, the policies and processes for introducing a new tool into the PEQ tool kit have been well practiced. This begins by defining the utility and efficacy of the new test in comparison with current standard tests, validating this equivalence through comparative trials and then integrating the new tool into the PEQ testing program. Change is necessarily conservative and measured, with due regard to the consequences of getting it wrong.

Like any other diagnostic technology applied in PEQ, the correct application of the test method is critical for confidence in the results. As such, using HTS in PEQ requires a high degree of prescription for each adopted protocol. This must include instruction on selecting the test platform, taking and handling the samples of plant tissue, extracting nucleic acids, selecting the sequence libraries, sequencing, reading and interpreting the sequences, confirming the organisms present and reporting this to the client. Quality management is essential, with critical controls and assurance indicators integrated throughout the testing protocol. These ensure the validity of the test result and underpins confidence in the outcome. The policy principles and framework in this report should enable a standard approach for documenting and assuring effective implementation of HTS methods as they are adopted.

The use of bioinformatics differentiates HTS from other diagnostic processes. As noted by the OIE "bioinformatic analysis assembling the pathogen genomic sequence from the raw data and the subsequent secondary analysis are the critical elements in HTS". As PEQ is a prescribed process, the bioinformatics software appropriate to the test purpose and test platform should be well documented and analysis undertaken by trained personnel. Test methodology should be captured in standardised protocols based on validation research, including the performance criteria of bioinformatics packages. A protocol for testing and validating new software packages in the suite of standard operating procedures, is advisable.

The transparency of the method to external scrutiny and evaluation will ensure that it remains fit for purpose in the short term, and open to innovation in the medium to longer term.

The policy framework proposed for HTS in PEQ is informed by the international standards and guidelines that are or have been developed in plant, animal and human health sectors. A number of these are referenced in Part 1 of this report. The key elements of these guides are, unsurprisingly, consistent as they focus on the factors for confidence in the outcomes achieved by using HTS methods. This was also a conclusion of the NBC workshop in 2020 [NBC, 2020]. Investment in developing a consistent and harmonised framework for the development, validation and use of HTS technologies across all health sectors will likely increase the collective confidence in HTS by users and provide a platform for ongoing innovation and efficiency. This will facilitate the evolution of the platform across even more applications and likely increase the cost:benefit of the technology.

A critical and costly component of HTS is data – collection, storage, analysis and interpretation, reference libraries and retention. While the cost of data storage is reducing, the security of and access to data and databases will be an ongoing and increasing imperative for HTS and other genomic-based tools and technologies. There is a strong case for sharing data-related resources across sectors, and particularly as it applies to managing biosecurity. Biosecurity tools, functions and applications are likely to be similar, which will enable harmonisation of data management.

As the use of HTS technology and applications expand across life sciences, so will the demand for specialist bioinformatics skills. This is another area where harmonisation of training and skills development across the sectors should be progressed as a priority. The alternative is the development of software platforms that remove the human element and complete the interpretation for the user. This will require an entirely new validation process.

Part 3 Implementation of HTS as a screening tool for viruses and viroids in PEQ

Integrating sRNA-seq HTS technology into PEQ testing protocols to detect viruses and viroids in selected high-risk hosts.

Background

Extensive research undertaken by DAWE and other research specialists during the last seven years has verified that HTS effectively and efficiently detects regulated viruses and viroids in a range of host plants that achieves a biosecurity outcome equivalent or better than current screening protocols.

The project MT18005 is progressing the operationalisation of sRNA-seq HTS for testing and screening viruses and viroids in plant PEQ for high-risk hosts from 1 January 2023. As outlined in Part 2 of this report, there is much to be done before then.

Scope

Part 3 of this report is intended to guide this implementation with reference to the policy framework, policy principles, documentation hierarchy and operational best practice. For illustrative purposes, this guidance will focus on integrating HTS for virus and viroid testing of imported stonefruit (*Prunus* spp.). Imported Prunus are subjected to many screening tests for bacteria, fungi, viruses, viroids and phytoplasmas. If sRNA-seq HTS were the primary screening tool for regulated viruses and viroids, most bioassays would no longer be required, and all the current PCR tests would only be required if their target were detected by HTS. A modified protocol is at Attachment 20.3. The final protocol will be determined by the outcomes of current validation research.

HTS technology is not yet ready for use as a confirmatory test of positive detections and identifications of viruses and viroids by other screening methods but may be in future. This would require the development of a separate but potentially overlapping policy for that purpose. Similarly, HTS platforms for detecting other organisms such as bacteria and fungi will require their own policy.

Timeline

As outlined in Part 2, many actions are required to integrate sRNA-seq HTS for the detection of regulated viruses and viroids into PEQ operations. These include but may not be limited to:

- Policy review and any legal determinations that may be required to allow changes to import testing requirements for (some/all) plants for planting undergoing PEQ at Mickleham based on a minimum equivalence with existing screening protocols
- Operational standards based on verification research.
- Changes to testing the manual
- Changes to instructional materials
- Staff training
- On-site proficiency verification
- Implementing IT and bioinformatics requirements
- Any outsourcing arrangements/contracts, procurement of machinery and consumables
- Any other administrative changes to import permits and to import conditions in BICON.

An indicative timeline for these actions to enable first HTS screening operations in January 2023 follows, noting that activities may overlap as determined by the HTS research program and be modified to accommodate the implementation program.

Action	2021					2022												2023			
	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	Q 1	Q 2	Q 3	Q 4
Policy framework																					
Prepare SOPs																					
IT sorted																					
Policy determination																					
PIO/BICON																					
Finalise SOPs/ instructional material																					
Review SOPs/ IM																					
Training																					
HTS screening commenced – Prunus spp																					
HTS other hosts																					
Stakeholder engagement																					

Implementation of sRNA-seq HTS in PEQ



Denotes new process and guidance required.

HTS Method 1 - Scope

Detecting regulated viruses and viroids by HTS small RNA-sequencing of plant extracts (sRNA-seq HTS) [in stonefruit (*Prunus* spp.) undergoing PEQ].

Policy

The policy principles for the adoption and implementation of HTS technologies into plant PEQ apply.

The current list of regulated pests for stonefruit [*Prunus* spp.] should be published in BICON as outlined in Attachment 20.4.

Legislation

The following provisions of the *Biosecurity Act 2015* [FRL, 2021] apply to the screening and testing of imported plant material in post-entry quarantine:

- Section 119—Goods brought into Australian territory are subject to biosecurity control
- Section 12—Inspecting goods and taking samples
- Section 137—Regulations that may provide other biosecurity measures
- Section 162—Release of goods from biosecurity control
- Section 554—Carrying out tests on samples

Purpose

This assay detects and enables the characterisation of regulated viruses and viroids in imported high-risk plants [stonefruit (Prunus spp.)] undergoing PEQ. A confirmatory test is required to verify the detection and identification.

Description

Detecting viruses and viroids in [stonefruit (Prunus spp.) plants using sRNA HTS requires extraction of sRNAs from a tissue sample from the target plant(s), which are then sequenced and biometrically analysed against a reference library of whole or partial virus and viral genomes, in accordance with official standards and processes adopted by DAWE. A second test is required to confirm the identity of regulated viruses and viroids detected by HTS technology.

Scope

This test should be used to detect and confirm the presence of regulated virus and viroids in high-risk plants for planting undergoing PEQ verification of health status at the PEQ facility in accordance with the relevant testing and screening protocols adopted by DAWE and published in the testing and screening manual.

This test should not be used for (i) the detection and identification of organisms other than viruses and viroids such as invertebrate pests, bacteria, phytoplasma/spiroplasma, fungi, nematodes, or seeds; (ii) open ended testing for all known and unknown plant viruses and viroids; (iii) screening of low and medium risk plants for planting and plant products.

This test should not be used unless it has been appropriately validated and is supported by an approved official methodology published in a Standard Operating Procedure (SOP).

Relevant reference materials

Manual:

- Testing and screening of high-risk plant material in post-entry quarantine [Stonefruit (*Prunus* spp.) testing manual extracted and modified to integrate sRNA-seq HTS [*not shown, available for internal use only*]
- Any quality systems manuals that may be developed for PEQ horticulture and technical operations.

Standard Operating Procedures:

Guideline:

- Management of plant pests or diseases detected at the Mickleham post-entry quarantine facility
- HTS quality, management and system controls – with reference to general and specific pipelines
- HTS reporting guideline - a standardised analysis of the HTS diagnostic test, and the results achieved

Work Instructions:

- Performing PCR testing on plant material
- Performing ELISA testing on plant material
- Performing bacterial isolations from plant material
- Performing fungal isolations from plant material
- Performing electron microscopy on plant material
- Performing biological testing on plant material
- Performing nematode extractions from plant material
- Performing shoot-tip grafting in citrus plants.

The following related material is available on the departmental website:

- Information on department-approved chemical disinfectant.

Method limitations

Application of RNA-seq to detect viruses and viroids in plant samples relies on adequate experimental design that takes into consideration the main goal of the HTS experiment. Qualitative (e.g., diagnosis) and quantitative (e.g., differential gene expression analysis host plant genes or measurement of the HTS method sensitivity) data analyses have some different requirements such as those related to the starting RNA amount, the number and type of replicates, library type and preparation, controls, sequencing platforms, throughput, coverage and depth, and read length [R Barrero, MT18005: Diagnosis of viruses and viroids using high throughput sequencing (HTS) technologies]. Methods, standards and limitations will be defined by the research program and documented in SOPs.

Methodology

The methods for collecting plant samples, extracting sRNA, processing analysis and generating sequence reads, completing biometric analysis and data management including standards, quality and sample controls and assurance are contained in respective standard operating procedures (SPOs) for each step of the process. The SOPs are developed from the comparative testing and validation of the sRNA-seq method and reflect best practice to achieve a level of diagnostic confidence equivalent to existing processes where they are in place, or to an acceptable level of biosecurity risk where they are not. Information that may be included in each SOP is outlined in Attachment 20.1. A draft SOP template is in Attachment 20.5. It is likely that the format for each SOP will evolve relevant to its content and use.

SOP list:

Number	Name
SOP 1	Sampling
SOP 2	Nucleic acid extraction
SOP 3	Library preparation
SOP 4	Sequencing
SOP 5	Analysis of raw reads
SOP 6	Identification of target(s)
SOP 7	Analysis of controls
SOP 8	Confirmation
SOP 9	Interpretation and reporting

SOP 10	Data management
SOP 11	Data sharing
SOP 12	Validations
SOP 13	Training

The National Pathology Accreditation Advisory Council, 2017, which is managed by the Australian Government Department of Health, has published minimum best practice requirements for medical pathology laboratories undertaking the performance and implementation of human genetic testing utilising massively parallel sequencing (MPS) for all applications (ie. single gene, panel of genes, somatic testing, whole exome or whole genome or non-invasive prenatal screening) also known as next generation sequencing. It takes a risk-based approach to defining standards for the implementation of these new technologies in human health that none-the-less will usefully inform any SOPs that are developed for HTS use in plant PEQ. They should be referenced when drafting the SPOs listed above.

Controls in HTS protocols

System and process controls contribute to overall assurance that the test has been undertaken correctly and underpin confidence in the diagnostic outcome. The research program is considering controls for key process steps as outlined below. These should be integrated into SOPs where appropriate and relevant. As noted above, there is potential for international standardisation of system controls, including novel synthetic sequences and universal alien viruses.

	Negative controls	Positive controls	Alien controls	Internal controls
Aim/Monitoring	Contamination	Contamination Monitor the analytical specificity of the test Monitor the analytical sensitivity of the test when used at low concentration	Contamination Monitor the analytical specificity of the test when use at high concentration Monitor the analytical sensitivity of the test when use at low concentration	Monitor the analytical sensitivity of the test
Description	Same matrix of the analysed samples but free of the target(s) or extraction buffer (NIC), or molecular grade water (NAC)	Same matrix and range of target(s) expected to be detected in the analysed samples and processed alongside the samples and preferably at low concentration (naturally infected or spiked)	A target not expected to be found in the analysed samples (i.e. alien target) and processed alongside the samples and not expected to be detected in the samples to be tested when used at high concentration and/or expected to be detected only in the alien control when used at low concentration	Non target nucleic acids not related to the sample targets naturally present (e.g. plant genes) or known target spiked at low concentration in the samples (e.g. synthetic nucleic acids, known target not expected to be found in the samples to be tested)
Analysis	Absence of target(s) Contamination target(s) below a set threshold (*) (**)	Presence of positive control targets Contamination targets below a set threshold (**)	Absence of the alien target in the analysed samples (when used at high concentration) (**) Contamination targets below a set threshold (**) Presence of expected alien target in the alien control (when used at low concentration)	Presence of internal control in each sample

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

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

(*) the absence of target sequences is practically nearly not possible in a negative control

(**) if an unexpected target is detected in any control or an alien target is detected in the samples, their presence should be quantified and compared with the controls and samples infected by the target

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

Results analysis and conclusions

A reporting guideline provides for documenting a standardised analysis of the HTS diagnostic test, and the results achieved. The report should provide the decision-maker with the information they need to determine that the test outcomes are valid. The guideline includes:

- The list of test targets
- The test platform used
- The critical controls applied and evidence of their application
- Their status (present or absent) in the plant sample as indicated by test results
- Any conclusions on the presence or absence of other regulated pests
- Any unknown organisms indicated
- The outcomes of confirmatory tests to verify the presence of regulated pests detected by the HTS test
- The level of confidence in the test outcome
- Any other supporting technical information and analysis

Decision-making

Whether to accept or not the results of this diagnostic process can be directed by a work instruction that takes the decision-maker through a series of analytical steps that are based on the critical control points of the test method. For sRNA-seq this might include but probably not be limited to:

- Assessing evidence that each of the critical controls in relevant SOPs has been met
- Policy for a positive or negative results has been applied correctly
- Any confirmatory testing of HTS results has been directed/initiated/completed
- Results have been documented
- Data has been recorded and stored
- The opinion and recommendation of the diagnostician has been documented
- Any further actions in relation to detection of an unknown virus or viroid has been analysed and a recommendation made on associated biosecurity risks

Communicating a biosecurity determination

A form for reporting the outcomes of diagnostic testing includes a section on HTS tests. This section outlines the test undertaken, the context in relation to other tests, analysis of the results and the contribution to the overall recommendation on release from PEQ of the imported plants or not.

Training

Proposed Training Package for DAWE staff on using sRNA-seq HTS for detecting regulated viruses and viroids in PEQ

-

Module 1 - Overview of HTS diagnostics and Introduction to VSD/VirReport/GA-VirReport

-

Module 2 - Introduction to Galaxy Australia platform

-

Module 3 – VSD/VirReport/GA-VirReport pipeline on Galaxy

-

Module 4 - Interpretation and archival of results

-

Module 5 - Processing real-time samples in Galaxy

-

Assessing staff competency

Capturing and interpreting side-by-side PEQ and HTS outputs 

Sharing information on HTS outcomes with other regulators was identified by the PH Quads Working Group:100 as an important means of expanding knowledge on the use of HTS, sharing results and supporting ongoing evolution of the technology.

Confluence results table - PEQ testing is an important validation method for HTS result.

To capture data and information that may be shared with researchers and other regulators.

Sample name	Host species	Virus detected	Method(s) of detection	If PCE test performed, was the same RNA/DNA extraction used as HTS?	If PCR test was performed, was the product confirmed by SANGER sequencing?	If method other than PCE was used for detection, was it performed at the same time as extraction for HTS? If not what date?	What kind of sample is it?	Other relevant information
				Y/N	Y/N	Y/N If N, date	Control, duplicate (if so, to which other sample), resampling, priority etc.	

Issues to be resolved for operationalisation

It is likely that issues will be identified during the research program that will need to be resolved before sRNA-seq HTS is used in PEQ. Some that have been raised to date include:

- Sampling time
- Seasonal variations in virus/viroid titre
- Growth stage/target tissue variations to virus/viroid titre
- Optimising sRNA titre and quality, target nucleotide length
- Reducing and/or removing contaminating DNA/RNA sequences
- Detecting and assessing endogenous viral sequences in plant genomes
- HTS service provider performance testing
- System quality and management controls including positive and negative control sequences

Discussion and conclusions

The research being undertaken through project MT18005 will inform the end-user guides and SPOs required to effectively implement sRNA-seq HTS into PEQ operations. Operational standards should be drafted with reference to available best practices that are outlined in Part 1 of this report, noting that these are continuing to evolve and reflect practices applying in human, animal and plant health and biosecurity sectors.

Adequate, appropriate and effective controls at all stages of the HTS process are critical to providing assurance that the diagnostic result is accurate so that the biosecurity decision-maker has confidence in the outcome.

This report has analysed international best practices and standards for the use of HTS technologies; the policy principles and framework for using HTS in plant PEQ and the system requirements for integrating sRNA-seq HTS into plant PEQ operations to detect and identify regulated viruses and viroids in imported plants. The latter task will be informed by current research, which will need to continue to be an iterative process between HTS service providers, the research collaborators and the PEQ end users. This approach simultaneously provides for a better outcome that is fit for its intended purpose and builds in significant confidence in the rigour of the adopted methodology.

References

1. FAO. 2016. The importance of pest diagnosis. Commission on Phytosanitary Measures Recommendation No. 7. Rome. Published by FAO on behalf of the Secretariat of the International Plant Protection Convention (IPPC). https://assets.ippc.int/static/media/files/publication/en/2019/10/R-07_En_2017-08-24_Combined.pdf
2. World Organisation for Animal Health. 2021. Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. https://www.oie.int/en/what-we-do/standards/codes-and-manuals/terrestrial-manual-online-access/
3. Matthijs, G., Souche, E., Alders, M. et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet 24, 2–5 (2015). https://doi.org/10.1038/ejhg.2015.226
4. United States Food and Drug Administration. 2018. Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing (NGS)-based In Vitro Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases. https://www.fda.gov/media/99208/download
5. International Standards Organization. 2020. Quality control metrics for DNA sequencing ISO/TS 22692:2020. https://www.iso.org/standard/73693.html
6. Lebas, B and S Massart. 2020. "Best practice" guidelines for validation and routine use of non-targeted techniques in diagnostic setting which could serve as a basis for a new EPPO Standard. Valitest. Validation of diagnostic tests to support plant health. EU Grant Agreement N. 773139. Unpublished.
7. FAO. 2019. Preparing to use high-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes. Commission on Phytosanitary Measures Recommendation No. 8. Rome. Published by FAO on behalf of the Secretariat of the International Plant Protection Convention (IPPC). https://assets.ippc.int/static/media/files/publication/en/2021/07/R-08_En_2019-05-06_Post-CPM-14_Fixed-BackCover.pdf
8. National Biosecurity Committee (NBC). 2020
9. FAO. 1997. International Plant Protection Convention. Rome. Published by FAO on behalf of the Secretariat of the International Plant Protection Convention (IPPC). https://www.ippc.int/en/core-activities/governance/convention-text/
10. Whattam, Mark, Adrian Dinsdale, and Candace E. Elliott 2021. "Evolution of Plant Virus Diagnostics Used in Australian Post Entry Quarantine" *Plants* 10, no. 7: 1430. https://doi.org/10.3390/plants10071430

11. Australian Government Department of Agriculture, Water and the Environment (DAWE). 2021. Testing and screening of high-risk plant material in post-entry quarantine. Version no.: 10. Published: 03/05/2021 [Internal document]

12. FAO. 2006. Diagnostic protocols for regulated pests. International Standard for Phytosanitary Measures No. 27. Rome. FAO on behalf of the Secretariat of the International Plant Protection Convention. https://assets.ippc.int/static/media/files/publication/en/2016/01/ISPM_27_2006_WithoutApp2_En_2015-12-22_PostCPM10_InkAmReformatted.pdf

13. FAO. 2017. Guidelines for the determination and recognition of equivalence of phytosanitary measures. International Standard for Phytosanitary Measures No. 24. Rome. FAO on behalf of the Secretariat of the International Plant Protection Convention. https://www.ippc.int/static/media/files/publication/en/2016/01/ISPM_24_2005_En_2015-12-22_PostCPM10_InkAmReformatted.pdf

14. Roberto A. Barrero, Kathryn R. Napier, James Cunnington, Lia Liefting, Sandi Keenan, Rebekah A. Frampton, Tamas Szabo, Simon Bulman, Adam Hunter, Lisa Ward, Mark Whattam & Matthew I. Bellgard. 2017. An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. BMC Bioinformatics 18, 26 (2017). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1428-4

15. Plant Health Quadrilaterals Working Group 100: Managing Regulatory Issues from New Diagnostic Technology. 2021 (in draft). Managing regulatory issues arising from new diagnostic technology: A QUADS position paper using high throughput sequencing as a model.

16. Rott, Michael; Xiang, Yu; Boyes, Ian; Belton, Mark; Saeed, Hanaa; Kesanakurti, Prasad; Hayes, Sheila; Lawrence, Tracy; Birch, Carrie; Bhagwat, Basdeo and Rast, Heidi. 2017. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Disease. 101. https://doi.org/10.1094/PDIS-03-17-0306-RE .

17. Al Rwahnih, M, Daubert, S, Golino, D, Islas, C, and Rowhani, A. 2015. Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. Phytopathology 105: 758-763. http://dx.doi.org/10.1094/PHYTO-06-14-0165-R

18. Massart, Sebastien, and Michela Chiumenti, Kris De Jonghe, Rachel Glover, Annelies Haegeman, Igor Koloniuk, Petr Komínek, Jan Kreuze, Denis Kutnjak, Leonidas Lotos, François Maclot, Varvara Maliogka, Hans J. Maree, Thibaut Olivier, Antonio Olmos, Mikhail M. Pooggin, Jean-Sébastien Reynard, Ana B. Ruiz-García, Dana Safarova, Pierre H. H. Schneeberger, Noa Sela, Silvia Turco, Eeva J. Vainio, Eva Varallyay, Eric Verdin, Marcel Westenberg, Yves Brostaux, and Thierry Candresse. 2017. Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies. Phytopathology 109: 488-497. http://doi.org/10.1094/PHYTO-02-18-0067-R

19. Federal Register of Legislation. 2021. Biosecurity Act 2015. Australian Government. https://www.legislation.gov.au/Details/C2019C00097

20. National Pathology Accreditation Advisory Council (NPAAC). 2017. Requirements for Human Medical Genome Testing Utilising Massively Parallel Sequencing Technologies https://www1.health.gov.au/internet/main/publishing.nsf/Content/npaac-pub-mps

Attachments

Attachment 20.1: Operational requirements for implementing HTS in plant PEQ

Attachment 20.2: The DAWE HTS development pathway

Attachment 20.3: Example of the stonefruit (*Prunus* spp.) regulated pest list published in BICON

Attachment 20.4: Draft Standard Operating Procedure (SOP) Template

Attachment 20.1

Operational requirements for implementing HTS in plant PEQ

Application of RNA-seq to detect viruses and viroids in plant samples relies on adequate experimental design that takes into consideration the main goal of the HTS experiment. Qualitative (e.g., diagnosis) and quantitative (e.g., differential gene expression analysis host plant genes or measurement of the HTS method sensitivity) data analyses have some different requirements such as those related to the starting RNA amount, the number and type of replicates, library type and preparation, controls, sequencing platforms, throughput, coverage and depth, and read length. [Ref. R Barrero; Overview of existing NGS / HTS regulatory policy documents (ppt presentation). MT18005 Policy Regulatory Meeting, Canberra, 13 April 2021]

Priority actions recommended by the NBC workshop in February 2020 are aligned with plant health standard operating procedures (SOPs) where relevant.

Operational requirements	SOP	NBC workshop priority actions
Wet Laboratory		
The national HTS standards framework (workshop recommendation) should: Establish a minimal set of internal quality assurance standards (including positive and negative controls) to be used in state and national network member laboratories Encourage laboratories with ISO or NATA accreditation to partner and mentor non-accredited laboratories conducting HTS diagnostics to transition to accredited operations		
Sample preparation Sampling (appropriate to hosts and targets) – growing cycle, season, material Sample information – specimen type, sampling date, specimen origin Sampling material – PEQ guidelines for individual plant commodities (i.e., leaves, petioles) Sample handling – ensure sample quality and integrity, subsampling, traceability, preservation, storage, transport, retention and disposal timelines	SOP 1 Sampling	Perform sampling consistently by identifying specific sampling requirements and minimum DNA quality standards for HTS library preparations
Replicates (technical and biological)	SOP 2 Nucleic acid extraction	Identify minimum negative controls required for HTS protocols including sequencing at final library quantification step
DNA / RNA testing guideline – ISO 17025 Nucleic acid extraction method – validated commercial kit(s) Assess nucleic acid quality – nucleic acid purity & integrity Establish minimal thresholds for quality and quantity (i.e., metrics suitable for PCR also suitable for HTS) Contamination management and prevention Spike-in controls, RNA isolation and construction, enrichment, cDNA Library construction, RNA/cDNA fragmentation, cDNA synthesis		

Operational requirements	SOP	NBC workshop priority actions
Library preparation Library prep (type, pooling, platform) NGS/HTS library construction – library type, library input amount, library size insert, library construction kit, quantification, multiplexing/pooling considerations, sample contamination Summary of library preparation metrics – metrics, value, performance criteria, optionality,	SOP 3 Library preparation	Adopt dual barcodes in multiplex assays to reduce inappropriate de-multiplexing of samples.
Sequencing information – Platform, depth, NGS/HTS instrument, read length, sequencing direction, instrument running mode Summary of sequencing related metrics – category, metrics, value, performance criteria, optionality Running quality – error rate, percentage of data quality > Q30 (99.9% accuracy) Adapters sequences and ligation	SOP 4 Sequencing	Standardise use and calculation of sequencing metrics terms such as coverage, depth, N50, L50, SNP and SNV.
Bioinformatics		
Analysis of raw reads – unambiguous base calls, remove redundant/noise Data quality measurement – total reads, read length, mean coverage, uniformity, duplication rate, on-target rate, Q30 rate Data analytics approaches – direct annotation, de novo assembly & reference mapping	SOP 5 Analysis of raw reads	Develop and conduct regular verification activities to ensure data analysis tools and processes are updated regularly.
Approaches (direct annotation, de novo assembly and reference mapping)		
Target identification (use of clean reads, controls, and samples)	SOP 6 Identification of target(s)	
Taxonomic identification – ref databases (ie., ICTV, Intl Com Taxonomy of Viruses), custom databases (i.e., NCBI Virus DB, PVirDB) Version control – software, scripts and database versions SOP for changes – software, scripts and/or database changes, reference control panel		
Bioinformatics Sample metadata information Mapping, assembling and visualizing mapped reads Pipeline implementation, optimization and validation – Accuracy, Precision (Reproducibility & Repeatability), Sensitivity, Index hopping Quality control and reads processing	SOP 7 Analysis of controls	
Functional assignment, whole genome assembly Variant calling & filtering – variant quality score, germline or somatic mutation, variant annotation		

Operational requirements	SOP	NBC workshop priority actions
Downstream analyses		
Pest ID, interpretation, and reporting		
Confirmation of pest(s) (plant health risk, regulated vs non-regulated pests)	SOP 8 Confirmation	Fund development of national biosecurity HTS networks to convene working parties to assess criteria for reporting on a species based on incomplete genomes
Use of distinct test(s) for confirmation (where possible use validated tests – NDP, IPPC, EPPO, etc.)		
Interpretation (biological relevance of HTS targets; sample, taxonomic and genome information; confirmation of results; causation/ aetiology, viability of the organism)		
Requirements for internal vs external	SOP 9 Interpretation and reporting	Adopt a common framework for reporting HTS data and diagnoses. Establish a parallel scoring system akin to existing medical standards and guidelines Prepare national guidelines on reporting of incidental findings and actions required in cases of high impact incidental findings
Report (general recommendations, inconclusive results, unexpected organisms, additional analysis, disclaimers)		
Key requirements of genomics reports Report details, NGS/HTS test description, summary of results, major findings, interpretive comment		
Summary comment		
Stakeholder consultation (Plant User Group, SPHD, NBC, NPBDN?)		
Data Management		
Sample and data management plan (primary materials, data transfer & storage, retention time & disposal) Minimum data standards	SOP 10 Data management	Establish Australian biosecurity database matching services, akin to Beacon and Matchmaker to allow deposits of unknown sequences, with minimal metadata to allow laboratories to share information Enable access to standardised national computer platforms with version-controlled tool sets and reference data Facilitate standardised use of reference databases to avoid individual laboratories having to maintain their own curated version of a public database Develop guidelines and procedures for file storage (onsite and/or offsite) prescribing a minimum of information to be kept that allows for the complete reproduction of a result from the raw files
IT and information systems to be used		
Database access		
Curation, storage, transfer, and maintenance		
Human resources required		
Accessibility, reporting and sharing arrangements		
Sensitivity of data/confidentiality		
Trusted network		
System to share		
Data applications to inform and support future policy development, e.g., PIO, PSaRA, PHP		

Operational requirements	SOP	NBC workshop priority actions
Data sharing		
Policy for variants established Variants database for each pest (i.e., multiplexing, cross-sample contamination, cross-experiment contamination) Establish local pests and variant laboratory databases Establish NGS/HTS data sharing guidelines to appropriate public databases, quality assurance and dissemination approval process Clear protocol for unsolicited data	SOP 11 Data sharing	Identify a secure national data transfer platform with clear access permission control to enable data sharing for surveillance and detection, in line with processes adopted by the Communicable Diseases Genomics Network
Validation		
Validation, interpretation, and data dissemination Validation of HTS diagnosis – independent testing approaches (i.e., PCR and Sanger sequencing), use validated molecular tests or design new assays, verify true and false positives, guideline for verification (i.e., same nucleic acid extract vs fresh/additional nucleic acid extracts) Scope of pest validation – pose plant health risk, regulated vs non-regulated pests, endemic pests of interest	SOP 12 Validation	
Training		
	SOP 13 - Training	
The national HTS standards framework (workshop recommendation) should: Identify a working party to harmonise and standardise biosecurity HTS use nationally (the NPAAC guide which covers staff competencies, laboratory procedures and bioinformatics could be used as a model with appropriate adaptions) Implement an on-going training program in bioinformatics Expand HTS staff competency by developing and delivering national HTS and bioinformatics workshops Support and fund existing biosecurity HTS networks to develop a certified education and training program (similar to the RCPA Education program) with a focus on demystifying HTS assays and reporting and supporting laboratory to adopt HTS Identify key HTS leads to champion development of networks with academia, state and federal biosecurity including key healthcare experts as they are further advanced and offer valuable learnings (the PHLN network is a model to review)		

The DAWE HTS development pathway

Prepared by Mark Whattam

DAWE's HTS Journey

Individuals within DAWE have been championing the use of HTS in PEQ for almost a decade. Between 2014-2018, DAWE was a key partner in a Plant Biosecurity Co-operative Research Centre project with Murdoch University and New Zealand Ministry of Primary Industry that demonstrated an early 'proof of concept' using small RNA-seq HTS for testing viruses and viroids in a broad range of imported plants in PEQ (Barrero et al, 2017). This project demonstrated strong correlation for major horticulture crops using side-by-side trials comparing existing PEQ diagnostics with small RNA-seq HTS methods.

In 2019, DAWE contributed funding to a four year $1.5M project with Horticulture Innovation, Queensland University of Technology (QUT), and Agriculture Victoria to 'Improve plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing'. The project has three clear objectives:

i) optimise HTS methodology and develop 'end to end' standard operating procedures
ii) compare HTS with current PEQ diagnostics
iii) develop a regulatory policy to adopt HTS in PEQ

A key output from the project will be publications of peered reviewed scientific papers with technical details of the HTS methodology and results. This will directly provide DAWE with the technical justifications and developing a science based and transparent adoption of a proven methodology.

In February 2020, DAWE helped co-ordinate a National Biosecurity Committee (NBC) workshop with the outcome being to develop a strategic roadmap to progress the development and adoption of HTS as a diagnostic tool across all Australian jurisdictions. A small working group was established including Dr Adrian Dinsdale as the DAWE representative to develop an implementation plan to use HTS as a biosecurity diagnostic tool (estimated @ July 2021). A national workshop with DAWE and State/Territory (via LEADRR and SPHDs) regulatory agencies will present the findings to help progress development of a broad national HTS policy (working group to facilitate by mid-2021) noting the DAWE HTS plant biosecurity policy will be a sub-component of this broader strategy.

In mid-December 2020, DAWE commenced funding a two-year non-ongoing bioinformatician position with QUT to work alongside DAWE diagnosticians to become operationally ready to use HTS in PEQ by December 2022.

Internationally DAWE has played an active role in progressing the global development of HTS for PEQ purposes. At the IPPC Commission on Phytosanitary Measures (CPM) 13th meeting in Rome (April, 2017), Mark Whattam presented a paper prepared by Australia, New Zealand and the European and Mediterranean Plant Protection Organization on 'High-throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes' (https://www.ippc.int/en/news/cpm-13-side-session-on-gene-sequencing-and-molecular-technologies/). Whilst there were some minor modifications, the recommendation was adopted at the CPM 14th meeting held in Rome (April, 2019) (https://www.ippc.int/static/media/files/publication/en/2019/07/2._CPM-14_2019_Update.pdf).

Whilst progress is slowly being made at the international level, given the unspecified timeframe and complexity to get all countries to progress a global HTS standard combined with the increasing need to implement more efficient and effective testing in Australia's PEQ, the project team continues to actively progress development of a HTS policy to use for testing plant viruses and viroids in PEQ.

In April 2021, PIC@PEQ facilitated a highly successful HTS workshop with project partners and DAWE representatives from SSG, PIO, PHP and PSRA. The workshop provided an update on the technical outputs from the project and progress by the National Biosecurity Committee and Subcommittee of Plant Health Diagnostics on national HTS frameworks. A review of existing HTS regulatory policy documents of relevance including animal and human health policies was provided. PHP and PSRA also provided presentations on the requirements to develop a regulatory policy to use HTS in PEQ. Actions from the workshop included:

i) Development of a HTS Regulatory Policy Working Group to be Chaired by David Dall
ii) Agreement to develop and implement a HTS policy for testing viruses and viroids of biosecurity concern in high-risk plants in PEQ for viruses/viroids by December 2022
iii) Development of a roadmap with specific actions, timeframes, and people responsible for completing outputs

What are other regulators doing?

HTS is now becoming a routine diagnostic tool in many Australian and overseas research laboratories. For instance, HTS technologies have been used by the Victorian State Department of Jobs, Precincts and Regions (DJPR) in collaboration with national and international groups to decipher the genomes of diverse bacterial pathogens using a pan-genomic approach to design new standard diagnostics tests such as polymerase chain reaction (PCR) since 2007. There is an opportunity to build upon this capability to move HTS from a research and development tool to a powerful diagnostic tool for rapid detection and accurate identification of viruses in imported horticultural species in PEQ.

Regulatory agencies are slowly progressing trialling and using HTS technologies in regulatory settings.

- NBC/SPHD: Progressing development of a national HTS framework
- USDA: Provisional release of plants from clean breeding programs for propagation with HTS negative results
- CFIA: Restricted release of imported plants for trials, developing policy/technical guidelines for release of imported plants using HTS
- CFIA: CLEan plAnt extractioN SEquencing Diagnostics (CLEANSED) for Clean Grapevines in Canada. HTS being developed to reduce quarantine time for grapevine from up to 3 years to 10 days
- EU: Formal HTS based regulation processes are being developed
- IPPC: Commission for Phytosanitary Measures recommendation "High throughput sequencing (HTS) technologies as a diagnostic tool for phytosanitary purposes"

Likewise, HTS is being used by animal and human health regulators which provides opportunities for development of this policy statement and not reinventing the wheel. HTS frameworks are captured in attached appendices.

References

Barrero, R.A., Napier, K.R., Cunnington, J., Liefting, L., Keenan, S., Frampton, R.A., Szabo, T., Bulman, S., Hunter, A., Ward, L., et al. (2017). An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. Bmc Bioinformatics 18.

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

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.

Maliogka, V.I., Minafra, A., Saldarelli, P., Ruiz-Garcia, A.B., Glasa, M., Katis, N., and Olmos, A. (2018). Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. Viruses-Basel 10.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1, 263-273.

[To be published in BICON]

Regulated pests of stonefruit (*Prunus* spp.)

Published: 3 May 2021

All diseases including:

Bacteria

Erwinia amylovora

Pseudomonas amygaldi

Pseudomonas syringae pv. *avii*

Pseudomonas syringae pv. *morsprunorum*

Pseudomonas syringae pv. *persicae*

Xylella fastidiosa

Fungi

Apiosporina morbosa

Ceratocystis spp.

Monilinia fructigena

Phomopsis amygdali

Monilinia polystroma

Phymatotrichum omnivorum

Phytophthora ramorum

Podosphaera clandestina (exotic strains)

Aerially dispersed *Phytophthora* spp.:

P. ramorum

P. kernoviae

P. pseudosyringae

P. nemorosa

Viruses

Apricot latent foveavirus

Apricot latent ringspot nepovirus

Arabis mosaic nepovirus

Asian prunus foveavirus (1,2,3)

Carnation Italian ringspot tombusvirus

Cherry leaf roll nepovirus

Cherry mottle leaf trichovirus

Cherry rasp leaf nepovirus

Cherry twisted leaf robigovirus

?Cherry twisted leaf foveavirus

Myrobalan latent ringspot nepovirus

Peach mosaic trichovirus

Peach rosette mosaic nepovirus

Petunia asteroid mosaic tombusvirus

Plum pox potyvirus

Raspberry ringspot nepovirus

Raspberry ringspot nepovirus – cherry strain

Strawberry latent ringspot nepovirus

?Strawberry latent ringspot sadwavirus

Stocky prune cheravirus

Tobacco necrosis alphanecrovirus – plum strain

Tomato black ring nepovirus

Tomato ringspot nepovirus

Viroids

Apple scar skin viroid (ASSVd)

Phytoplasmas

Candidatus Phytoplasma asteris (16SrI)

Candidatus Phytoplasma aurantifolia (16SrII)

Candidatus Phytoplasma fraxini (16SrVII)

Candidatus Phytoplasma mali (16SrX-A)

Candidatus Phytoplasma phoenicium (16SrIX)

Candidatus Phytoplasma pruni (16SrIII)

Candidatus Phytoplasma prunorum (16SrX-B)

Candidatus Phytoplasma pyri (16SrX-C)

Candidatus Phytoplasma solani (16SrXII-A)

Candidatus Phytoplasma trifolii (16SrVI)

Candidatus Phytoplasma ulmi (16SrV-A

Candidatus Phytoplasma ziziphi (16SrV-B)

Draft Standard Operating Procedure Template

Content

1. **Scope**
2. **Definitions and abbreviations**
3. **Personnel pre-qualifications**
4. **Procedure**
   - 4.1 Method
   - 4.2 Equipment and materials
   - 4.3 Detailed instructions
   - 4.4 Data
   - 4.5 Quality management
   - 4.6 Waste management and other safety precautions
5. **Related documents**
6. **Rationale for change of SOP version**

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Scope

Definitions and abbreviations

Personnel pre-qualifications

- *3.1 Competencies*
- *3.2 Training and accreditation*

Procedure

- *4.1 Method*
- *4.2 Equipment and materials*
- *4.3 Detailed instructions*
- *4.4 Data*
  - *4.4.1 Recording*
  - *4.4.2 Analysis and interpretation*
  - *4.4.3 Reporting*
- *4.5 Quality management*
  - *4.5.1 Controls, including critical controls*
  - *4.5.2 Assurance – measuring and reporting against controls*
- *4.6 Waste management and other safety precautions*

Related documents

Rationale for change of SOP version

Appendix 21: Validation of the NGS pipeine for culturable bacteria

Introduction

Culturing pathogenic bacteria for whole genome sequencing is a means to isolate pathogen DNA from host and environmental DNA. In the previous milestone, we described a high throughput sequencing (HTS) pipeline to identify plant pathogenic bacteria from acquisition of bacterial cultures to bioinformatics data analysis as outlined in Figure 21.1. This includes protocols for DNA extraction, library preparation, sequencing and data analysis. Furthermore, we showed that shallow short read sequencing (~300 MB) using the Illumina platform or long read sequencing using the Oxford Nanopore Technology (ONT) MinION sequencing of genomic DNA extracted from cultured bacteria is often sufficient for identification of many pathogenic species down to the subspecies/pathovar level.

In this report, we further validate this pipeline by applying it to bacterial isolates from post entry quarantine (PEQ) samples supplied by the Science and Surveillance Group (SSG). To do so, PEQ sourced pure genomic DNA of 44 cultured bacteria isolated from various infected plants were run through the HTS pipeline. PEQ provided bacterial identification for these isolates which was undisclosed. Progress is outlined in this report.

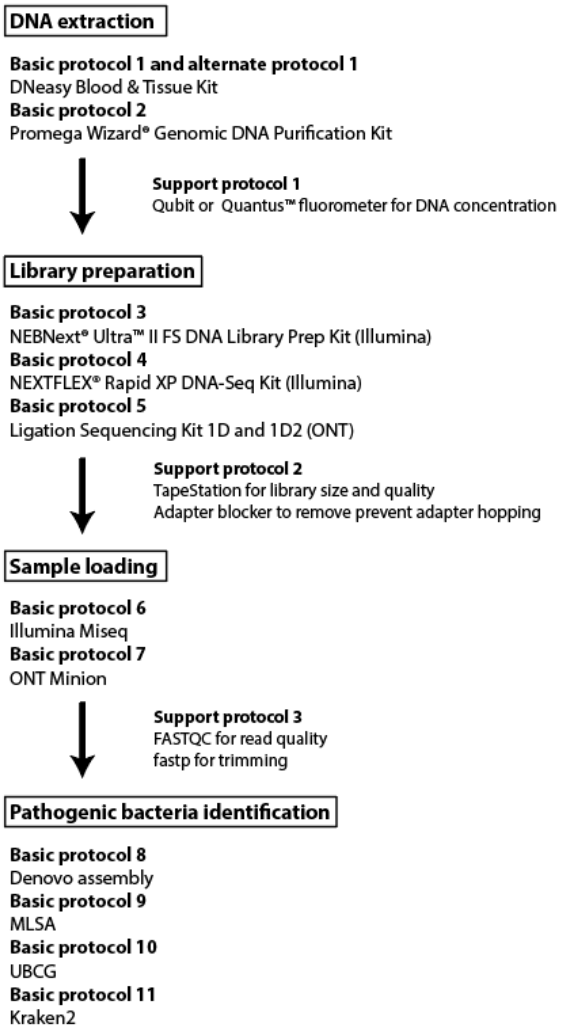


Figure 21.1 The high throughput sequencing pipeline used to identify culturable plant pathogenic bacteria.

Material and Methods

PEQ samples

Ten microlitre aliquots of genomic DNA of 44 cultured, pre-identified bacterial isolates (Table 21.1) were supplied by PEQ and the SSG community for pipeline validation.

Table 21.1 Post entry quarantine (PEQ) bacterial genomic DNA samples that were used to validate the high

throughput sequencing pipeline, including plant host and PEQ identification (ID).

Sample ID	Plant host	PEQ Molecular ID	DNA concentration (nanogram (ng)/microlitre (µl))
PEQ-1	*Sansevieria*	*Pantoea stewartii*	50.6
PEQ-2	*Solanum*	*Sphingomonas* sp.	102
PEQ-3	*Actinidia*	*Pseudomonas citronellolis* or *Hyphomicrobium facile*	1.48
PEQ-4	*Prunus*	*Paenibacillus* sp.	12.2
PEQ-5	*Actinidia*	*Staphylococcus epidermidis*	37.2
PEQ-6	*Prunus*	*Bacillus pumilus*, *invictae* or *safensis*	22.2
PEQ-7	*Solanum nigrum*	*Pseudomonas viridiflava*	20.6
PEQ-8	*Apium graveolens*	*Pectobacterium carotovorum*	32.4
PEQ-9	*Fragaria*	*Dyella* sp.	25.8
PEQ-11	*Juglans*	*Pantoea agglomerans* / *vagans*	6.92
PEQ-12	*Solanum*	*Paenibacillus* sp.	0.226
PEQ-13	*Solanum*	*Microbacterium paludicola*	3.16
PEQ-14	*Solanum*	*Chryseobacterium taeanense*	4.82
PEQ-15	*Solanum*	*Rhizobium radiobacter*	0.456
PEQ-16	*Solanum*	*Methylobacterium* sp. (*M. aquaticum*, *M. indicum*, *M. platani*, *M. currus*)	too low
PEQ-17	*Solanum*	*Alcaligenes faecalis*	1.34
PEQ-18	*Epipremnum*	*Stenotrophomonas maltophila*	15
PEQ-20	*Solanum melongena*	*Bacillus pumilus* or *B. safenis*	25.6
PEQ-21	*Solanum melongena*	*Bacillus* sp, *B. Licheniformis*, *B. subtilis* or *B. tequilensis*	18.2
PEQ-22	*Solanum melongena*	*Bacillus pumilus* or *B. safenis*, *B. australimari* or *B. invictae*	1.51
PEQ-23	*Brassica oleracea*	*Bacillus subtilis*	39.8
PEQ-24	*Ficus*	*Acinetobacter* sp.	12.5
PEQ-25	*Lilium*	*Rahnella aquatilis*	14.2
PEQ-26	*Phalaenopsis*	*Dickeya* sp.	106
PEQ-27	*Cynodon*	*Stenotrophomonas maltophilia*	19.6
PEQ-28	*Phalaenopsis*	*Dickeya fangzhongdai*	39.2
PEQ-29	*Fragaria*	*Luteibacter* sp. or *Dyella* sp.	24.6
PEQ-30	*Fragaria*	*Luteibacter yeojuensis*	0.4

PEQ-31	Sugar snap peas	*Pseudomonas syringae*	14.3
PEQ-32	Sugar snap peas	*Pseudomonas syringae*	14.7
PEQ-33	*Phalaenopsis*	*Dickeya fangzhongdai*	31.2
PEQ-34	*Phalaenopsis*	*Dickeya fangzhongdai*	130
PEQ-35	*Phalaenopsis*	*Dickeya fangzhongdai*	77.2
PEQ-36	King oyster mushrooms (multiple)	*Ewingella americana*	5.74
PEQ-37	King oyster mushrooms (multiple)	*Ewingella americana*	177
PEQ-38	King oyster mushrooms (multiple)	*Ewingella americana*	118
PEQ-39	Chrysenthemum	*Pseudomonas* sp.	32.8
PEQ-40	*Phalaenopsis*	*Dickeya fangzhongdai*	118
PEQ-41	*Phalaenopsis*	*Dickeya fangzhongdai*	137.6
PEQ-42	*Zamioculcas zamiifolia*	*Dickeya dadantii*	18.8
PEQ-43	*Phalaenopsis*	*Dickeya* sp.	11.2
PEQ-44	*Echinocactus grusonii*	*Dickeya dadantii*	21
PEQ-45	*Sansevieria trifasciata*	*Dickeya zeae*	43.2
PEQ-46	*Sansevieria cylindrica*	*Dickeya zeae*	36

Illumina library preparation and sequencing

Using the Quantus™ fluorometer, the concentration of each DNA sample was measured and ~8 ng DNA was used for each library preparation. For samples with low concentrations such as PEQ-12, PEQ-15 and PEQ-30 (0.226-0.456 ng/µL) and no detection of DNA for PEQ-16, 8 µL of sample was used. Libraries were generated using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) with PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A no template control library was also generated. Libraries were then paired end sequenced (2 × 250 base pairs (bp)) using Illumina NovaSeq.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen et al., 2018) and subsequently to the k-mer-based analysis tool, Kraken2 (Wood et al., 2019) for OTU classification using a custom database containing all completed bacterial reference genomes in NCBI (19/03/2021). Reads were *de novo* assembled using SPAdes v3.15.2 (Bankevich et al., 2012). Assembled contigs were compared to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990). Depending on bacterial genera and the quality of data, 35 PEQ samples were chosen for up-to-date bacterial core gene (UBCG) analysis (Na et al., 2018).

Kraken2 and BLASTn results detected mixed bacterial genera in some samples. Sequences of the identified bacterial genus in mixed samples were isolated from the *de novo* assembled contigs using an in-house Perl script and BLASTn results. Isolated sequences were used for UBCG phylogenetic analysis.
N50 is the length (bp) of the shortest contig in the top 50% of total contig length. This number is used to describe the quality of an assembly, where the higher the number, the better the assembly. N50 was calculated using an open source n50 program on miniconda (https://docs.conda.io/en/latest/index.html).

Results

Sequencing output

Information of trimmed reads and contigs for each sample are listed in Table 21.2. A small number of reads were observed in the negative control (Blank). Sequencing depth for each sample ranged between 275 MB to 1.57 GB except for PEQ-12 and PEQ-16 which had 176 MB and 21 MB respectively. Samples with low contig numbers (below 1,270 contigs) were shown to have a higher N50 value (above 65,000 bp).

Table 21.2 Sequencing output for post entry quarantine (PEQ) genomic DNA samples.

Sample ID	Sequence (MB)	Total reads	Contigs	N50	Largest contig (base pairs)
Blank	0.241	270	2	131	131
PEQ-1	1200	10756220	18049	633	8024
PEQ-2	762	7006942	21127	507	5808
PEQ-3	607	5421992	15070	645	7325
PEQ-4	577	5391712	10161	1977	17164
PEQ-5	389	3394528	83	146754	217815
PEQ-6	510	4465548	26303	541	4882
PEQ-7	586	5446496	14702	1026	10936
PEQ-8	551	4996890	88	196221	630970
PEQ-9	587	5364106	15233	675	7609
PEQ-11	618	5706106	15378	683	9090
PEQ-12	176	1490704	3524	6379	42289
PEQ-13	710	6471434	41315	521	3969
PEQ-14	602	5284776	19710	567	6698
PEQ-15	422	3763484	10540	3803	37607
PEQ-16	21	189342	6491	539	6343
PEQ-17	626	5804108	27407	569	6018
PEQ-18	576	5050458	11	2343140	2343140
PEQ-20	398	3581744	32	970881	1311081
PEQ-21	428	3785320	10980	892	8831
PEQ-22	540	4799350	14300	1079	9622
PEQ-23	275	2410648	12073	781	8468
PEQ-24	504	4514220	19313	1636	21892
PEQ-25	363	3142136	11067	1394	14278
PEQ-26	526	4725898	10397	2917	24347
PEQ-27	289	2830906	9593	1338	35940

					18803
PEQ-28	406	3820630	5128	3180	
					7152
PEQ-29	526	4787732	19714	586	
					41893
PEQ-30	508	4600396	7000	4619	
					8143
PEQ-31	535	4852388	15084	951	
					1274174
PEQ-32	791	6871406	64	764793	
					18749
PEQ-33	340	3063316	6550	2166	
					13577
PEQ-34	354	3170166	9302	1408	
					333749
PEQ-35	290	2719566	1049	65430	
					6746
PEQ-36	399	3533330	20938	692	
					22171
PEQ-37	559	5292286	5260	3167	
					11968
PEQ-38	856	7565444	45170	692	
					7128
PEQ-39	861	7696370	32778	544	
					860408
PEQ-40	666	8523194	357	243950	
					860464
PEQ-41	1572	20295704	1270	235164	
					705483
PEQ-42	757	9664202	389	245826	
					677711
PEQ-43	804	10238804	118779	479	
					1969058
PEQ-44	1227	15774710	910	332711	
					364579
PEQ-45	658	8433280	417	207181	
					635072
PEQ-46	518	6612838	331	145941	

Bioinformatic analysis, molecular identification, and comparison to previous identification

The HTS pipeline for culturable bacteria, Kraken2 for genus identification followed by BLASTn and UBCG phylogenetic analysis, gave bacterial identifications similar to PEQ (Table 21.3): *Pantoea stewartii* closely related to *Pantoea stewartii* subsp. *stewartii* (PEQ-1; Figure 21.2), *Pseudomonas viridiflava* (PEQ-7; Figure 21.5), *Bacillus subtilis* (PEQ-23; 4), *Dickeya fangzhongdai* (PEQ-28, 33, 34, 35, 40 and 41; Figure 21.11), *Pseudomonas syringae* closely related to *Pseudomonas syringae dysoxyl* (PEQ-31 and 32; Figure 21.5), *Pseudomonas* sp. closely related to *Pseudomonas fluorescence* (PEQ-39, Figure 21.5), *Dickeya dadantii* (PEQ-42; Figure 21.11) and *Dickeya zeae* (PEQ-45; Figure 21.11).

The HTS pipeline for culturable bacteria further clarified some PEQ samples which had multiple species identification per sample (Table 21.3) with results giving *Pantoea eucalypti* for PEQ-11 (Figure 21.2), *Bacillus safensis* subsp. *safensis* for PEQ-6 and 20 (Figure 4), *Bacillus subtilis* for PEQ-21 (Figure 4), *Luteibacter* sp. for PEQ-29 (Figure 21.7) and *Dickeya dadantii* subsp. *dadanti* for PEQ-44 (Figure 21.11).

Within this study, the HTS pipeline for culturable bacteria also reclassified the species for some PEQ identifications (Table 21.3), reclassifying as *Sphingomonas melonis* for PEQ-2 (Figure 21.3), *Pectobacterium versatile* for PEQ-8 (Figure 21.6), *Rahnella* sp. closely related to *Rahnella aqualitis* for PEQ-25 (Figure 21.9), *Luteibacter* sp. for PEQ 30 (Figure 21.7) and *Dickeya oryzae* for PEQ-46 (Figure 21.11).

The HTS pipeline for culturable bacteria also reclassified the genus for some PEQ identifications (Table 21.3), reclassifying as *Luteibacter* sp. for PEQ-9 (Figure 21.7), *Curtobacterium flaccumfaciens* sp. for PEQ-17 (Figure 21.8), *Burkholderia contaminans* for PEQ-26 (Figure 21.10) and *Pseudomonas* sp. for PEQ-36 (Figure 21.5). Kraken2 analysis using the HTS pipeline for culturable bacteria managed to identify extra bacterial genera in

addition to those identified by PEQ. Among those samples were PEQ-15 and 22 where the highest proportion of reads were *Agrobacterium* (76%) and *Bacillus* (95%), respectively. The second bacterial genera identified was *Achromobacter* with 19% reads for PEQ-15 and *Stenotrophomonas* with 3.55% reads for PEQ-22 (Table 21.3). BLASTn results was in line with Kraken2 results, identifying both genera per sample. Contigs of *Agrobacterium* and *Bacillus* were isolated using BLASTn results. Both full contigs (PEQ-15 and PEQ-22) and isolated *Agrobacterium* (PEQ-15_Agro) and *Bacillus* (PEQ-22_Bacillus) contigs of PEQ-15 and 22, respectively, were subjected to UBCG phylogenetic analysis. Results showed that isolated *Agrobacterium* and *Bacillus* contigs gave better identification in comparison to full contig analysis identifying PEQ-15 as *Rhizobium* (*Agrobacterium*) *rhizogenes* (Figure 21.12) and PEQ-22 as *Bacillus safensis* subsp. *safensis* (Figure 21.4). With the altered pipeline, results reclassified the genus for PEQ-15 and further clarified the species for PEQ-22. Extra bacterial genera were also detected in PEQ-24 and 43 where the highest proportion of reads were *Sphingomonas* (42%) and *Pseudomonas* (40%), respectively. The second bacterial genera identified was *Curtobacterium* with 10.92% reads for PEQ-24 and *Dickeya* with 18% reads for PEQ-43 (Table 21.3). BLASTn results was in line with Kraken2 results, identifying both genera per sample. Contigs of all four detected genera were isolated using BLASTn results. Isolated contigs, *Sphingomonas* (PEQ-24_Sphingomonas), *Curtobacterium* (PEQ-24_Curtobacterium), *Pseudomonas* (PEQ-43_Pseudomonas) and *Dickeya* (PEQ-43_Dickeya) contigs were subjected to UBCG phylogenetic analysis. Results showed that all four isolated genera, irregardless of read %, were identified; *Sphingomonas* sp. closely related to *Sphingomonas paucimobilis* (Figure 21.3) and *Pseudomonas* sp. (Figure 21.5) for PEQ-24 and *Curtobacterium* sp. closely related to *Curtobacterium citreum* (Figure 21.8) and *Dickeya fangzhongdai* (Figure 21.11) for PEQ-43. With the altered pipeline, results reclassified the genus for PEQ-24 and classified a new bacterial genus while classifying a new bacterial genus and finding the PEQ identified genus of PEQ-22 as the second bacterial genus of the sample and reclassifying the species.

A problem occurred while using the HTS pipeline for culturable bacteria on PEQ-37 and 38. Kraken2 analysis for PEQ-37 classified the genus as *Rahnella* with a low read percentage of 9.65% while for PEQ-38, two bacterial genera was identified, *Pseudomonas* (25.38% reads) and the second genus, *Rahnella* with (8.72% reads). BLASTn identification was in line with Kraken2 results. Contigs of all three detected genera were isolated using BLASTn results. Full contigs of PEQ-37 and isolated *Rahnella* (PEQ-37_Rahnella) along with isolated contigs of *Pseudomonas* (PEQ-38_Pseudomonas) and *Rahnella* (PEQ-38_Rahnella) from PEQ-38 were subjected to UBCG phylogenetic analysis. Results from UBCG contradicted Kraken2 and BLASTn results, classifying the *Rahnella* genus for both PEQ-37 and 38 as *Ewingella americana* (Figure 21.9), similar to PEQ identification. UBCG results identified the *Pseudomonas* from PEQ-38 as a *Pseudomonas* sp. closely related to *Pseudomonas libanensis* (Figure 21.5). Results also showed that there was no difference between isolated *Rahnella* contigs and full contigs for PEQ-37 even though read percentage was low.

High proportion of unclassified reads (above 20%) were observed PEQ-9, 12, 13, 14, 24, 29, 30, 37, 43 and 48. This data corelates with high no-hit contigs from BLASTn (above 18%). These samples also showed low percentage of reads (below 54%) for genera identified by Kraken2 and has some association with UBCG identification to genus only, though it was also observed in PEQ-36 and 39 which had above 97% reads. Using the BLASTn search results, no-hit contigs were isolated from the SPAdes assembled genome and then subjected to another comparison to the non-redundant database at NCBI using a less stringent BLASTn algorithm. Most hits had less than 50% query coverage indicating they are either unidentified organisms or the *denovo* assembly was of poor quality, which could have happened as assembly was based on isolates instead of metagenomic sequencing (data not shown).

Table 21.3 Identification of PEQ samples from PEQ, k-mer analysis (Kraken2), BLASTn analysis and UBCG phylogenetic analysis. Kraken2 was used to classify each sequence read identifying bacterial genera with the highest read percentages, BLASTn was used to compare assembled contigs to the non-redundant database at NCBI and UBCG phylogenetic analysis was used to extract 92 genes from assembled contigs for phylogenetic analysis with reference genomes for identification.

		KRAKEN2			BLASTN		UBCG	
SAMPLE	**PEQ identification**	**Unclassified reads**	**Kraken2 (genus)**	**Reads (%)**	**BLASTn**	**No-hit contigs (%)[1]**	**Phylogeny**	**No. of genes out of 92**
BLANK	-	13.33	*Clostridium*	15.56		2/2 (100%)		
PEQ-1	*Pantoea stewartii*	4.26	*Pantoea*	92.88	*Pantoea stewartii* subsp. *stewartii*	1100/18049 (6.09%)	*Pantoea stewartii* subsp. *stewartii*	80
PEQ-2	*Sphingomonas* sp.	7.34	*Sphingomonas*	88.29	*Sphingomonas melonis*	966/21127 (4.57%)	*Sphingomonas melonis*	68
PEQ-3	*Pseudomonas citronellolis* or *Hyphomicrobium facile*	12.38	*Hyphomicrobium*	85.24	*Hyphomicrobium* sp.	1087/15070 (7.21%)		72
PEQ-4	*Paenibacillus* sp.	7.82	*Paenibacillus*	91.44	*Paenibacillus* sp.	601/10161 (5.91%)		86
PEQ-5	*Staphylococcus epidermidis*	1.93	*Staphylococcus*	97.04	*Staphylococcus epidermidis*	6/83 (7.23%)		92
PEQ-6	*Bacillus pumilus*, *invictae* or *safensis*	0.39	*Bacillus*	99.2	*Bacillus safensis*	154/26303 (0.59%)	*Bacillus safensis* subsp. *safensis*	77
PEQ-7	*Pseudomonas viridiflava*	3.54	*Pseudomonas*	96.03	*Pseudomonas viridiflava*	261/14702 (1.78%)	*Pseudomonas viridiflava*	77

PEQ-8	*Pectobacterium carotovorum*	1.57	*Pectobacterium*	96.16	*Pectobacterium carotovorum*	6/88 (6.82%)	*Pectobacterium versatile*	92
PEQ-9	*Dyella* sp.	41.67	*Luteibacter*	21.02	*Luteibacter rhizovicinus*	5423/15233 (35.6%)	*Luteibacter* sp.	76
PEQ-11	*Pantoea agglomerans / vagans*	6.02	*Pantoea*	87.69	*Pantoea agglomerans / vagans*	552/15378 (3.59%)	*Pantoea eucalypti*	77
PEQ-12	*Paenibacillus* sp.	57.62	*Paenibacillus*	29.41	*Paenibacillus* sp.	1602/3524 (45.46%)		90
PEQ-13	*Microbacterium paludicola*	32.73	*Microbacterium*	26.9	*Microbacterium* sp.	14326/41315 (34.68%)		70
PEQ-14	*Chryseobacterium taeanense*	39.2	*Chryseobacterium*	53.01	*Chryseobacterium* sp.	6396/19710 (32.45%)		66
PEQ-15_AGRO[2]	*Rhizobium radiobacter*	2.81	*Rhizobium* (*Agrobacterium*)[3]	75.98	*Rhizobium* (*Agrobacterium*) *rhizogenes*[4]	154/10540 (1.46%)	*Rhizobium rhizogenes*	81
PEQ-15_ACHRO[3]	-		*Achromobacter*	19.00	*Achromobacter xylosoxidans*			
PEQ-16	*Methylobacterium* sp. (*M. aquaticum, M. indicum, M. platani, M. currus*)	12.68	*Methylobacterium*	54.66	*Methylobacterium aquaticum*	749/6491 (11.54%)		
PEQ-17	*Alcaligenes faecalis*	7.17	*Curtobacterium*	86.29	*Curtobacterium flaccumfaciens* pv. f*laccumfaciens*	1173/27407 (4.28%)	*Curtobacterium flaccumfaciens* sp.	72
PEQ-18	*Stenotrophomonas maltophila*	3.98	*Stenotrophomonas*	93.65	*Stenotrophomonas maltophilia*	4/11 (36.36%)		92

PEQ-20	*Bacillus pumilus* or *B. safenis*	0.8	*Bacillus*	98.69	*Bacillus safensis*	4/32 (12.5%)	*Bacillus safensis* subsp. *safensis*	92
PEQ-21	*Bacillus* sp., *B. licheniformis*, *B. subtilis* or *B. tequilensis*	0.63	*Bacillus*	98.63	*Bacillus subtilis*	347/10980 (3.16%)	*Bacillus subtilis*	78
PEQ-22_ BACILLUS[2]	*Bacillus pumilus* or *B. safenis*, *B. australimari* or *B. invictae*	0.72	*Bacillus*	95.13	*Bacillus safensis*	177/14300 (1.24%)	*Bacillus safensis* subsp. *safensis*	83
PEQ-22_STENO[3]	-		*Stenotrophomonas*	3.55	*Stenotrophomonas maltophilia*			
PEQ-23	*Bacillus subtilis*	0.75	*Bacillus*	98.67	*Bacillus subtilis*	139/12073 (1.15%)	*Bacillus subtilis*	85
PEQ-24_ BACILLUS[2]	*Acinetobacter* sp.	27.2	*Sphingomonas*	42.08	*Sphingomonas* sp.	3662/19313 (18.96%)	*Sphingomonas* sp.	74
PEQ-24_ CUTOBACTERIUM[3]	-		*Curtobacterium*	10.92	*Curtobacterium* sp.		*Curtobacterium* sp.	84
PEQ-25	*Rahnella aquatilis*	12.85	*Rahnella*	71.49	*Rahnella aquatilis*	825/11067 (7.45%)	*Rahnella* sp.	85
PEQ-26	*Dickeya* sp.	1.68	*Burkholderia*	94.69	*Burkholderia contaminans*	72/10397 (0.69%)	*Burkholderia contaminans*	85
PEQ-27	*Stenotrophomonas maltophilia*	4.37	*Stenotrophomonas*	86.02	*Stenotrophomonas maltophilia*	583/9593 (6.08%)		92
PEQ-28	*Dickeya fangzhongdai*	0.56	*Dickeya*	97.57	*Dickeya fangzhongdai*	94/5128 (1.83%)	*Dickeya fangzhongdai*	85

PEQ-29	*Luteibacter* sp. or *Dyella* sp.	41.97	*Luteibacter*	20.92	*Luteibacter* sp.	7447/19714 (37.78%)	*Luteibacter* sp.	70
PEQ-30	*Luteibacter yeojuensis*	43.61	*Luteibacter*	18.2	*Luteibacter rhizovicinus/pinsoli*	2341/7000 (33.44%)	*Luteibacter* sp.	92
PEQ-31	*Pseudomonas syringae*	1.16	*Pseudomonas*	98.67	*Pseudomonas syringae*	146/15084 (0.97%)	*Pseudomonas syringae*	79
PEQ-32	*Pseudomonas syringae*	0.98	*Pseudomonas*	98.88	*Pseudomonas syringae*	7/64 (10.94%)	*Pseudomonas syringae*	91
PEQ-33	*Dickeya fangzhongdai*	0.3	*Dickeya*	98.26	*Dickeya fangzhongdai*	64/6550 (0.98%)	*Dickeya fangzhongdai*	85
PEQ-34	*Dickeya fangzhongdai*	0.31	*Dickeya*	98.27	*Dickeya fangzhongdai*	55/9302 (0.59%)	*Dickeya fangzhongdai*	81
PEQ-35	*Dickeya fangzhongdai*	0.35	*Dickeya*	98.1	*Dickeya fangzhongdai*	80/1049 (7.63%)	*Dickeya fangzhongdai*	89
PEQ-36	*Ewingella americana*	1.4	*Pseudomonas*	97.64	*Pseudomonas* sp	123/20938 (0.59%)	*Pseudomonas* sp.	72
PEQ-37	*Ewingella americana*	43.34	*Rahnella*	9.65	*Rahnella* sp.	1423/5260 (27.05%)	*Ewingella americana*	88
PEQ-38_ RAHNELLA[2]	*Ewingella americana*	27.46	*Rahnella*	8.72	*Rahnella* sp./*aqualitis*	8155/45170 (18.05%)	*Ewingella americana*	79
PEQ-38_ PSEUDOMONAS[3]	-		*Pseudomonas*	25.38	*Pseudomonas libanensis*		*Pseudomonas* sp.	69
PEQ-39	*Pseudomonas* sp.	1.87	*Pseudomonas*	97.61	*Pseudomonas* sp.	304/32778 (0.93%)	*Pseudomonas* sp.	85
PEQ-40	*Dickeya fangzhongdai*	0.09	*Dickeya*	99.1	*Dickeya fangzhongdai*	4/357 (1.12%)	*Dickeya fangzhongdai*	72

PEQ-41	*Dickeya fangzhongdai*	0.07	*Dickeya*	99.13	*Dickeya fangzhongdai*	11/1270 (0.87%)	*Dickeya fangzhongdai*	89
PEQ-42	*Dickeya dadantii*	2.53	*Dickeya*	94.22	*Dickeya sp.*	10/389 (2.57%)	*Dickeya dadantii* sp.	89
PEQ-43_ DICKEYA[2]	*Dickeya* sp.	34.05	*Dickeya*	17.78	*Dickeya* sp.	85123/118779 (71.67%)	*Dickeya fangzhongdai*	90
PEQ-43_ PSEUDOMONAS[3]	-		*Pseudomonas*	40.22	*Pseudomonas* sp.		*Pseudomonas* sp.	92
PEQ-44	*Dickeya dadantii*	0.66	*Dickeya*	98.18	*Dickeya dadantii*	66/910 (7.25%)	*Dickeya dadantii* subsp. *dadantii*	91
PEQ-45	*Dickeya zeae*	1.1	*Dickeya zeae*	92.99	*Dickeya zeae*	6/417 (1.44%)	*Dickeya zeae*	89
PEQ-46	*Dickeya zeae*	1.29	*Dickeya zeae*	93.07	*Dickeya zeae*	2/331 (0.6%)	*Dickeya oryzae*	92

[1] Contigs unidentified by BLASTn
[2] Genus with highest read percentage in PEQ sample
[3] Extra genus detected in PEQ sample with the second highest read percentage
[4] *Agrobacterium rhizogenes* has been reclassified as *Rhizobium rhizogenes*

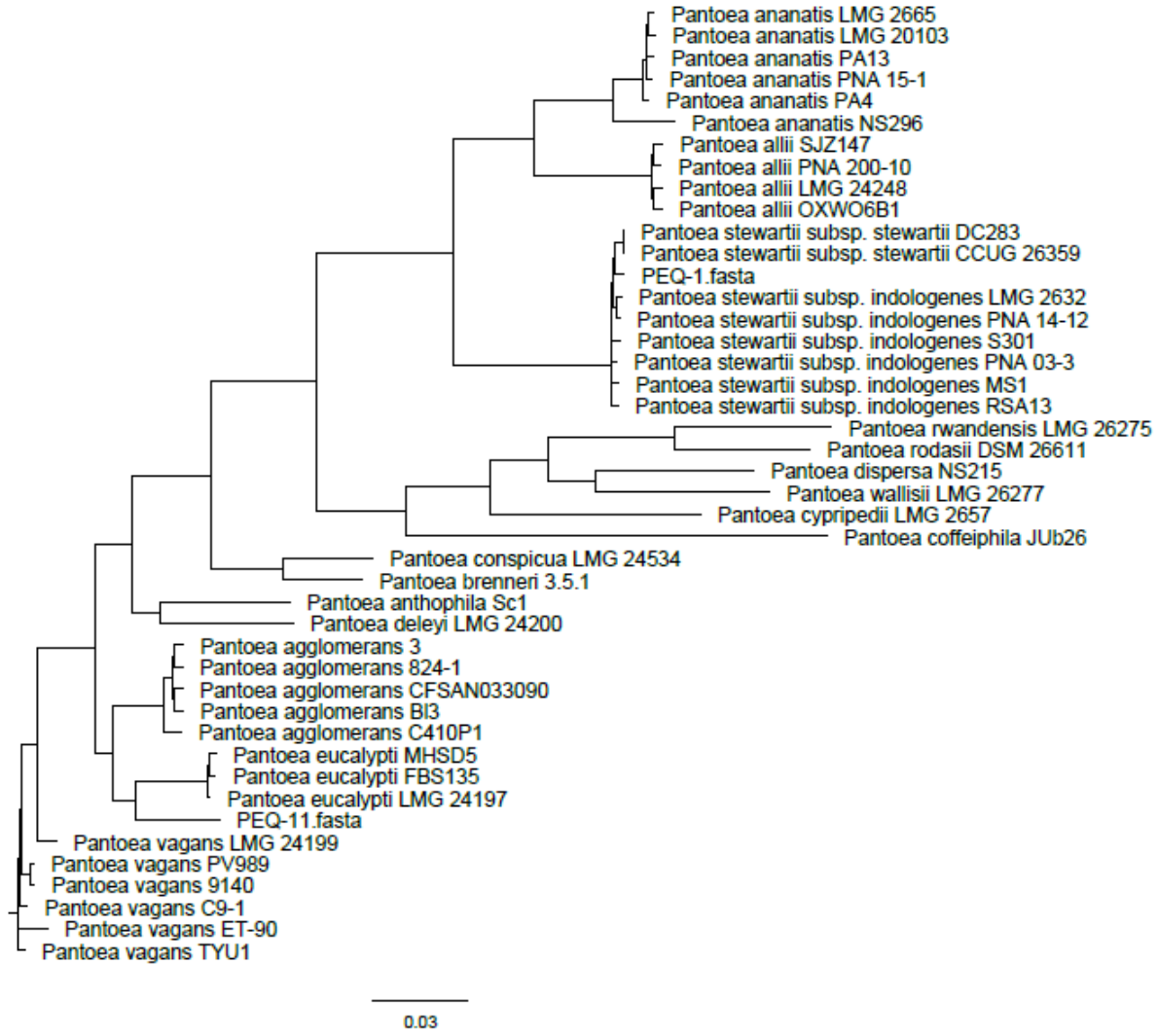


Figure 21.2 Phylogenetic position of PEQ genomes (PEQ-1 and 11) within the relevant *Pantoea* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

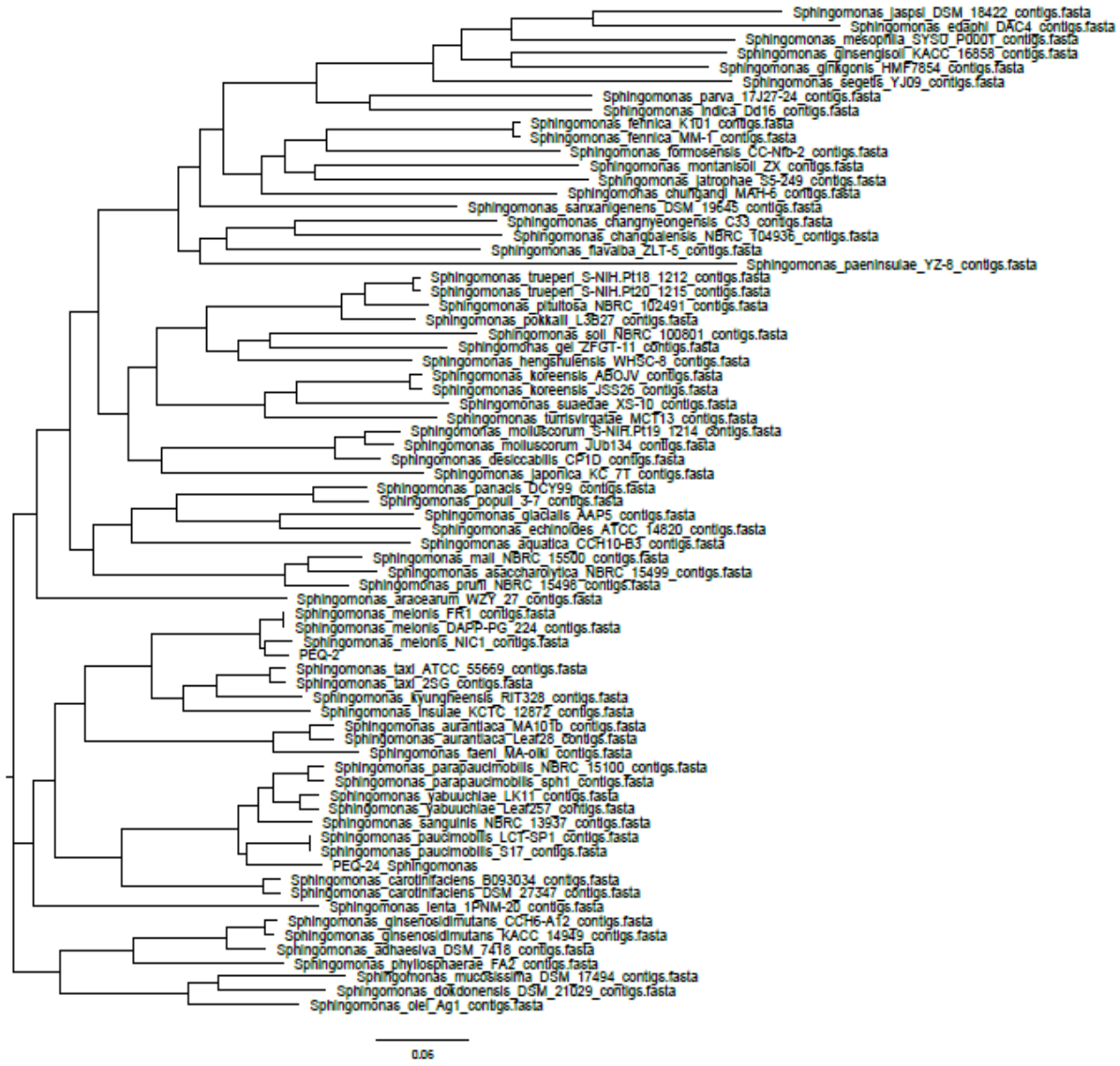


Figure 21.3 Phylogenetic position of PEQ genomes (PEQ-2) and the isolated *Sphingomonas* contigs (PEQ-24_Sphingomonas) within the relevant *Sphingomonas* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

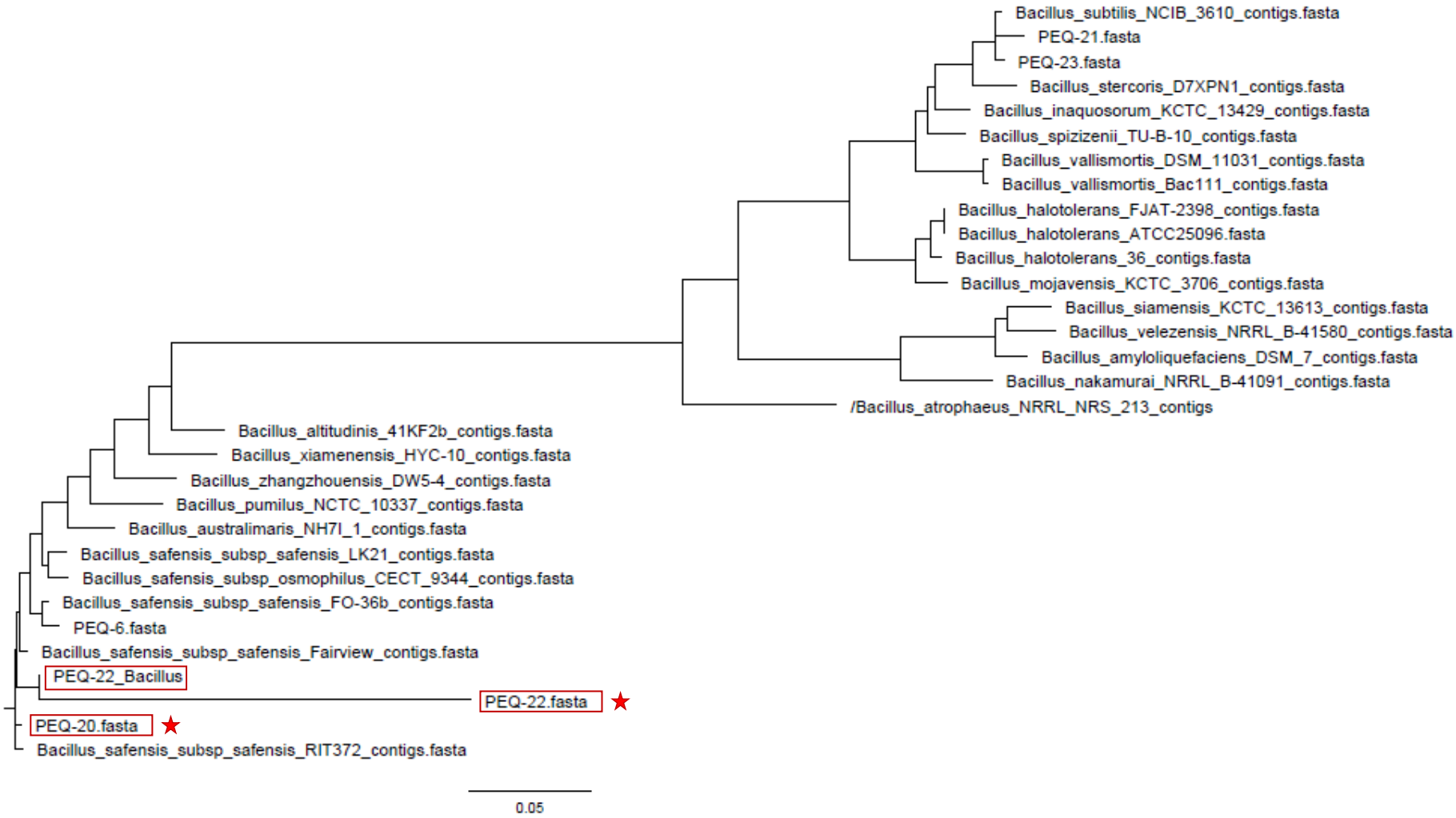


Figure 21.4 Phylogenetic position of PEQ genomes (PEQ-6, 20, 21, 22 and 23) and isolated *Bacillus* contigs (PEQ-22_Bacillus) within the relevant *Bacillus* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

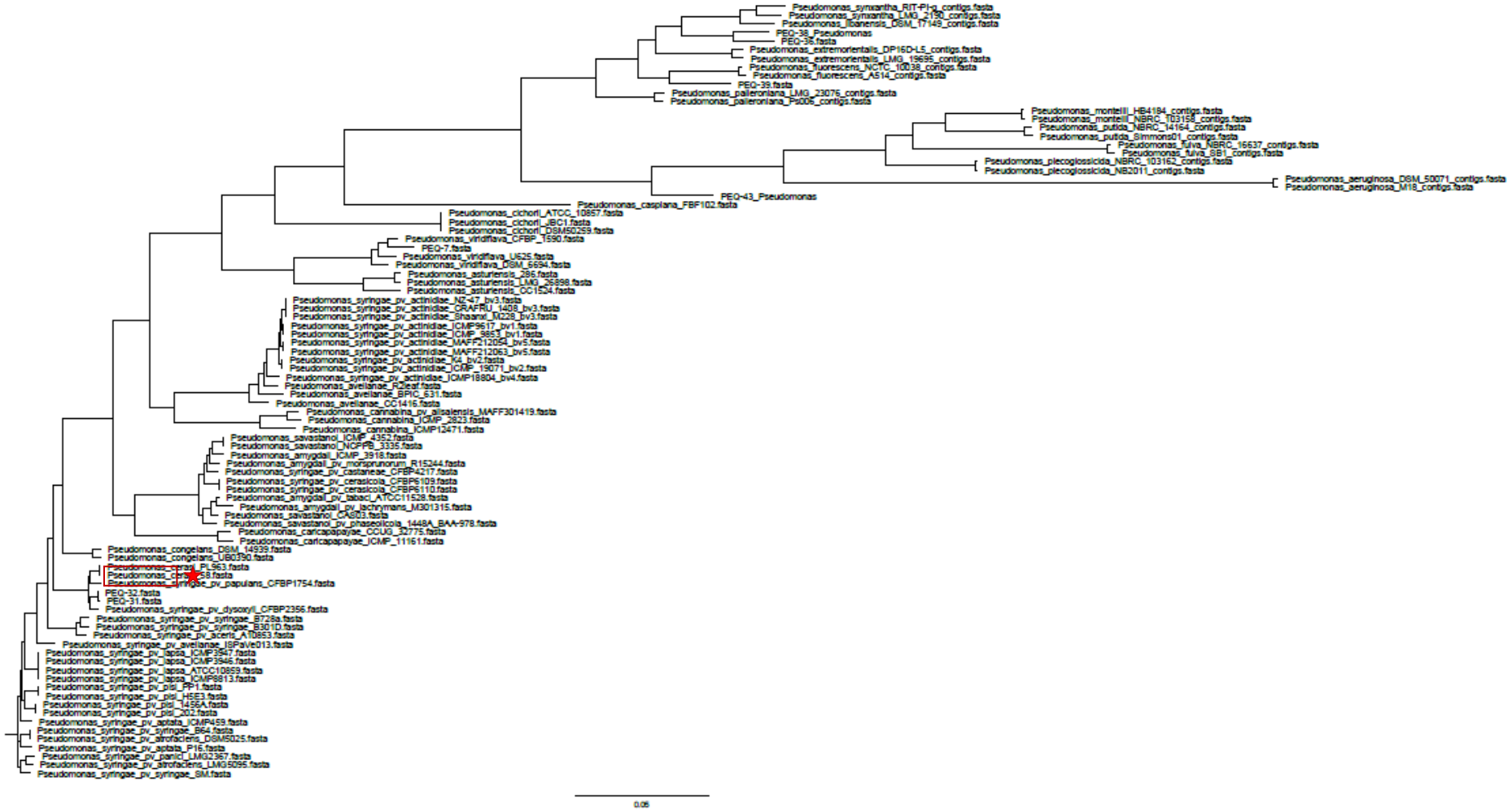


Figure 21.5 Phylogenetic position of PEQ genomes (PEQ-7, 31, 32, 36 and 39) and isolated *Pseudomonas* contigs (PEQ-38_Pseudomonas and PEQ-43_Pseudomonas) within the relevant *Pseudomonas* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

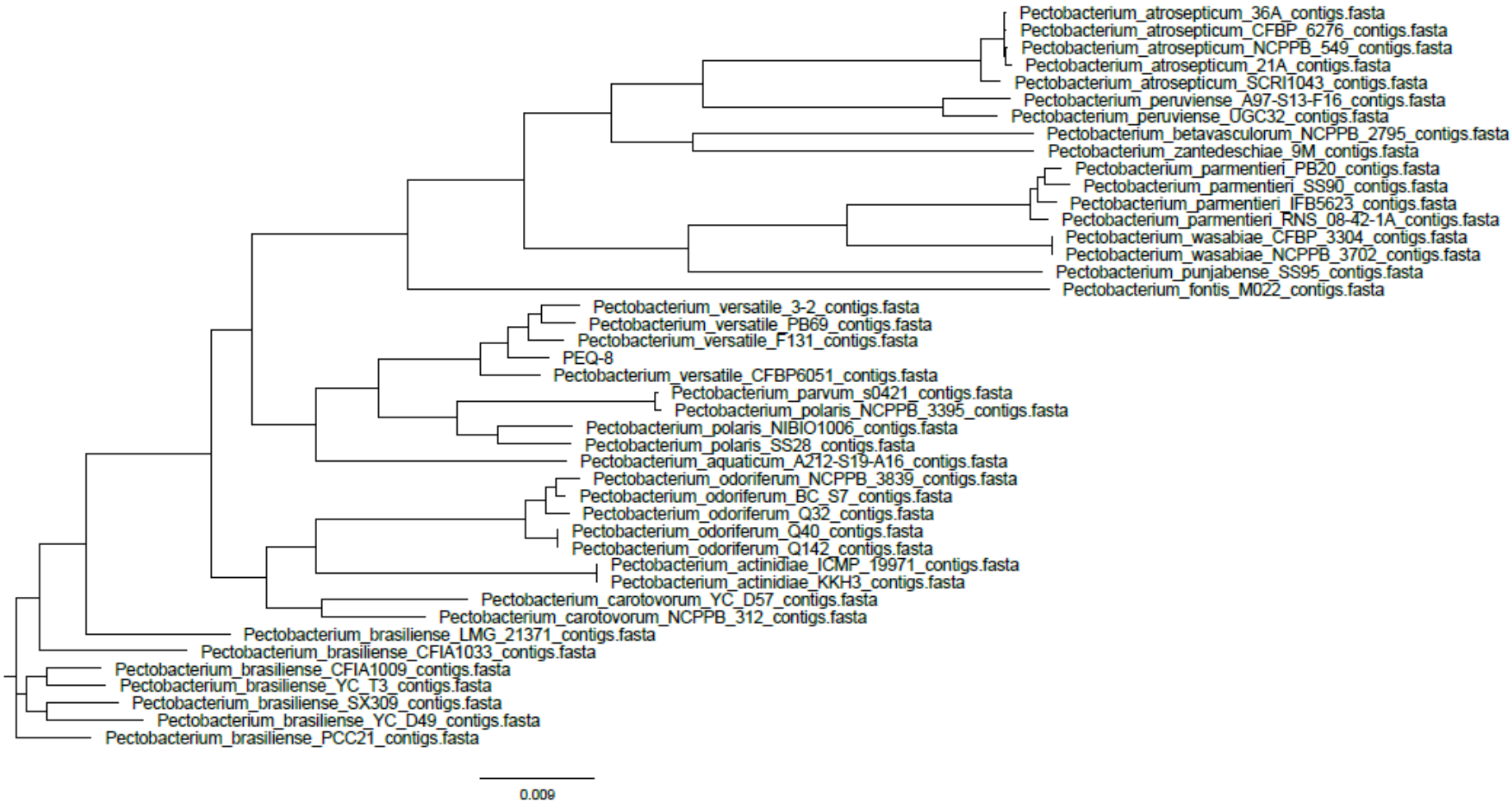


Figure 21.6 Phylogenetic position of PEQ genomes (PEQ-8) within the relevant *Pectobacterium* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

.

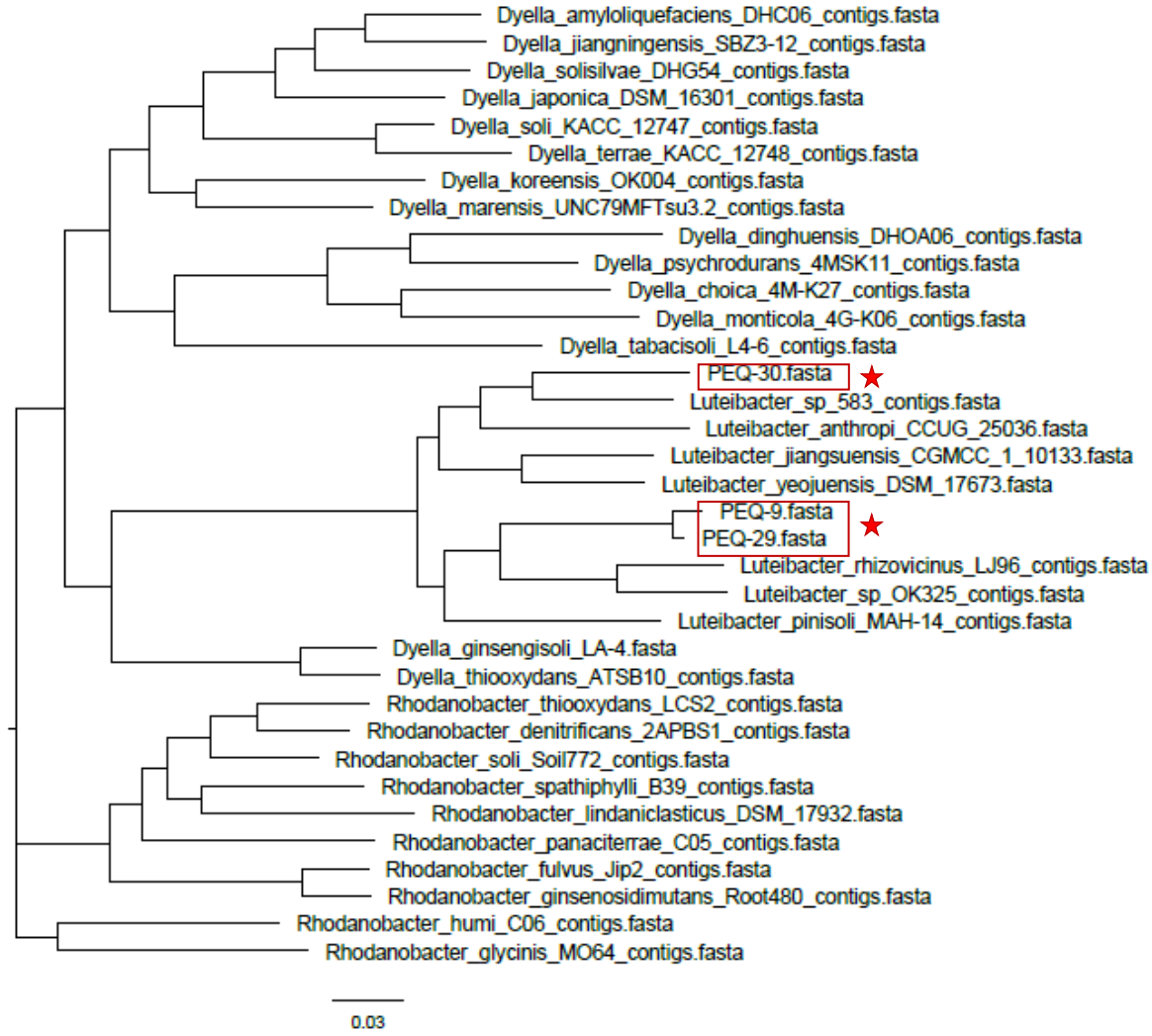


Figure 21.7 Phylogenetic position of PEQ genomes (PEQ-9, 29 and 30) within the relevant *Luteibacter* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

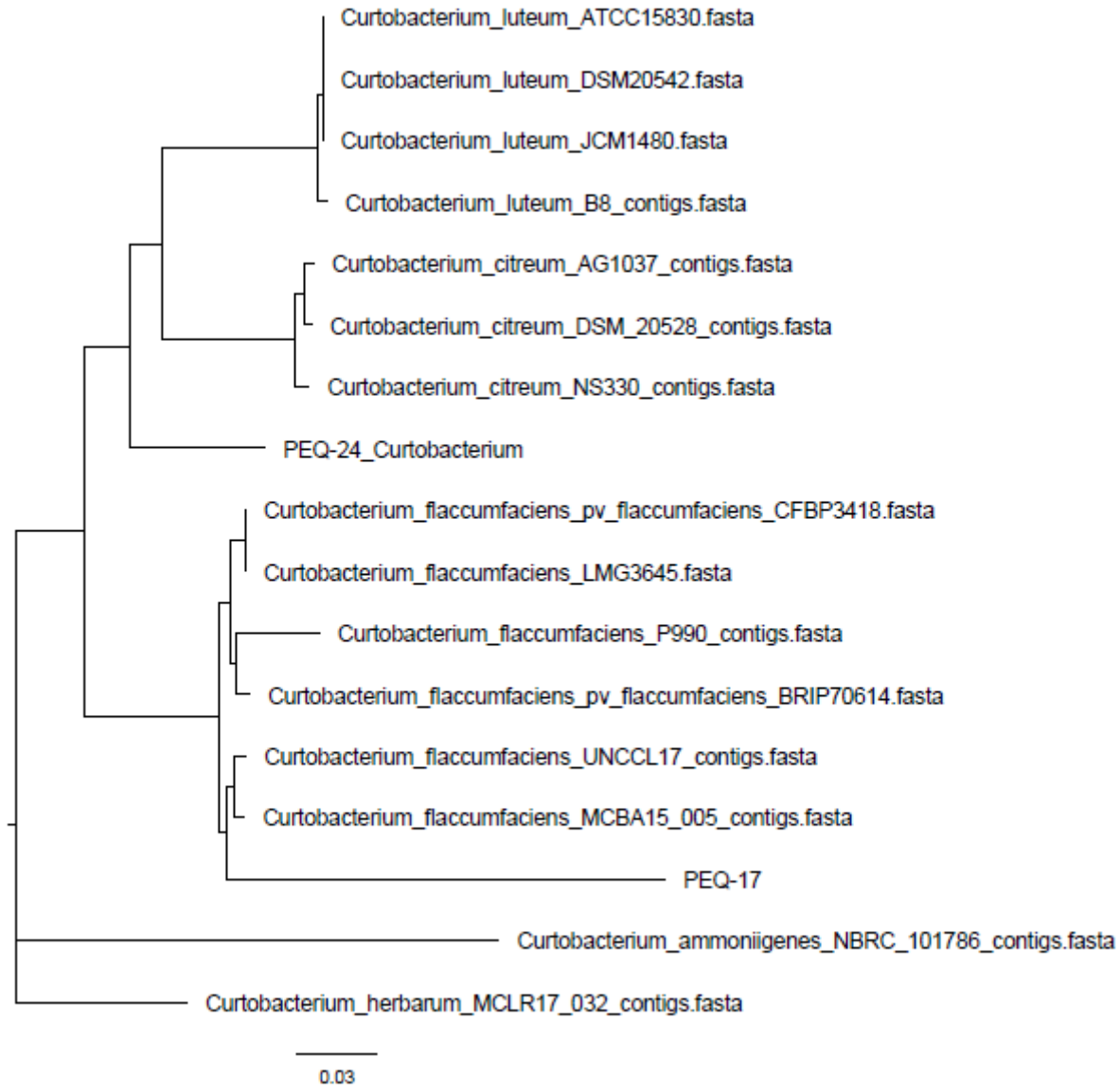


Figure 21.8 Phylogenetic position of PEQ genomes (PEQ-17) and isolated *Curtobacterium* contigs (PEQ-24_Curtobacterium) within the relevant *Curtobacterium* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

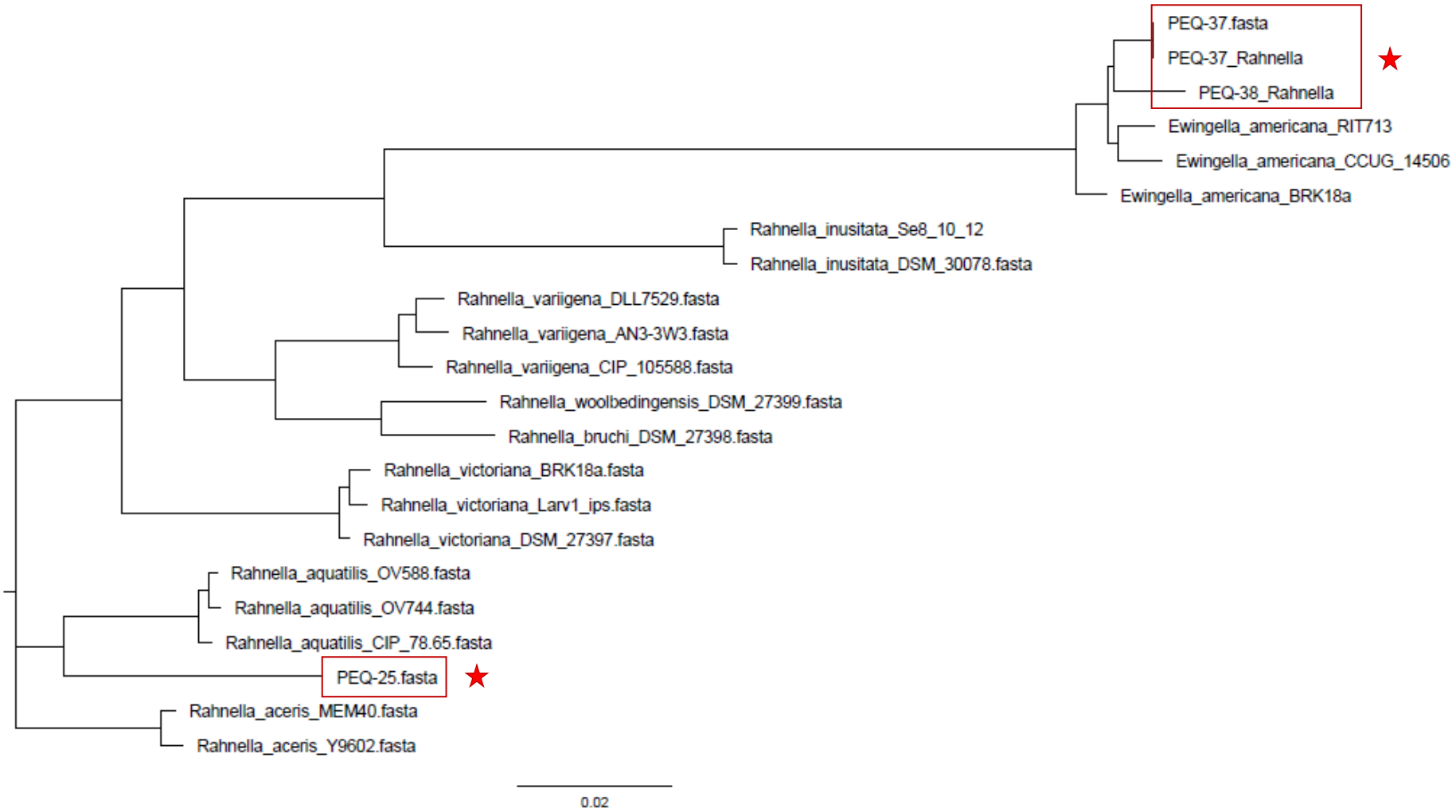


Figure 21.9 Phylogenetic position of PEQ genomes (PEQ-25 and 37) and isolated *Rahnella* contigs (PEQ-37_Rahnella and PEQ-38_Rahnella) within the relevant *Rahnella* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

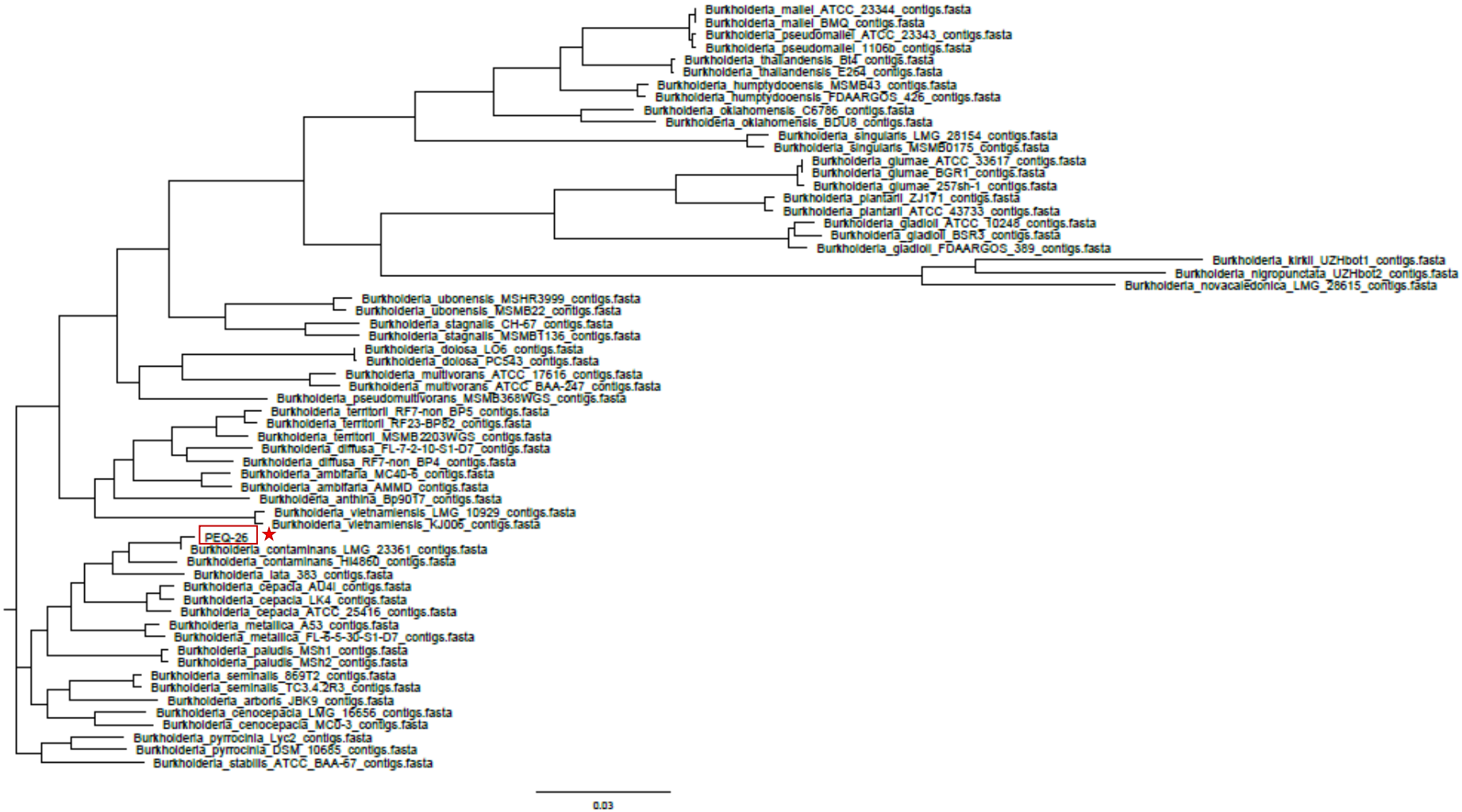


Figure 21.10 Phylogenetic position of PEQ genomes (PEQ-26) within the relevant *Burkholderia* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

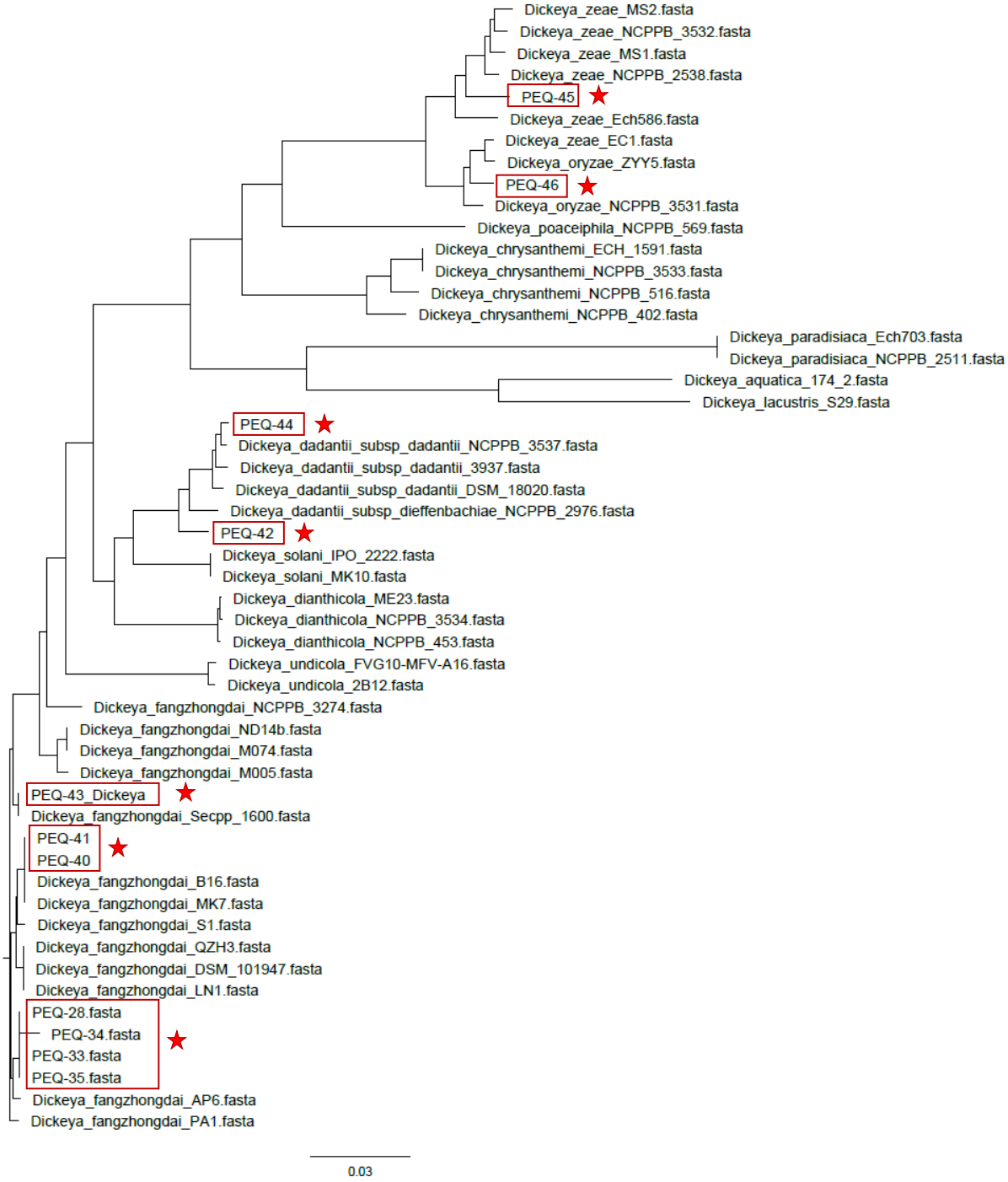


Figure 21.11 Phylogenetic position of PEQ genomes (PEQ-28, 33, 34, 35, 40, 41, 42, 44, 45 and 46) and isolated *Dickeya* contigs (PEQ-43_Dickeya) within the relevant *Dickeya* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

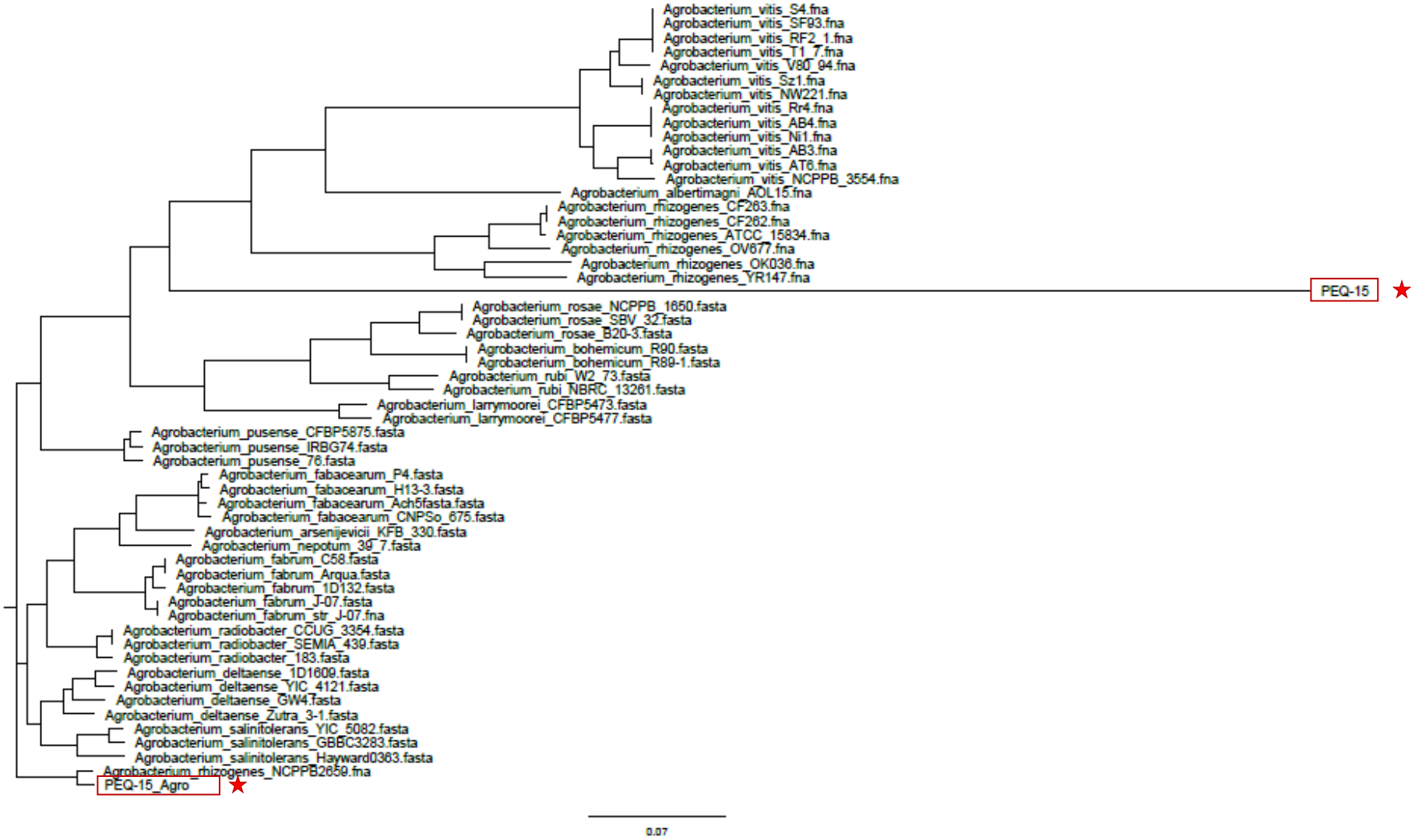


Figure 21.12 Phylogenetic position of PEQ genomes (PEQ-15) and isolated *Agrobacterium* contigs (PEQ-15_Agro) within the relevant *Agrobacterium* strains, based on 92 core gene sequences. Phylogenetic tree was constructed using UBCG (Na et al., 2018). The PEQ isolates are highlighted with a red box and a red star.

Discussion

In this report, genomic DNA of bacterial isolates provided by PEQ were identified through the HTS pipeline from Milestone 104.Sequencing was done using the Illumina NovaSeq and trimmed reads were then subjected to Kraken2 analysis to identify the genus of PEQ bacterial isolates. Contigs were also assembled from all isolates and compared to the non-redundant database at NCBI using the BLASTn algorithm. UBCG java script was run on 35 assemblies out of the 44 to extract 92 core genes and phylogenetic trees were constructed with the relevant genus strains based on Kraken2 results. UBCG was not used for the remaining 9 assembles due to either low sequencing depth and/or genera with no known pathogenic strains such as *Stenotrophomonas* and *Chryseobacterium*.

Input DNA can play a major role in sequencing output. Low concentration samples, PEQ-12, 15 and 30 (0.226-0.456 ng/µL), along with PEQ-16 that had no DNA detected, all generated sequencing data and the data was able to be used to identify the genus and species, based on the available data in NCBI, although the accuracy of identification may be low if the depth and coverage of the core genome is low. Ultimately, when using half-reaction of the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit, a minimum of 1.5 ng input DNA is required in generating adequate sequencing output as seen by PEQ-12 generating 176 MB with 1.6 ng input DNA.

The initial bioinformatic pipeline uses Kraken2 to identify genomes to the genus level and this information is used to pull out the genome database associated with that genus for phylogenetic analysis. Within this report, Kraken2 also identified contamination of other bacterial genus within multiple samples which were confirmed by BLASTn results. These bacterial contaminations, even when at low % reads, interfered with UBCG phylogenetic analysis as seen with PEQ-15 and 22 with the genus *Agrobacterium* and *Bacillus*, respectively. This interference is due to the extra genes of the bacterial contaminants detected by UBCG. To overcome this issue, the BLASTn results can be used to identify contigs associated with the main genus for UBCG phylogenetic analysis, as was done in this study.

Kraken2 analysis also identified a high proportion of unclassified reads for multiple samples which corelates with percentage of no-hit contigs from BLASTn (above 18%). High proportions of unclassified reads might be due to genomes of a genus or species which was not in the used databases of Kraken2 and BLASTn. Using the HTS pipeline, the unclassified reads showed no effect on the identification of the sample as long as no other bacterial genera were detected at high enough quantity such as PEQ-15 and 22. This can be seen in PEQ-37 with 43.34% unclassified reads and 9.65% of *Rahnella* reads, giving similar results between isolated and non-isolated *Rahnella* contigs.

In the previous milestone, it was stated that roughly 300 Mb would be sufficient to generate enough data for phylogenetic analysis. However, due to the possibility of contamination, it would be best that samples be sequenced at a higher sequencing depth such as 0.5 - 1 Gb. This can be observed for PEQ-37 and 38. For these samples, above 500 Mb of data was generated but only 8-10% of those reads generated can be used for identification. After isolating contigs for the genus *Rahnellla*, UBCG java script pulled out 88 and 79 genes out of 92 genes for PEQ-37 and 38, which were sufficient to identify *Ewingella americana* as the species.

When looking at the number of genes generated using the UBCG java script, the amount of data generated for the isolated sample or a high percentage of reads of the genus does not necessarily equate to high numbers of UBCG genes out of the 92. For example, though PEQ-2 had generated 762 Mb of data and Kraken2 indicated 88.29% of reads were the genus S*phingomonas*, the UBCG java script only generated 68 genes out of 92 genes, the lowest number of genes to be extracted in this report. This could be due to truncated genes as N50 was low (507 bp) indicating majority of PEQ-2 contigs were small. Even though it had low number of UBCG genes, UBCG phylogenetic analysis still managed to identify PEQ-2 to the species level, *Sphingomonas melonis*. Hence, identification through UBCG phylogenetic analysis can still be done if the core gene set is not complete. It would be useful to identify the minimum number of genes to accurately identify a species and subspecies using a UBCG analysis, although this is likely to be species dependent.

Kraken2 and BLASTn misidentified the genus for PEQ-37 and 38 as *Rahnella* instead of *Ewingella* that PEQ and UBCG identified. This misclassification reflects the content of the databases used rather than the ability of the actual program. This also shows the importance of phylogenetic analysis to detect new species or genetic variants that may give false positives or may not be detected using some of the database comparison methods. It should be noted that the accuracy and quality of information derived from using a UBCG core gene phylogenetic approach is also dependant on including appropriate references for comparison, further affirming the need for up-to-date databases for each genus.

As seen in this report, using core genes in a phylogenetic analysis, as outlined for UBCG, enabled identification for all 35 PEQ samples that could be analysed at least to genus level. A number of isolates identified only to species level by PEQ and by the BLASTn analysis were identified to subspecies level, however, as previously mentioned in Milestone 104, BLASTn results can be highly erroneous when identifying new or unidentified species or pathovars/subspecies as the database is incomplete, such as that seen in this report with the genus *Ewingella* mentioned previously. Similarly with UBCG, lack of reference genomes would lead to identification only to the genus level. However, reference genomes used within this study included all known pathogenic species, and for PEQ purposes, ruling out pathogenic strains would be sufficient.

In conclusion, minor modifications are necessary in the HTS pipeline. This includes detection of extra bacterial genera that can interfere with UBCG phylogenetic analysis and being attentive to UBCG results that may indicate an incomplete Kraken2 and/or BLASTn database. Furthermore, this report shows that even with low input DNA, sequencing data can still be generated, however should be avoided when possible.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Na, S.-I., Kim, Y. O., Yoon, S.-H., Ha, S.-m., Baek, I., & Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *Journal of Microbiology, 56*(4), 280-285. doi:10.1007/s12275-018-8014-6

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Appendix 22: Testing and optimisation of hybridization probes for enrichment and detection of fastidious bacteria

Introduction

High throughput sequencing (HTS) has been an essential tool for metagenomic studies, the study of genetic material from environmental samples (Ghosh et al., 2019). It has the capability to perform unbiased identification of known and unknown plant pathogens as total DNA samples are sequenced. However, one of the main limitations of metagenomic sequencing for pathogen detection is that the titre of a target pathogen's DNA can be very low compared to that of its host genome. This means that to detect the pathogen, very deep sequencing may be required, which can be expensive, and in some cases the pathogen still may not be detected.

Targeted HTS has the potential to be more sensitive that, metagenomics for pathogen detection, and potentially as sensitive as polymerase chain reaction (PCR) assays whilst providing core genome information for accurate pathogen identification. The use of hybridisation probes prior to HTS enables capture of target nucleic acids from a broader pool, essentially enriching the sample for a specific target and reducing the amount of sequencing required. Hybridisation probes can be used for a range of applications including diversity studies, molecular diagnostics and breed/cultivar improvement (Donaldson et al., 2017; Mascher et al., 2013; Muraya et al., 2015; Pankin et al., 2014; Whiston et al., 2017). In Milestone 105, we designed and synthesized a hybridization probe capture assay for the detection of pathogenic bacteria within the National Priority Plant Pests (NPPP) list (2019), many of which can be present in low titre in their plant hosts and difficult to accurately identify with routine diagnostics (Table 22.1). Multiple genomic targets are included for each bacterial pathogen, to facilitate accurate identification, but the design strategy varies depending on genome sequence availability. They were also chosen based on their use for species, sub-species or sub-group differentiation for taxonomic purposes. Consequently, a total of 5,901 dsDNA hybridization probes were designed for *Xylella fastidiosa* (Xf), *X. taiwanensis* (Xt)*, Xanthomonas citri* subsp. *citri* (Xcc), *Erwinia amylovora* (Ea), '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso), and 16SrI group phytoplasmas genomes (Table 22.1). Probes were designed with 4 X tiling to increase capture rate of low titre pathogens and low complex regions were removed to reduce off-target binding. Each probe consists of a 120 mer dsDNA oligonucleotide with a biotin molecule at one end to capture targeted genes using Streptavidin bound to magnetic beads.

Table 22.1 The genes that were chosen for hybridization capture for detection and identification of *Erwinia amylovora, Xanthomonas citri* subsp. *citri, Xylella fastidiosa, Xylella taiwanensis*, *Candidatus* Liberibacter species and phytoplasmas in the 16SrI group.

Unique regions				MLST genes+ unique regions		MLST genes			
Erwinia amylovora		***Xanthomonas citri* subsp. *citri***		***Xylella fastidiosa* / *Xylella taiwanensis*** [1]		***Candidatus* Liberibacter sp.**[2]		**Phytoplasma 16SrI**	
Gene annotation	Gene length (bp)	Gene annotation	Gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)
eop2	1991	AMDFNCPH_00526	1518	group_806	2257	*rplL*	396	rbfA	342
WP_004154715.1_7	6255	AMDFNCPH_00529	1185	group_1256	5311	*rplJ*	519	nusA	1074
WP_004155245.1_424	3285	AMDFNCPH_01080	2025	*mshA*	2568	*metG*	1536	rpsC-rpsS	1434
amsE	801	AMDFNCPH_01855	1185	group_7658 (subsp. *pauca* only)	2093	*adk*	606	rplF	555
amsD	1056	AMDFNCPH_01857	1380	*copB*	933	*mutS*	2703	secY	1242
Total length	13388	AMDFNCPH_01859	2988	group 1350	1378	*grpE*	636	map	744
		AMDFNCPH_02036	2109	*dacC*	1446	*dnaG*	1935	secA	2508
		AMDFNCPH_02047	2118	group_852	2217	*glnA*	1404	Total length	7899
		AMDFNCPH_02358	10173	*gltT*	1313	Total length	9735		
		AMDFNCPH_02724	1911	*lacF*	881				
		AMDFNCPH_03192	1218	*nuoL*	2142				
		Total length	27810	*holC*	425				
				leuA	1559				
				cysG	1429				
				Total length	25951				

[1] Includes *Xylella taiwanensis*, *X. fastidiosa* subsp. *fastidiosa*, *X. fastidiosa* subsp. *multiplex*, *X. fastidiosa* subsp. *pauca* and *X. fastidiosa* subsp. *sandyi*
[2] Includes *Candidatus* Liberibacter solanacearum, *Ca.* Liberibacter africanus, *Ca.* Liberibacter americanus and *Ca.* Liberibacter asiaticus

Hybridization probe process

Hybridization probes are biotinylated DNA or RNA fragments, either less than 120 bp for Illumina libraries or less than 500 bp for long read sequencing, that are designed to bind to targeted gene or genomic regions (Karamitros et al., 2015; Samorodnitsky et al., 2015). After addition to a sample and once the targets are hybridised, these probes can be captured using streptavidin-coated magnetic beads, effectively capturing target DNA while background DNA is washed away. Alternatively, the process can be subtractive, and the targeted probes can remove unwanted regions of DNA (e.g. host or rRNA).

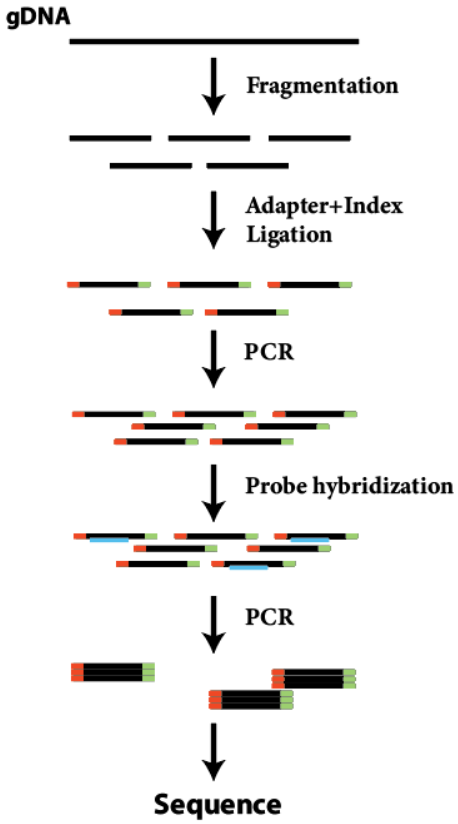


Figure 22.1 Hybridization enrichment flowchart.

In this work the technology was assessed using hybridization probes that were supplied from Twist Bioscience. Twist hybridization probes are 120 bp dsDNA designed to target selected genomic regions of high priority bacterial pathogens (Table 22.1). Figure 1 shows the hybridization enrichment flowchart of DNA that enables detection of targets using HTS. Prior to hybridization with the probes, genomic DNA is sheered to 150-200 bp fragments and subjected to library preparation. Up to eight libraries can be pooled to a maximum of 1.5 µg of DNA libraries, to increase efficiency (time and cost) before starting the hybridization process. During the hybridization process, the target genomic regions are bound by adding the biotinylated hybridization probe panel, the probes are captured using streptavidin-coated magnetic beads and PCR amplification is used to increase concentration of targeted DNA, which is purified for HTS.

In this study, the detection limit of HTS sequencing using hybridization probes was compared to metagenomic (non-probed) samples and current diagnostic approaches including qPCR, LAMP and conventional PCR (cPCR). Additionally, the level of taxonomic identification (e.g. species, subspecies, haplotype and sequencing type (ST)) that can be determined from the HTS data was also assessed.

Material and Methods

DNA extraction

Nucleic acids were extracted from olive, lemon and apple leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Olive, lemon and apple trees are known to be plant hosts of Xf, Xcc and Ea respectively, hence, used as background DNA for positive samples used in validation experiments.

Positive samples for hybridization probe capture validation

Positive samples use for hybridisation probe capture validation are listed in Table 22.2. Genomic DNA for each species was tested using PCR and/or qPCR to confirm identity, except for the phytoplasma DNA, which was validated in another AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria). The concentration of each sample was measured using the Quantus™ fluorometer following manufacturer's instructions.

Table 22.2 Bacterial pathogen isolates used for validation of targeted high throughput sequencing using hybridization capture.

Sample	Species	Host background	Origin
ICMP 15971	*Xylella fastidiosa* subsp. *fastidiosa*	Pure culture	USA
CLbr-Plant	*Candidatus* Liberibacter brunswickensis	*Solanum melongena*	Australia
CLbr-Insect	*Candidatus* Liberibacter brunswickensis	*Accizia solanicola*	Australia
CLso-Plant	*Candidatus* Liberibacter solanacearum	*Solanum tuberosum*	New Zealand
CLso-Insect	*Candidatus* Liberibacter solanacearum	*Bactericera cockerelli*	New Zealand
BRIP 70839 (CLaf)	*Candidatus* Liberibacter africanus	*Citrus sinensis* (L.) Osbeck	South Africa
Xcc 83	*Xanthomonas citri* subsp *citri*	Pure culture	Australia
Ea322A	*Erwinia amylovora*	Pure culture	Attenuated strain generated and supplied by Cornell University (USA). Original Ea322 strain from France
StLL (16SrXXXVII)	*Ca.* phytoplasma stylosanthis	*Solanum tuberosum*	Australia
WF-B (16SrII)	*Ca.* phytoplasma australasiae	*Solanum tuberosum*	Australia

Preparation for sensitivity test and metagenomic comparisons

Olive DNA with various concentrations of spiked Xf ST 2 DNA was prepared in Milestone 105 Appendix III and used in this study to compare the sensitivity of HTS diagnostic techniques for detection of Xf with currently used molecular tests like qPCR and LAMP. Nucleic acids were extracted from olive leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and were shown to be Xf negative using species-specific PCR tests (as outlined in section 2.2). DNA of Xf isolate ICMP15971, at a concentration of 4.75 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of $4.75x10^{-8}$ ng/µL. A volume of 50 µL of each *Xylella* dilution was then spiked into 950 µL of 25 ng/µL olive DNA to reach a final concentration range of 0.2375 ng/µL to $2.375x10^{-8}$.

In order to compare the sensitivity of HTS diagnostic techniques on Ea to currently used molecular tests like PCR, apple DNA was spiked with various concentrations of Ea DNA. Nucleic acids were extracted from apple leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and were shown to be Ea negative using species-specific PCR tests (as outlined in section 2.5). DNA of Ea isolate 322A, at a concentration of 1.94 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of $1.94x10^{-8}$ ng/µL. A volume of 10 µL of each Ea dilution was then spiked into 37.5 µL of 51.8 ng/µL apple DNA to reach a final concentration range of 0.409 ng/µL to $4.09x10^{-8}$ ng/µL.

In order to compare the sensitivity of HTS diagnostic techniques on Xcc to currently used molecular tests like PCR, lemon DNA was spiked with various concentrations of Xcc DNA. Nucleic acids were extracted from lemon leaves using Qiagen DNeasy® 96 Plant Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions and were shown to be Xcc negative using species-specific PCR tests (as outlined in section 2.5). DNA of Xcc isolate 83, at a concentration of 1.4 ng/µL, was serially diluted 1:10 giving a total of eight dilutions and a final concentration of $1.4x10^{-8}$ ng/µL. A volume of 10 µL of each Xcc dilution was then spiked into 21.3 µL of 65.8 ng/µL lemon DNA to reach a final concentration range of 0.448 ng/µL to $4.48x10^{-8}$ ng/µL.

DNA samples extracted from host samples for the *Ca.* Liberibacter and phytoplasma species were sourced from other projects. CLaf was provided by Biosecurity Queensland (Department of Agriculture and Fisheries) which was sourced from CRI facilities, Nelspruit, South Africa. Phytoplasmas were provided by an AVR and Hort Innovation funded project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria). CLso samples were provided by Plant and Food Research, New Zealand. *Ca.* Liberibacter brunswickensis samples were provided by an AVR PhD project. Bacterial spiked plant DNA (8 dilutions of Ea, Xf and Xcc spiked in plant DNA), 3 phytoplasma samples and five *Ca.* Liberibacter samples (CLaf, 2 CLso and 2 *Ca.* Liberibacter brunswickensis) were used for probe and no-probe comparisons in duplicates.

qPCR

Table 22.3 lists the primers and probes used to assess DNA quality and the presence of the target pathogens in DNA extracts.

Successful plant DNA extraction was assessed using qPCR targeting the *cox* gene. Plant DNA was further diluted 1:10 as high concentration of contaminants could inhibit PCR reactions. Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR (Promega), 1 µL cox primer mixture (COX primers and final concentrations listed in Table 22.3), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water) and a positive plant control. The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

The identity of *Ca.* liberibacter DNA samples were confirmed using qPCR. Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL HLB primer mixture (HLB primers and final concentrations listed in Table 22.3), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water). The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 10 seconds and 58°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

DNA extracted from olive leaf tissue was tested for the presence of Xf using qPCR. The sensitivity of qPCR was also tested on the Xf serial dilutions in olive DNA (Section 2.3). Each Xf qPCR reaction was performed in triplicate, in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL qPCR primer mixture (primers and final concentrations listed in Table 22.3), 7 µL molecular biology grade water and 2 µL of template DNA. Each test included a negative template control (water) and a positive control (Xf ICMP15971 DNA). The cycling conditions were 50 °C for 2 minutes, 95 °C for 2 minutes, followed by 40 cycles of 95 °C for 10 seconds and 62 °C for 30 seconds on a QuantStudio 3 (Thermo Fisher). Where multiple replicates were run, the mean Ct (threshold cycle) was calculated and recorded.

Table 22.3 List of primers and final reaction concentrations (Conc.) used in qPCR tests to confirm presence of plant DNA, Candidatus Liberibacter asiaticus' (CLas), 'Ca. Liberibacter americanus' (CLam), 'Ca. Liberibacter africanus' (CLaf) and 'Ca. Liberibacter solanacearum' (CLso) DNA.

Primer	Gene Target	Species	Primer sequence (5'-3')	Conc. (µM)	Reference
COX-F			CGTCGCATTCCAGATTATCCA	0.25	
COX-R	*cox*	Plants	CAACTACGGATATATAAGAGCCAAAACTG	0.25	Weller et al. (2000)
COX-P			TGCTTACGCTGGATGGAATGCCCT	0.1	
HLBaf	16S rRNA	CLaf	TCGAGCGCGTATGCAATACG	0.25	Li et al. (2006)
HLBso	16S rRNA	CLso	GTCGAGCGCTTATTTTTAATAGGA		Li et al. (2009)
HLBr	16S rRNA	CLaf, CLam, CLas, CLso	GCGTTATCCCGTAGAAAAAGGTAG	0.25	Li et al. (2009); Martoni et al. (2021)
HLBp	16S rRNA	CLaf, CLam, CLas, CLso	AGACGGGTGAGTAACGCG	0.1	Li et al. (2009)
XF-F			CACGGCTGGTAACGGAAGA	0.3	
XF-R	*rimM*	Xf	GGGTTGCGTGGTGAAATCAAG	0.3	Harper et al. (2010)
XF-P			TCGCATCCCGTGGCTCAGTCC	0.1	
Ea1-F	*eop2*		GACACCGGCTATCTGCATCA	0.4	
Ea1-R		Ea	AAATCCTTTGATCGGCGGGT	0.4	Mann et al. (in prep)
Ea2-F	*PKS*		AACGCATCGTCCGGGCAAATGT	0.4	
Ea2-R			AGCGCTGTCTGGAAACCCATGC	0.4	
XF-F3			CCGTTGGAAAACAGATGGGA	0.2	
XF-B3			GAGACTGGCAAGCGTTTGA	0.2	
XF-FIP			ACCCCGACGAGTATTACTGGGTTTTTCGCTACC GAGAACCACAC	1.6	
XF-BIP	*rimM*	Xf	GCGCTGCGTGGCACATAGATTTTTGCAACCTTT CCTGGCATCAA	1.6	Harper et al. (2010)
XF-LF			TGCAAGTACACACCCTTGAAG	0.8	
XF-LB			TTCCGTACCACAGATCGCT	0.8	

cPCR

Neat and a 1:10 dilution of DNA extracted from apple leaf tissuewas tested for the presence of Ea using cPCR. The sensitivity of Ea cPCR was also tested on the Ea serial dilutions in apple DNA (Section 2.3). Each Ea PCR reaction was performed in a final volume of 25 µL using primer sets Ea1-F&R and Ea2-F&R (Table 22.3). The reaction mixture contained 1 x PCR buffer (Invitrogen, Netherlands), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM of each forward and reverse primer, 1 U of Platinum Taq polymerase (Invitrogen, Netherlands) and topped up to 25µL using molecular biology grade water. The PCR conditions for both primer sets were as follows: denaturation at 95°C for 2 mins, 35 cycles of 95°C for 30 s, 64°C for 30 s and 72°C for 30 s, and a final hold step at 72°C for 7 mins. Each test included a negative template control (water) and a positive control (Ea322A DNA). PCR products for Ea PCR reactions were visualised by agarose gel electrophoresis (1% agarose gel) with SYBR safe DNA gel stain (Invitrogen, Netherlands).

DNA extracted from lemon leaf tissue, neat and 1:10 dilution, was tested for the presence of Xcc using an unpublished PCR method targeting the R8 and R22 genes (NSW DPI, in prep). Sensitivity of Xcc PCR was also tested on the Xcc serial dilutions in lemon DNA (Section 2.3). Each test included a negative template control (water) and a positive control (Xcc 83 DNA). PCR products for Xcc PCR reactions were visualised by agarose gel electrophoresis (2.5% agarose gel) with SYBR safe DNA gel stain (Invitrogen, Netherlands).

Xf LAMP

LAMP reaction was performed on Xf serial dilutions in olive DNA (Section 2.3) in duplicate, in a final volume of 25 µL. The reaction mixture contained 15 µL Isothermal Master Mix (ISO-004), 2 µL LAMP primer mixture (primers and final concentrations listed in Table 22.3), 6 µL molecular biology grade water and 2 µL of template DNA. Each test included a negative template control (water) and a positive control (Xf ICMP15971 DNA). The reactions were run for 65 °C 30 mins, anneal 98 °C to 80 °C (rate 0.05 °C/sec) on either a Genie III (OptiGene) or a QuantStudio 3 (Thermo Fisher). If target DNA was present, dsDNA was amplified and an increase in fluorescence from the dsDNA intercalating dye was visualised on the Genie III and the time taken recorded (amplification time). Where multiple replicates were run, the mean amplification time was recorded.

Target enrichment: library preparation, hybridization, enrichment and sequencing

DNA concentration of all samples was measured using a Quantus™ fluorometer. Approximately 50 ng of each serial dilution of Xf, Ea and Xcc from Section 2.3 was used for library preparation. Approximately 3 ng of DNA from plant or insect samples naturally infected with phytoplasma or *Ca.* Liberibacter species was used for library preparation with the exception of the *Ca.* Liberibacter solanacearum infected tomato potato psyllid sample CLso-I, where 5 µL containing ~ 0.8 ng of DNA was used as the sample volume was limited. Libraries with ~50 ng starting DNA were generated using half reaction volumes and libraries with ~3 ng starting DNA were generated using full reaction volumes of the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit and PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A total of 128 libraries were prepared, which included four replicate libraries for each sample, a plant only library and a no template control library.

The DNA concentration of each library was quantified using a Quantus™ fluorometer. Two of the four replicates for each sample were used for further hybridization probing steps while the remaining two were used as metagenomic controls. Hybridization probing steps starts with pooling eight libraries into one hybridization capture reaction pool and a total of eight hybridization capture reaction pools were prepared. Pooled libraries were subjected to hybridization capture using the Twist Target Enrichment Protocol (Twist Bioscience, USA) according to manufacturer's instructions.

Libraries for the non-captured CLso-I duplicates gave 0.164 ng/µL and 0.14 ng/µL, lower than all other libraries. The 2200 TapeStation system (Agilent technologies) was used to assess library quality and size distribution of libraries. Libraries for the non-captured CLso-I duplicates were not detected on the TapeStation, however, they were still added to the sequencing run using 15 µL. Hybridization captured and metagenomic libraries were paired end sequenced (2 x 250 bp) using the Illumina NovaSeq on an SP flow cell.

Bioinformatic analysis

All Illumina raw sequence reads were subjected to adapter trimming and quality filtering using fastp (Chen et al., 2018). Trimmed reads were analysed using Kraken 2 (Wood et al., 2019) and Bracken analysis was conducted on Kraken2 results to obtain fraction of reads assigned as the pathogen. Sequences of all genes used for probe design, including an extra 1 kbp sequence of the 5' and 3' end were concatenated into one file and used as a reference for mapping trimmed reads, using the mapping tool, Bowtie2 (Langmead et al., 2012) except for phytoplasma samples. Full phytoplasma genomes of StLL and WF-B were provided by the AVR and Hort Innovation funded project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria) and was used as a reference for mapping trimmed reads of phytoplasma samples. The probed *Candidatus* Liberibacter reads were assembled using SPAdes v3.6.0 (Bankevich et al., 2012) and assembled contigs were used for MLSA phylogenetic analysis using MEGA (version 7.0.26) (Kumar et al., 2016). MLSA functions to differentiate species by generating a phylogenetic tree from several concatenated essential protein-encoding gene sequences (housekeeping genes), in which for this report was the probed-out genes *rplL/rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA*. MLSA genes used are full genes. A k-mer based mapping tool, StringMLST v. 0.6.3 (Gupta et al., 2017), was used to determine if the sequence type (ST) could be assigned for each Xf dilution at the varying sequence depths.

Results

Sequencing output

Reads for each data set were quality trimmed and the average data output from replicates is shown in Table 22.4. Sequencing output for probed samples were highly variable, compared to the metagenomic samples, as concentration of pathogenic DNA may affect sequencing output (Table 22.4).

Table 22.4 The volume of sequence data (megabyte; Mb) and the total number of reads generated by metagenomic sequencing of uninfected plant controls and the pathogen spiked or infected samples and the probed pathogen spiked or infected samples. All pathogen spiked or infected samples were sequenced in duplicate and the average volume and number of reads is presented. XF = *Xylella fastidiosa* dilutions, EA = *Erwinia amylovora* dilutions, XCC = *Xanthomonas citri* subsp *citri* dilutions.

Uninfected plant host control			Input DNA concentration (ng/µl)
Sample	Size (Mb)	Reads	
Apple	565	4806268	51.8
Lemon	423	3711988	65.8
Olive	2440	13493506	20

Metagenomic pathogen spiked or infected samples			Probed pathogen spiked or infected samples			Input DNA concentration (ng/µl)
Sample	Size (Mb)	Reads	Sample	Size (Mb)	Reads	
XF-2-0	791	9118330	P-XF-2-0	6,281	60447357	19
XF-2-1	636	8169016	P-XF-2-1	2,194	20865534	19
XF-2-2	715	7576157	P-XF-2-2	998	9298002	19
XF-2-3	764	8294788	P-XF-2-3	736	7181871	19
XF-2-4	682	7482041	P-XF-2-4	690	6276356	19
XF-2-5	746	8010559	P-XF-2-5	693	6664122	19
XF-2-6	660	7552909	P-XF-2-6	575	5377575	19
XF-2-7	710	8054240	P-XF-2-7	457	4458065	19
EA-2-0[2]	797	8771775	P-EA-2-0	1,683	14888763	49.2
EA-2-1	929	9898833	P-EA-2-1	1,031	8947778	49.2
EA-2-2	906	9304453	P-EA-2-2	163	1458692	49.2

EA-2-3	762	7942867	P-EA-2-3	168	1468257	49.2
EA-2-4	749	7420874	P-EA-2-4	189	1705558	49.2
EA-2-5	894	9001673	P-EA-2-5	225	1990704	49.2
EA-2-6	840	8594938	P-EA-2-6	834	7552006	49.2
EA-2-7	810	8769276	P-EA-2-7	912	8203255	49.2
XCC-2-0[3]	847	9501042	P-XCC-2-0	923	8408121	62.5
XCC-2-1	956	9956700	P-XCC-2-1	122	1056719	62.5
XCC-2-2	636	6861413	P-XCC-2-2	571	4908793	62.5
XCC-2-3	863	9098928	P-XCC-2-3	813	7256933	62.5
XCC-2-4	777	8385552	P-XCC-2-4	674	5977758	62.5
XCC-2-5	870	8587869	P-XCC-2-5	515	4608088	62.5
XCC-2-6	947	9905996	P-XCC-2-6	208	1801999	62.5
XCC-2-7	785	8226234	P-XCC-2-7	180	1578096	62.5
CLso-I	900	8157829	P- CLso-I	204	1824880	2.64
Clso-P	947	8545992	P- Clso-P	3,263	30269491	1.42
Clbr-I	0.11	1040	P- Clbr-I	43	408381	0.163
Clbr-P	1,039	9217579	P- Clbr-P	49	438745	1.11
Laf	961	8671376	P-Laf	281	2546838	0.49
StLL	750	6360077	P-StLL	383	3241615	5.08
WF-B	752	6503547	P-WF-B	1,115	9471393	7.06

Genomic DNA validation

The *cox* qPCR assay was used to confirm DNA extraction on olive, apple and lemon DNA detected the plant *cox* gene of olive DNA and both neat and diluted apple and lemon DNA (1:10). PCR and qPCR results also showed that apple, olive and lemon DNA were negative for the exotic pathogens Ea, Xf and Xcc, respectively to confirm that the plant samples were clean. Species-specific PCR or qPCR confirmed the presence of Ea, Xf and Xcc in the genomic DNA of Ea322A, Xf ICMP 15971 and Xcc 83 (data not shown).

The identity of *Ca* Liberibacter species in naturally infected plant and insect DNA were confirmed using species-specific qPCR. Positive qPCR results are indicated by cyle threshold (Ct) below 40 (data not shown). The presence and specific species of phytoplasma in DNA of infected hosts was validated in another AVR project (CMI 106078: VG16086 Area wide management of vegetable diseases: viruses and bacteria).

Gene presence

Bowtie2 was used for mapping reads to all genes used for probe design to identify gene presence and confirm full or partial coverage. Results showed that all five Ea genes were partially mapped for Ea metagenomic samples, however, probed samples only captured WP_004154715.1_7 with full gene coverage (Table 22.5). Similarly, all seven phytoplasma genes were fully mapped to for StLL and WF-B metagenomic samples, however, no phytoplasma genes were detected for probed StLL and WF-B samples. For Xcc, two Xcc genes were partially mapped for metagenomic samples while six out of eleven genes were mapped for probed samples where one was partially mapped to (Table 22.5).

All genes for Xf, CLaf, and CLso metagenomic samples were detected with a few partially mapped genes except CLso which had full coverage for all genes. Similarly, all Xf, CLaf, and CLso probed samples were detected with all genes having full coverage. None of the CLbr genes were detected for metagenomic CLbr samples, however, probed CLbr samples had all genes had partial mapping.

Table 22.5 Detection of hybridization capture genes (full and partial) of metagenomic and probed samples through mapping the highest concentration pathogen spiked samples; *Erwinia amylovora* (EA-2-0), *Xanthomonas citri* subsp. *citri* (XCC-2-0) and *Xylella fastidiosa* (XF-2-0); and infected samples; *Candidatus* Liberibacter solanacearum (CLso), *Ca* Liberibacter africanus (CLaf), *Ca* Liberibacter brunswickensis (CLbr) and Phytoplasma samples (StLL and WF-B).

Metagenomic					
Species	**Total genes**	**Full genes**	**Total**	**Partial genes**	**Total**
EA-2-0	5/5		0	eop2, WP_004154715.1_7, WP_004155245.1_424, amsE, amsD	5/5
XCC-2-0	2/11		0	AMDFNCPH_02358, AMDFNCPH_00526	2/11
XF-2-0	13/13	group_1256	1/13	group_806, mshA, group_7658, copB, group_1350, dacC, group_852, gltT, lacF, nuoL, holC, leuA, cysG	12/13
CLso (plant)	7/7	rplL/rplJ, metG, adk, mutS, grpE, dnaG, glnA	7/7		
CLaf	7/7	metG, rplL/rplJ	2/7	glnA, dnaG, adk, grpE, mutS	5/7
CLbr (plant)	Nil				
StLL/WF-B	7/7	rbfA, nusA, rpsC-rpsS, rplF, secY, map, secA	7/7		
Probed					
Species	**Total genes**	**Full genes**	**Total**	**Partial genes**	**Total**
EA-2-0	1/5	WP_004154715.1_7	1/5		
XCC-2-0	6/11	AMDFNCPH_00526, AMDFNCPH_00529, AMDFNCPH_02036, AMDFNCPH_02047, AMDFNCPH_02358	5/11	AMDFNCPH_01859	1/11
XF-2-0	13/13	group_806, group_1256, mshA, group_7658, copB, group_1350, dacC, group_852, gltT, lacF, nuoL, holC, leuA, cysG	13/13		
CLso (plant)	7/7	rplL/rplJ, metG, adk, mutS, grpE, dnaG, glnA	7/7		
CLaf	7/7	rplL/rplJ, metG, adk, mutS, grpE, dnaG, glnA	7/7		
CLbr (plant)	7/7		0	rplL/rplJ, metG, adk, mutS, grpE, dnaG, glnA	7/7
StLL/WF-B	Nil				

Xylella fastidiosa

Kraken2 and Bracken software were used to classify individual reads to an operational taxonomic unit for probed and metagenomic serial diluted Xf samples using full genome databases. As expected, total Xf reads and percentage of Xf reads decreased as Xf concentration decreased for both probed and metagenomic samples (Table 22.6). This was also observed for reads mapped to probed Xf regions. Mapping of Xf probed regions showed that all Xf genes were mapped to.

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples up to Xf concentration of 2.4 x 10^{-06} ng/µL. This was also observed for reads mapped to probed Xf regions. Xf concentration of 2.4 x 10^{-03} ng/µL gave the highest fold increase; 5,736.3 X more when using Bracken with full genome databases and 122,206.3 X more when mapped to probed regions. Xf concentration of 2.4 x 10^{-06} ng/µL gave the lowest fold increase; 93.2X more when using Bracken and 1,963.5X more when mapped to probed regions (Table 22.6).

Regarding sensitivity, the LAMP and qPCR assays detected Xf at femtogram levels, 2.4 x 10^{-05} and 2.4 x 10^{-06} ng/µL, respectively (Table 22.7). Using Bracken, Xf was detected confidently up to 2.4 x 10^{-02} ng/µL for metagenomic Xf samples, 1,000 X and 10,000 X less sensitive than LAMP and qPCR respectively. Xf was detected confidently up to 2.4 x 10^{-06} ng/µL for probed Xf samples, 10 X more sensitive than LAMP and equally as sensitive as qPCR.

Using StringMLST on trimmed reads, the sequence type (ST) was able to be identified in the first dilution, 0.24 ng/µL, of metagenomic Xf samples. StringMLST was also able to identify some MLST genes at Xf concentration 2.4 x 10^{-02} ng/µL, however, it was insufficient to allocate a ST. StringMLST was unable to allocate a ST for all probed Xf samples due to the absence of hybridization probes targeting the gene *petC*. Nevertheless, StringMLST was still able to detect the rest of the six MLST genes up to Xf concentration 2.4 x 10^{-05} ng/µL. StringMLST was also able to identify some MLST genes at Xf concentration 2.4 x 10^{-06} ng/µL, however, not all six genes were detected.

Table 22.6 Bracken and mapping result average for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to assign Xf reads based on full genome databases and Bowtie2 was used for mapping Xf reads to probed Xf regions. Fold enrichment = fold increase of Xf percentage of total reads from metagenomic to probed samples.

			Metagenomic			Probed			
	Sample	Concentration of *Xylella* DNA (ng/µL)	Xf reads	Total paired reads	Xf percentage of total reads (%)	Xf reads	Total paired reads	Xf percentage of total reads (%)	Fold enrichment
K-mer (Bracken)-full genome	XF-2-0	0.24	18030.5	4559165	0.39548	29163112	30223678.5	96.49094	244
	XF-2-1	2.4×10^{-02}	982.5	4084508	0.02405	7550430.5	10432767	72.37227	3,008.7
	XF-2-2	2.4×10^{-03}	134	3788078.5	0.00354	943358.5	4649001	20.29164	5,736.3
	XF-2-3	2.4×10^{-04}	45.5	4147394	0.00110	99788.5	3590935.5	2.77890	2,533
	XF-2-4	2.4×10^{-05}	21.5	3741020.5	0.00057	10622	3138178	0.33848	589
	XF-2-5	2.4×10^{-06}	22.5	4005279.5	0.00056	1745	3332061	0.05237	93.2
	XF-2-6	2.4×10^{-07}	18	3776454.5	0.00048	35.5	2688787.5	0.00132	2.8
	XF-2-7	2.4×10^{-08}	22	4027120	0.00055	51	2229032.5	0.00229	4.2
Mapping (Bowtie2)-probed region	Sample	Concentration of *Xylella* DNA (ng/µL)	On-target reads	Total reads	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
	XF-2-0	0.24	1430.5	9118330	0.01569	53180973	60447357	87.98	5,608
	XF-2-1	2.4×10^{-02}	64	8169016	0.00078	13792088	20865534	66.1	84,370.4
	XF-2-2	2.4×10^{-03}	11.5	7576157	0.00015	1724774	9298002	18.55	122,206.3
	XF-2-3	2.4×10^{-04}	9	8294788	0.00011	183880	7181871	2.56	23,597.2
	XF-2-4	2.4×10^{-05}	0.5	7482041	6.68E-06	19208.5	6276356	0.306	45,796.9
	XF-2-5	2.4×10^{-06}	2	8010559	2.5E-05	3267	6664122	0.049	1,963.5
	XF-2-6	2.4×10^{-07}	0	7552909	0	0.5	5377575	9.3E-06	*
	XF-2-7	2.4×10^{-08}	1.5	8054240	1.86E-05	0.5	4458065	1.12E-05	0.6

* Can't be calculated as fold enrichment is not divisible by zero

Table 22.7 LAMP and qPCR results from Milestone 105, and sequencing results for serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA. Bracken was used to assign Xf reads and calculate fraction of reads. ST2 = sequence type of *Xylella* isolate used in this experiment, UD = Undetermined, Ct = threshold cycle of amplification in qPCR.

Sample	Concentration of *Xylella* DNA (ng/µL)	qPCR (Ct)	LAMP (amp time)	StringMLST		Xf reads		Xf percentage of total reads (%)	
				Metagenomic	Probed[2]	Metagenomic	Probed	Metagenomic	Probed
XF-2-0	0.24	21.09	9.9	ST2	XF	18030.5	29163112	0.39548	96.49
XF-2-1	2.4×10^{-02}	24.52	11.2	XF[1]	XF	982.5	7550430.5	0.02405	72.37
XF-2-2	2.4×10^{-03}	27.87	12.9	-	XF	134	943358.5	0.00354	20.29
XF-2-3	2.4×10^{-04}	31.17	14.7	-	XF	45.5	99788.5	0.00110	2.78
XF-2-4	2.4×10^{-05}	34.21	20.4	-	XF	21.5	10622	0.00057	0.338
XF-2-5	2.4×10^{-06}	37.45	UD	-	XF[1]	22.5	1745	0.00056	0.0524
XF-2-6	2.4×10^{-07}	UD	UD	-	-	18	35.5	0.00048	0.00132
XF-2-7	2.4×10^{-08}	UD	UD	-	-	22	51	0.00055	0.00229
Negative	0	UD	UD	-	-	35		0.00011	

[1] Incomplete gene detection/typing

[2] *petC* was not included in the probe set

Xanthomonas citri subsp. citri

Kraken2 and Bracken software were used to classify individual reads to an operational taxonomic unit for probed and metagenomic serial diluted Xcc samples using full genome databases. As expected, total Xcc reads and percentage of Xcc reads decreased as Xcc concentration decreased for both probed and metagenomic samples (Table 22.8). This was also observed for reads mapped to probed Xcc regions. Mapping of Xcc probed regions showed that six genes out of the 11 Xcc genes were mapped to; AMDFNCPH_00526, AMDFNCPH_00529, AMDFNCPH_01859, AMDFNCPH_02036, AMDFNCPH_02047 and AMDFNCPH_02358 (Table 22.5).

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples up to Xcc concentration of 4.5 x 10^{-03} ng/µL. For reads mapped to probed Xcc regions, Bracken percentage of total reads were higher for probed samples up to Xcc concentration of 4.5 x 10^{-04} ng/µL (Table 22.8).

Regarding sensitivity, the cPCR assay detected Xcc at picogram levels, 4.5 x 10^{-03} (Figure 2). Xcc was not able to be detected confidently for metagenomic Xcc samples using both Bracken and mapping analysis. Xcc was detected confidently up to 4.5 x 10^{-03} ng/µL for probed Xcc samples using Bracken and Bowtie2, equally as sensitive as cPCR.

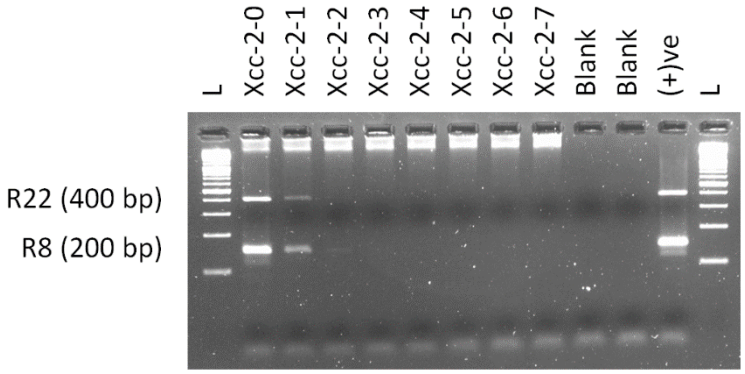


Figure 22.2 PCR analysis of serially diluted Xcc samples (1:10) in lemon DNA. Lane L: GeneRuler 1 kb Plus (Thermo Scientific™), (+)ve: Positive control (Ea322A).

Table 22.8 Bracken and mapping result average for serially diluted *Xanthomonas citri* subsp. *citri* (Xcc) DNA into olive DNA. Bracken was used to assign Xcc reads based on full genome databases and Bowtie2 was used for mapping Xcc reads to probed Xcc regions. Fold enrichment = fold increase of Xcc percentage of total reads from metagenomic to probed samples, + = right sized band present on gel.

				Metagenomic			Probed			
	Sample	Concentration of Xcc DNA (ng/µL)	PCR results	Xcc reads	Total paired reads	Xcc percentage of total reads (%)	Xcc reads	Total paired reads	Xcc percentage of total reads (%)	Fold enrichment
K-mer (Bracken)-full genome	XCC-2-0	0.45	+	91	4750521	0.00192	50968.5	4204060.5	1.21236	632.9
	XCC-2-1	4.5×10^{-02}	+	16	4978350	0.00032	539.5	528359.5	0.10211	317.7
	XCC-2-2	4.5×10^{-03}	+ [1]	8.5	3430706.5	0.00025	335	2454396.5	0.01365	55.1
	XCC-2-3	4.5×10^{-04}	-	11	4549464	0.00024	12	3628466.5	0.00033	1.4
	XCC-2-4	4.5×10^{-05}	-	5	4192776	0.00012	11.5	2988879	0.00038	3.2
	XCC-2-5	4.5×10^{-06}	-	11	4293934.5	0.00026	12.5	2304044	0.00054	2.1
	XCC-2-6	4.5×10^{-07}	-	11	4952998	0.00022	5	900999.5	0.00055	2.5
	XCC-2-7	4.5×10^{-08}	-	12.5	4113117	0.00030	3	789048	0.00038	1.3
	Sample	Concentration of Xcc DNA (ng/µL)	PCR results	On-target reads	Total reads	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
Mapping (Bowtie2)-probed region	XCC-2-0	0.45	+	4	9501042	4.21E-05	1006783	8408121	11.974	284,412.2
	XCC-2-1	4.5×10^{-02}	+	1	9956700	1.004E-05	10661.5	1056719	1.0089	100,455.6
	XCC-2-2	4.5×10^{-03}	+ [1]	0	6861413	0	4867	4908793	0.0992	*
	XCC-2-3	4.5×10^{-04}	-	0	9098928	0	426 [2]	7256933	0.00587	*
	XCC-2-4	4.5×10^{-05}	-	0	8385552	0	6.5	5977758	0.00011	*
	XCC-2-5	4.5×10^{-06}	-	0	8587869	0	0	4608088	0	0.0
	XCC-2-6	4.5×10^{-07}	-	0	9905996	0	0	1801999	0	0.0
	XCC-2-7	4.5×10^{-08}	-	0	8226234	0	0	1578096	0	0.0

[1] Faint band on gel

[2] Only detected on one replicate

* Can't be calculated as fold enrichment is not divisible by zero

Erwinia amylovora

Kraken2 and Bracken software were used to classify individual reads to an operational taxonomic unit for probed and metagenomic serial diluted Ea samples using full genome databases. As expected, total Ea reads and percentage of Ea reads decreased as Ea concentration decreased for both probed and metagenomic samples (Table 22.9). This was also observed for reads mapped to probed Ea regions. Mapping of Ea probed regions showed that one gene out of the five Ea genes were mapped to; WP_004154715.1_7 (Table 22.5).

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples up to Ea concentration of 4.1 x 10^{-04} ng/µL. This was also observed for reads mapped to probed Ea regions (Table 22.9). Ea concentration of 4.1 x 10^{-02} ng/µL gave the highest fold increase; 845.9 X more when using Bracken with full genome databases and 72,538.9 X more when mapped to probed regions. Ea concentration of 4.1 x 10^{-04} ng/µL gave the lowest fold increase, 183.9 X more when using Bracken (Table 22.9).

Regarding sensitivity, the cPCR assay detected Ea at picogram levels, 4.1 x 10^{-05} (Figure 22.3). Ea was detected confidently up to 4.1 x 10^{-03} ng/µL for metagenomic Ea samples using Bracken and Bowtie2, 100 X less sensitive than cPCR. Ea was detected confidently up to 4.5 x 10^{-04} ng/µL for probed Xf samples using Bracken and Bowtie2, 10 X less sensitive than cPCR.

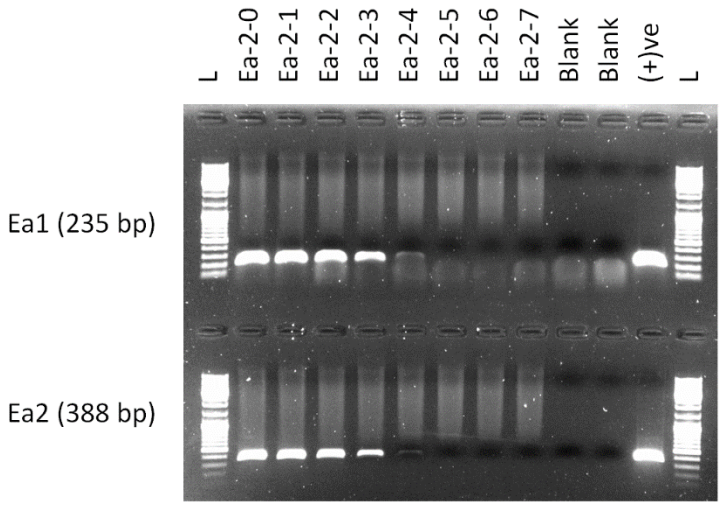


Figure 22.3 PCR analysis of serially diluted Ea samples (1:10) in apple DNA. Lane L: GeneRuler 1 kb Plus (Thermo Scientific™). (+)ve: Positive control (Ea322A).

Table 22.9 Bracken and mapping result average for serially diluted *Erwinia amylovora* (Ea) DNA into olive DNA. Bracken was used to assign Ea reads based on full genome databases and Bowtie2 was used for mapping Ea reads to probed Ea regions. Fold enrichment = fold increase of Ea percentage of total reads from metagenomic to probed samples, + = right sized band present on gel.

				Metagenomic			Probed			
	Sample	Concentration of Ea DNA (ng/µL)	PCR results	Ea reads	Total paired reads	Ea percentage of total reads (%)	Ea reads	Total paired reads	Ea percentage of total reads (%)	Fold enrichment
K-mer (Bracken)-full genome	EA-2-0	0.41	+	3400	4385887.5	0.07752	2912740.5	7444381.5	39.12670	504.7
	EA-2-1	4.1×10^{-02}	+	394	4949416.5	0.00796	301248.5	4473889	6.73348	845.9
	EA-2-2	4.1×10^{-03}	+	47.5	4652226.5	0.00102	6003	729346	0.82307	806.1
	EA-2-3	4.1×10^{-04}	+	16.5	3971433.5	0.00042	561	734128.5	0.07642	183.9
	EA-2-4	4.1×10^{-05}	+ [1]	15	3710437	0.00040	73	852779	0.00856	21.2
	EA-2-5	4.1×10^{-06}	-	15.5	4500836.5	0.00034	5	995352	0.00050	1.5
	EA-2-6	4.1×10^{-07}	-	18	4297469	0.00042	5.5	3776003	0.00015	0.3
	EA-2-7	4.1×10^{-08}	-	8.5	4384638	0.00019	6	4101627.5	0.00015	0.8
	Sample	Concentration of Ea DNA (ng/µL)	PCR results	On-target reads	Total reads	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
Mapping (Bowtie2)-probed region	EA-2-0	0.41	+	31	8771775	0.0003534	2536308.5	14888763	17.03505187	48,202.5
	EA-2-1	4.1×10^{-02}	+	4	9898833	4.041E-05	262278	8947778	2.931208173	72,538.9
	EA-2-2	4.1×10^{-03}	+	1	9304453	1.075E-05	5963	1458692	0.408790889	38,035.8
	EA-2-3	4.1×10^{-04}	+	0	7942867	0	501	1468257	0.034122092	*
	EA-2-4	4.1×10^{-05}	+ [1]	0	7420874	0	60	1705558	0.00351791	*
	EA-2-5	4.1×10^{-06}	-	2	9001673	2.222E-05	2	1990704	0.000100467	4.5
	EA-2-6	4.1×10^{-07}	-	0.5	8594938	5.817E-06	0	7552006	0	0.0
	EA-2-7	4.1×10^{-08}	-	0	8769276	0	0	8203255	0	*

[1] Faint band on gel

* Can't be calculated as fold enrichment is not divisible by zero

***Candidatus* Liberibacters**

Mapping of *Candidatus* Liberibacter probed regions showed that reads were mapped to all *Candidatus* Liberibacter genes for both CLaf and CLso samples (Figure 22.3). As CLbr genes were not added to the probing list, CLbr genes did not show full gene coverage for probed CLbr samples. Probed CLbr samples also showed reduced mapped reads for the *adk* gene and minor increase for the *glnA* gene when compared to metagenomic samples (Figure 22.4).

Bracken percentage of total reads were higher for probed samples in comparison to metagenomic samples for all *Candidatus* Liberibacter samples. This was also observed for reads mapped to probed *Candidatus* Liberibacter regions (Table 22.10). CLaf gave the highest fold increase; 470.6 X more when using Bracken with full genome databases and 22,481.36 X more when mapped to probed *Candidatus* Liberibacter regions.

Mapped reads of probed and metagenomic *Candidatus* Liberibacter samples were extracted from BAM files and was *denovo* assembled using SPAdes. Samples that produced full *Candidatus* Liberibacter gene datasets were: probed CLaf (P-Laf), probed CLso-Insect (P-CLso-I), probed CLso-Plant (P-CLso) and metagenomic CLso-Plant (CLso). Choosing one replicate, P-Laf, P-CLso and CLso were, all eight genes (*rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA*) were extracted from assembled reads, concatenated, and aligned against multiple *Candidatus* Liberibacter species. Phylogenetic tree was then constructed and shown in Figure 22.5. From the phylogenetic tree, P-CLso and CLso were classified as *Candidatus* Liberibacter solanacearum haplotype A and P-Laf as *Candidatus* Liberibacter africanus closely related to the CLaf isolate PTSAPSY.

Phytoplasma

Mapping of reads from phytoplasma samples (StLL and WF-B) to their respective genomes showed that there was no increase of reads mapped to any of the phytoplasma genomes in the metagenomic and probed samples. This observation was backed by bracken percentage of total reads of reads from probed phytoplasma samples being lesser than metagenomic phytoplasma samples (Table 22.11).

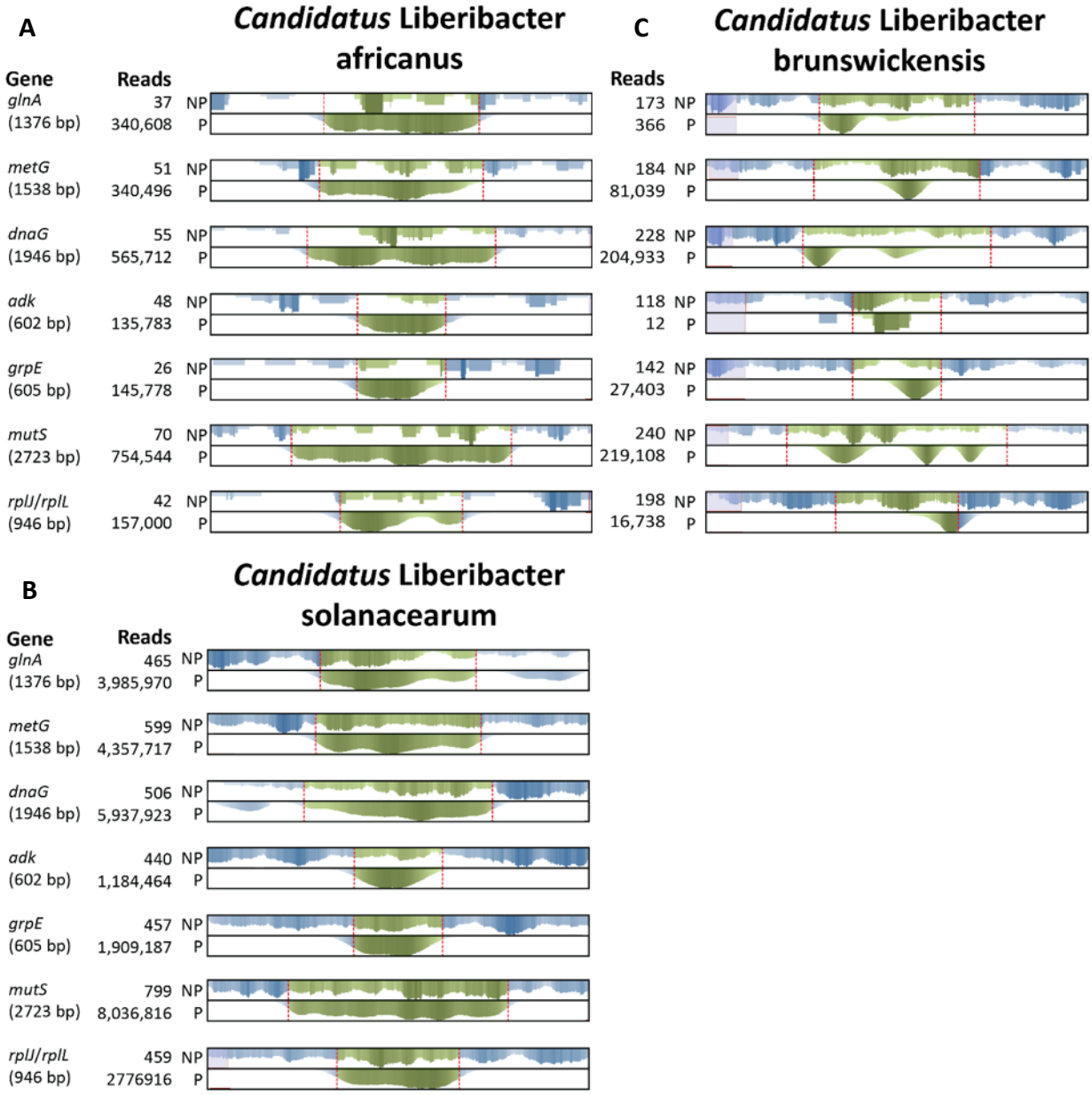


Figure 22.4 Mapping of metagenomic (NP) and probed (P) reads to *Candicatus* Liberibacter genes for (A) *Ca.* Liberibacter aricanus, (B) *Ca.* Liberibacter solanacearum (plant host) and (C) *Ca.* Liberibacter brunswickensis (insect host). Green regions within the red dotted lines indicate gene location while blue regions are 1 kb sequences 5' and 3' of the gene.

Table 22.10 Bracken and mapping result average for *Candidatus* Liberibacter brunswickensis (CLbr) in host insect and plant DNA, *Ca*. Liberibacter solanacearum (CLso) in host insect and plant DNA, and *Ca*. Liberibacter africanus (CLaf) in host plant DNA. Bracken was used to assign *Ca*. Liberibacter reads based on full genome databases and Bowtie2 was used for mapping *Ca*. Liberibacter reads to probed *Ca*. Liberibacter regions. Fold enrichment = fold increase of *Ca*. Liberibacter percentage of total reads from metagenomic to probed samples.

		Metagenomic			Probed			
K-mer (Bracken)-full genome	Sample	*Ca.* Liberibacter reads	Total paired reads	*Ca.* Liberibacter percentage of total reads (%)	*Ca.* Liberibacter reads	Total paired reads	*Ca.* Liberibacter percentage of total reads (%)	Fold enrichment
	Clbr-Insect	13106.5	8157829	0.1607	216498	1824880	11.8637	73.84
	Clbr-Plant	168	9217579	0.0018	517	438745	0.1178	64.65
	Clso-Insect	29	1040	2.7885	144866.5	408381	35.4734	12.72
	Clso-Plant	100018.5	8545992	1.1704	14923358	30269491	49.3016	42.13
	CLaf	8031	8671376	0.0926	1110086	2546838	43.5868	470.62
Mapping (Bowtie2)-probed region	Sample	On-target reads	Total yield (reads)	On-target percentage (%)	On-target reads	Total reads	On-target percentage (%)	Fold enrichment
	Clbr-Insect	1377	8157829	0.0169	538619	1824880	29.5153	1,748.59
	Clbr-Plant	2	9217579	0.00002	1232	438745	0.2808	12,941.52
	Clso-Insect	0	1040	0	288231	408381	70.5789	*
	Clso-Plant	3638	8545992	0.0426	29505607	30269491	97.4764	2,289.81
	CLaf	334.5	8671376	0.0039	2208676	2546838	86.7223	22,481.36

* Can't be calculated as fold enrichment is not divisible by zero

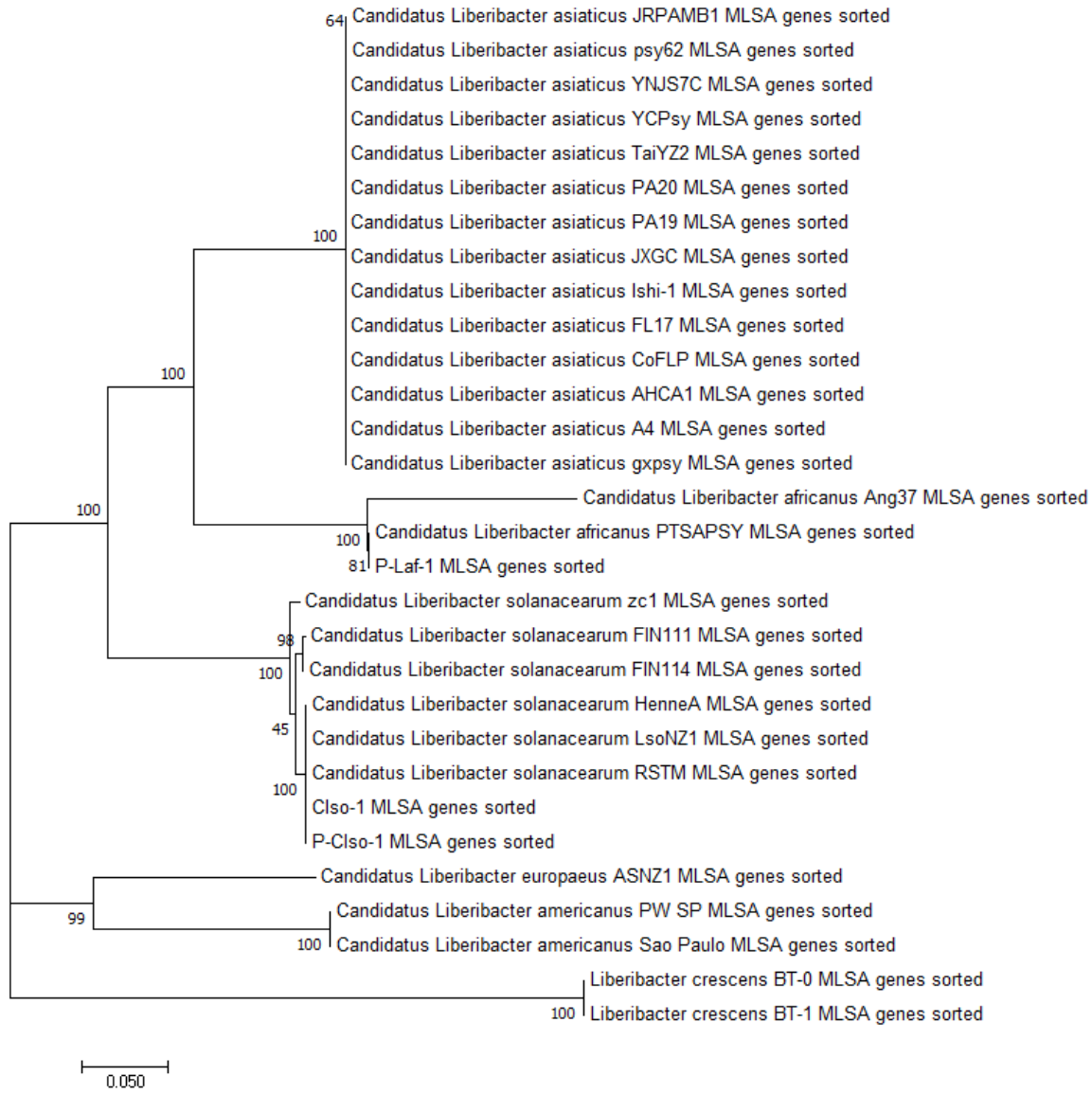


Figure 22.5 Maximum likelihood tree based on *rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA* gene sequences of probed CLaf (P-Laf), probed CLso-Plant (P-CLso) and metagenomic CLso-Plant (CLso) and their closest phylogenetic neighbours.

Table 22.11 Bracken and mapping result average for Phytoplasma (StLL and WF-B) in host plant DNA. Bracken was used to assign Phytoplasma reads based on full genome databases and Bowtie2 was used for mapping Phytoplasma reads to their respective genomes. Fold enrichment = fold increase of Phytoplasma percentage of total reads from metagenomic to probed samples.

		Metagenomic			Probed			
K-mer (Bracken) – full genome	Sample	Phytoplasma reads	Total paired reads	Phytoplasma percentage of total reads (%)	Phytoplasma reads	Total paired reads	Phytoplasma percentage of total reads (%)	Fold enrichment
	StLL	966.5	6360077	0.0152	242.5	3241615	0.0075	0.49
	WF-B	5008.5	6503547	0.0770	5374	9471393	0.0567	0.74
Mapping (Bowtie2) – full genome	Sample	On-target reads	Total yield (reads)	On-target percentage (%)	On-target reads	Total yield (reads)	On-target percentage (%)	Fold enrichment
	StLL	10429.5	6360077	0.1640	2432	3241615	0.0750	0.46
	WF-B	149950	6503547	2.3057	140715.5	9471393	1.4857	0.64

Discussion

In a previous report (Milestone 105), we showed that metagenomic HTS is not as sensitive for detection of low titre bacterial pathogens as qPCR and LAMP that are currently used in laboratory testing. A similar observation was observed in this report showing metagenomic samples were 1,000 – 10,000 X less sensitive than LAMP, qPCR and cPCR for Xf and Ea serial dilutions. Metagenomic HTS for serially diluted Xcc showed that Xcc was unable to be detected confidently through both Bracken and mapping analysis. However, it should be noted that the qPCR and LAMP used in this analysis could only identify Xf to the species level and metagenomic sequencing, though less sensitive, could provide species, subspecies and even ST level identification at some concentrations.

When probes were applied to metagenomic libraries, detection of pathogens increased to or close to qPCR/cPCR levels for Xf, Xcc and Ea serial dilutions. Though serial dilutions were not conducted for non-culturable bacteria, all *Candidatus* Liberibacter samples also saw in increase of probed reads in comparison to metagenomic reads no matter which bioinformatic tool was used (mapping and k-mer analysis). This, however, was not observed for phytoplasma samples. This could be due to the use of a 16Sr1 group phytoplasma reference genome for designing phytoplasma probes, as we have targeted pathogens on Australia's National Priority Plant Pest list. This phytoplasma group is distinct from StLL, group 16SrXXXVII, and WF-B, phytoplasma group 16SrII and the specificity of the probes to the 16SrI phytoplasmas may result in a lack of hybridization due to a high degree of genetic variation at the targeted region. Lack of genes have been considered, however, after analysis of both StLL and WF-B genomes, all MLSA genes were detected. The high variability in nucleotide sequence identity among some genes from different phytoplasma 16Sr groups and StLL has been reported (Rodrigues Jardim et al., 2021). To overcome this problem, more genomes from different phytoplasma 16Sr groups should be added as references for hybridization probe design. It may be that specific phytoplasmas groups will require specific probes to enable detection using targeted HTS.

Kraken2 and Bracken is a fast tool that can be used as an early indicator for pathogen detection. When comparing metagenomic samples to probed samples, Bracken was able to detect the increase of reads. It also gave an indication of fold enrichment, which is the fold increase of the percentage of pathogen detected between metagenomic samples and probed samples. As expected, the percentage of pathogen associated sequence reads decreases as pathogen concentration decrease, however, the fold enrichment gives an indication how the pathogen concentration affects the probes' ability to capture targets. When looking at Xf results, the peak concentration for the most fold enrichment was 2.4×10^{-03} ng/µL, or 0.00354% of Xf, giving 5,736.3 X increase in read percentage. This was also observed for Ea where Ea concentration 4.1×10^{-02} ng/µL, or 0.00796% of Ea, gave the highest fold enrichment of 845.9X. This would indicate that the ideal percent of reads to see the highest fold enrichment is around 0.003% - 0.008%. Additionally, the highest fold enrichment for Xf was around 6.7 X higher than Ea. As only one gene was captured for Ea in comparison to the 13 captured by Xf, there could be a correlation between length of probed regions to fold enrichment, where more probed regions would lead to higher fold enrichment.

Gene targets for Ea included genes associated with biochemical pathways in Ea and unique genes. However, mapping analysis indicated that reads were only mapped to one gene, WP_004154715.1_7. A possible reason for this is that the non-mapped genes may be highly variable among Ea isolates. Nevertheless, Ea reads were still detected in higher percentage than metagenomic samples. The lack of target regions for Ea may have reduced probe detection sensitivity as detection using cPCR was 10 X more sensitive than probed samples, irrespective of which bioinformatic tool was used for detection. Ultimately, hybridization targets for Ea would need to be updated if it were to be used for Ea detection.

Bowtie2 was used to map trimmed reads to regions that were to be probed out with an extra 1kb buffer in the 5' and 3' to detect extra peripheral reads. Looking at fold enrichment using mapping results is a bit more difficult as samples with low titre pathogens may not be detected when mapped to probed regions leading to an undefined number as the equation for fold enrichment is divided by zero. However, when pathogen reads for metagenomic samples aren't zero, we get an increase between 30-85X more than Bracken's fold enrichment. This increase in fold enrichment and the discrepancy between mapped and Bracken reads is probably due to the localized regions being used for mapping instead of a full genome. Nevertheless, the fold enrichment of mapped reads was proportionate to Bracken's fold enrichment.

StringMLST is similar to Kraken2 except it classifies trimmed reads based on an MLST allele library in order to rapidly assign an ST to bacteria. In this report, it was used to assign an ST to Xf samples. ST was only assigned to the first dilution of the XF dilution series, 0.24 ng/µL, of metagenomic Xf samples. Due to an error, the gene *petC* was not added to the probe list impeding the software from assigning an ST for probed Xf samples. This gene will be added in the next validation run. Though unable to assign an ST, there are indications that when included, ST might be assigned up to Xf concentration 2.4 x 10^{-05} ng/µL due to the completion of the remaining genes resulting with an assigned ST at CT 34.21.

The starting material for the Xcc serial dilution may not have been pure culture as Xcc was only detected in the first three serial dilutions using cPCR, Bracken and mapped reads. An indication of the contamination can be seen in Figure 2 where large bands were detected at the top of the gel indicating they were unaffected by the Xcc primers. Though concentration of Xcc was lower than expected, detection of Xcc was not hindered. Xcc was detected confidently up to 4.5 x 10^{-03} ng/µL for both cPCR and probed Xcc samples using Bracken analysis. When trimmed reads were mapped to probed regions, only six genes out of the 11 Xcc genes were detected. Nevertheless, using the mapping tool, Xcc was detected until the Xcc concentration 4.5 x 10^{-03} ng/µL based on these six genes. More than 850 reads were also detected at Xcc concentration 4.5 x 10^{-04} ng/µL. However, Xcc was only detected in one replicate at this dilution and nearly all reads were mapped to the gene AMDFNCPH_02358, which is more than 10 kb long, the longest gene to be probed out in this report. It was also one of the six genes to be mapped to in more concentrated Xcc concentrations. A reason for this could be due to contamination in that replicate from more concentrated Xcc samples, however, it does not justify why only one gene was detected. Alternatively, there is a probability that increasing the length of a probed region could increase the sensitivity of hybridization probes.

The lack of hybridization capture for multiple Ea and Xcc genes would need to be further assessed as variability between isolates may be a problem. Unlike MLSA genes which uses housekeeping genes that can be found in almost every bacterial genome, utilizing unique genes specific to a species or subspecies may prove to be risky. Though it should be noted that variation of MLSA genes can still happen specifically for unculturable bacteria where full genomes for many species remain unavailable, such as phytoplasmas. However, this is a problem for all forms of targeted HTS sequencing. Further analysis to confirm the limit of variation before losing binding capabilities of probes should be considered, which includes testing different isolates on the same probes.

This is due to the probability of high variation or even the absence of the gene within a new isolate. However, this is a common problem for all targeted HTS methods as even primer binding sites could be different. Essentially, multiple strains of each species or subspecies should be compared.

Visualization of CLso and CLaf mapped reads to probed regions showed that reads may even be probed out in the peripheral regions surrounding the gene. This extra information may prove to be useful as the peripheral sequences may aid in differentiating different haplotypes/subtypes. Alternatively, gene length for probed regions can be reduced, leading to lesser, hence, cheaper probes. From Figure 4, especially for CLso, extra regions close to probed regions were also detected. Further analysis would be needed to determine what those regions are.

CLbr is a *Candidatus* Liberibacter species closely related to Claf. To date, there have been no reports of CLbr causing any pathogenic symptoms to host plants, hence it was not included in as a reference for hybridization probe design. CLbr DNA was added in to assess off-target probing with highly related species. Initial analysis of sequenced CLbr showed that there was an increase in sequence reads identified in probed CLbr samples in comparison to metagenomic CLbr samples. Visualization of CLbr mapped reads to probed regions indicated that even though there was a fold enrichment using reads, it may only be due to highly similar regions within probed regions. Unlike CLaf, none of the CLbr genes sequences were complete after targeted HTS using probes indicating that near full gene sequences are required for accurate identification and partial gene sequences may be indicative of another species.

Mapped reads of probed and metagenomic *Candidatus* Liberibacter samples were extracted from BAM files and was *denovo* assembled using SPAdes. All probed CLso and CLaf samples produced all eight *Ca.* Liberibacter genes. Due to the high percentage of CLso in its host plant, metagenomic CLso also produced all eight *Ca.* Liberibacter genes. These genes were then used for MSLA phylogenetic analysis which identified them to the right Liberibacter species. CLso samples gave further

information, identifying it as haplotype A as it clustered with CLso isolate NZ1 and HenneA (Thompson et al., 2015). There are 5 CLso haplotypes identified (Nelson et al., 2013; Thompson et al., 2015). In this report, MLSA phylogenetic analysis separated haplotype B and C accordingly while genomes for haplotype D and E will be added when full genomes become publicly available.

The hybridization probe protocol required samples to be pooled before probing and an amplification step post probing. Due to this, sequencing output per sample cannot be controlled in each pool leading to higher amounts of sequencing data for samples with higher pathogenic DNA, which was observed in this report. This could interfere with pathogen detection when low titre pathogens are mixed with high titre pathogens in the same pool. This potential problem will be addressed in future experimentation.

Results from this report showed that probes can increase HTS pathogen detection sensitivity to qPCR/cPCR levels. However, if possible, options to increase sensitivity should be trialled. This report has indicated that longer probed regions might increase sensitivity as seen for Xcc. More probed regions were also indicated to increase fold enrichment, which would lead to better identification due to the extra reads and better gene assembly for phylogenetic assembly. Another method would be to modify the hybridization enrichment protocol. After probe capture using streptavidin-coated magnetic beads, a PCR amplification is done to increase concentration of targeted DNA (Figure 1). Due to the protocols being designed for accumulatively longer probed regions such as whole exome capture, there is a probability that increasing the cycles of the final PCR may increase pathogen detection.

In conclusion, this protocol shows potential for pathogen detection using probe based targeted HTS. although further improvements and validations are necessary. This includes increasing the number of reference genomes representing a broad range of strains of all pathogens, for hybridization probe design. These draft protocols need to be tested with PEQ to further validify the pipeline and identify any improvements needed.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Bronzato Badial, A., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., & King, J. (2018). Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. *Plant Dis, 102*(8), 1648-1652. doi:10.1094/pdis-04-17-0488-re

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., . . . Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology, 68*(2), 229-238. doi:10.1111/ppa.12957

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics, 6*(348). doi:10.3389/fgene.2015.00348

Ghosh, A., Mehta, A., & Khan, A. M. (2019). Metagenomic Analysis and its Applications. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 184-193). Oxford: Academic Press.

Gupta, A., Jordan, I. K., & Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics, 33*(1), 119-121. doi:10.1093/bioinformatics/btw586

Harper, S. J., Ward, L. I., & Clover, G. R. G. (2010). Development of LAMP and Real-Time PCR Methods for the Rapid Detection of Xylella fastidiosa for Quarantine and Field Applications. *Phytopathology®, 100*(12), 1282-1288. doi:10.1094/PHYTO-06-10-0168

Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., & Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. *Phytobiomes Journal, 3*(2), 92-101. doi:10.1094/pbiomes-01-19-0004-r

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, *35*(6), 1547.

Nelson, W. R., Sengoda, V. G., Alfaro-Fernandez, A. O., Font, M. I., Crosslin, J. M., & Munyaneza, J. E. (2013). A new haplotype of "Candidatus Liberibacter solanacearum" identified in the Mediterranean region. *European Journal of Plant Pathology*, *135*(4), 633-639

Jardim, B. R., Kinoti, W. M., Tran-Nguyen, L. T., Gambley, C., Rodoni, B., & Constable, F. E. (2021). 'Candidatus Phytoplasma stylosanthis', a novel taxon with a diverse host range in Australia, characterised using multilocus sequence analysis of 16S rRNA, secA, tuf, and rp genes. *International Journal of Systematic and Evolutionary Microbiology*, *71*(1)

Thompson, S. M., Johnson, C. P., Lu, A. Y., Frampton, R. A., Sullivan, K. L., Fiers, M. W., ... & Smith, G. R. (2015). Genomes of 'Candidatus Liberibacter solanacearum'haplotype A from New Zealand and the United States suggest significant genome plasticity in the species. *Phytopathology*, *105*(7), 863-871

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Appendix 23: Evaluation of performance of different small RNASeq library kits

Background

In our previous report, we carried out a RING test which highlighted the importance of using the right combination of library preparation and sequencing technology to minimise cross-contamination between samples caused by index misassignment (cross-talk). The shift from NextSeq 500 to NovaSeq 6000 sequencing machine greatly increased the rate of index hopping we observed in the samples processed with the NEBNext kit by the sequencing provider AGRF. This is expected as random clustering like the old NextSeq are less prone to index misassignment than patterned flow cells used by modern Illumina instruments such as NovaSeq. With these instruments, index misassignment is expected to occur at ~2% frequency (~1.2M reads at 60M depth). This problem is further compounded by the fact that NovaSeq has an increased data yield compared to the NextSeq. Random clustering allowed for variations in data output so you could reduce the sequencing capacity as required if needed, whereas the NovaSeq flow-cells now have relatively fixed outputs. Furthermore, the NEBNext kit caps the number of samples that can be processed to 48 whereas other kits are more flexible.

With the move to patterned flow cells in modern Illumina machines that suffer significantly more the risk of index hopping, it is therefore recommended to use unique dual indexed sequencing adapters (UDIs) and/or unique molecular identifiers (UMIs) to minimize index misassignment [52]. Unfortunately, the NEBNext kit use single-indexed sequencing adapters when pooling samples prior to sequencing, It does not offer at the moment the possibility to incorporate UDIs which could mitigate the index hopping issues observed. In our RING test evaluation, we found that QIAseq kit with both UMIs and UDIs incorporation was our top contender. Key observed features include relatively even distribution of raw sequencing output, detection of all targeted viruses and viroids but no false positive (index hopping events).

To further validate our findings we performed a larger side-by-side comparison of samples processed using either the NEBNext library kit (AGRF) or the QIAseq library kit (Ramaciotti). In partnership with the SSG team and DPI, a total of 48 samples (which were part of Batch 3 and 4) were selected from a range of commodities, which included 26 citrus, 10 apples, 3 clonal grasses, flowering plants, and strawberries and a single raspberry, cactus and tree (**Table 23.1**). Total RNA extraction was performed using an automated approach on the Maxwell® Rapid Sample Concentrator Instrument with a SimplyRNA Tissue kit (AS1340, Promega). The same RNA extraction from each plant was used for the preparation of sRNA-Seq libraries by each sequencing providers (Ramaciotti Centre for Genomics at the UNSW in Sydney and the Australian Genome Research Facility in Melbourne).

Table 23.1. Description of commodities represented in the comparative study. See also Tables 23.1A and 23.B for additional details.

Commodity	**Sample count**
Apple	10
Cactus	1
Citrus	26
Clonal grass	3
Flowering plant	3
Raspberry	1
Strawberry	3
Tree	1
Total	**48**

Table 21.1A: Batch3 sample metadata

Sample ID	Commodity	Genus	Species	Common name	Plant tissue collected	Collection date (Month, Year)	Total RNA extraction kit
MT212	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT213	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT214	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT215	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT216	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT217	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT218	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT219	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT220	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT221	Pome fruit	Malus	domestica	Apple	50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT222	Citrus	Citrus		lime	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT223	Citrus	Citrus		orange	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT224	Citrus	Citrus		orange	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT225	Citrus	Citrus		lime	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue

MT226	Citrus	Citrus	mandarin	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT227	Citrus	Citrus	orange	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT228	Citrus	Citrus	lime	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT229	Citrus	Citrus	lemon	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT230	Citrus	Citrus	lemon	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT231	Citrus	Citrus	citron	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT232	Citrus	Citrus	mandarin	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT233	Citrus	Citrus	citron	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT234	Citrus	Citrus	orange	50 mg leaf/midrib	Nov-20	ReliaPrep RNA Miniprep Tissue
MT235	Citrus	Citrus	orange	50 mg leaf/midrib	Nov-20	ReliaPrep RNA Miniprep Tissue
MT236	Citrus	Citrus	mandarin	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT237	Citrus	Citrus	citron	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT238	Citrus	Citrus	orange	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue

MT239	Citrus	Citrus		citron	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT240	Citrus	Citrus		lemon	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT241	Citrus	Citrus		lime	50 mg leaf/midrib	Jan-21	ReliaPrep RNA Miniprep Tissue
MT242	Table Grapes	Vitis	vinifera	Grapevine	50 mg leaf/midrib	Apr-21	Maxwell RSC Pure RNA
MT243	Blueberry	Vaccinium	corymbosum	highbush Blueberry	50 mg leaf/midrib	Apr-21	Maxwell RSC Pure RNA
MT244		Matucana	sp	Cactus	20 mg tissue	Jul-21	Maxwell RSC Pure RNA
MT245new	Strawberry	Fragaria	ananassa	Tulameen	50 mg petiole	Jul-21	Maxwell RSC Pure RNA
MT246	Citrus	Citrus	reticulata	Mandarin	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT247	Citrus	Citrus	reticulata	Mandarin	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT248	Citrus	Citrus	reticulata	Mandarin	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT249	Citrus	Citrus	reticulata	Mandarin	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT250	Citrus	Citrus	reticulata	Mandarin	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT251	Strawberry	Fragaria x	ananassa	Strawberry	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT252	Strawberry	Fragaria x	ananassa	Strawberry	50 mg leaf/midrib	May-21	Maxwell RSC Pure RNA
MT253	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT254	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT255	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT256	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA

MT257	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT258	Stonefruit	Prunus	sp		50 mg leaf/midrib	Dec-20	Maxwell RSC Pure RNA
MT259	Clonal Grass	Cynodon	dactylon	Bermuda grss	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT260	Clonal Grass	Zoysia	japonica hybrid	Zoysia	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT261	Clonal Grass	Stenotaphrum	secundatum	Stenotaphrum	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT262	Clonal Grass	Zoysia	spp. willd	Zoysia	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT263	Clonal Grass	Zoysia	spp. willd	Zoysia	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT264	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT265	Hydrangea	Hydrangea	paniculata	Hydrangea	50 mg leaf/midrib	Jul-21	Maxwell RSC Pure RNA
MT266	Hydrangea	Hydrangea	paniculata	Hydrangea	50 mg leaf/midrib	Jun-21	Maxwell RSC Pure RNA
MT267	Lagerstroemia	Lagerstroemia	indica x fauriei	Crepe Myrtle (Hot Pink)	50 mg leaf/midrib	Jun-21	Maxwell RSC Pure RNA
MT268	Strawberry	Fragaria x	ananassa	Strawberry	50 mg petiole	Jul-21	Maxwell RSC Pure RNA
MT269		Viburnum	plicatum	Japaneese snowball	50 mg leaf/midrib	Jun-21	Maxwell RSC Pure RNA
MT270	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT271	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT272	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT273	Citrus	Citrus	limon	lemon	50 mg leaf/midrib	Jul-21	Maxwell RSC Pure RNA
MT274	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA
MT275	Strawberry	Fragaria	ananassa	Strawberry	50 mg petiole	Jun-21	Maxwell RSC Pure RNA

MT276	Clonal Grass	Pennisetum	purpureum	Pennisetum	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT277	Clonal Grass	Pennisetum	purpureum	Pennisetum	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT278	Clonal Grass	Pennisetum	purpureum	Pennisetum	50 mg leaf blade	Jun-21	Maxwell RSC Pure RNA
MT279	rosa	rosa	multiflora	Rose	50 mg symptomatic leaf	Nov-20	Maxwell RSC Pure RNA

Table 23.1B. Batch 4 Sample metadata

Sample ID	Commodity	Genus	Species	Common name	Plant tissue collected	Collection date (Month, Year)	Total RNA extraction kit
MT280	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT281	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT282	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT283	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT284	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT285	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT286	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT287	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT288	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT289	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT290	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA

MT291	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT292	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT293	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT294	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT295	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT296	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT297	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT298	Strawberry	Fragaria	Ananassa	Strawberry	50 mg petiole	Sep-21	Maxwell RSC Pure RNA
MT299	Potato	Solanum	tuberosum	Potato	50 mg symptomatic leaf	Sep-21	Maxwell RSC Pure RNA

Major findings

Citrus samples

- Twenty five out of the 26 viruses and viroids identified by molecular assays were both recovered in the NEBNext and the QIAseq samples. One Citrus tristeza virus detection was inaccurately predicted to be a false positive by our contamination flag in sample MT232 due to the high variability in read abundance for this virus in this dataset. In comparison the range of abundance recovered for this virus in the QIAseq data was much narrower **(Table 23.2)**.

- A low titer viroid present in sample MT232, Citrus virioid I low sequence similarity (CVd-I-LSS), was detected by molecular testing but was not recovered by VirReport in the QIAseq data, even after additional sequencing was performed. Only 7 reads aligned to the top CVd-I-LSS reference hit, which explains why the Velvet de novo assembly failed to recover a contig for this viroid.

While CVd-I-LSS was detected in the NEBNext-derived MT232 sample, it was flagged as a candidate false positive in this sequencing batch as it was recovered at a similar abundance to that of other confirmed false positive CVd-I-LSS identified in other samples of this batch (**Table 23.2**). The reads are predicted to originate from a high titer CVd-I-LSS present in sample MT233.

- MT233 is an inoculum of mandarin (MT232) onto etrog citron and thus these two plants are expected to be infected by the same CVd-I-LSS. Thus investigation of the nucleotide composition of the viroid sequence in M232 will not permit to distinguish reads that might truly originate from MT232 to those that have been misassigned as they will all share the same SNPs (**Figure 23.1**).

- Two additional pathogens were identified via HTS in sample MT233 by both NEBNext and QIAseq methods: a citrus vein enation virus and a citrus bent leaf viroid (**Table 23.2**). DPI will run RT-PCR to independently confirm this result.

- Twenty additional viruses/viroids were only detected in the NEBNext samples (CEVD was detected across 9 samples, CVd-I-LSS into 8 plants, CVdVII into 2 plants and CTV into one) and were flagged as contamination events. These detections were not detected through DPI molecular assays and were not predicted in the QIAseq samples further giving support that these are false positives (**Tables 23.2 and 23.3**).

- No false positives were detected in the QIAseq samples (**Tables 23.2 and 23.3**).

Figure 23.1. The high titer CVd-I-LSS virus derives siRNAs recovered in MT233 (represented in the top panel) is likely to be at the origin of all sequencing read products detected in other samples present in this batch, as all of them share the same SNPs. Sample MT232 is represented in the middle panel and sample MT236 at the bottom.

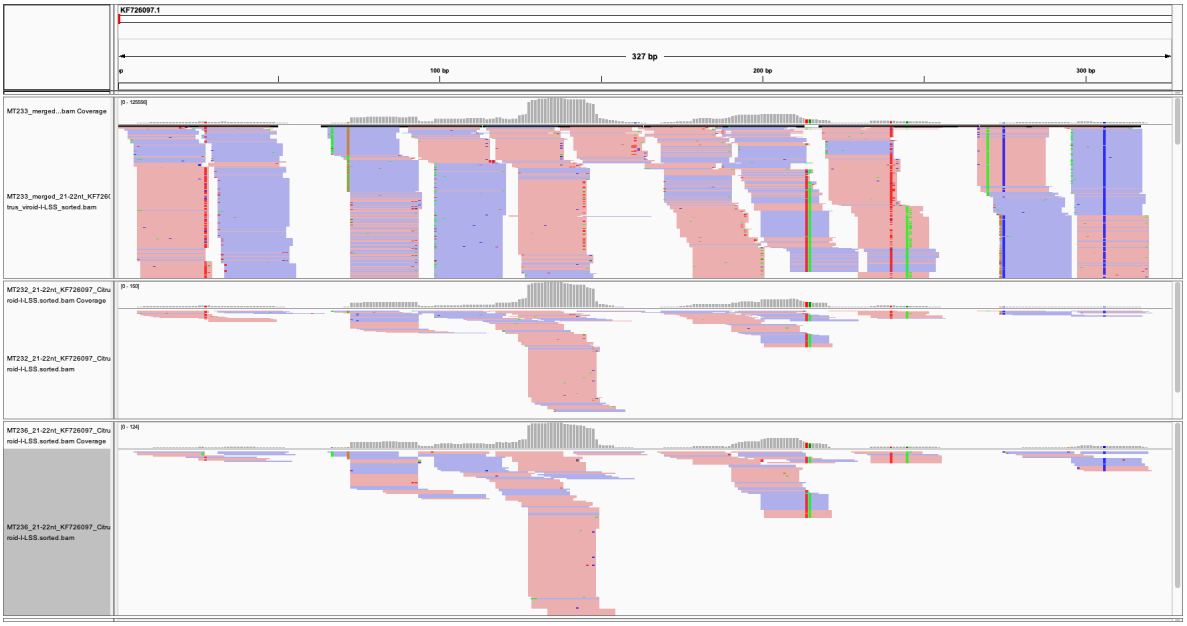


Table 23.2. Top BLASTN results of assembled virus and viroid sequences, coverage statistics and implementation of a contamination flag to identify false positives. The batch of samples prepared using NEBNext are located on the left-hand side (in orange) and the batch of samples prepared using QIAseq are located on the right-hand side (in blue). Predicted false positives are represented in red. Additional findings via HTS are represented in blue. sacc = subject (reference) accession sequence, slen = subject (reference) length, coverage = percent of the reference genome covered by the assembled sequence(s), av-pident = average sequence identity percent of assembled sequence(s) to reference, CF = contamination flag, DPI = Detection confirmed via molecular testing, FP = false positive, Mean cov = mean coverage in bases to the genome/sequence of the best homology match, PCT_1X = fraction of bases that attained at least 1X sequence coverage, RPKM = Reads Per Kilobase of transcript, per Million mapped reads. Species abbreviations: CBLVd = Citrus bent leaf viroid; CDVd = Citrus CBLVd = Citrus bent leaf viroid; CDVd = Citrus dwarfing viroid; CEVd = Citrus exocortis viroid; CiVA = Citrus virus A; CTV = Citrus tristeza virus; CTLV/ASGV = Citrus tatter leaf virus/Apple stem grooving virus; CVd-VI = CVd-VI; CVd-VII = Citrus viroid VII; CVEV = Citrus vein enation virus; CVd-I-LSS = Citrus viroid I-low sequence similarity; HSVd = Hop stunt viroid.

Sample	Virus/ Viroid	DPI	sacc	slen	av-pident	Mean cov	RPKM	PCT_1X	CF	sacc	slen	av-pident	Mean cov	RPKM	PCT_1X	CF
MT222	CTV	Y	MH051718	19254	99.4	124.1	112	1		MH051718	19254	99.6	106.8	1,667	0.99	
MT222	CEVd		KJ810547	375	100.0	77.3	40	0.96	FP							
MT223	CTV	Y	MH279618	19259	98.2	154.0	210	0.97		MK779711	19260	98.2	130.2	2,055	0.94	
MT223	CEVd		KJ810543	372	100.0	70.3	43	0.97	FP							
MT224	CTV	Y	MH051718	19254	99.1	234.5	700	1		MH051718	19254	99.4	167.1	4,338	1	
MT224	CEVd		KJ956802	384	100.0	74.5	46	0.95	FP							
MT225	CTV	Y	FJ525435	19273	99.6	243.5	1155	1		FJ525435	19273	99.4	217.4	5,330	1	
MT225	CEVd		KJ956803	372	100.0	58.2	45	0.97	FP							
MT226	CTV	Y	MH051718	19254	99.4	66.1	55	0.97		MH051718	19254	99.5	127.0	1,252	0.97	
MT226	CEVd		KX084708	374	100.0	57.6	37	0.87	FP							
MT227	CTV	Y	MH051719	19253	98.1	149.0	199	0.99		EU857538	19251	99.0	160.3	2,832	0.99	
MT227	CEVd		KX084708	374	100.0	72.9	49	0.88	FP							
MT228	CTV	Y	FJ525435	19273	99.6	228.5	501	1		FJ525435	19273	99.5	183.9	3,530	1	
MT228	CEVd		FJ773256	372	99.3	108.8	90	0.97	FP							
MT229	CDVd	Y	EU564176	296	100.0	239.0	1012	0.97		EU564176	296	99.0	238.7	14,110	0.99	
MT229	CTV	Y	EU857538	19251	99.6	204.2	375	1		EU857538	19251	99.6	163.6	2,617	1	
MT229	HSVd	Y	KY829130	302	99.2	220.7	2285	0.96		KY654683	299	99.0	248.5	53,732	1	
MT229	CEVd		KJ956802	384	100.0	74.2	49	0.95	FP							
MT230	CTV	Y	KU883267	19280	99.4	140.2	208	0.99		GQ454870	19269	99.0	103.4	2,835	0.99	
MT230	CVEV	Y	MN187044	5982	99.2	55.5	34	0.95		MN187043	5983	99.5	65.3	974	0.94	
MT230	CVd-VII	Y	KX013549	368	100.0	175.5	381	0.94		KX013549	368	100.0	205.3	16,961	0.97	
MT230	CEVd		KJ956802	384	100.0	67.4	39	0.95	FP							
MT231	CEVd	Y	KY110722	370	100.0	238.1	103335	0.99		KY110721	372	100.0	235.7	895,840	1	
MT231	CTV		FJ525435	19273	100.0	2.0	1	0.66	FP							
MT232	CTV	Y	**FJ525435**	**19273**	**99.5**	**12.5**	**6**	**0.9**	FP	FJ525435	19273	99.0	194.7	4,183	1	
MT232	CVd-VII		KX013549	368	100.0	5.0	2	0.67	FP							
MT232	CVd-I-LSS	Y	**KF726097**	**326**	**100.0**	**20.2**	**9**	**0.92**	FP							
MT233	CBLVd	Y	**KY654680**	**327**	**97.4**	**152.4**	**3476**	**0.75**		**KY110715**	**318**	**100.0**	**197.8**	**122,817**	**0.91**	
MT233	CTV	Y	KU883267	19254	99.2	127.6	243	0.99		KU883267	19280	99.0	161.0	2,327	1	
MT233	CVEV		**MN187044**	**5982**	**99.1**	**32.5**	**39**	**0.91**		**MN187040**	**5983**	**99.7**	**96.4**	**802**	**0.97**	
MT233	CVd-I-LSS	Y	AB054641	327	100.0	235.3	29,305	0.98		KF726097	326	100.0	247.1	554,747	1	
MT234	HSVd	Y	KC677729	297	100.0	62.0	33	0.78		KX156934	303	99.4	237.3	43,309	1	
MT234	CVd-I-LSS		KF726095	328	98.2	33.9	19	0.94	FP							
MT235	ASGV/CTLV	Y	MK929792	6475	98.8	45.3	20	0.87		MK929792	6475	99.0	72.4	727	0.92	
MT235	CDVd	Y	LN681197	297	100.0	202.6	116	0.96		JX259432	294	99.4	215.2	6,349	0.96	
MT235	CTV	Y	MH051719	19253	99.3	86.6	62	0.99		MH051719	19253	99.1	54.7	590	0.95	
MT235	CVd-I-LSS		AB054639	325	100.0	17.9	7	0.89	FP							
MT236	CVd-V	Y	MN640600	294	100.0	69.8	34	0.94		MN640600	294	100.0	139.7	1,370	0.94	
MT236	CVd-VI	Y	KC769203	331	100.0	46.5	22	0.94		KC769203	331	100.0	98.3	997	0.95	
MT236	CVd-I-LSS		KF726097	326	100.0	18.0	9	0.87	FP							
MT237	CVd-VII	Y	KX013551	368	100.0	234.1	5674	0.96		KX013551	368	99.6	240.1	822,149	0.98	
MT237	CVd-I-LSS		KF726097	326	100.0	33.9	34	0.91	FP							
MT238	CiVA RNA1	Y	MG764565	6691	98.7	42.6	21	0.86		MK273077	6690	97.2	52.8	1,478	0.91	
MT238	CiVA RNA2	Y	MG764566	2740	96.4	50.6	24	0.87		MZ463040	2686	98.2	67.6	1,774	0.9	
MT238	CVd-I-LSS		KF726097	326	98.1	17.2	8	0.9	FP							
MT239	CVd-I-LSS		KF726097	326	100.0	17.9	7	0.91	FP							
MT240	CVd-I-LSS		KF726098	323	100.0	12.0	9	0.79	FP							
MT241	CVd-VII		KX013549	368	100.0	10.0	4	0.87	FP							
MT241	CVd-I-LSS		KF726095	328	98.1	27.3	12	0.91	FP							

Table 23.3. Summary of HTS results highlighting which detections were shared or private to NEBNext and QIAseq. The detections represented in red are predicted to be false positives. Additional detections that need to be validated by molecular testing are represented in blue. Species in italics have been independently validated via molecular testing via DPI or SSG. ND = no detections. Species abbreviations: CBLVd = Citrus bent leaf viroid; CDVd = Citrus CBLVd = Citrus bent leaf viroid; CDVd = Citrus dwarfing viroid; CEVd = Citrus exocortis viroid; CiVA = Citrus virus A; CTV = Citrus tristeza virus; CTLV/ASGV = Citrus tatter leaf virus/Apple stem grooving virus; CVd-VI = CVd-VI; CVd-VII = Citrus viroid VII; CVEV = Citrus vein enation virus; CVd-I-LSS = Citrus viroid I-low sequence similarity; HSVd = Hop stunt viroid; ACLSV = Apple chlorotic leaf spot virus; AGCaV = Apple green crinkle associated virus; AHVd = Apple hammerhead viroid; ApMV = Apple mosaic virus; ARWV-1 = Apple rubbery wood virus 1; ARWV-2 = Apple rubbery wood virus 2; ASGV = Apple stem grooving virus; ASPV = Apple stem pitting virus; CCGaV = Citrus concave gum associated virus; CMMoV = Cactus mild mottle virus; CTV = Citrus tristeza virus; CVEV = Citrus vein enation virus; CVX = Cactus virus X; HSVd = Hop stunt viroid; PiVX = Pitaya virus X; SchVX = Schlumbergera virus X; RCNaV = Rattail cactus necrosis-associated virus; ZyVX = Zygocactus virus X.

Sample	Viral detections by both NEBNext and QIASeq	Viral detections by NEBNext only	Viral detections by QIASeq only
MT212	ACLSV, AHVd, ASPV		
MT213	ACLSV, AHVd, ARWV-1, ASGV, ASPV, CCGaV		
MT214	ACLSV, AHVd, ApMV, ASGV, ASPV, CCGaV		
MT215	ACLSV, AHVd, ARWV-1, ASGV, ASPV		
MT216	ACLSV, AHVd, ApMV, ARWV-1, ASGV, ASPV, CCGaV		
MT217	ACLSV, AHVd, ApMV, ARWV-2, ASGV, CCGaV		ASPV, Cherry symptomless virus
MT218	ACLSV, AGCaV, AHVd, ApMV, ARWV-1, ASGV, ASPV		
MT219	ACLSV, AHVd, ARWV-1, ASGV, ASPV		
MT220	AHVd		
MT221	ACLSV, AHVd, ASGV, ASPV		
MT222	CTV	CEVd	
MT223	CTV	CEVd	
MT224	CTV	CEVd	
MT225	CTV	CEVd	
MT226	CTV	CEVd	
MT227	CTV	CEVd	
MT228	CTV	CEVd	
MT229	CDVd, CTV, HSVd	CEVd	
MT230	CTV, CVEV, HSVd	CEVd	
MT231	CEVd	CTV	
MT232	CTV	CVd-I-LSS*, CVd-VII	
MT233	CTV, CVd-I-LSS, **CVEV, CbIVd**		
MT234	HSVd	CVd-I-LSS	
MT235	ASGV, CDVd, CTV	CVd-I-LSS	

MT236	CVd-V, CVd-VI	CVd-I-LSS	
MT237	CVd-VII	CVd-I-LSS	
MT238	CiVA	CVd-I-LSS	
MT239		CVd-I-LSS	
MT240		CVd-I-LSS	
MT241		CVd-VII, CVd-I-LSS	
MT244	CMMoV, CVX, Mytcor virus 1, PiVX, Potexvirus ST5, SchVX, ZyVX,	AHVd	RCNaV
MT245	RYNV	AHVd	
MT246	CTV	AHVd	
MT247	CTV, CVEV	AHVd	
MT248	**ND**		
MT249	CTV, HSVd	AHVd	
MT250		AHVd	
MT251		AHVd	
MT252		AHVd	
MT265		AHVd	
MT266		AHVd	
MT267			
MT269		AHVd	
MT273		AHVd	
MT275	**ND**		
MT276		AHVd	
MT277	**ND**		
MT278		AHVd	
Total	**90**	**35**	**3**

PEQ samples

- Ten samples originated from an apple orchard in Tasmania and were not client plants undergoing quarantine testing at PEQ. SGS was interested to get the profile of locally established plants. Thus all virus and viroid detected via NGS are being reported here for these samples regardless of their regulatory status. SGS will not be conducting diagnostic testing on these extracts. All 10 apple samples were predicted to be infected with at least one virus/viroid. Nine virus species (Apple stem pitting virus, Apple chlorotic leaf spot virus, Apple rubbery wood virus 1, Apple rubbery wood virus 2, Apple stem grooving virus, Citrus concave gum-associated virus, Apple mosaic virus, Apple green crinkle associated virus and Cherry symptomless virus) and 1 viroid species (Apple hammerhead viroid) were detected across the 10 samples. Up to 8 viruses/viroids were predicted to be found in the same plant (i.e. MT217). Fifty pathogens were detected by both the NEBNext and the QIAseq samples. Two additional viruses were only detected by QIAseq: Apple stem pitting virus and Cherry symptomless virus in sample MT217. The Apple hammerhead viroid was detected in 14 other commodities in the NEBNext samples and were flagged as contamination likely originating from sample MT219. This viroid was

not detected in these sample in the QIAseq batch processed at Ramaciotti, suggesting that they are indeed false positives (**Tables 23.3 and 23.4**).

- MT244 was a Matunaca (cactus) sample that was intercepted at the border and thus was not a client plant quarantined at PEQ. Diagnostic testing conducted at PEQ with in-house assays detected Cactus virus X. This virus was also detected in both NEBNext and QIAseq-derived sample, along with an additional six pathogens. A Rattail cactus necrosis-associated virus was only detected in the QIAseq-derived sample. An AHVd was only detected in the AGRF dataset. This detection was flagged by our pipeline as an index hopping event from an apple sample (see the apple section above) and the absence of the viroid in the Ramaciotti dataset supports this hypothesis (**Tables 23.3 and 23.4**).

- Amongst the six citrus plants provided by SGS, 3 plants (MT248, MT250 and MT273 had no detections through both PEQ and HTS methods aside from the AHVd false positive detection (see the apple section above). The other three plants (MT246, MT247 and MT249) had positive detections but only CTV was detected through PEQ assays. The other detections (CVEV in MT247 and HSVd in MT249) will need to be validated by SGS (**Tables 23.3 and 23.4**).

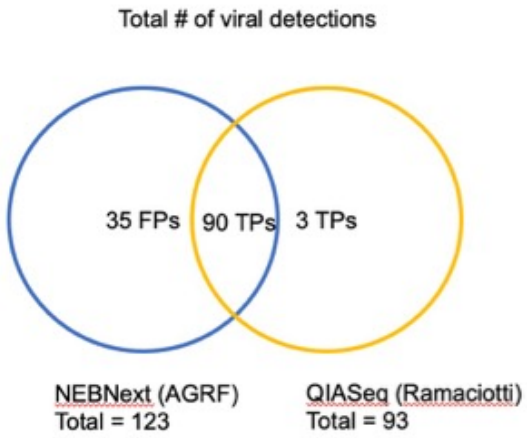
- With the exception of the AHVd false positive, there were no other detections found in the clonal grasses (MT276, MT277 and MT278), lagerstroemia (MT267), strawberry (MT251, MT252, MT275), and flowering plants tested (MT265, MT266, MT267, MT269); and only an endogenous RYNV was detected in the raspberry 'Tulameen' sample MT245 (**Tables 23.3 and 23.4**).

Table 23.4. Top BLASTN results of assembled virus and viroid sequences for the PEQ samples, coverage statistics and implementation of a contamination flag to identify false positives. Samples indicated with an asterisk are pending testing at SGS/PEQ. See Table 2 for legends details. Species abbreviations: ACLSV = Apple chlorotic leaf spot virus; AGCaV = Apple green crinkle associated virus; AHVd = Apple hammerhead viroid; ApMV = Apple mosaic virus; ARWV-1 = Apple rubbery wood virus 1; ARWV-2 = Apple rubbery wood virus 2; ASGV = Apple stem grooving virus; ASPV = Apple stem pitting virus; CMMoV = Cactus mild mottle virus; CCGaV = Citrus concave gum associated virus; CTV = Citrus tristeza virus; CVEV = Citrus vein enation virus; CVX = Cactus virus X; HSVd = Hop stunt viroid; PiVX = Pitaya virus X; SchVX = Schlumbergera virus X; RCNaV = Rattail cactus necrosis-associated virus; RYNV = Rubus yellow net virus; ZyVX = Zygocactus virus X.

Sample	Commodity	Virus/Viroid	PEQ_Stat	sacc	slen	av-	Mean	RPKM	PCT_1	CF	sacc	slen	av-	Mean	RPKM	PCT_1	CF
MT212*	Apple	ACLSV	NA	AM498050	793	95.6	4.5	2	0.66		MK929794	7511	100.0	8.3	32	0.75	
MT212*	Apple	AHVd	NA	MK188696	435	100.0	218.7	1431	0.95		LC579906	434	100.0	239.2	51326	0.97	
MT212*	Apple	ASPV	NA	MK836301	9270	98.1	8.6	48	0.81		MK836301	9270	95.7	40.2	145	0.95	
MT213*	Apple	ACLSV	NA	KP772645	358	93.7	36.3	10	0.99		MZ126521	7542	99.0	13.3	128	0.82	
MT213*	Apple	AHVd	NA	MK690629	374	98.5	246.2	4220	0.99		MK947213	436	100.0	240.1	151821	1	
MT213*	Apple	ARWV-1	NA	MK936226	1595	97.9	117.8	629	0.92		MK936225	1392	97.2	125.0	4390	0.91	
MT213*	Apple	ASGV	NA	MK923757	6470	98.9	23.9	162	0.88		LT160740	6458	99.1	38.9	383	0.86	
MT213*	Apple	ASPV	NA	KY081174	1157	94.1	10.4	9	0.81		MZ148046	9324	98.6	16.3	157	0.9	
MT213*	Apple	CCGaV RNA1	NA	MK940540	6652	98.9	19.5	106	0.84		MK940540	6652	99.1	32.2	333	0.86	
MT213*	Apple	CCGaV RNA2	NA	MK940543	2700	98.4	146.6	964	0.96		MK940542	2700	99.3	170.3	7746	0.99	
MT214*	Apple	CCGaV RNA1	NA	MH534886	582	99.0	21.8	8	0.93		MZ126515	7539	98.4	17.6	160	0.87	
MT214*	Apple	CCGaV RNA2	NA	MK188700	433	96.7	210.5	2110	0.99		MK947213	436	96.1	233.5	136168	1	
MT214*	Apple	ApMV RNA1	NA	KY965059	3477	99.7	19.4	50	0.87		MN822137	3274	99.5	34.4	314	0.93	
MT214*	Apple	ApMV RNA2	NA	KY883319	2929	100.0	10.9	26	0.53		MN822138	2983	99.1	25.7	258	0.87	
MT214*	Apple	ApMV RNA3	NA	KY971019	1991	99.5	81.3	146	0.97		KY971019	1991	99.6	147.4	2730	1	
MT214*	Apple	ASGV	NA	MK923757	6470	97.0	27.6	168	0.90		MN786531	6487	97.5	55.6	544	0.95	
MT214*	Apple	ASPV	NA	KF321967	9324	93.3	4.0	24	0.67		KF321967	9324	97.4	9.3	85	0.85	
MT214*	Apple	CCGaV RNA1	NA	MK940540	6652	98.9	26.5	127	0.89		MK940541	6647	99.2	49.0	518	0.91	
MT214*	Apple	CCGaV RNA2	NA	MH038043	2704	99.0	153.9	871	0.97		MK940543	2700	99.0	181.7	8382	0.99	
MT215*	Apple	ACLSV	NA	MK929796	7521	100.0	5.2	28	0.50		LC533838	7557	98.9	23.8	197	0.86	
MT215*	Apple	AHVd	NA	MK188688	434	97.3	245.7	1120	1.00		MK188687	434	94.8	246.6	72526	1	
MT215*	Apple	ARWV-1	NA	MF062136	7219	98.9	24.9	119	0.84		MK936227	7171	97.0	38.0	368	0.83	
MT215*	Apple	ASGV	NA	MK923757	6470	98.5	22.0	108	0.88		MZ126543	6488	98.9	51.4	438	0.92	
MT215*	Apple	ASPV	NA	KJ522472	9262	95.8	4.4	24	0.62		MZ148054	9351	96.6	15.2	125	0.89	
MT216*	Apple	ACLSV	NA	MH534887	582	100.0	22.7	8	0.83		MK929796	7521	96.2	22.9	185	0.89	
MT216*	Apple	AHVd	NA	MK188696	435	96.4	222.4	2083	0.94		MG662376	423	98.6	246.1	149978	1	
MT216*	Apple	ApMV RNA1	NA	KY965059	3477	98.0	10.8	25	0.79		KY965059	3477	99.5	41.5	365	0.92	
MT216*	Apple	ApMV RNA2	NA								MN822138	2983	99.1	22.8	246	0.85	
MT216*	Apple	ApMV RNA3	NA	KY971019	1991	98.9	61.1	115	0.98		KY971019	1991	98.3	159.5	2722	1	
MT216*	Apple	ARWV-1	NA	MF062137	1613	99.8	92.0	246	0.91		NC_055390	7219	99.2	21.8	156	0.78	
MT216*	Apple	ASGV	NA	MK929791	6478	99.0	10.7	56	0.84		MK923757	6470	99.0	44.7	326	0.93	
MT216*	Apple	ASPV	NA	KF319056	9284	91.1	4.5	25	0.63		MZ147984	9345	96.5	31.6	238	0.91	
MT216*	Apple	CCGaV RNA1	NA	MK940540	6652	98.9	19.7	87	0.88		MK940541	6647	98.9	68.4	612	0.94	
MT216*	Apple	CCGaV RNA2	NA	MK940542	2700	98.9	142.6	652	0.97		MK940543	2700	98.9	196.7	12658	0.98	
MT217*	Apple	ACLSV	NA	MK929795	7520	100.0	9.4	41	0.64		MK929796	7521	94.5	17.7	157	0.68	
MT217*	Apple	AHVd	NA	MK947213	436	98.5	237.6	2229	0.99		MN786530	434	95.3	179.3	45260	0.9	
MT217*	Apple	ApMV RNA1	NA	KY965059	3477	98.0	13.1	29	0.82		KY965059	3477	98.5	28.0	226	0.89	
MT217*	Apple	ApMV RNA2	NA								MN822138	2983	99.3	19.2	157	0.84	
MT217*	Apple	ApMV RNA3	NA	KY971019	1991	99.3	70.7	125	0.95		KY971019	1991	98.6	123.2	1608	0.98	

MT246	Citrus	AHVd	NA		MK690630	374	99.0	19.2	5	0.93	FP						
MT246	Citrus	CTV	Y	R	MK491895	19252	99.3	96.2	1527	1.00		AB046398	19302	99.5	86.5	654	0.98
MT247	Citrus	AHVd	NA		MK690630	374	97.4	26.6	6	0.99	FP						
MT247	Citrus	CTV	Y	R	MH279618	19259	98.8	55.2	883	0.99		MK779711	19260	98.6	60.6	700	0.98
MT247	Citrus	CVEV	N	R	MN187044	5982	99.5	91.5	376	0.98		MN187044	5982	99.5	151.0	1814	0.99
MT248	Citrus	ND	ND									ND					
MT249	Citrus	AHVd	NA		MK947213	436	98.0	29.9	8	0.99	FP	ND					
MT249	Citrus	CTV	Y	R	MK491895	19252	99.2	76.2	1101	0.99		AB046398	19302	99.3	112.3	659	0.98
MT249	Citrus	HSVd	N	R	MH995526	226	100.0	137.8	24	1.00		MT917186	302	100.0	212.9	2701	1
MT250	Citrus	AHVd	NA		MK690630	374	99.0	24.0	5	0.99	FP	ND					
MT251	Strawberry	AHVd	NA		MK188696	435	100.0	23.3	5	0.90	FP	ND					
MT252	Strawberry	AHVd	NA		MK947213	436	98.0	19.0	5	0.95	FP	ND					
MT265	Hydrangea	AHVd	NA		MK690630	374	96.4	26.2	6	0.97	FP	ND					
MT266*	Hydrangea	AHVd	NA		MK947213	436	97.9	27.5	6	0.94	FP	ND					
MT267*	Lagerstroemia	ND	NT									ND					
MT269*	Japanese	AHVd	NT		MH049330	433	97.7	19.7	5	0.90	FP	ND					
MT273	Lemon	AHVd	NA		MH049330	433	98.0	17.5	5	0.92	FP	ND					
MT276*	Clonal Grass	AHVd	NA		MG662375	419	97.1	17.7	6	0.86	FP	ND					
MT278*	Clonal Grass	AHVd	NA		MK188696	435	100.0	20.0	6	0.88	FP	ND					
MT217*	Apple	ARWV-2	NA		MN163134	1594	98.4	48.6	51	0.84		MN163134	1594	98.3	60.7	574	0.88
MT217*	Apple	ASGV	NA		MK929792	6475	99.3	34.2	143	0.93		MK929792	6475	98.9	62.9	559	0.95
MT217*	Apple	ASPV	NA									MZ148059	9279	97.1	2.8	22	0.49
MT217*	Apple	Cherry symptomless virus	NA									MK770441	7460	94.9	2.2	17	0.19
MT217*	Apple	CCGaV RNA1	NA		MH038042	6675	99.2	22.1	76	0.86		MK940541	6647	99.3	37.8	318	0.89
MT217*	Apple	CCGaV RNA2	NA		MK940543	2700	99.0	115.0	317	0.94		MK940543	2700	98.9	136.5	3160	0.97
MT218*	Apple	ACLSV	NA		MK929796	7521	100.0	10.3	43	0.74		MK929796	7521	99.4	10.8	80	0.81
MT218*	Apple	AGCaV	NA		LC476839	5573	90.0	6.2	18	0.65		LC476839	5573	91.5	9.1	67	0.72
MT218*	Apple	AHVd	NA		MK188698	433	98.0	174.7	1896	0.97		MN786530	434	100.0	247.9	199256	1
MT218*	Apple	ApMV RNA1	NA		KY965060	3328	97.9	5.7	10	0.65		KY965059	3477	98.7	10.8	80	0.78
MT218*	Apple	ApMV RNA3	NA		KY971019	1991	98.9	53.3	77	0.94		KY971019	1991	99.0	80.0	680	0.96
MT218*	Apple	ARWV-1	NA		MF062136	7219	98.1	34.3	163	0.85		MK936227	7171	96.9	30.5	250	0.82
MT218*	Apple	ASGV	NA		MK929792	6475	98.4	34.1	164	0.92		MK929792	6475	99.0	50.2	402	0.94
MT218*	Apple	ASPV	NA		KJ522472	9262	97.7	8.3	41	0.69		MZ147999	9279	90.0	16.3	124	0.77
MT219*	Apple	ACLSV	NA		AB060963	1724	100.0	32.3	36	0.77		MZ126491	7542	97.7	24.3	227	0.85
MT219*	Apple	AHVd	NA		MK947213	436	99.4	238.6	4909	1.00		MK188703	433	100.0	221.5	107052	0.94
MT219*	Apple	ARWV-1	NA		MF062137	1613	99.0	170.2	1125	0.95		MK936227	7171	97.2	43.3	310	0.84
MT219*	Apple	ASGV	NA		LT160740	6458	99.3	44.8	245	0.87		MK923757	6470	99.3	76.9	602	0.96
MT219*	Apple	ASPV	NA		JX673835	1157	95.7	22.3	13	0.92		MZ148044	9298	97.3	24.0	158	0.91
MT220*	Apple	AHVd	NA		KR605506	434	97.0	210.7	774	0.94		MZ344573	432	95.9	247.5	30515	1
MT221*	Apple	ACLSV	NA		KX579123	7544	97.2	10.9	50	0.75		MK929793	7505	97.2	21.5	179	0.89
MT221*	Apple	AHVd	NA		MK188687	434	100.0	227.5	1113	1.00		MK188698	433	99.5	246.3	89471	1
MT221*	Apple	ASGV	NA		LT160740	6458	96.1	32.3	170	0.89		MN786531	6487	97.9	86.2	848	0.96
MT221*	Apple	ASPV	NA		KY798310	9291	88.5	11.1	52	0.68		MZ148079	9291	97.4	38.4	326	0.96
MT244	Cactus	AHVd	NA		MK188696	435	100.0	23.9	5	0.86	FP						
MT244	Cactus	CMMoV	NA		EU043335	6449	88.9	39.6	1197	0.37		EU043335	6449	93.2	35.4	2094	0.3
MT244	Cactus	CVX	Y		KX883791	6610	97.7	219.1	2844	1.00		KX883791	6610	97.3	185.3	8974	1
MT244	Cactus	Mytcor virus 1	NA		MG210801	5689	98.5	226.1	2690	1.00		MG210801	5689	93.0	0.2	2	0.04
MT244	Cactus	PiVX	NA		JF930327	6677	98.8	224.8	2929	1.00		JF930327	6677	98.6	163.8	6131	1
MT244	Cactus	Potexvirus ST5	NA		EU676008	663	97.7	174.7	114	1.00		EU676008	663	97.7	138.8	7811	0.98
MT244	Cactus	RCNaV	NA									KY348772	6459	71.6	16.5	1001	0.13
MT244	Cactus	SchVX	NA		KP090203	6615	95.1	184.3	1364	0.99		AY366207	6633	95.0	125.3	4258	0.99
MT244	Cactus	ZyVX	NA		AY366208	6624	97.2	143.7	1242	1.00		AY366208	6624	97.0	116.4	5326	0.99
MT245*	Raspberry	AHVd	NA		MK947213	436	97.0	20.7	5	0.97	FP						
MT245*	Raspberry	RYNV	NT	R	MN245240	7844	99.7	9.9	41	0.45		MN245240	7844	99.4	21.5	187	0.46

Figure 23.2: Total number of virus/viroid detected by both approaches.



Conclusions and recommendations

90 viruses and viroids were recovered by both approaches. However a quarter of the NEBNext detections (35 out of 123) were predicted to be false positives. In comparison, no false positive was detected in the QIAseq dataset (Figue 2). These findings were in agreement with our RING study pilot performed in 2021. We have thus decided to use the Qiagen QIAseq kit as our default sRNA library kit.

Appendix 24: HTS testing for viruses and viroids across diverse plant commodities undergoing quarantine testing at PEQ

Background

Out of the 195 samples sent by PEQ for sequencing, 179 were plant consignments and are discussed in this section. Raw reads for each sample were analysed using our qc and VirReport pipelines.

Major Findings

- 38 out of the 169 imported plants tested positive for the presence of viruses and viroids.
- Molecular testing is pending for these batches of samples, but 7 of the HTS detections are in agreement with molecular testing.
- 6 plants tested positive for 6 endemic viruses and viroids and 16 plants tested positive for 16 regulated viruses and viroids (**Table 24.1**).
- We failed to detect a Cherry Virus F (CVF) present at low titer in a Prunus sample (MT258) but this virus was detected by another bioinformatics tool, VirusDetect, which uses reference-guided assembly on top of de novo assembly to identify viruses. MT258 is a re-sample of plant MT094 and MT098 and CVF presence was molecularly validated in all extracts using Nepovirus B PCR followed by sanger sequencing. This suggests that the read abundance was too low in this sample to overlap so that our de novo assembler Velvet could successfully assemble them into a contig.
- An additional 14 plants tested positive for Rubus yellow net virus (RYNV) (MT306-MT317, MT320 and MT321). In our previous report, we already described how raspberries are known to harbour up to six endogenous RYNV integrated sites in their genome and thus most RYNV detection in Rubus are likely to be matches to these integrated sequences. By mapping our contigs to a local database incorporating endogenous RYNVs, which have been provided by Dr Thien Ho, we found that the contigs showed highest homology to these endogenous RYNVs.
- Four Apple hammerhead viroids were detected in clonal grasses but these detections were flagged as false positives originating from apple samples present in the same sequencing batch (**Appendix 21**, **Table 21.1A**). While 3 additional viruses were flagged as potential false positives (**Table 24.1**), we are awaiting for the matching molecular test data to establish what is the relationship of some of the samples that were sequenced. This might explain the difference in titer observed (e.g. different tissue tested, shoot tip grafting).
- Contigs matching to Nectarine marafivirus M were detected in 4 prunus samples. This virus is not currently tested at SGS and its status is to be confirmed. Additionally, short contigs matching to grapevine red globe virus were also detected at low depth in 3 Prunus persica var. nectarina samples. This virus is not known to infect prunus plants and so these detections will need to be validated by an independent method.
- Three clonal grasses harboured sequences that showed 85-95% homology to several badnaviruses of the family Caulimoviridae which were present at low titer (mean coverage < 10 and % bases covered at 1X < 40). Endogenous badnaviruses from this family occur in banana (Banana streak virus) and rice (Rice tungro bacilliform virus). Based on the low number of reads recovered that match to these viruses, we infer that our detections are likely to be also resulting from viral integrations present in the clonal grass genomes; and that it is unlikely that full ORFs are transcribed and subsequently translated, to cause disease.

Table 24.1. Top BLASTN results of assembled virus and viroid sequences, coverage statistics and implementation of a contamination flag to identify false positives. Confirmed false positives are represented in red. sacc = subject (reference) accession sequence, slen = subject (reference) length, coverage = percent of the reference genome covered by the assembled sequence(s), av-pident = average sequence identity percent of assembled sequence(s) to reference, CF = contamination flag, DPI = Detection confirmed via molecular testing, FP = false positive, Mean cov = mean coverage in bases to the genome/sequence of the best homology match, PCT_1X = fraction of bases that attained at least 1X sequence coverage, FPKM = Fragments Per Kilobase of transcript, per Million mapped reads.

Sample	Batch	Species	PEQ	Virus/Viroid	Type	sacc	slen	av-pident	Mean cov	FPKM	% bases 1X	CF
MT242	3 (A)	Vitis	NT	Grapevine virus B	Regulated	EF583906	7599	89.19	10.98	41	0.53	
MT242	3 (A)	Vitis	Y	Grapevine yellow speckle viroid 1	Endemic	MK804769	366	100	189.3	117	1	
MT242	3 (A)	Vitis	NT	Hop stunt hostuviroid	Regulated	MF497538	296	100	237.5	200	1	
MT246	3 (A,R)	Citrus	Y	Citrus tristeza virus	Regulated	AB046398	19302	99.58	86.5	654	0.98	
MT247	3 (A,R)	Citrus	Y	Citrus tristeza virus	Regulated	MK779711	19260	98.22	60.6	700	0.98	
MT247	3 (A,R)	Citrus	NT	Citrus vein enation virus	Regulated	MN187044	5982	99.39	151.0	1814	0.99	
MT249	3 (A,R)	Citrus	Y	Citrus tristeza virus	Regulated	AB046398	19302	99.26	112.3	659	0.98	
MT249	3 (A,R)	Citrus	NT	Hop stunt viroid	Regulated	MT917186	302	100	212.9	2701	1	
MT258	3 (A)	Prunus	Y	ND [missed CVF, detected with VirusDetect]	Regulated							
MT259	3 (A)	Clonal grass	NT	Apple hammerhead viroid	Not_regulated	MK188696	435	100	27.9	10	0.89	FP
MT260	3 (A)	Clonal grass	NT	Banana streak UL virus	Not_regulated	KF545135	1239	85.96	7.04	91	0.32	
MT260	3 (A)	Clonal grass	NT	Banana streak UM virus	Not_regulated	HQ593110	7532	81.34	1.3	17	0.07	
MT260	3 (A)	Clonal grass	NT	Banana_streak_VN_virus	Not_regulated	KF318350	575	83.33	8.66	111	0.2	
MT260	3 (A)	Clonal grass	NT	Dioscorea bacilliform virus	Not_regulated	KY555511	528	88.64	5.16	67	0.27	
MT260	3 (A)	Clonal grass	NT	Sugarcane bacilliform virus	Not_regulated	JN377535	7613	88.89	0.77	10	0.05	
MT261	3 (A)	Clonal grass	NT	Apple hammerhead viroid	Not_regulated	MH049330	433	97.92	15.66	3	0.85	FP
MT262	3 (A)	Clonal grass	NT	Apple hammerhead viroid	Not_regulated	MK690630	374	99.29	27.87	254	0.99	FP
MT262	3 (A)	Clonal grass	NT	Banana streak UM virus	Not_regulated	KF545127	1239	80.14	5.86	52	0.26	
MT262	3 (A)	Clonal grass	NT	Sugarcane bacilliform virus	Not_regulated	KM214357	7764	82.85	1.01	9	0.05	
MT263	3 (A)	Clonal grass	NT	Apple hammerhead viroid	Not_regulated	MT993976	435	99.32	23.57	399	0.88	FP
MT263	3 (A)	Clonal grass	NT	Banana streak UL virus	Not_regulated	HQ593109	7401	82.76	0.27	5	0.05	
MT263	3 (A)	Clonal grass	NT	Banana streak UM virus	Not_regulated	KF545119	1239	83.33	1.6	27	0.17	
MT263	3 (A)	Clonal grass	NT	Cacao red vein banding virus	Not_regulated	MH785300	6846	95.92	0.48	8	0.02	
MT279	3 (A)	Rose	Y	Prunus necrotic ringspot virus RNA1	Endemic	AF278534	3332	98.99	83.19	181	0.99	
MT279	3 (A)	Rose	Y	Prunus necrotic ringspot virus RNA3	Endemic	LC382468	1952	98.5	188.8	674	0.99	
MT306	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	92.17	76.26	1248	0.75	
MT307	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	91.57	82.01	1208	0.75	
MT308	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	97.72	62.59	301	0.72	
MT309	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	98.2	78.64	764	0.74	
MT310	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	87.76	78.48	920	0.54	
MT311	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	89.74	74.45	985	0.51	
MT312	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	96.61	135.6	1515	0.87	
MT313	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	96.07	124.5	1543	0.86	
MT314	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	95.4	128.3	2314	0.88	
MT315	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	96.93	135.5	1648	0.86	
MT316	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	89.64	26.44	216	0.39	
MT317	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	87.08	25.2	161	0.36	
MT320	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	97.65	42.31	267	0.61	
MT321	5	Rubus	NT	Rubus yellow net virus	Regulated	MN245240	7844	99.51	0.03	0	0.01	FP
MT331	5	Prunus	NT	Grapevine red globe virus	Regulated	MZ451069	6850	84.02	1.9	4	0.21	
MT332	5	Prunus	NT	Grapevine red globe virus	Regulated	KX109927	6863	93.1	2.25	3	0.24	
MT338	5	Prunus	NT	Prunus geminivirus A	Regulated	MF504038	3174	99.86	110.9	221	0.98	
MT341	5	Prunus	NT	Prunus geminivirus A	Regulated	MF504038	3174	99.93	50.75	63	0.92	
MT349	5	Prunus	NT	Grapevine red globe virus	Regulated	KX109927	6863	92.3	2.81	3	0.23	
MT350	5	Prunus	NT	Cherry virus A	Endemic	KY510908	7431	99.5	132.8	495	0.99	
MT350	5	Prunus	NT	Prune dwarf virus RNA1	Regulated	MZ220991	3345	98.73	147.5	409	1	
MT350	5	Prunus	NT	Prune dwarf virus RNA2	Regulated	MZ220875	2518	99.55	102.5	315	0.99	
MT350	5	Prunus	NT	Prune dwarf virus RNA3	Regulated	MZ221014	2039	99.17	231.4	2955	1	
MT357	5	Prunus	NT	Nectarine marafivirus M	To be confirmed	KT273412	6471	98.57	71.16	309	0.99	
MT365	5	Prunus	NT	Nectarine marafivirus M	To be confirmed	KT273411	6421	99.12	63.48	492	1	
MT368	5	Prunus	NT	Nectarine marafivirus M	To be confirmed	KT273412	6471	97.83	95.7	727	0.99	
MT378	5	Prunus	NT	Nectarine marafivirus M	To be confirmed	KT273412	6471	98.43	0.23	1	0.04	FP
MT383	6	Prunus	NT	Cherry virus A	Endemic	KY510909	7418	98.93	152.8	707	1	
MT383	6	Prunus	NT	Prune dwarf virus RNA1	Regulated	MZ291947	3350	99.55	171.8	803	1	
MT383	6	Prunus	NT	Prune dwarf virus RNA2	Regulated	MZ220875	2518	99.35	128.9	657	1	
MT383	6	Prunus	NT	Prune dwarf virus RNA3	Regulated	MZ221007	2081	99.17	236.5	5426	1	
MT385	6	Prunus	NT	Cherry virus A	Endemic	KY510909	7418	94.09	0.03	0	0.01	FP
MT385	6	Prunus	NT	Prune dwarf virus RNA1	Regulated	MT013233	3376	98.73	0	0	0	FP
MT385	6	Prunus	NT	Prune dwarf virus RNA2	Regulated	KY883329	2556	96.87	0.02	0	0.02	FP
MT385	6	Prunus	NT	Prune dwarf virus RNA3	Regulated	KY883331	2051	98.38	0	0	0	FP
MT411	6	Vitis	NT	Grapevine yellow speckle viroid 1	Endemic	KX966259	366	99.43	0	0	0	
MT411	6	Vitis	NT	Hop stunt viroid	Regulated	HM357802	297	100	0.07	0	0.07	

Appendix 25: Draft policy documents for operational adoption of HTS technology for routine plant virus/viroid diagnosis

Australian Government
Department of Agriculture and Water Resources

WORK INSTRUCTION

Using high-throughput sequencing of small RNA to detect viruses and viroids in plants undergoing testing at the Mickleham post entry quarantine facility

Direction to staff

This document is instructional material for the Department of Agriculture and Water Resources (the department) under its Practice Statement Framework. All staff must comply with it.

Summary of main points

This document outlines the procedures that plant health diagnosticians at the Post-Entry Quarantine (PEQ) Facility need to effectively and safely conduct high-throughput sequencing (HTS) testing on imported plant material to detect and identify regulated viruses and viroids.

In this document

This document contains the following topics.

Purpose of this document 2
Definitions 2
Legislative framework 2
Roles and responsibilities 3
Work health and safety 3
Essential equipment 3
HTS testing 4
Sampling and RNA extraction 4
Batching samples for third-party sequencing 4
Third party cDNA library preparation and sequencing 5
Processing of sequencing results 6
The following table outlines how to analyse sequencing data when it returns from the sequencing provider 6
Confirmation of HTS result with secondary test & reporting 6
Record keeping 7
Related material 7
Document information 7
Version history 7

Purpose of this document

The purpose of this document is to outline the steps that plant health diagnosticians undertake to safely and effectively complete high-throughput sequencing (HTS) testing on imported plant material at the Mickleham Post-Entry Quarantine Facility using small RNA-Seq protocol.

Definitions

The following table defines the terms used in this document.

Term	Definition
High-throughput sequencing (HTS)	A technique that enables the sequencing of nucleic acids present in an organism in a rapid and cost effective manner It is used in Post-Entry Quarantine as a diagnostic test to detect all viruses and viroids present in an imported plant with a single test.
Laboratory Information Managements System (LIMS)	A software package for managing and recording laboratory data, equipment, methods and results.
Post-Entry BiosecuritySystem (PEBS)	An import system for managing bookings of plants and animals needing post entry quarantine at the Mickleham Facility., The system facilitates bookings, testing status and invoicing of consignments of imported plant material under quarantine at PEQ.
Post-Entry Quarantine (PEQ)	Where high-risk imported animals and plants are held for a specified period in a quarantined environment, subject to satisfactory completion of testing for unwanted organisms before release into Australia.
sRNA-Seq	small RNA sequencing using HTS. Profiles the genetic information of short molecular size fragments present in a test sample.
GenBank	GenBank ® is the genetic sequence database of the American National Institute of Health – an annotated collection of all publicly available DNA sequences.
BLAST	Basic Local Alignment Search Tool (BLAST) is an algorithm for comparing primary biological sequence information.
VirReport/ Bioinformatics workflow	A bioinformatics pipeline for the diagnosis of plant viruses and viroids using sRNA-Seq data.
Plant material	Any type of tissue collected from a plant.

Legislative framework

The following list outlines the legislation that applies to testing of imported biological material by high-throughput sequencing:

Biosecurity Act 2015:

- Section 12 – Inspecting goods and taking samples.
- Section 554 – Carrying out tests on samples.

Roles and responsibilities

The following table outlines the roles and responsibilities undertaken in this work instruction.

Role	Responsibility
Assistant Director, Operational Science & Surveillance, Mickleham	• Ensure all plant health diagnosticians undertaking these tests are trained to follow this work instruction before conducting HTS testing. • Provide advice and direction to officers when required. • Verify procedures as per business assurance framework.
Plant health diagnosticians, Operational Science & Surveillance, Mickleham	• Follow this work instruction when conducting HTS tests of plants at the post-entry quarantine facility. • Take all reasonably practicable steps to ensure they do not take any actions that create a risk (in relation to both Biosecurity and WHS), or increase an existing risk, to their own health or safety, or that of other people.

Work health and safety

- Some of the procedures included in Performing HTS testing pose a work health and safety risks due to:excessive noise generated by the use of a tissue lyser
- the use of sharp blades during sub-sampling of plant tissue with razor blade
- risk of chemical exposure during RNA extraction
- the use of ultra low temperature freezers during RNA extraction and storage

Important: All authorised operators must:

- be trained in the safe operation of each piece of equipment in this procedure and be deemed competent before following the work instruction without supervision.
- wear appropriate personal protective equipment (PPE) when working in the laboratory
  - lab coat and gloves at all times
  - safety glasses when there is a risk of chemical splash to eyes
  - ear protection and eye protection during use of Tissue Lyser
  - Cryogloves when handling frozen tubes or when opening Ultra low temperature freezer
- wash hands after handling goods

Essential equipment

The following list of equipment is required to complete HTS tests:

- tissue lyser
- RNA extraction kit
- RNA quantification device/s centrifuges
- heat block
- pipettes

- aerosol-resistant pipette tips
- ultra low temperature freezer (-80°C)
- balance
- excision blades
- lab book

HTS testing

Sampling and RNA extraction

The following table outlines the procedure for sampling and RNA extraction.

Step	Action
1.	Collect adequate tissue (e.g., petiole, leaves, budwood) from the plant being tested.
2.	Perform RNA extraction in 9.141 PCR Lab 1 (Prep).
3.	Perform appropriate quality checking of RNA extraction using either Nanodrop and/or TapeStation.
4.	Enter QC values for each sample into LIMS/extraction worksheet
5.	Store RNA extracts at -80°C and log location in LIMS/extraction worksheet

If sample…	Then…
passes QC check controls for RNA extraction	**go to step 4.**
fails QC check controls for RNA extraction	**do not proceed.**

Batching samples for third-party sequencing

The following table outlines the procedure for preparing a batch of samples for shipment to a third-party sequencing provider for cDNA library preparation and sequencing.

Step	Action
1.	Obtain quote from third-party sequencing provider.
2.	Organise a work order under Standing Offer Number 3470177 (Deed of Standing Offer for Genome Services between DAWE and the service provider); PGPA approval and a purchase order.
4.	Organise courier shipment of nucleic acid samples on dry ice to the sequencing provider.
5.	Prepare samples for shipment in format required by sequencing service provider
6.	Enter sample metadata into the submission portal of the sequencing provider and obtain a submission ID.
7.	Record Project name, submission ID, quote number, PO number, date of submission, importer contact details for invoicing into the appropriate database (Master Plant Testing Workbook and/or LIMS).

Third party cDNA library preparation and sequencing

The following table outlines how to review the QC report of the sequencing provider prior to their proceeding with cDNA synthesis and sequencing using unique dual indexes.

<table>
<tr><th>Step</th><th>Action</th></tr>
<tr><td>1.</td><td>Ship sample to accredited HTS third party provider as per instructions above.</td></tr>
<tr><td>2.</td><td>Review QC report summary for library preparation QC
<table>
<tr><th>If...</th><th>Then…</th></tr>
<tr><td>sample passes QC check</td><td>confirm to sequencing provider that they can proceed to cDNA library preparation step.</td></tr>
<tr><td>sample fails QC check</td><td>do not proceed.</td></tr>
</table></td></tr>
<tr><td>3.</td><td>Sequence samples (indexed with unique dual identifiers)</td></tr>
<tr><td>4.</td><td>Assess quality of reads</td></tr>
<tr><td>5.</td><td>Perform assembly and annotation</td></tr>
<tr><td>6.</td><td>Interpret results. Perform homology search of sequences obtained against publically available sequences.
<table>
<tr><th>If...</th><th>Then…</th></tr>
<tr><td>pathogen detection meets minimum confidence checkpoints</td><td>• direct PEQ Plant to isolate the plant pending confirmation of HTS result
• proceed to confirmation of HTS result with secondary test (PCR and/or amplicon sequencing).</td></tr>
<tr><td>pathogen detection does not meet minimum confidence checkpoints</td><td>• proceed to confirmation of HTS result with secondary test (PCR and/or amplicon sequencing)</td></tr>
</table></td></tr>
</table>

Processing of sequencing results

The following table outlines how to analyse sequencing data when it returns from the sequencing provider.

Step	Action
1.	• Assess quality of reads
2.	• Perform assembly and annotation
3.	• Interpret results. Perform homology search of sequences obtained against publically available sequences.

If	Then
Pathogen detection meets minimum confidence checkpoints	• Direct PEQ Plant team to isolate the plant pending confirmation of sequencing result with secondary test • Inform PIO of preliminary detection • Proceed to confirmatory testing
Pathogen detection does not meet minimum confidence checkpoints	• Detection is declared a false positive

Confirmation of HTS result with secondary test & reporting

The following table outlines the procedure for confirmation and reporting of results.

Step	Action
1.	Conduct confirmatory secondary test

If	Then
Secondary test is positive	• HTS result is entered into Master Plant Testing Workbook/LIMS as “of Biosecurity Concern” • Inform PEQ – Plant team • Enter detection into PEBS
Secondary test is negative	• Detection is declared a false positive and plant can continue PEQ

Record keeping

The following table outlines the record keeping requirements of this work instruction.

Record type	Requirements
HTS records	Entered into the laboratory record book and/or LIMS, Master Plant Spreadsheet and PEBS by the plant healthdiagnostician performing the test. Testing records must include: • name of sample (including species/variety or cultivar where relevant), or specific identifying code, from test material/organism • date the test was conducted • results of the HTS test. • Result of the secondary test

Record type	Requirements
All records	The laboratory manager is responsible for appropriately classifying, storing and handling records in accordance with the department's information and records management policies. This includes: • storage in accordance with the Departmental Policy: *Recordkeeping*, which covers the creation, storage and disposal of all departmental records • retention for seven years and disposal consistent with the requirements under the *Archives Act 1993* and departmental disposal guidelines.

Related material

The following related material is available on the Instructional Material Library (IML):

- Departmental Policy: *Recordkeeping*
- Reference: *Post-entry quarantine pathogen testing manual*
- *Add additional relevant guidelines*

Document information

The following table contains administrative metadata.

Instructional Material Library document ID	Instructional material owner
IMLS-XX-XXX	Director, Operational Science & Surveillance Group.

Version history

The following table details the published date and amendment details for this document.

Version	Date	Amendment details
1.0	21/11/2022	New instruction on performing HTS testing on plant material.

Standard Operating Procedure for sampling plants for High Throughput Sequencing (HTS_SOP1_20220729)

1. Purpose

This SOP provides the procedure for sampling plants at PEQ for High Throughput Sequencing

2. Scope

The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of the Science and Surveillance Group during the sampling of any plant commodity at PEQ to be used for High Throughput Sequencing and subsequent validation through molecular and/or conventional assays. This sampling plan is designed to meet Australian Standards for General requirements for the competence of testing and calibration laboratories AS ISO/IEC 17025:2018 section 7.3 Sampling.

3. Definitions and abbreviations

HTS – High Throughput Sequencing

SSG – Science and Surveillance Group

PEQ – Post Entry Quarantine facility

Competence—staff attain competence through demonstrating the successful completion of a procedure several times in the presence of a qualified member of staff.

Index (mother) plant – Plant established from the original imported budwood, cutting or tissue-culture plantlet. Also known as the 'mother plant' selected as the test plant.

Sessile leaves – Leaves that grow directly from the plant stem (i.e. there is no petiole)

4. Personnel pre-qualifications

All personnel sampling should be inducted to and have access to the glasshouses. All staff members performing sampling are required to be trained and supervised by a competent SSG molecular biologist/ plant pathologist and assessed for competence prior to conducting the procedure independently.

5. Equipment and materials

Gloves, marker pen, sterile RNase-free surgical scalpel blades (to be used only when necessary), small sharps container, resealable plastic bags, Petri dishes (if collecting stem samples), ice box and pre-cooled ice pack(s)

6. Procedure

Step 1- Preparations before tissue sampling

- Print list of plants ready for sampling. As per the document "**Testing and screening of high-risk plant material in post-entry quarantine"** plants are to be sampled "during active growth".
    - A plant that is "actively growing" and ready for testing within the context of Post Entry Quarantine at Mickleham:
        - Is grown in a pot, under greenhouse-controlled environment conditions at approx. 22ºC, 40-60% relative humidity (RH) with natural or supplemental lighting
        - Is not a tissue culture plant, germinating seed, dormant scion material, dormant bare rooted material or bare rooted material devoid of foliage, however, depending on cultivar, may be physiologically immature
        - Is established, well-rooted, with good foliage development and is of a suitable size such that any sampling will not hinder plant development or cause death
            - Herbaceous plants (e.g. potato, sweet potato, hops, grasses) should be established for approximately 3 months whilst woody plants should be established for approximately 6 months, ideally tested post dormancy (likely 8-10+ months).
        - Has mature as well as developing and expanding foliage, developing bark (depending on genus), is increasing in stem diameter, and is not senescing due to seasonal changes
        - Does not present as nutritionally deficient, pest infested, or water stressed
    - Active growth may be "forced" in green houses. For instance, for late plant arrivals, supplemental electric lighting may be used to accelerate growth (which particularly occurs with Prunus). However, seasonal conditions generally dictate periods of active growth
- Identify and document location of plants in the glasshouse so that sampling can be completed in a timely fashion
- Label the resealable bags and/or petri dishes with the collector's name, index numbers and variety/cultivar name, and date of sampling before attending the glasshouse.
- Bring adequate gloves and sterile scalpel blades such that each plant can be sampled with fresh gloves and scalpel
- Bring marker pen, labelled bags, and the list of plants to be collected on the day along with the ice box.
- Whilst in the glasshouse premises follow the standard protocols before entering each room (see related documents section below).

Step 2- Sampling of plant tissue

- Identify plants to be sampled in each glasshouse room and place labelled bag next to each pot of interest
- Before sampling, verify that sample bag, plant tag and pot and your sampling list have the same information. If there are any discrepancies, do not sample and consult with Senior Plant Pathologist and/or PEQ horticultural staff to reconcile these differences.
- Don a fresh pair of gloves before handling each plant.
- Remove three leaves preferably with gentle downward pressure with gloved fingers (including both blade and petiole) from different zones and sides of the plant (top, bottom, and middle) and place directly into the labelled sample bag and store in cooled ice box
    - Exceptions
        - Sessile leaves – pluck leaf blade.
        - Monocots and other plants with fewer than 10 leaves – pluck 2 leaves from different areas.
        - For plants with exceptionally small or thin leaves – sample adequately to make sure you have enough material (>50mg) for nucleic acid extractions.
    - Use a scalpel blade only where necessary and dispose of blades in between plants (refer to section 7 below).
- Dispose of the used gloves properly (see section 7 below).
- Use a fresh pair of gloves each time you sample a new plant.

- Once you finish sampling, take the ice box to the extraction room (PCR1 prep – 9.141). and store the labelled resealable bags with samples at 4ºC.
- Proceed to tissue subsampling and extraction of nucleic acids as soon as possible to prevent degradation of RNA.

7. **Waste management and other safety precautions**
   - Used gloves should be disposed of in the waste bin in each room before exiting a glasshouse room.
   - Scalpel blades should be discarded into sharps containers.

8. **Related documents**
   - Testing and screening of high-risk plant material in post-entry quarantine V15 document.
   -
9. **Rationale for change of SOP version**

   N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					

Laboratory area:	No of copies:	Reason for change:

Australian Government Department of Agriculture, Fisheries and Forestry	**Standard Operating Procedure- RNA extraction of Plant Material for High Throughput Sequencing**	
	Document ID	HTS_SOP2
	Revision	In draft
	Page	< # > of 6
Document owner:	Science and Surveillance Group	
Document approver:	Candace Elliott, Senior Scientist	
Purpose:	To define the procedure for extracting total RNA from plant samples at PEQ for High Throughput Sequencing.	
Scope:	The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of the Science and Surveillance Group at PEQ for the extraction of total RNA from plants sampled as per HTS_SOP1 for the purpose of High Throughput Sequencing and subsequent validation through molecular and/or conventional assays.	

This is an uncontrolled copy when printed

1. Definitions and abbreviations

HTS – High Throughput Sequencing

SSG – Science and Surveillance Group

PEQ – Post Entry Quarantine facility

RNA – Ribonucleic acid

Competence – Staff attain competence through demonstrating the successful completion of a procedure several times in the presence of a qualified member of staff.

PVP – Polyvinylpyrrolidone is a water-soluble polymer that forms hydrogen bonds with polyphenols and negatively charged polysaccharides thus removing them from the RNA preparation and preventing inhibition of downstream enzymatic processes.

DTT – Dithiothreitol is a reducing agent that denatures disulphide bonds in proteins, thus facilitating their removal from the nucleic acid solution.

DNase I – Deoxyribonuclease I is an endonuclease that degrades double-stranded or single-stranded DNA.

2. Personnel pre-qualifications

All staff performing RNA extractions must be inducted by qualified member of staff before working in the Extraction room (PCR1 prep – 9.141). All staff performing RNA extractions are required to be trained and

supervised by a competent SSG molecular biologist and assessed for competency prior to conducting the procedure independently.

3. Occupational Health and Safety

- **Chemicals and The Maxwell Kit –** The Maxwell kit contains a variety of reagents. Risk assessments are contained in the general laboratory and in Chemwatch. Be sure to familiarise yourself with the risks each pose, the PPE required, the method of disposal and the method of clean up in case of spills. Use the spill kit if required to contain any large spills.
- **Temperature –** Temperatures utilised in this method can burn the skin. When handling cold frozen goods, use freezer gloves.
- **Mechanical hazards –** Moving parts in instruments (Maxwell instrument, centrifuge, TissueLyser, and vortex) can be a mechanical hazard. Training is essential prior to using any specialised instrument. Never place yourself in harm's way, be careful of moving parts.
- **Electrical hazards –** Electrical appliances (fridges and freezers, Maxwell instrument, centrifuge, TissueLyser, vortex and balance) can cause electrical hazards. Report any defects in instrumentation and do not use instrumentation with obvious defects. Keep liquids away from power cables and electrical appliances. Spill kits should be readily available near electrical appliances.
- **Environmental hazards –** Live specimens can pose an environmental hazard. Specimen and specimen by-products need to be disposed of as biosecurity waste. Decontaminate workspace after each use.
- **Sharps –** All sharps (tips, scalpel blades and syringe needles) should be handled with care and discarded into sharps containers.

4. Equipment and materials

Plant material sampled according to HTS_SOP1, weighing boats, gloves, surgical scalpel blades, porcelain tile, electronic balance, 2 ml microcentrifuge tubes with safe-lock lid, permanent marker pen, stainless steel beads (7 mm), Qiagen TissueLyser LT, Promega Maxwell RSC Simply RNA Tissue Kit (AS1430), Maxwell RSC instrument, fridge and freezer, fume hood, micropipettes, microcentrifuge, vortex, Kimwipes, 80% ethanol, nuclease free water, PVP, DTT, RNase AWAY™, DNA AWAY™

5. Procedure

5.1. Extraction worksheet preparation

- Prepare extraction worksheet (see related documents) if not already prepared during plant sampling and bring printed copy to Extraction room (PCR1 prep – 9.141).
- Post extraction, keep a hardcopy of the extraction sheet in the extraction room lab book.

5.2. Extraction set-up

- Decontaminate workspace as per **Decontamination SOP**.
- Place **2 mL** microcentrifuge tubes (one per sample) in a tube rack.
- Label the lids and sides of the tubes with the extraction number using a permanent marker pen.
- Aseptically insert **two 7 mm stainless steel beads** into the tubes using the TissueLyser Bead Dispenser.
- Place microcentrifuge tubes, two empty racks and the TissueLyser rotor in the -80ºC freezer to chill prior to subsampling.

5.3. Subsampling

- Decontaminate workspace as per **Decontamination SOP**.
- Turn on the balance and make sure it is level adjusting as required by referring to the levelling bubble.
- Cover a white porcelain tile with a new paper towel to use as a cutting board.
- Remove a weigh boat/petri dish aseptically from packaging, place on balance and press 'Tare'
- Dissect plant tissue from a minimum of three leaves using a sterile scalpel blade to weigh a maximum of **50 mg** per sample.

- Refer to commodity specific instructions in SOP: SubsamplingXX
- Cut the weighed tissue into smaller pieces that can fit into the tube and transfer aseptically to the pre-chilled tubes.
- Discard scalpel blade into sharps container, then gloves, weighing boat, and the paper towel covering the tile into the Biosecurity waste bin between samples.
- Subsample no more than three samples before transferring them to the -80ºC.
- Decontaminate workspace in between samples if necessary.
- Hold samples at **-80ºC** for a minimum of **30 min** before proceeding.

5.4. Reagent and instrument preparation

- Decontaminate workspace as per **Decontamination SOP**.
- Pre-chill the centrifuge to 4ºC (press the "fast temperature" button to bring down the temperature of the centrifuge to to 4ºC).

5.4.1. Prepare homogenisation solution and label with date of preparation

- Add **2.5%** PVP (0.75 g dry PVP) to **30 mL Promega Maxwell RSC Simply RNA Tissue Kit** Homogenisation solution bottle.
- Mix vigorously.
- Add **600 mL** of 1-Thioglycerol to the **Homogenisation Solution** and mix.
- Discard after 1 month.

5.4.2. Prepare fresh DNase I and label with date of preparation

- Add **275 µL** nuclease free water to **Promega Maxwell RSC Simply RNA Tissue Kit** DNase I. Mix by inverting and turning tube to dissolve, do not vortex.
- Add **5 µL** of the **Promega Maxwell RSC Simply RNA Tissue Kit** blue dye.
- Store at -80ºC and discard after 10 freeze-thaw cycles.

5.4.3. Prepare Maxwell instrument

- Turn on the Maxwell instrument and turn on the tablet.
- Open the Maxwell software.
- Sanitise the instrument for 1 min.
- Press start.
- Choose the extraction program (**AS1340** for RNA extraction).
- Select the cartridge position.
- Take out the tray.
- Load the tray with the number of Maxwell cartridges required, one per sample, press down until the cartridges click in.
- Remove the sealing foil from each of the cartridges aseptically.
- Aseptically insert the **pipette plungers** into the **front well** of each cartridge (the well closest to the tray holes).
- Aseptically place **elution tubes** into the holes in the **tray** for each cartridge. The lids must be open and opening away from the cartridges.
- Aliquot **100 µL** of **nuclease free H_2O** from the Promega Maxwell RSC Simply RNA Tissue Kit into each elution tube.
- Add to **5 µL** of **DNase I** to the **well #4** in each cartridge (yellow reagent).

5.5. Sample processing

5.5.1. Homogenisation

- Remove samples and TissueLyser rotor from -80ºC keeping samples in pre-chilled racks throughout RNA extraction
- Add **200 µL** of chilled **Homogenisation Solution** into the microcentrifuge tubes.

- Place the tubes into the rotor and fasten the rotor in place on the **TissueLyser**.
- Operate the TissueLyser for **1 min at 50 Hz**.
- Spin tubes very briefly in chilled microcentrifuge using Pulse button.
- Add **400 µL** of **Homogenisation Solution** into each tube and operate TissueLyser for **1 min at 50 Hz**.
- Spin tubes very briefly in chilled microcentrifuge using Pulse button.

5.5.2. Lysis

- Add **300 µL** of **Lysis Buffer** into each tube.
- Vortex for 10-15 seconds then spin tubes for a minimum of 5 min at 20,000 G.

5.5.3. Run Maxwell Instrument

- Add **4 µL** of spike-in **Norgen miR39 (3.3 fmol/µl)** to the **back well** (the big well) of each cartridge prepared in **4.4.3.**
- Add **500 µL** of **supernatant of your sample preparation** to the **back well** (the big well) of the corresponding cartridge.
- Place the tray back onto the Maxwell. Orient the tray with the elution tubes at the front.
- Do a visual scan of your tray to confirm:
  - the elution tubes are open and contain H2O,
  - your sample is aliquoted in the back well,
  - DNase I is added into the 4th well,
  - plungers are placed in the front well
  - Start the extraction.
  - Promptly proceed with step **5.6** once the run is complete.

5.6. Post-RNA extraction procedure

- Label two breakable 96 well plates with extraction number, commodity, RNA, date and initials of operator.
- Mark first and last sample wells to specify the orientation of the plate.
- Once extraction is complete, remove the tray from the Maxwell.
- Transfer the elutes from the elution tubes to the prepared plates, with 25 µL aliquoted into one plate for TapeStation and HTS submission (Plate 1) and the remainder into the second plate for long term freezer storage (Plate 2).
- Store Plate 1 in the fridge if conducting TapeStation quality checks on the same day as extraction or in the -80ºC freezer until TapeStation analysis will be done.
- Move Plate 2 into the Template room and store in the designated place in -80ºC freezer.

5.7. Final sanitation

- Gently unclick each Maxwell cartridge out of the tray and dispose of the cartridges and the elution tubes.
- Wipe the tray and rods with 80% ethanol using a Kimwipe.
- Place the tray back into the Maxwell instrument.
- Sanitise the instrument for 1 min.
- Close the Maxwell software.
- Turn off the Maxwell instrument.
- Turn off the tablet from the desktop.
- Decontaminate all workspaces including TissueLyser rotor and Centrifuge using procedures described in Decontamination SOP.

6. Reference documents

- Refer to most current version of the following technical manuals

- Qiagen TissueLyser LT – https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/instruments-equipment/TissueLyser-lt/

- Promega Maxwell RSC Simply RNA Tissue Kit - https://www.promega.com.au/products/nucleic-acid-extraction/rna/maxwell-rsc-simplyrna-tissue-kit/?catNum=AS1340

7. Related documents

- HTS_SOP1: sampling
- HTS_SOP3: Tapestation
- Extraction worksheet
- Subsampling SOP
- Decontamination SOP

1. Rationale for change of SOP version

N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Standard Operating Procedure for RNA quantification, quality assessment and submission for High Throughput Sequencing (HTS_SOP3_20221024)

1. **Purpose**

This SOP provides the procedures for quantification and qualification of total RNA extracted from plant samples at PEQ and for submitting RNA to external sequencing provider Ramaciotti Centre for Genomics for High Throughput Sequencing (HTS).

2. **Scope**

The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of the Science and Surveillance Group at PEQ intending to quantify, qualify and submit total RNA from a previously extracted plant following the HTS_SOP2 for the purpose of High Throughput Sequencing and subsequent validation through molecular and/or conventional assays.

3. **Definitions and abbreviations**

HTS – High Throughput Sequencing

SSG – Science and Surveillance Group

PEQ – Post Entry Quarantine facility

RNA – Ribonucleic acid

RIN - RNA Integrity Number

4. **Personnel pre-qualifications**

All personnel performing RNA extractions should perform adequate induction training to gain access to the Extraction room (PCR1 prep – 9.141). New staff members are required to be trained by and supervised by a trained SSG molecular biologist/plant pathologist prior to attempting the procedure on their own.

5. **Equipment and materials**

RNA extracted following HTS_SOP2, Agilent 4200 TapeStation system, microcentrifuge, Agilent RNA ScreenTape, Agilent RNA ScreenTape sample buffer, Agilent loading tips, Vortex mixer with 96-well sample plate adapter, 96-well sample plates, 96-well Plate Foil Seal, 96-well Plate transparent seal, micropipettes, filtered micropipette tips, ice bucket, thermal cycler, permanent marker, cryo-dots/cryo-pen, Kimwipes.

6. **Procedure**

Preparations before RNA quantification and qualification

- Clean your work bench according to the decontamination SOP.
- Place the tubes/plates with extracted RNA to be quantified in an ice box in the extraction room. If the RNA is frozen, thaw on ice before quantification.
- Spin down all samples using the pulse/short button on the microcentrifuge.
- Store RNA samples on ice throughout the procedure.

Quantification and Qualification using the TapeStation device

Step 1- Prepare TapeStation reagents and equipment

- Retrieve the Agilent sample buffer and the Agilent ScreenTape device from 2–8°C and allow 30 min to equilibrate to room temperature.
- Turn on the heat block and set the temperature to 72°C.
- Launch the 2200 TapeStation Controller Software.
- If using a new ScreenTape device:
  - 'Flick' ScreenTape device twice to eliminate bubbles in the separation channel, which could interfere with sample loading.
  - Write the date the packet was opened on the cover. Screen tape devices expire after a month from opening.
- Open the lid of the instrument and load the RNA ScreenTape device so that the barcode in the screen tape is facing away from you and the green bar is facing toward you, to your right.
- TapeStation device should automatically read the barcode and load the protocol 'RNA Protocol (Eukaryotic RNA or Prokaryotic RNA)' in the TapeStation controller software.
- Load tips into the 2400 TapeStation instrument.

Step 2 - Prepare sample for TapeStation analysis

- Prepare each sample by mixing 5 µL of RNA Sample Buffer with 1 µL of RNA in Agilent strip tubes or 96-well plates. If strip tubes are used, seal with strip caps and if 96-well plates are used, seal with foil.
- Spin down, then vortex using IKA vortexer and adaptor at 2000 rpm for 1 min.
- Spin down to bring down all the liquid to the bottom of the tube.
- Heat the samples at 72°C for 3 min in the heat block.
- Place samples on ice for 2 min.
- Spin down to bring down all the liquid to the bottom of the tube.

Step 3 - Sample analysis on TapeStation

- Load samples into the 2400 TapeStation instrument.
- Right click and select e-ladder in the 2200 TapeStation Controller Software.
- Select the required samples on the 2200 TapeStation Controller Software and label the wells in the e-gel from well A1 onward.
- Click 'Start' and specify a filename with which to save the analysis results.

Step 4 - Interpretation of TapeStation data

- Scale the e-gel image 'to sample' (Figure 1).
- Take note of the RIN^e score and the assigned concentration for each of your sample.
- Export the data to a folder renamed with the extraction ID.

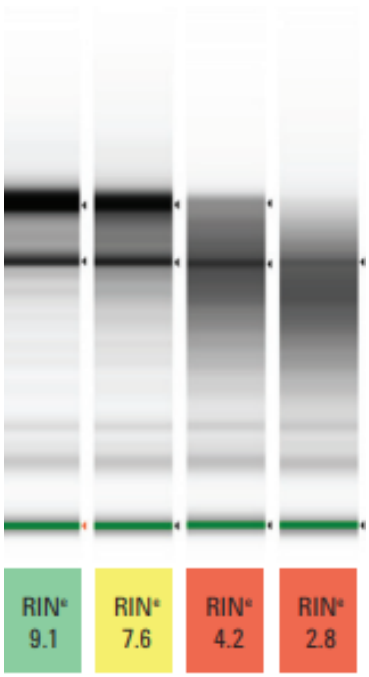


Figure 1. Example of the Tape station report with e-gel and RIN^e scores

- Intact total RNA will have sharp, clear 28S and 18S ribosomal RNA (rRNA) bands on the e gel. The 28S rRNA band should be approximately twice as intense as the 18S rRNA band (Figure 1, lane 1). Degraded RNA will have a smeared appearance, will lack the sharp rRNA bands, or will not exhibit the 2:1 ratio of high-quality RNA (Figure 1, lane 4).
- RIN^e values range from 1 to 10. A high RIN^e indicates intact RNA, while a low RIN^e indicates highly degraded RNA.
- If the RIN^e score is >7, the quality of the sample is optimal.
- If the RIN^e score is <7 but >=6, the quality is suboptimal.
- If the RIN^e score is <=5, and the e-gel lacks sharp rRNA bands, re-extract RNA.
- Record the RIN^e score and the assigned concentration [ng/µl] in the quantification tab in the extraction worksheet.
- Fill out the master HTS metadata sheet on SharePoint.

Submitting sample RNA to external sequencing service provider Ramaciotti Centre for Genomics (RCG), NSW

Step 1 – Online submission of samples

- Create a profile with RCG for sample submission.
- Request a quote from RCG for the number of samples that passed internal QC, specifying required flow cell and sequencing chemistry (see **Appendix 1**).
    - Technical specification for small RNA-Seq data generation: samples to be prepared using the QIAseq miRNA Library Prep kit [QIAGEN] using unique dual indexes (UDIs) and sequenced with an Illumina NovaSeq instrument. A minimal sequencing depth of ~40 million reads per sample is recommended.
- Review and accept the quote (**Appendix 2**).
- Generate a Purchase Order (PO) for the quote and get it approved by the supervisors and forward this to the RCG project manager.
- Ship samples within 30 days of accepting the quote.
- Log into the RCG sample submission portal and navigate to ‘resources’.

- Download the sample submission form for 96-well plates/tubes as appropriately. To avoid upload errors, download a new copy of the submission form each time you submit.
- Fill the sample metadata required by RCG.
- Provide the DAWE email address(es) of all relevant persons to be contacted by RCG when the sequencing data is ready for delivery.
- Make sure the PEQ index numbers are de-identified for client samples.
- Create a 'new project' in the RCG submission portal with the quote and work order IDs.
- Fill in the required information and in the Description, specify the flow cell type, number of samples, platform and service type.
- Upload the saved sample submission form.
- The window should now be populated with your sample metadata.
- Sign the declaration and submit the samples and your submission will receive an RCG LIMS ID.

Step 2 – Preparing plates/tubes for RCG

- Do not use screw cap or strip tubes, or seal tubes with parafilm, foil or tape when submitting the samples.
- Use cryo-pens or cryo-dots with permanent markers when labelling tubes/plates.

Submitting samples in tubes

- Projects with less than 10 samples can be submitted in 1.5 ml tubes. Clearly label your tubes on both the top and side of the tube, with the corresponding sample LIMS ID.

Submitting samples in plates

- Projects with 10 or more samples must be submitted in 96-well PCR plates (clear wells, semi or fully skirted).
- Label plates with the unique LIMS ID issued during online submission.
- Plates must also be labelled with the container name e.g. plate 1, 2, 3.
- Make sure that plates are well sealed using an aluminium sealing film first and then a clear adhesive film. Aluminium seals are compatible with cold storage and the clear seal is a secondary protectant to the aluminium seal.
- Print out the sample submission confirmation and put inside a Ziplock bag.
- Put your tubes/plates inside a separate Ziplock bag.
- Tape the two Ziplock bags together to make sure they stay together.

- Wrap everything using bubble wrap and write 'Top' to indicate the storage direction and the LIMS ID on bubble wrap.

- Freeze the wrapped samples overnight before the courier arrives.

Step 3 – Organising the courier pick up

- Request quote from a courier service to send samples on dry ice.
- Organise the courier to pick up the samples to be delivered to RCG early during the week. Avoid mid or late week courier pick-up to ensure the samples are delivered to RCG by mid-week so they are attended to appropriately.
- RCG will send an email notification upon sample receipt. RCG will also perform their own internal QC on the samples and seek advice where sample quality is suboptimal (see SOP4).
- Save the Quote, Purchase order, RCG metadata sheet, sample submission confirmation, internal TapeStation QC results and the Courrier order form to an appropriate local/SharePoint folder.

7. **Waste management and other safety precautions**

- **Chemicals –** The handling of solvents, samples and reagents can hold health and safety risks. When using/handling the ScreenTape device and working with these substances observe appropriate safety procedures (for example by wearing goggles, safety gloves and protective clothing). The volume of substances should be reduced to the minimum required for the analysis. Risk assessments are contained in the specimen laboratory for all reagents in the kit. Be sure to familiarise yourself with the risks each pose, the PPE required, the method of disposal and the method of clean up in case of spills. Use the spill kit if required to contain any large spills.
- **Temperature** – Temperatures are utilised in this method can burn the skin. When handling cold frozen goods, use cold protective gloves.
- **Mechanical hazards –** Mechanical hazards exist due to moving parts in the instruments used. Ensure you have training in the use of any instrument you are using. Mechanical hazards exist in the TapeStation instrument, centrifuge, thermal cycler, and vortex. Never place yourself in harm's way, be careful of moving parts.
- **Electrical hazards** – Electrical hazards exist with many of the apparatus used in the method. These include: fridges and freezers, TapeStation instrument, centrifuge, thermocycler, vortex. Report any defects in instrumentation, do not use instrumentation with obvious defects, like exposed wires. Instrumentation should be off at the power point unless required. Attempt to keep liquids away from power cables, power boards, and instrumentations. Have a liquid spill in or around instrumentation cleaned, with OHS direction with the power off on the instrument. Do not put yourself at risk, report hazard to your manager.
- **Environmental hazards** – Potential environmental hazards exist in specimens being processed. Specimen and specimen by-products need to be disposed of as quarantine waste, in a biohazard bag, and placed in a large yellow bin for commercial disposal, or autoclaved before deep burial. Clean up bench space with or RNaseZAP followed by 80% alcohol. When using spray bottle, spray directly into towel, away from face so as not to breathe in aerosols.

8. **Related documents**

- User guide for Agilent TapeStation https://www.agilent.com/cs/library/usermanuals/public/RNA_QuickGuide.pdf
- RCG sample submission guide https://www.ramaciotti.unsw.edu.au/sample-submission/how-submit-samples

9. **Rationale for change of SOP version**

N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					

Laboratory area:	No of copies:	Reason for change:

Appendix 1. Example of email to RCG to request quote:

From: XXXX<XXXXXX>;
To: Ramaciotti Centre for Genomics <ramaciotti@unsw.edu.au>

Subject: Small RNA-seq

Dear XXX,

We are keen to sequence X samples using small RNA-seq as per the standard process optimised with QUT and DAFF consisting of i) ribosomal depletion using FastSelect (QIAGEN), and ii) library prep using the QIAseq miRNA kit with UMIs and UDIs. We require a sequencing depth of 40-45M reads per sample.

Please provide a quote at your earliest convenient. Additionally, advise when it will be best to ship the samples.

Thank you,
XXXXX

Appendix 2. Example of quote provided by RCG

UNSW SYDNEY Ramaciotti Centre for Genomics

The Ramaciotti Centre for Genomics
Level 2, Biosciences South Building (E26)
UNSW Sydney, NSW 2052

Tel: (02) 9385 1241
Email: ramaciotti@unsw.edu.au
Web: www.ramaciotti.unsw.edu.au

ABN: 57 195 873 179

QUOTATION

Quote Number:	Q0009405
Quote Date:	22/07/2022
Quote Expiry:	30 days from quote date
Client:	Roberto Barrero Gumiel
Prepared By:	Jess Gurr
Job Number:	OPP186215

CLIENT INFORMATION

Roberto Barrero Gumiel
Queensland University of Technology
roberto.barrero@qut.edu.au

SERVICE

Description	Quantity	Unit	Rate (AUD)	Extended Price (ex GST)
QIAseq miRNA library prep (using UDIs)	28.00	EACH	$230.00	$6440.00
NovaSeq 6000 S1 100cycle flowcell (run at 1x75bp to read the miRNA plus the UMI)	1.00	EACH	$6650.00	$6650.00
			Total AUD Excl. GST	$13090.00

QUOTATION SUMMARY

Data output: Up to 1.6B reads (~50M reads/sample with 28 samples); Includes PhiX spike-in

Quality Control: >85% bases higher than Q30 at 1x100 bp

Deliverables: demultiplexed fastQ files

Turnaround Time: 20-30 business days (once kits in stock)

Sample requirement: The input for the QIAseq prep varies (from 1ng to 300ng input and standard requirement is total RNA at 40-100ng/ul and 15-20ul volume). Keep volume to 12-13ul per sample if you have less than 10ng in total (results are not guaranteed for input under 10ng of total RNA). Please get in contact prior to submission if amount is an issue and/or if you have serum/plasma samples. It is your responsibility that you retain miRNA in your extraction as we cannot verify this prior to start of prep

STANDARD SERVICE TERMS

Follow this link to review our Standard Service Terms.

SAMPLE SUBMISSION

Follow this link for information on online sample submission and shipping.

NATA

Page 1 of 1 Commercial in Confidence

Standard Operating Procedure for cDNA library preparation and sequencing by external sequencing provider

(HTS_SOP4_20220929)

1. **Purpose**

   This SOP provides the procedure performed by an external sequencing provider (Ramaciotti Centre for Genomics - RCG) for sample QC checking, cDNA library preparation and High Throughput Sequencing to deliver sequencing data for the detection of viruses and viroids using sRNA data. Technical details for data generation, deliverables and turn-around-time should match what was specified in the quote provided by the sequencing provider.

2. **Scope**

   The following operating procedure is to be performed by an external sequencing provider. SSG staff needs to be trained in interpreting quality control (QC) checks performed on RNA, downloading and storage of HTS data.

3. **Definitions and abbreviations**

   HTS – High Throughput Sequencing

   PEQ – Post Entry Quarantine facility

   RCG – Ramaciotti Centre for Genomics

   RNA – Ribonucleic acid

   RIN - RNA Integrity Number

   SSG – Science and Surveillance Group

   sRNA – small RNA

   QC – quality control

4. **Personnel pre-qualifications**

   NA

5. **Equipment and materials**

NA

6. **Procedure**

Quantification and Qualification of RNA

- The RCG sequencing provider uses spectrophotometry to determine the RNA concentration and purity. RCG uses either NanoDrop or an Epoch Spectrophotometer to measure the absorbance of nucleic acids and potential contaminants.
- Nucleic acids and proteins have absorbance maxima at 260 and 280 nm, respectively. A 260/280 ratio of ~1.8 for DNA and a 260/280 ratio of ~2.0 for RNA are generally accepted as "pure".
- Similarly, the absorbance ratio at 260/230 is an indication of contamination with other agents such as phenol or salts. The 260/230 values for “pure” nucleic acid are often higher than the respective 260/280 values. Expected 260/230 values are commonly in the range of 2.0–2.2.
- RCG to assess the integrity of RNA samples run aliquot of total nucleic acid extracts on the Agilent 4200 TapeStation using the RNA ScreenTape system. An RNA integrity number (RIN) is assigned to individual samples based on their overall integrity. Samples with high quality RNA have a RIN score >= 8.0.
- SSG will be notified if the samples pass or do not pass QC and are advised of the next steps and possible solutions. Samples can fail QC for several reasons including: insufficient RNA concentration, RNA degradation, contamination with protein or other extraction protocol reagents (i.e., phenol).
- SSG will need to review the QC report summary provided by sequencing provider to confirm whether samples can proceed to cDNA library preparation steps and subsequent HTS. The SSG team takes the final decision on whether to process or not all samples based on the QC report and advice.
- The RCG does not guarantee the successful sequencing of samples that do not pass their recommended RNA integrity score (RIN >= 8) or quantity specifications (> 20 ng/μl).

cDNA library preparation

- The recommend commercial kit for small RNA-Seq HTS data generation is the QIAseq miRNA Library kit (QIAGEN) using unique dual indexes (UDIs) and unique molecular identifiers (UMIs) provided as part of the QIAseq miRNA Index kit (QIAGEN). We do not recommend using other commercial kit solutions as they did not yield good performance or have not been thoroughly tested.
- RCG follows manufacturer’s recommendations. The volume is adjusted to input exactly 100 ng for each sample for the library prep into 5 μl. If the concentration is too high, the volume is topped up with nuclease-free water accordingly. If the concentration is too low the SpeedVac will be used to concentrate the RNA.
- A positive control (Human XpressRef Universal Total RNA) and a negative control (water) are included with each library prep batch to assess for library preparation failure.
- Adapters are ligated sequentially to the 3' and 5' ends of small RNAs. Subsequently, universal cDNA synthesis with UMI assignment, cDNA clean-up, library amplification and library clean-up are performed.
- Quality control of individual library is performed using the Quant-iT PicoGreen dsDNA assay kit and the Perkin Elmer LabChip High Sensitivity DNA chip.
- Libraries are normalised and pooled in equimolar ratios. A final clean-up is performed with Qiagen QMN beads.
- Quality control on the clean pool is performed using Qubit dsDNA HS assay and Agilent TapeStation High Sensitivity D1000 assay.

Sequencing

- Samples are sequenced using the Illumina NovaSeq 6000 instrument with either SP 100c, S1 100c, or S2 100c flow cells depending on the number of samples in the batch and the desired sequencing depth per sample. We recommend generating at least 45 million small RNA reads per sample. Raw Illumina reads are typically single end 100 bases long sequences encoding a small RNA (i.e., miRNA, siRNA) plus UMI barcode.
- 1% PhiX spike-in is included for quality control to check the error rate of the sequencing run.
- Demultiplexing and base quality of reads are performed using bcl2fastq followed by delivery of FASTQ files. Prior to releasing the data, the quality and output are reviewed to ensure they pass specifications. Any analysis that fails to meet the QC metrics set by the manufacturer will be repeated free of charge. The only exception being if SSG was advised to proceed with a sample that has not passed their quality or quantity specifications, or deviation from standard protocol was requested.

Data delivery and storage

- RCG sends an email communication to the sample submitter and other relevant contacts including DAWE's IT team to indicate the completion of the sequencing and instructions for downloading the data.
- DAWE's IT has established an automated data ingestion process that is triggered by a JSON file attached to the RCG email communication. Data is automatically stored at DAWE's secure Azure Cloud storage space.
- Metadata of sequenced samples is stored along with FASTQ raw sequencing files. Files are archived after 30 days and maintained as per data management regulations.

7. **Waste management and other safety precautions**

- NA

8. **Related documents**

- Illumina sequencing output (https://sapac.illumina.com/systems/sequencing-platforms/novaseq/specifications.html)
- RCG sample submission and standard service terms (https://www.ramaciotti.unsw.edu.au/sample-submission/standard-service-terms & https://www.ramaciotti.unsw.edu.au/sample-submission/quality-control-overview).
- Interpreting QC data provided by RCG (https://www.ramaciotti.unsw.edu.au/sites/default/files/2019-03/RAMAC_QC_Datasheet_2015.pdf)

9. **Rationale for change of SOP version**

- **NA**

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Standard Operating Procedure for Bioinformatics Analysis of Small RNA Sequencing Data using PEQ-HiTS website (HTS_SOP5_20221111)

1. **Purpose**

This SOP provides the procedures for retrieving and analysing of small RNA sequencing data using either the VirReport pipeline on the cloud or the Galaxy Australia platform for the purpose of detection of viruses and viroids in plants processed at the Post Entry Quarantine Facility.

2. **Scope**

The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of Science and Surveillance Group at Post Entry Quarantine Facility intending to retrieve, analyse small RNA sequencing data generated by Ramaciotti Centre for Genomics following the HTS_SOP4 using either the VirReport pipeline on the cloud or the Galaxy Australia platform.

3. **Definitions and abbreviations**

GA – Galaxy Australia

HTS – High Throughput Sequencing

SSG – Science and Surveillance Group

sRNA – small RNA

ORF – Open Reading Frame

PEQ – Post Entry Quarantine facility

RNA – Ribonucleic acid

RCG – The Ramaciotti Centre for Genomics

4. **Personnel pre-qualifications**

All personnel performing data analysis should be trained by a qualified bioinformatician.

5. **Equipment and materials**

sRNA sequencing data generated following HTS_SOP4

Pre-requirements for analysing data using GA platform

- You should have access to the shared GA profile at (*For internal use only*)
- Basic data analysis will be conducted in GA.
- Make sure the required GA-workflows and datasets are shared with the GA profile. GA-VirReport analysis requires the following:

- Sequence dataset of the latest custom virus viroid database (PvirDB) in .fasta format
- Latest version of the Non-informative dataset in .fasta format
- Sequence dataset for the spike in control in .fasta format
- Latest version of GA-VirReport workflow
- Latest version of GA-VirReport-Stats workflow that assesses genome coverage, sequencing depth and generates a consensus genome.

- Request access to the PEQ-HiTS web page (*For internal use only*) from DAFF IT. Storing, summarising GA outputs and generating summary reports will be done in Azure cloud platform. Data analysis will be conducted in GA. Launching of these tasks in Azure will be done using the PEQ-HiTS web page.

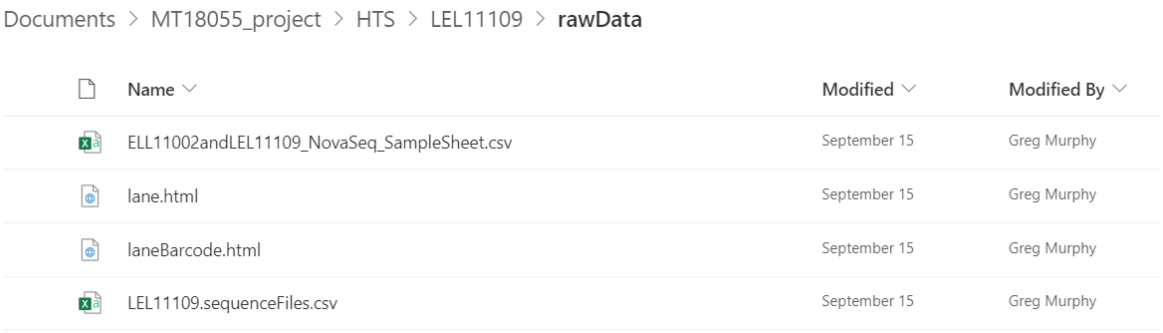
Pre-requirements for analysing data using VirReport on the cloud

- Data analysis will be done in Azure cloud platform using the VirReport Netxflow pipeline. Batches will be analysed automatically once sequencing data is made available by the sequencing provider but can also be triggered manually via the PEQ-HiTS web page. Request access to the PEQ-HiTS web page (*For internal use only*) from DAFF IT if you need to manually reanalyse batches.
- For the VirReport pipeline to run, the DAWE IT team needs to provide access to the necessary on-demand cloud compute resources to run the bioinformatics analyses, as well as install the following resources:
  - Nextflow (https://www.nextflow.io/docs/latest/getstarted.html)
  - VirReport (https://github.com/eresearchqut/VirReport)
  - PVirDB database
  - Public NCBI NT and NR databases which will be upgraded at 6-months intervals
  - Bowtie indices for the RNA source analysis and the filtering of the non-informative reads
- For a given batch, the pipeline will require several files.
  - The sampleSheet.csv and index.csv files which will be automatically generated by DAWE IT based on the data they receive from the RCG.
  - A sample_info.txt file provided by the SSG team upon sample submission.

6. **Procedure**

Automated processing of sequencing data

- Once Ramaciotti performs QC of the samples they will send the report through a nominated email (*For internal use only*).
- The attachments in the email will be pulled and made available in the ~HTS/BATCHID/QC Report folder.
- Once the sequencing data are available from Ramaciotti they will notify through a nominated email (*For internal use only*).
- Upon notification of data availability from the Ramaciotti, sequencing data will be automatically downloaded to the active storage of the Azure cloud platform.
- A snapshot of the downloaded compressed file from Ramaciotti will be created under the ~HTS/BATCHID/Ramaciotti folder as BATCHID.zip.txt along with the BATCHID.json file with data retrival information from Ramaciotti.
- Then the downloaded data will be automatically unzipped in Azure and the SampleSheet.csv with sample names and adapter information, lane.html and lane barcode.html files with flowcell cluster information and sequenceFiles.csv with file names of sequencing files, size and unzipped date and time will be posted to the ~HTS/BATCHID/rawData folder.



GA-VirReport process:

- Then the data will be automatically processed through the GA-VirReport workflow in GA with creating individual histories for each sample ID with the history name – **BatchID_SAMPLEID-Date-time-attempt #**. Azure will attempt a maximum of 5 attempts before failing.
- After completion of the GA-VirReport pipeline, the read_size_distribution image, Cap3 outputs (contigs), getorf outputs (ORFs), eXpress outputs (spike in read counts) and blastn and tblastn outputs will be pulled from GA into Azure and will also be available on SharePoint under ~HTS/BATCHID/SAMPLEID/Results folder.
- Then the blastn outputs will be automatically summarized in Azure and the summary outputs will be posted into the respective Results folder in SharePoint.
- If there are any entries in the summary outputs, for each '**sacc**', the GA-VirReport-Stats workflow will be triggered in a new history with the history name –B**atch ID_ SAMPLEID-sacc-Date-time-attempt #**. Azure will attempt a maximum of 5 attempts before failing.
- Once the GA-VirReport-Stats pipeline is completed, the bamQC output (read coverage statistics) and consensus sequence generated for each sacc will be posted to the respective Results folder.
- Once all the data are available you can use SOP6 to report and interpret the data.

VirReport process:

- The data will also be automatically processed through the VirReport Nextflow pipeline on Azure. The current process automatically creates the index.csv file required by the pipeline to run based on the fastq.gz file names.
- An email will be sent when the pipeline has finished running and all the results are available on the SharePoint. The results will be posted in the HTS/BATCHID/VirReport folder on SharePoint, under a folder named after the batch run.
- Once all the data are available you can use SOP6 to report and interpret the data.

Manual processing of individual steps using PEQ-HiTS website

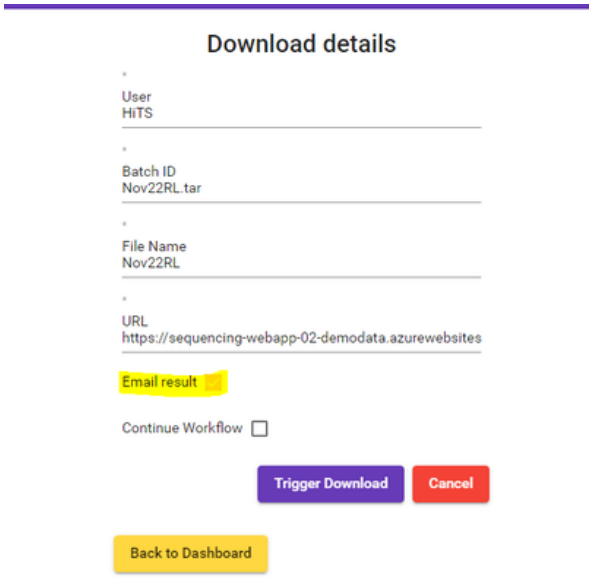
GA-VirReport process:

Use the following steps as required if you want to manually trigger the analysis process for any batch/sample.

Step 1- Retrieving data from the Ramaciotti Centre for Genomics to Azure storage

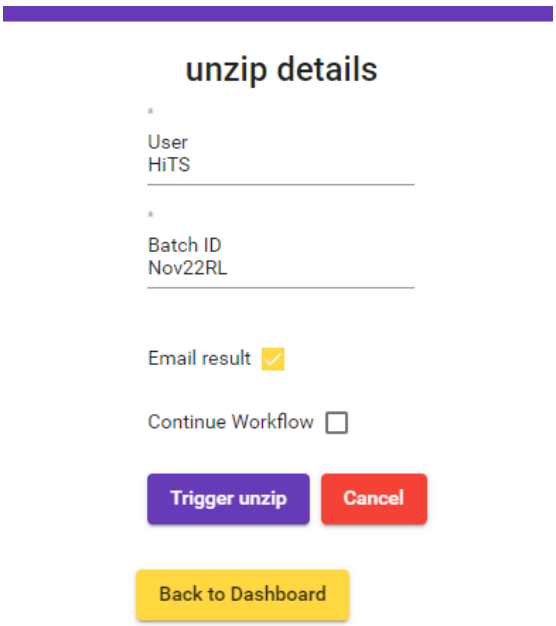
- Ramaciotti retains data in their storage for 3 months upon delivery. Downloading of data into Azure is automated. If you want to download the data again within the 3-month period, follow the procedure below.

- You will need the URL from Ramaciotti for data download. This should be available in the PEQ-HiTS shared mailbox.
- Go to the PEQ-HiTS website (https://sequencing-webapp-01.azurewebsites.net/postresult)
- Under **Galaxy Australia options** go to **Download**
- Enter **HiTS** as **UserID**.
- Type the **Batch Id**.
- If you check the '**Continue workflow**' check box, the rest of the workflow will be automatically triggered after downloading.
- Click **Trigger Download.**



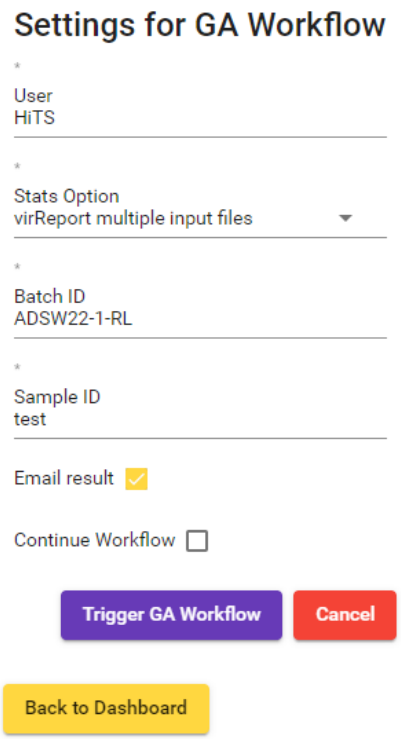
Step 2- Unzipping data

- Use this function to unzip any pre-downloaded data from a batch.
- Go to https://sequencing-webapp-01.azurewebsites.net/ to access the new PEQ-HiTS website
- Under **Galaxy Australia options** go to **RawData** and then **Unzip**.
- Enter **HiTS** as **UserID** in the popup menu.
- Type **Batch Id**.
- You can opt to email the results which will send an email upon completion of unzipping.
- If you check the '**Continue workflow**' check box, the rest of the workflow will be automatically triggered after unzipping.
- Click **Trigger unzip**.



Step 3- Run GA-VirReport workflow

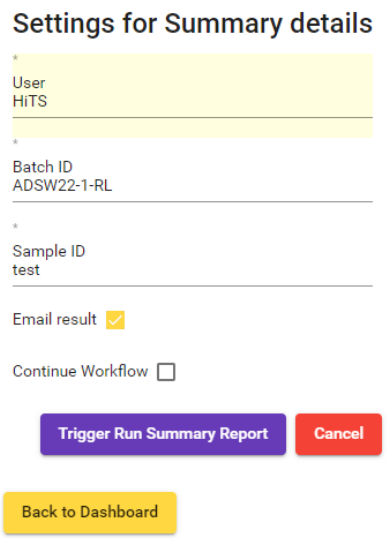
- Use this step to process any unzipped data with the GA-VirReport workflow.
- Go to https://sequencing-webapp-01.azurewebsites.net/ to access the new PEQ-HiTS website.
- Under **Galaxy Australia options** go to **Datasets.**
- Select **Run GA Workflow**
- Under Settings for GA Workflow select **virReport - multiple input files** as the **Stats Option**.
- Type **Batch Id**
- Enter **Sample Id**
- You can opt to email the results which will send an email upon completion of unzipping.
- If you check the '**Continue workflow'** check box, the rest of the workflow will be automatically triggered after unzipping.
- Click **Trigger GA Workflow**.



- This will trigger GA-VirReport workflow in GA creating individual histories for each sample ID with the history name – **SAMPLEID-Date-time-attempt#.**
- Once this step is completed you should see outputs of read size distribution, Cap3, getorf, eXpress, blastn and tblastn under **~HTS/BATCHID/SAMPLEID/Results** folder in SharePoint.

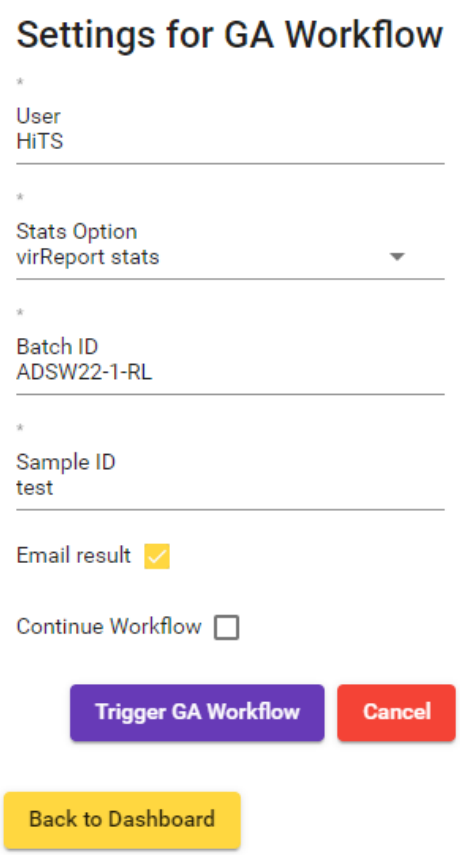
Step 3- Summarise Blastn outputs

- Use this step to summarise blastn outputs from the GA-VirReport workflow.
- Go to https://sequencing-webapp-01.azurewebsites.net/ to access the new PEQ-HiTS website
- Under **Galaxy Australia options** go to **Datasets**.
- Select **Run VirReport Summary**
- In **Settings for summary details**,
- Type **HiTS** as **User**
- Type **Batch Id**
- Enter **Sample Id**
- You can opt to email the results which will send an email upon completion of GA-VirReport-multiple input file workflow.
- If you check the '**Continue workflow'** check box, the rest of the workflow will be automatically triggered after summarising.
- Click **Trigger Run Summary Report**.
- This will trigger summarising of all blastn outputs within a batch.
- Upon completion, **summary_SAMPLEID _viruses_viroids.txt** file will be posted to the respective results folder.



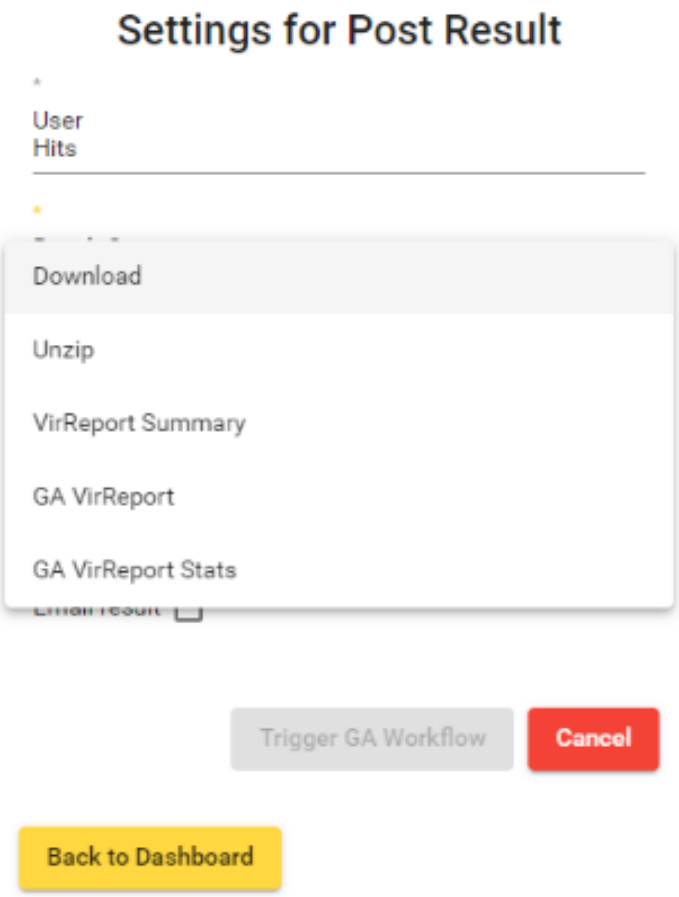
Step 4- Run GA-VirReport-Stats workflow

- Use this step to run the GA_VirReport Stats workflow for any sample.
- Go to https://sequencing-webapp-01.azurewebsites.net/ to access the new PEQ-HiTS website
- Under **Galaxy Australia options** go to **Datasets**.
- Select **Run GA Workflow**
- **Under Settings for GA Workflow** select **virReport - stats** as the Stats Option.
- Type Batch Id.
- Enter Sample Id.
- You can opt to email the results which will send an email upon completion of GA-VirReport-Stats workflow.
- Click '**Trigger GA Workflow'**.
- This will perform Stats workflow in GA for all detections in your summary output for a particular sample creating a new history with the history name **– SAMPLEID-sacc-Date-time-attempt #.**
- Upon completion, bamQC output (read coverage statistics) and Consensus sequence for the sacc will be posted to the respective results folder.



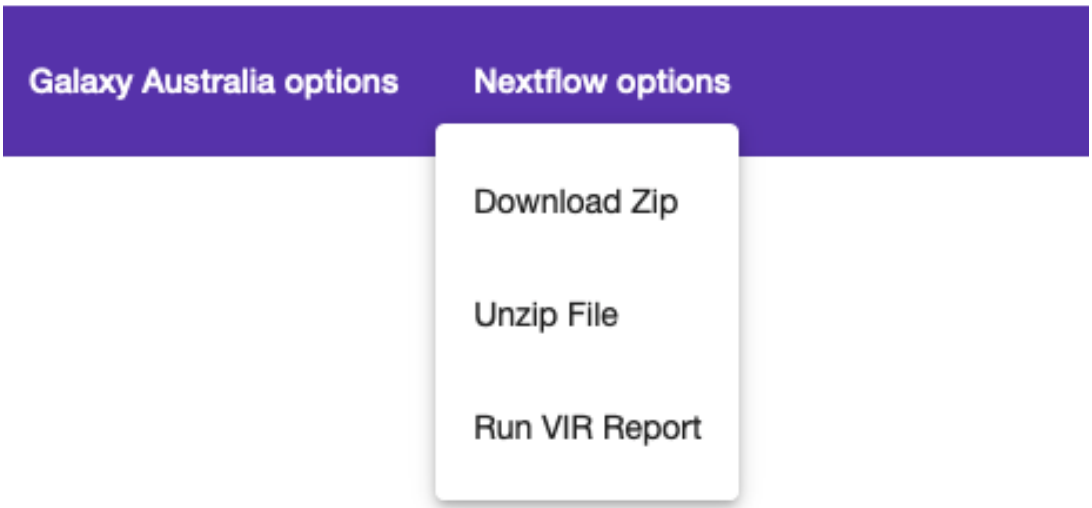
Step 5 - Post results

- Use this step if you ever want to post results from any of the above steps to SharePoint.
- Go to https://sequencing-webapp-01.azurewebsites.net/ to access the new PEQ-HiTS website
- Under **Galaxy Australia options** go to **Datasets**.
- Select **Post Results**
- Under **Settings for Post Results** select any type of results that you want to post to SharePoint.
- Selecting Download and unzip will recreate the ~HTS/BATCHID/Ramaciotti and ~HTS/BATCHID/Ramaciotti/rawData folders respectively.
- Type Batch Id.
- Enter Sample Id.
- You can opt to email the results which will send an email upon completion of Post-Results.
- Click 'Trigger GA workflow'.



VirReport process:

- Similarly you will be able to trigger the following three steps on PEQ-HiTS:
  - retrieval of sequencing data from a batch sequenced by the RCG,
  - unzipping any pre-downloaded data from a batch,
  - running the VirReport pipeline on a batch of samples after unzipping has been performed



7. **Related documents**

- User guide for GA-VirReport

8. **Rationale for change of SOP version**

N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Standard Operating Procedure for reporting the detection of viruses and viroids using bioinformatics analysis of small RNA sequencing data processed through VirReport (HTS_SOP6_VirReport_20221111)

1. **Purpose**

This SOP provides the procedures for reporting the detection of viruses and viroids using bioinformatics analysis of small RNA sequencing data in plants processed at the Post Entry Quarantine facility.

2. **Scope**

The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of Science and Surveillance Group at Post Entry Quarantine Facility intending to interpret bioinformatics analysis of small RNA sequencing data to report the detection of viruses and viroids in plants processed at the Post Entry Quarantine facility.

3. **Definitions and abbreviations**

HPC – High Performance Computing

HTS – High Throughput Sequencing

PEQ – Post Entry Quarantine facility

RCG – Ramaciotti Centre for Genomics

RNA – Ribonucleic acid

SSG – Science and Surveillance Group

sRNA – small RNA

4. **Personnel pre-qualifications**

All personnel performing data analysis and interpretation should be trained by a qualified bioinformatician and plant diagnostician.

5. **Equipment and materials**

- Results from sRNA sequencing data analysis, generated using VirReport following HTS_SOP5, Procedure for Bioinformatics Analysis of Small RNA Sequencing Data using PEQ HiTS website

- Get access to SharePoint location where the VirReport results will be saved

- NGS_report_template.xlsx

- Get access to the Detections.xlsx and PEQ_2022_Spring_Metadata.xlsx files in the VirReport Sharepoint

- QC reports from the RCG and SSG

6. Procedure

Pre-requirements for analysing data using VirReport (HPC version)

- Download the run_qc_report_[date].txt (Appendix 1), the run_read_size_distribution_[date].pdf (Appendix 2), the read_origin_pc_summary_[date].txt and the read_RNA_source.pdf (Appendix 3) files to check the quality of the sequencing run.
- Download the synthetic_oligo_summary_[*date*].txt (Appendix 4)
- If performing homology searches against NT, download the VirReport_detection_summary_21-22nt_ncbi_[date].txt and the VirReport_detection_summary_collapsed_21-22nt_ncbi_[date].txt file (Appendix 5) to review the virus and viroid detections for the run.
- If performing homology searches against the PVirDB, download the VirReport_detection_summary_21-22nt_viral_db_[date].txt and the VirReport_detection_summary_collapsed_21-22nt_ viral_db _[date].txt file to review the virus and viroid detections for the run.
- Make a copy of the current report template excel file (NGS_report_template.xlsx), rename with following name convention:
  - The suffix (PEQ2223)
  - the number of the batch being sequenced (e.g. batch_01),
  - the name of the batch being sequenced (e.g. LEL11291),
  - and the NGS report suffix (report.xlsx),
  - Each element should be separated by underscores (e.g. PEQ2223_batch_01_LEL11291_NGS_report.xlsx).
- Save the report into the relevant results folder (using the same naming convention as above).

Step 1 - Populate the sample metadata

- Record all relevant information that needs to be captured at SSG in the LIMS (currently captured in first tab of the report called "**RUN_INFORMATION**"): Sequencing run name, Project name, Date submitted, Name of submitter, Quote number, Purchase order number, UNSW code, Shipping company, Shipping job number, Date sequenced, Sequencing provider, Sequencing provider LIMS ID, Library kit, Sequencing machine, Flow cell, Demultiplexed raw sequence read length (bp), Adapter sequence, Name of person analysing the data, date the data was analysed, name of the person performing the final interpretation of the results. Some of the information has been pre-filled by default:

#####################################	
1) RUN INFORMATION	
#####################################	
Sequencing run name	
Project name	
Date submitted	
Submitter_name	
Quote number	
Purchase order number	
UNSW code	
Shipping company	LabCabs
Shipping job number	
Date sequenced	
Sequencing provider	The Ramaciotti Centre for Genomics
Sequencing provider LIMS ID	
Library kit	QIAseq miRNA Library Kit with UDIs and UMIs - QIAGEN
Sequencing machine	NovaSeq 6000
Flow cell	
Demultiplexed raw sequence read length (bp)	75 bp
Adapter sequence	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
Name of person performing bioinformatics analysis	
Date analysed	
Name of person performing final interpretation of results	

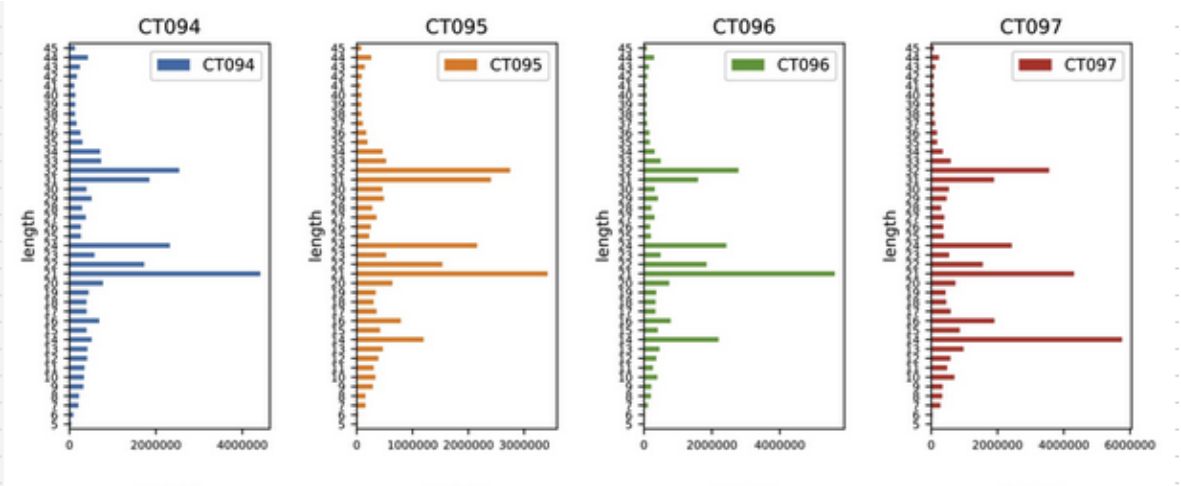
Step 2 - Check the quality of the sequencing run

- Review the correspondence with the sequencing provider to familiarise yourself with any issues that might have arisen with the batch during cDNA library preparation and sequencing (e.g. sample failing library preparation, sample for which there was a quality issue).
- Update the sample status in the PEQ metadata table (*For internal use only*) to "**Sequenced**".
- Extract the required information from the PEQ metadata table (*For internal use only*) and paste in the "**SAMPLE_QC**" tab of the report.

##########													
2) SAMPLE QC													
##########													
Sample	PEQ index number	LIMS ID	Commodity	Species	Common name	Plant tissue collected	Collection date (Month, Year)	Concentration Epoch/Nanodrop ng/µl	260:280 Epoch	260:230 Epoch	RIN Bioanalyser/TapeStation	RIN quality tier	QC_flag
##########													
Notes from and correspondence with sequencing provider	Date												

- Fill in the RIN quality tier:
    - RIN score >=7 are flagged as 1
    - RIN score >= 6 and <7 are flagged as 2
    - RIN score < 6 are flagged as 3
- Samples with a RIN score < 6 should be flagged as poor quality in the "QC_flag" column. Depending on the commodity, and the sample RNA quality and RNA content, consider re-sequencing or performing a new RNA extraction

- In the "**TAPESTATION**" tab, paste images of both internal and the RCG TapeStation results for review. QC reports are in the respective HTS batch folder on SharePoint. Check that samples originating from the same commodity display a similar gel profile.
- Open the run_qc_report.txt file (Appendix 1) and paste content into the "**NGS_QC**" tab of the report. This table contains the following information:
    - Unique dual index 1 (UDI1)
    - Unique dual index 2 (UDI2)
    - raw reads count
    - umi cleaned reads count
    - quality filtered reads > 18bp count
    - total filtered bases
    - q20 bases
    - q30 bases
    - percent gc content
    - informative reads count
    - informative reads 18-25 nt count
    - informative reads 21-22 nt count
    - informative reads 24 nt count
    - percent UMI incorporation
    - percent quality filtered
    - percent informative reads 18-25 nt
    - percent informative reads 21-22 nt
    - flags
- A sample will be flagged in the "flags" column if:
    - it has fewer than 30 million raw starting reads
    - it has a low % of reads with UMIs (<50% of the original raw reads)
    - it has a low % of quality filtered reads (<50% of the original raw reads)
    - it has fewer than 4 million 21-22 nt long informative reads
- Depending on the sample RNA quality and RNA content, consider re-sequencing or performing a new RNA extraction
- Open the .html report of the FastQC report and check that the per base sequence quality recovered post quality filtering is above 30 for the first 25 bp of the reads.
- Insert the run_read_size_distribution.pdf file (Appendix 2) into the **"READ_SIZE_DISTRIBUTION_IMAGE"** tab of the report and review to ensure that the read distribution profile for a given sample shows typical dominant peaks at 21, 22 and 24 nt long reads after adapter and UMI removal. Note if there are unusually high peaks for reads shorter than 18 bp which suggest that the sample has suffered some degradation (e.g. sample CT097 below).
- Check whether samples originating from the same commodity display a similar peak profile.

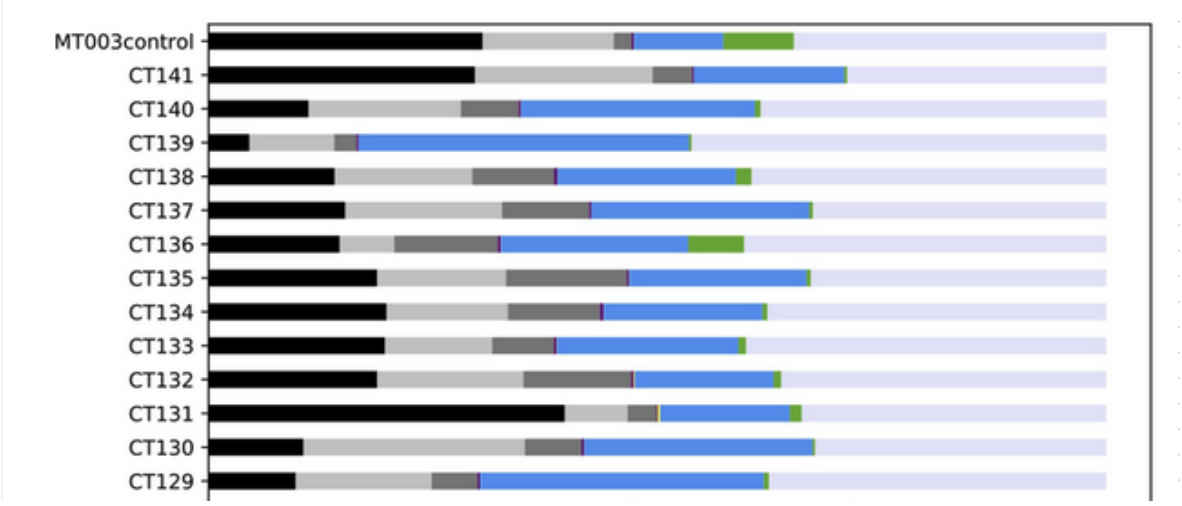


- Paste the read_RNA_source.txt file (Appendix 3) into the "**RNA_SOURCE**" tab.

4) RNA SOURCE

Sample	PEQ index number	LIMS ID	Commodity	Species	Plant tissue collected	rRNA_total	plant_tRNA	mitochondrial_and_plastid_genes	plant_non_coding_RNA	plant_transposons	PhiX_and_artefacts	miRNA_total	virus_and_viroids	leftover	rRNA_and_tRNA	miRNA/vsiRNA	rRNA/tRNA_flag	miRNA/vsiRNA_flag

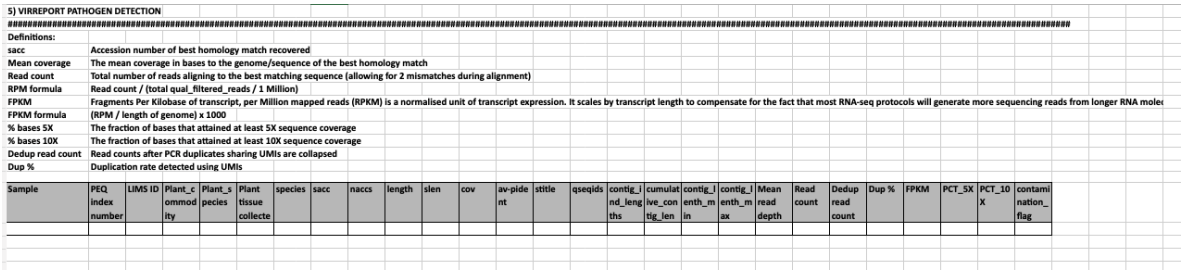
- Check whether there are samples with high amount of rRNA and tRNA content (these will be flagged in the **rRNA_and_tRNA_flag** column if the number of 18-25 nt usable reads mapping to rRNA and tRNA is higher than 50%). This suggests a sample contains a lot of degraded rRNA and tRNA, which will waste a lot of sequencing reads and possibly hinder the detection of viruses and viroids. Depending on the sample RNA quality and RNA content, consider performing a new RNA extraction.
- Samples with low amount of miRNA and viral RNA will be flagged in the **miRNA/vsiRNA flag** column if the number of 18-25 nt usable reads mapping to miRNA and viral small interfering RNA is less than 10%.
- Insert the read_RNA_source.pdf (Appendix 3) into the "**READ_SOURCE_IMAGE**" tab



- Paste the synthetic_oligo_summary.txt (Appendix 4) into the "**SYNTHETIC_OLIGOS**" tab of the report. Check the FPKM counts recovered for the synthetic oligo to make sure we recover signal in all samples that were spiked in.

Step 3 - Review the virus and viroid detections for the run against PVirDB or NT

- Paste the VirReport_detection_summary_21-22nt_ncbi/viral_db_[date].txt file (Appendix 5) for the blastn results against NT or/and the PVirDB into the "**VIRREPORT_ DETECTIONS_PVIRDB**" and/or "**VIRREPORT_ DETECTIONS_NCBI**" tabs of the report. Review detections.



5) VIRREPORT PATHOGEN DETECTION

Definitions:	
sacc	Accession number of best homology match recovered
Mean coverage	The mean coverage in bases to the genome/sequence of the best homology match
Read count	Total number of reads aligning to the best matching sequence (allowing for 2 mismatches during alignment)
RPM formula	Read count / (total qual_filtered_reads / 1 Million)
FPKM	Fragments Per Kilobase of transcript, per Million mapped reads (RPKM) is a normalised unit of transcript expression. It scales by transcript length to compensate for the fact that most RNA-seq protocols will generate more sequencing reads from longer RNA molec
FPKM formula	(RPM / length of genome) x 1000
% bases 5X	The fraction of bases that attained at least 5X sequence coverage
% bases 10X	The fraction of bases that attained at least 10X sequence coverage
Dedup read count	Read counts after PCR duplicates sharing UMIs are collapsed
Dup %	Duplication rate detected using UMIs

Sample	PEQ index number	LIMS ID	Plant_community	Plant_species	Plant tissue collecte	species	sacc	naccs	length	slen	cov	av-percent	stitle	qseqids	contig_lengths	cumulative_contig_len	contig_lenth_min	contig_lenth_max	Mean read depth	Read count	Dedup read count	Dup %	FPKM	PCT_5X	PCT_10X	contamination_flag

- Make sure that all expected viruses and viroids are detected in the control sample (currently MT003). This is a control sample which is run each time along the diagnostics samples to ensure the bioinformatics pipeline ran properly. It is infected with two viruses (CTV, CVEV) and two viroids (HSVd and CDVd).
- Check the contamination flag column for the presence of possible false positives. These detections are either reflecting the presence of a low titer virus or potential contamination due to index misassignment from a virus/viroid present at very high titer in another sample. These detections will need to be confirmed via an independent method.
- Review the blastn results against the NT database. Manually blast the longest contigs obtained with either blastn to NCBI to confirm viral prediction. Make sure the top matches are consistently viral in the Graphics Summary. If not, this could indicate an integrated virus/viroid, a misassembly or an erroneous annotation in the public database
- For detections which only have a single contig match <= 250 bp, perform a blastn search instead of megablast as this algorithm might give better results for short contigs.
- Check the read distribution profile (uniform versus fragmented) along the genome of the closest best hit by loading the bam file on IGV
- Check the derived consensus sequence to see if the reference-guided assembly improved the length of recovered contigs compared to the *de novo* approach
- All HTS detections are categorised into 4 categories based on three criteria:

    - The average percent identity of all *de novo* contigs aligning to a reference via blastn search
    - The percentage of read mapping to the reference genome at 10X depth
    - The minimum *de novo* contig length showing homology to reference

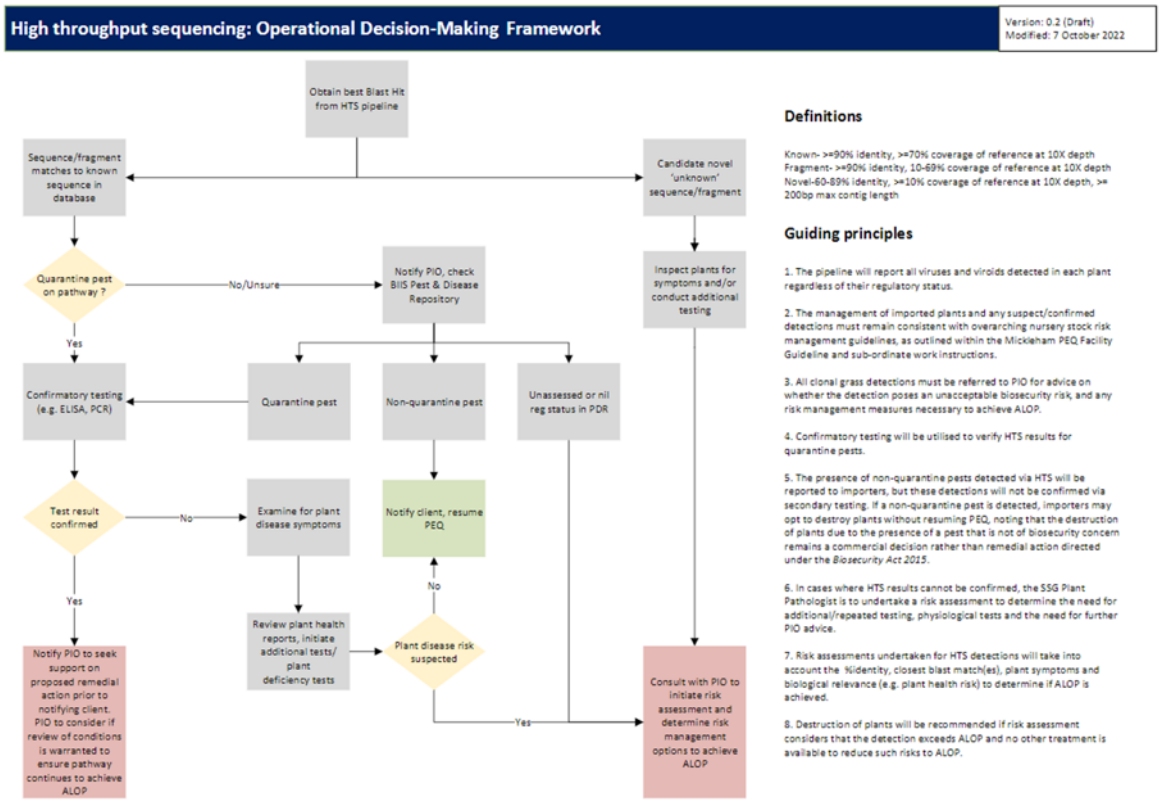
- These categories are:
    - **KNOWN** (high confidence)
        - Average percent identity to top reference (blastn search) >=85%
        - 10X genome coverage >=70%
    - **KNOWN FRAGMENT** (low confidence)
        - Average percent identity to top reference (blastn search) >=85%
        - 10X genome coverage 10-69%
    - **CANDIDATE NOVEL**
        - Average percent identity to top reference (blastn search) ranging between 60-84%
        - 10X genome coverage >=10%
        - Minimum contig length >= 200 bp
    - **EXCLUDE**
        - Detections that do not fit criteria for the above 3 categories

- For the **'CANDIDATE NOVEL'** detections**:**
- If several contigs match different viruses/viroids at lower homology (ie <85%) and share the same depth, it is possible that they originate from the same divergent virus and the top hit recovered changes depending on the genomic region considered. Scan the top hits from the blast results obtained from NCBI to see if they share similar % identity to some of the other viruses recovered in the sample.
- Check whether the read coverage is uniform or fragmented along the genome of the closest best hit by loading the bam file on IGV (However, note that with a novel virus, the

reads might not align well using the closest reference if the homology is too low; only highly conserved regions might align well resulting in a fragmented read distribution profile)

- Check the derived consensus sequence to see whether several regions (especially the 5' end and 3' end') of the genome are represented (However, note that this approach might not work well if the homology is too low)
- If long *de novo* contigs are obtained (>500 bp) and the *de novo* assembled contigs cover > 75% of the closest relative, align reads directly to *de novo* contigs and derive mean coverage and 10X coverage.
- Check the blastx results for the presence of open reading frames matching conserved proteins such as the RNA polymerase or the coat protein. Check for the presence of known conserved domains (e.g. Graphics Summary on Blast results, Expasy). Compare domain organisation to closest relatives and whether these domains are known to play a role in viral pathogenicity.
- Check literature to find out what is known about related virus pathogenicity and host distribution.
- If required, use the derived sequences (*de novo* and consensus) to design primers
- Consult PIO for decision

- Follow the reporting decision making process as depicted below, depending on the status of the detection ('**Quarantinable'** versus '**Higher plant viruses**'):



- All detections that are not categorised as '**EXCLUDE'** and are also categorised as "**quarantinable"** should be validated independently by a molecular assay.
- All detections that are not categorised as '**EXCLUDE'** and are also categorised as "**non-quarantinable (defined as higher plant viruses in the PVirDB)** will be reported to the client but do not require independent validation using a molecular assay.

- All detections **from a clonal grass sample** that are not categorised as '**EXCLUDE'** need to be independently validated using a molecular assay and reported to PIO. PIO will assess risk of each detection on a case-by-case manner. The detections will also be reported to the client.
- Any detection that is categorised as '**EXCLUDE'** is inconclusive and is not required to be reanalysed unless a positive virus-viroid detection was obtained from an independent method or RNA was poor quality and the number of 21-22 nt long informative reads was fewer than 4 million. Respective plant can be released if it does not test positive for other pathogens (e.g. bacteria, fungi) and if no suspect symptoms are found.
- Any sample with no detections in the blastn summary output is considered a negative and is not required to be reanalysed unless a positive virus-viroid detection was obtained from an independent method or RNA was poor quality and the number of 21-22 nt long informative reads was fewer than 4 million. Respective plant can be released if it does not test positive for other pathogens (bacteria, fungi) or if no suspect symptoms are found.
- Paste the VirReport_detection_summary_**collapsed**_21-22nt_ncbi/viral_db_[date].txt in the "**RESULTS_SUMMARY**" tab and edit accordingly. Record the results in the appropriate location (currently add the detections in the Detections.xlsx document).
- Record positive detections in the SSG Master plant testing workbook and store a copy of the populated reporting template in the SharePoint with output files for the respective sequencing batch.

- **Related documents**

  N/A

- **Rationale for change of SOP version**

  N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Standard Operating Procedure for RNA quantification, quality assessment and submission for High Throughput Sequencing (HTS_SOP3_20220206)

- **Purpose**

  This SOP provides the procedures for quantification and qualification of total RNA extracted from plant samples at PEQ and for submitting RNA to external sequencing provider- Ramaciotti Centre for genomics for High Throughput Sequencing.

- **Scope**

  The following operating procedure is to be adhered to by trained molecular biologists or plant pathologists of Science and Surveillance Group at PEQ intending to quantify, qualify and submit total RNA from a previously extracted plant following the HTS_SOP2 for the purpose of High Throughput Sequencing and subsequent validation through molecular and/or conventional assays.

- **Definitions and abbreviations**

  HTS – High Throughput Sequencing

  SSG – Science and Surveillance Group

  PEQ – Post Entry Quarantine facility

  RNA – Ribonucleic acid

  RIN - RNA Integrity Number

- **Personnel pre-qualifications**

  All personnel performing RNA extractions should be inducted to and have access to the Extraction room (PCR1 prep – 9.141). New staff members are required to be trained by and supervised by a trained SSG molecular biologist/ plant pathologist prior to attempting the procedure on their own.

- **Equipment and materials**

  RNA extracted following HTS_SOP2, NanoDrop One UV-Vis spectrophotometer, Agilent 4200 TapeStation system, microcentrifuge, Agilent RNA ScreenTape, Agilent RNA ScreenTape sample buffer, Agilent loading tips, Vortex mixer with 96-well sample plate adapter, 96-well sample plates, 96-well Plate Foil Seal, 96-well Plate transparent seal, micropipettes, filtered micropipette tips, Ice bucket, thermal cycler, permanent marker and cryo-dots/cryo-pen, Kimwipes

- **Procedure**

Preparations before RNA quantification and qualification

- Wear gloves at all times.
- Clean your work bench with RNASEZap followed by 80% alcohol.
- Place the tubes/plates with extracted RNA to be quantified in an ice box in amplification room/validation room/extraction room (Room no). If the RNA is frozen thaw on ice before quantification.
- Vortex and centrifuge all samples before use.
- Store RNA samples on ice throughout the analysis procedure.

Quantification and Qualification using the TapeStation device

Step 1- Prepare TapeStation reagents and equipment

- Retrieve the Agilent sample buffer and the Agilent ScreenTape device from 2 – 8 °C and allow 30 minutes to equilibrate to room temperature.
- Turn on the thermocycler and select the cycle with 72 °C hold.
- Launch the 2200 TapeStation Controller Software.
- If using a new ScreenTape device
    - 'Flick' ScreenTape devices to eliminate bubbles in the separation channel, which could interfere with sample loading.
    - Write the date the packet was opened.
- Open the lid of the instrument and load RNA ScreenTape device and loading tips into the 2200 TapeStation instrument.
- Select RNA Protocol (Eukaryotic RNA or Prokaryotic RNA).

Step 2 - Prepare sample for TapeStation analysis

- Prepare sample by mixing 5 µL RNA Sample Buffer with 1 µL RNA sample in Agilent strip tubes or 96 well plates. If strip tubes are used, seal with strip caps and if 96-well plates are used seal with foil.
- Spin down, then vortex using IKA vortexer and adaptor at 2000 rpm for 1 minutes.
- Spin down to position the sample at the bottom of the tube.
- Heat the samples at 72 °C for 3 minutes in the thermocycler.
- Place ladder and samples on ice for 2 min.
- Spin down to position the sample at the bottom of the tube.

Step 3 - Sample analysis on Tape station

- Load samples into the 2200 TapeStation instrument.
- Select the first well as e-ladder 2200 TapeStation Controller Software.
- Select the required samples on the 2200 TapeStation Controller Software and label the wells in the e-gel from well B1 onward.
- Click Start and specify a filename with which to save your results.

Step 4 - Interpretation of Tapestation data

- Scale the e-gel image 'to sample' (Figure 1).
- Take note of the RIN^e score and the assigned concentration for each of your sample.
- Export the data to a folder renamed with the extraction ID.
- Intact total RNA will have sharp, clear 28S and 18S rRNA bands on the e gel. The 28S rRNA band should be approximately twice as intense as the 18S rRNA band (Figure 1, lane 1). Degraded RNA will have a smeared appearance, will lack the sharp rRNA bands, or will not exhibit the 2:1 ratio of high-quality RNA (Figure 1, lane 4).

- RIN^e is calculated at a scale from 1 to 10. A high RIN^e indicates highly intact RNA, and a low RINe a strongly degraded RNA sample.
- If the RIN^e score is >8, the quality of the sample is optimal
- If the RIN^e score is <8 but >=6, the quality is suboptimal
- If the RIN^e score is <=5, and the e gel lacks sharp rRNA bands re-extract RNA

- Record the RIN^e score and the assigned concentration [ng/µl] in the quantification tab in the extraction worksheet.

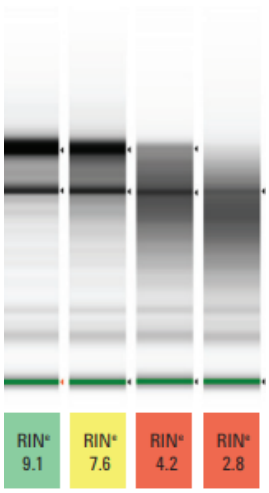


Figure 12. Example of the Tape station report with e-gel and RIN^e scores

Quantitation and qualification with the NanoDrop One UV-Vis spectrophotometer

Step 1 – Initializing the device and blanking

- Transfer RNA extracts in tubes/plates to an ice box.
- Turn on the NanoDrop One UV-Vis spectrophotometer in the PIC@PEQ lab (room No).
- Wait for the 'initialization process' to complete.
- Do NOT raise the sampling arm during initialization.
- Once the initialization is completed, use the touch screen to select 'RNA' as your application in the home screen.
- Raise the sampling arm and clean the upper and lower pedestals using a kimwipe.
- Pipette out 2 uL of nuclease free water into the lower pedestal and click 'blank'.

Step 2 – Measuring RNA samples

- Once the blanking is completed, clean the pedestals with a kimwipe.
- Pipette 2 uL of sample into the lower pedestals and click 'measure'.
- Clean both pedestals with a kimwipe between samples and repeat measuring each sample.
- Once the measuring is completed, connect a dedicated non-secure USB drive into the device using the USB port.
- Click on 'End experiment' and then 'export data'.
- Select '.csv' format and '.nanodrop' formats for exportation.
- Click the export data to the 'front USB' and create a new folder with a recognisable identifier (e.g.extraction ID).
- Select 'export data'.
- Remove the USB from the device.
- Clean both pedestals with a kimwipe and lower the sampling arm.

Step 3 - Interpretation of NanoDrop data

- Email the NanoDrop data into your DAWE email address and download the files to the QC folder with the extraction ID.

- Record the concentrations, A260/280 and A260/230 ratios in the QC tab of the respective extraction sheet.

- **Absorbance ratio at A260/280**
  - for RNA, a ratio of ~2.0 is generally accepted as "pure".
  - Low 260/280 ratios can be caused by residual phenol or other reagents from nucleic acid extraction or residual protein not removed during the extraction
- **Absorbance ratio at 260/230**
  - is an indication of contamination with other agents such as phenol or salts. The 260/230 values for "pure" nucleic acid are often higher than the respective 260/280 values. Expected 260/230 values are commonly in the range of 2.0–2.2.
  - Low 260/230 ratios are the result of a contaminant absorbing at 230 nm or less and can be caused by carbohydrate carryover, residual phenol, guanidine or other reagents from nucleic acid extraction.

- If the ratios are not within the expected range, try spinning down the samples 1 minute in high-speed to precipitate any magnetic beads that may have been carried over during the Maxwell extraction.
- You may see black beads precipitated at the bottom of the well.
- Transfer the supernatant to a fresh tube without disturbing the precipitate.
- Re-quantify using the NanoDrop.
- If the ratios are still not within the expected range, re-extract RNA.

Submitting sample RNA to external sequencing service provider (Ramaciotti centre of Genomics, NSW)

Step 1 – Online submission of samples

- Create a profile with Ramaciotti for sample submission.
- Request a quote from Ramaciotti for the number of samples that passed internal QC, flow cell and sequencing chemistry.
- Accept the quote.
- Generate a working order for the quote and get it approved by the supervisors and forward this to the Ramaciotti project manager.
- Ship samples within 30 days of accepting the quote.
- Log into the Ramaciotti sample submission portal and navigate to 'resources.
- Download the sample submission form for 96 well plates/ tubes as appropriately. To avoid upload errors, download a new copy of the submission form each time you submit.
- Fill the sample metadata required by Ramaciotti.
- Provide the DAWE email address for data delivery.
- Make sure the PEQ index numbers are de-identified for samples. (Use other sample IDs).
- Create a 'new project' in the Ramaciotti submission portal with the quote and work order IDs.
- Fill in the required information and in the description specify the flow cell type, number of samples, platform and service type.
- Upload the saved sample submission form.
- Your sample metadata should populate the window now.
- Sign the declaration and submit the samples and your submission will receive a Ramaciotti LIMS ID.

- Save the sample submission proof document to the QC folder of the respective RNA extraction.

Step 2 – Preparing plates/tubes for Ramaciotti

- Do not use screw cap or strip tubes, or seal tubes with parafilm, foil or tape when submitting the samples.
- Use cryo-pens or cryo dots with permanent markers when labelling tubes/plates.

Submitting samples in tubes

- Projects with less than 10 samples can be submitted in 1.5 ml tubes. Clearly label your tubes on both the top and side of the tube, with the corresponding sample LIMS ID.

Submitting samples in plates

- Projects with 10 or more samples must be submitted in 96-well PCR plates (clear wells, semi or fully skirted).
- Label plates with the unique LIMS ID issued during online submission.
- Plates must also be labelled with the container name e.g. plate 1, 2, 3.
- Make sure that plates are well sealed first using an aluminium sealing film and then a clear adhesive film. Aluminium seals are compatible with cold storage and the clear seal is a secondary protectant to the aluminium seal.

- Print out the sample submission confirmation and put inside a Ziplock bag.
- Put your tubes/plates inside a separate Ziplock bag.
- Tape the two Ziplock bags together to make sure they stay together.
- Wrap the everything using bubble wrap and write 'Top' to indicate the storage direction and the LIMS ID on bubble wrap.
- Freeze the wrapped samples overnight before the courier arrives.
- Request quote from a courier service to send samples on dry ice.
- Organize the courier to pickup the samples to be delivered to Ramaciotti early during the week. Avoid mid or late week courier pick-up to make sure the samples are delivered to Ramaciotti by mid-week to allow the samples to be attended appropriately.
- Upon the receipt of the samples Ramaciotti will notify via email. Ramaciotti will also perform internal QC on the samples and seek advice where sample quality is suboptimal.

- **Waste management and other safety precautions**

- **Chemicals –** The handling of solvents, samples and reagents can hold health and safety risks. When using/handling the ScreenTape device and working with these substances observe appropriate safety procedures (for example by wearing goggles, safety gloves and protective clothing). The volume of substances should be reduced to the minimum required for the analysis. Risk assessments are contained in the specimen laboratory for all reagents in the kit. Be sure to familiarise yourself with the risks each pose, the PPE required, and the method of disposal, method of clean up in case of spills. Use the spill kit if required to contain any large spills.

- **Temperature-** Temperatures are utilised in this method can burn the skin. When handling cold frozen goods, use cold protective gloves.

- **Mechanical hazards**- Mechanical hazards exist due to moving parts in the instruments used. Ensure you have training in the use of any instrument you are using. Mechanical hazards exist in the TapeStation instrument, centrifuge, thermal cycler, and vortex. Never place yourself in harm's way, be careful of moving parts.

- **Electrical hazards-** Electrical hazards exist with many of the apparatus used in the method. These include, fridges and freezers, Tapestation instrument, centrifuge, thermocycler, Nanodrop instrument, vortex. Report any defects in instrumentation, do not use instrumentation with obvious defects, like exposed wires. Instrumentation should be off at the power point unless required. Attempt to keep liquids away from power cables, power boards, and instrumentations. Have a liquid spill in or around instrumentation cleaned with OHS direction. With the power off on the instrument. Do not put yourself at risk, report hazard to your manager.

- **Environmental hazards –** Potential environmental hazards exist in specimens being processed. Specimen and specimen by-products need to be disposed of as quarantine waste, in a biohazard bag, and placed in a large yellow bin for commercial disposal, or autoclaved before deep burial. Clean up bench space with or RNASEZap followed by 80% alcohol. When using spray bottle, spray directly into towel, away from face so as not to breathe in aerosols.

- **Related documents**
  - Technical manual for NanoDrop
    https://assets.thermofisher.com/TFS-Assets/CAD/manuals/NanoDrop-2000-User-Manual-EN.pdf#page=6&zoom=100,0,0
  - User guide for Agilent TapeStation
    https://www.agilent.com/cs/library/usermanuals/public/RNA_QuickGuide.pdf
  - Ramaciotti sample submission:
    https://www.ramaciotti.unsw.edu.au/sample-submission/how-submit-samples

- **Rationale for change of SOP version**
  - N/A

	Compiled by	Examined by	Approved by	Replaced	New version
Name				Code:	Code:
Date					
Signature					
Laboratory area:		No of copies:	Reason for change:		

Appendix 26: Submitted NGS protocol for the diagnosis of all known viruses and viroids in horticulture crops to be submitted to the Subcommittee on Plant Health Diagnostics (SPHD)

Submitted to the SPHD for review on 12 January 2022.

[INSERT Image of plant pest/disease in colour]

Prepared by:

Dr Marie-Emilie Gauthier and

Dr Roberto Barrero Gumiel

12/01/2022

Queensland University of Technology

Mailing address: P Block, Gardens Point Campus, 2 George St, QLD 4000 | GPO Box 2434, Brisbane, QLD 4001:
Ph #: 07 3138 4027
Fax. #: 07 3138 4097

1 INTRODUCTION

This diagnostic protocol provides a guideline for the detection and identification of plant viruses and viroids that present a biosecurity concern (including endemic pathogens of interest by industries), using small RNA (sRNA) sequencing technology.

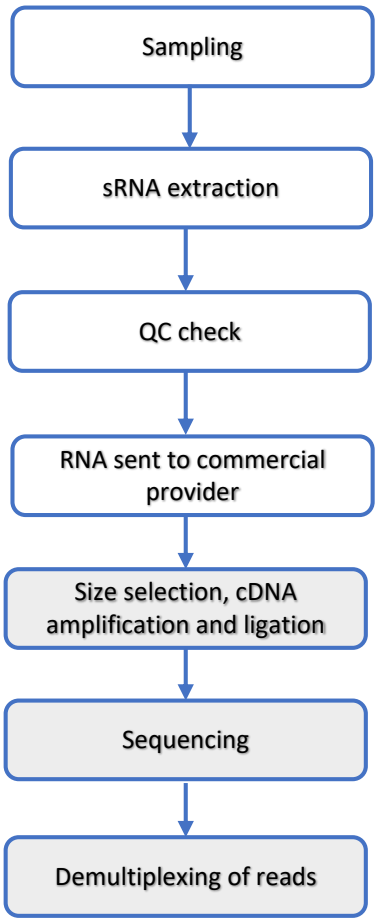
2 TAXONOMIC INFORMATION OF VIRUSES AND HOSTS

A list of the viruses and viroids present in specific plant hosts, are listed in **Table 1**. The hosts targeted that harbour these pathogens are:

- Avocado (Persea spp.)
- Blueberry and Cranberry (Vaccinium spp.)
- Cassava (Manihot spp.)
- Citrus (Citrus spp.)
- Clonal grasses
- Cotton (Gossypium spp.)
- Currants and Gooseberry (Ribes spp.)
- Elms (Ulmus, Planera and Zelkova spp.)
- Figs (Ficus carica)
- Grains (barley, maize, millet, oats, rye, sorghum, triticale and wheat)
- Grapevines (Vitis spp.)
- Hazelnuts (Corylus spp.)
- Hibiscus (Hibiscus spp.)
- Hops (Humulus spp.)
- Kiwifruit (Actinidia spp.)
- Legumes (chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean)
- Mulberry (Morus spp.)
- Olives (Olea spp.)
- Papaya (Carica papaya)
- Passionfruit (Passiflora spp.)
- Peanuts (Arachis spp.)
- Pome (Malus, Pyrus and Cydonia species)
- Potato (Solanum tuberosum)
- Raspberry and Blackberry (Rubus spp.)
- Rice (Oryza spp.)
- Stonefruit (Prunus spp.)
- Strawberry (Fragaria spp.)
- Sunflower (Helianthus spp.)
- Sweet potato (Ipomoea batatas)
- Taro (Colocasia spp.)
- Walnuts (Juglans spp.)

3 DETECTION

3.1 Diagnostic flow chart. The steps coloured in white are performed in-house, the steps coloured in blue are outsourced.



3.2 Symptoms

This protocol is for the detection and identification of viruses and viroids post-entry quarantine and thus can be applied to both symptomatic or asymptomatic specimens that require screening, and the plants tested can be tested regardless of whether they show symptoms.

3.3 Sample collection

The sampling will be dependent both on the host and the virus/viroid tested. Plant tissue targeted can be leaf, mid-rib, petiole, blade and spathe. It is recommended to undertake sampling during spring or early summer when temperatures are mild to warm (18-28 °C) (Glasa *et al*., 2003).

- **Personel pre-qualifications:**

All personnel sampling should be inducted to and have access to the glasshouses. New staff members are required to be trained by and supervised by a trained SSG molecular biologist/ plant pathologist prior to attempting the procedure on their own.

- **Equipment and materials:**

Glove box, marker pen, surgical scalpel blades as adequate, resealable plastic bags, Ice box

- **Procedure:**

Step1 - Preparations before sampling

- Make a list of plants that you want to sample and ensure all plants are in active growth stage at the time of sampling
- Write down or print the index numbers and their location in the glasshouse.
- Label the resealable bags with the index numbers, plant species and date of sampling

- Put adequate gloves, blades, marker pen, labelled bags, and the list of plants you prepared into an ice box.
- Once you are at the glasshouse premises follow the standard protocols before entering each room.
- Locate your plant of interest.

Step2 - Sampling

- Always sample from the index plant and sample from symptomatic areas of the plant if symptoms are visible.
- Wear a fresh pair of gloves.
- Except in the case of below exceptions, for each plant, pluck three leaves (both blade and petiole) from three parts of the plant from different maturity stages (top, bottom, middle and alternate sides).
    - Exceptions
        - Sessile leaves – pluck leaf blade
        - Monocots and other plants with leaves less than 10 – pluck 2 leaves from different areas
        - For plants with light leaves – sample adequately to make sure you have enough material (>50mg) for nucleic acid extractions.
- Use a scalpel blade where necessary and dispose the blades in between plants. (see waste management below)
- Once a plant has been adequately sampled, put the leaf material into the respectively labelled resealable bag and keep in the ice box until you finish sampling.
- Dispose the used gloves properly (see wate management below).
- Use a fresh pair of gloves each time you sample a new plant.
- Once you finish sampling, take the ice box to the extraction room and store the labelled resealable bags with samples at 4 ºC.

3.4 RNA extraction and sequencing

- **Personnel pre-qualifications:**

All personnel performing RNA extractions should be inducted to and have access to the Extraction room. New staff members are required to be trained by and supervised by a trained SSG molecular biologist/ plant pathologist prior to attempting the procedure on their own.

- **Equipment and materials**

Labelled resealable bags with plant material sampled according to HTS_SOP1, Weighing boats, gloves, surgical scalpel blades, porcelain tile, electronic balance, 2 ml Eppendorf tubes, marker pen, Stainless steel beads, Qiagen Tissuelyser LT, Promega Maxwell RSC Simply RNA Tissue Kit, Maxwell instrument, Fridge and freezer, laminar flow, micropipettes, centrifuge, vortex, Kim wipes, 80% ethanol, Nuclease free water, PVP, DTT

- **Procedure**

Step1 - Prepare the extraction worksheet

- Create a new extraction worksheet at SSG\Molecular pathology\Extraction Worksheets and enter the details of the samples you want to extract RNA from.
- Print out the extraction work sheet and paste in the respective lab book in the extraction room.

Step 2 - Clean the workbench and prepare 2 ml Eppendorf tubes for grinding the samples

- Clean your work bench with 80% ethanol.
- Take one of 2 ml round bottom Eppendorf tube per sample and place in a tube rack.
- Label the lid and the side of with the Eppendorf tubes with the corresponding plant index number in the extraction worksheet using a marker pen.
- Place **7mm stainless steel beads** into the Eppendorf tubes.
- Use one bead per tube for soft plant tissues and 2 beads per tube for petioles and other hard plant tissues.

Step 3 - Weigh 50mg of sample tissue into the tubes and freeze.

- Turn on the laminar flow.
- Wipe the inside of the laminar flow and the balance with 80% ethanol.
- Locate the samples in labelled resealable bags in the extraction room.
- If the samples were temporarily stored in the fridge take them out and move to the laminar flow.
- Wear a fresh pair of gloves.
- Cover a white porcelain tile with a wipe and use as your cutting board.
- Turn on the balance. Take a sterile weighing boat and place it on the balance. Press Tare.
- Dissect a combination of leaf blade and midrib from all leaves present in the bag using a sterile scalpel blade to weigh a maximum of 50 mg per sample.
    - Exceptions – For Rubus spp. use only the petiole and make sure you avoid the leaf blade.

- Cut the weighed tissue into smaller pieces that can fit in the lid of Eppendorf tube.
- Transfer the cut pieces of leaf tissue into the respectively labelled Eppendorf tube with bead/s using the static of the scalpel blade.
- Discard gloves, weighing boat, scalpel and the wipe covering the tile appropriately in between samples.
- Wipe the working space and balance in between samples.
- Once the weighing is completed, transfer the Eppendorf tubes with samples and beads into the Tissuelyser rotor, and then place sample and rotor into the -75°C freezer for 30 minutes.

Step 4 – Prepare Maxwell reagents and elution tubes

- Use either a pre opened **Promega Maxwell RSC Simply RNA Tissue Kit** or open a new kit depending on how many cartridges you need.
- If you are opening a new kit record the batch No. and the date you opened the kit in the logbook.
- **Prepare homogenisation solution**
    - Check whether there is a bottle of prepared homogenisation solution in the fridge.
    - Check the date of preparation and if the homogenisation solution is expired (>1 month old), discard the bottle and prepare a fresh batch.
    - Add 2.5% PVP (0.75g dry PVP), to **Promega Maxwell RSC Simply RNA Tissue Kit** 30mL Homogenization solution Bottle
    - Mix vigorously
    - Add 600uL of 1-thioglycerol to the Homogenisation solution and mix.
    - Write down the date in the bottle if you made a fresh batch.
- **Prepare DNAse 1**
    - DNase 1 is dyed blue and stored at -20°C. Thaw as required.
    - You will need 5µl DNase 1 per sample. Check whether you have enough.
    - To prepare a fresh DNase 1 batch when you open a kit, follow the instructions below.
    - Add 275µl nuclease free water to **Promega Maxwell RSC Simply RNA Tissue Kit** DNase I. Mix/Invert/turn tube to dissolve, do not vortex.
    - Add 5µl of **Promega Maxwell RSC Simply RNA Tissue Kit** blue dye.
    - Store solution at -20°C.
    - Thaw as required, and then refreeze.
- **Label 0.5 ml elution tubes**
    - Label one elution tube per sample from the **Promega Maxwell RSC Simply RNA Tissue Kit** with the sample index number on both the top and the side of the tube.

Step 5 – Prepare the frozen plant tissue for RNA extraction

- Get the cold Rotor with the frozen tissues from the -75°C freezer and insert into the **Tissuelyser**.
- Operate Tissuelyser for 2 minutes at 50 Hz. Check that a powder has formed, if not, repeat the freezing step and lyser step.
- Working as quickly as possible, add 600µl of **Homogenisation solution preparation** (in 4°C fridge), into the Eppendorf tube with ground tissue.
- Add 300µl **lysis buffer** from **Promega Maxwell RSC Simply RNA Tissue Kit**
- Mix by vortexing.
- Spin sample tube for 5 min at 20 000g.
- If gluggy take off supernatant and spin supernatant a second time.

Step 6 – Prepare the Maxwell instrument for RNA extraction

- Turn on the touch screen, button top right and turn on the Power at the back of the instrument.
- Press the Maxwell icon on the screen.
- Press the Maxwell icon again on the next screen, (A self-test will occur here)
- Press the start button.
- Choose the extraction program, **AS1340** for RNA extraction
- Press proceed.
- Press proceed again.
- Then ‘ok’, the door will open.
- Take out the tray and rest it on the bench.
- Load the tray with the number of Maxwell cartridges required, one per sample, press down until the cartridges click in.

- Remove the sealing foil from each of the cartridges aseptically.
- Aseptically insert the **pipette plungers** into the **front well** of each cartridge, (the well closest to the tray holes).
- Aseptically place **elution tubes** into the holes in the **tray** for each cartridge. The lids must be open, opening away from the cartridges.
- Aliquot **100µl** of **nuclease free H_20** from the Promega Maxwell RSC Simply RNA Tissue Kit into each elution tube.
- Add to **5µl** of thawed **Blue DNase I** to the **well with the yellow reagent** in each cartridge, (well #4).
- Finally add **500µl** of **supernatant of your sample preparation** to the **back well**, (the big well), of the corresponding cartridge.
- The tray is ready, it can be placed back onto the Maxwell. Orient the tray with the elution tubes at the front.
- Do a visual scan of your tray as you put it in and check whether
    - you have H_20 in the elution tubes
    - the tubes open
    - your sample is aliquoted in the back well
    - you add blue DNase I into yellow reagent well
    - you put in the plunger
- If all is ok, Press start. (The instrument door will close, the instrument will initialise and the run will start. The progress of the run can be seen on the bar on the computer screen)

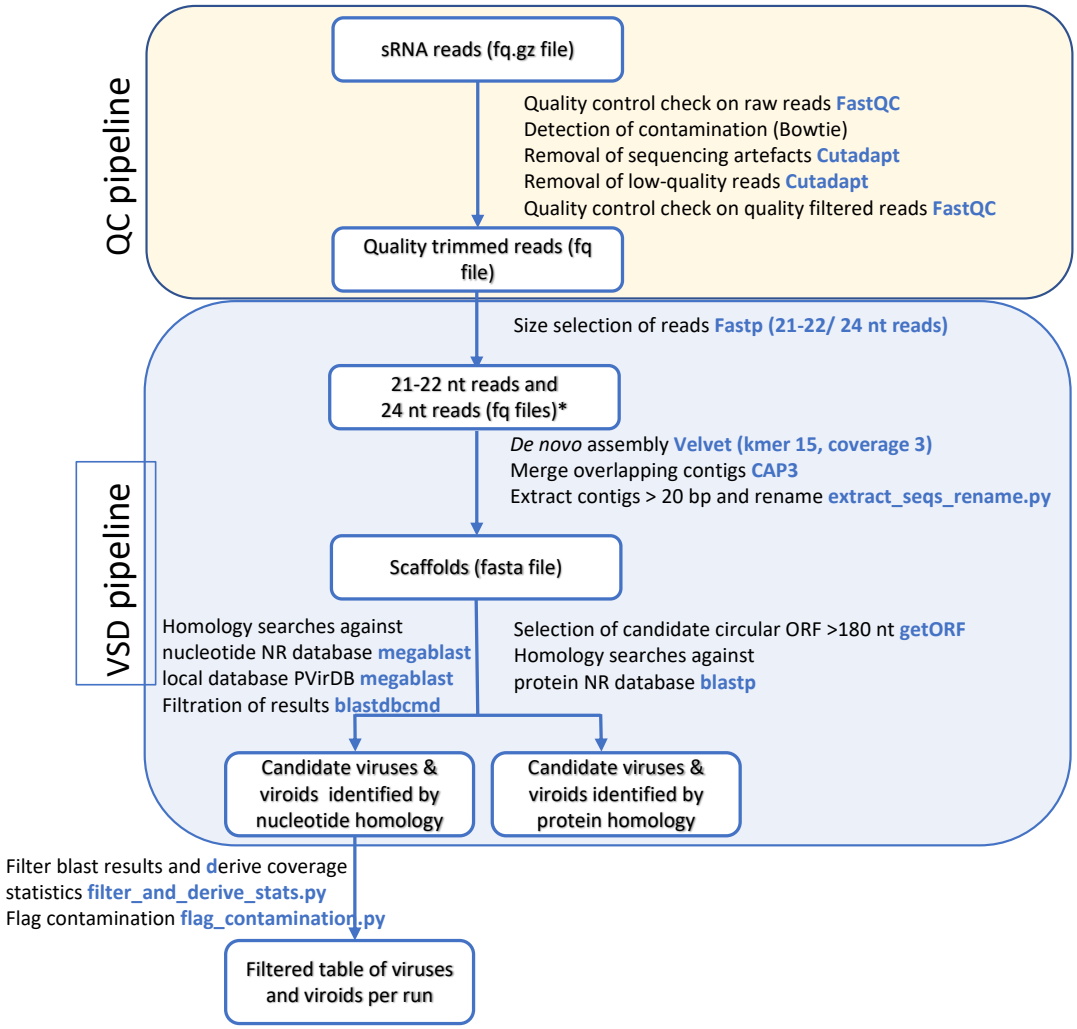
Step 7 - Post-RNA extraction procedure

- Once the instrument has completed, the bar on the screen will read 100%
- Press 'Open'
- Lift the tray out of the Maxwell and place on the bench.
- Close the lids of the elution and move to a storage box labelled appropriately.
- Store the extracts
    - If you are performing quantification on the same day, store the box in the fridge (4°C) temporally.
    - If the quantification will be done on a later day store the box in the -75°C freezer in the extraction room.
    - Note: if extracts contain too many beads, your extract will look murky, a strong magnet or centrifuging the extract, will help. Separate the supernatant from the beads and store
- Unclick each Maxwell cartridge out of the tray gently and place into a biohazard waste bin.
- Wipe the tray with 80% ethanol using a Kimwipe.
- Place tray back into Maxwell.
- Exit the Extraction program by arrowing, on the screen, back twice.
- You will reach a screen that has a Sanitise option. Press 'Sanitise'.
- Sanitation will take 1 minute.
- On completion of sanitation, exit the Maxwell program with the Arrows and return to the home screen.
- Turn off instrument computer, (Top right button)
- Turn off the instrument, (button at the back of the Maxwell)

The quality of the extracted RNA is assessed using Agilent Bioanalyser to measure both RNA integrity and level of degradation. A RIN score of 6 or above is recommended. Downstream sample preparation and sequencing is outsourced. The standard operating procedure of the Australian Genome Research Facility (AGRF) is followed to prepare and sequence small RNAs in batches of 48 barcoded samples. Complementary DNA amplification is performed using the NEB small RNA library kit and the Pippin prep (Sage Science) for sRNA size selection. Dual indices are used. 300-350 pM is loaded onto NovaSeq flow cells.

4 BIOINFORMATICS METHOD FOR IDENTIFICATION

The processing of the sequenced data includes quality control, read trimming, *de novo* assembly, and searches against reference database. The pipeline steps are described in detail in the flow chart below:



4.1 Quality control checks

Basic QC statistics for the raw reads are retrieved using FastQC (http://dx.doi.org/10.5772/intechopen.69250). Parameters to take into consideration are the amount of reads per sample, general read and base qualities, mean read length, read length distribution, G + C content, presence of unclipped adapters and unexpected repetitive sequences. This general overview will indicate if library construction and sequencing were properly performed, or if errors like contaminants, poor ribosomal RNA depletion or low sequencing output will demand a new round of experiments.

Quality trimming of raw reads for small RNA-seq libraries are recommended to be performed using cutadapt.

The sequenced virus reads are expected to be interspersed with background reads that are not related to the target, such as host sequences (e.g. ribosomal sequences, mitochondria, chloroplasts, miRNAs), phage sequences and/or contaminants (environmental or lab-derived). Therefore, the quality-trimmed reads can be mapped to these different categories of genetic material using tools such as Bowtie to evaluate the extent to which they contribute to the sequencing effort in a given run. The amount of leftover reads not mapping to the background permits to give an indirect indication of the actual number of reads that are likely to contain the viral sequences. If background sequences are overrepresented in a sequencing run, virus reads might be too few to permit successful detection and cause the sequencing run to fail. This overrepresentation might be an indication of inadequate size selection of sequences or RNA degradation.

4.2 Assembly and homology search

Target identification is performed using the bioinformatics tool called Virus Surveillance and Diagnosis (VSD) implemented into Nextflow. The current version of VSD is open-access, and builds upon an older framework of the tool that was developed using Yabi (34). The input required by the pipeine is the quality-trimmed FASTQ file. The bioinformatics pipeline relies on the *de novo* assembler Velvet 1.2.10, the scaffolding tools CAP3 (117) and data filtering and manipulation uses getORF (EMBOSS) (http://emboss.open-bio.org/wiki/Appdocs), as well as in-house python, perl,

python and bash scripts. The *de novo* assembly strategy uses kmer 15 and coverage of 3 which has been shown to maximise virus contig recovery. The homology search can be performed against NT or a local database of choice (here our own PVirDB).

It is recommended to install VSD using conda:

```
#conda can be installed following the steps described here:
https://genomics.sschmeier.com/ngs-tools/index.html

#All required tools to run VSD can be specified in a single yml file:

name: vsd-0.2
channels:
  - bioconda
  - conda-forge
  - defaults
  - r
dependencies:
  - blast=2.11.0
  - cap3=10.2011
  - spades=3.14.0
  - emboss=6.6.0
  - openjdk=8.0.152
  - fastp=0.20.1
  - biopython=1.76
  - numpy=1.16.5
  - matplotlib=2.2.3
  - velvet=1.2.10

#Create and activate conda environment
conda env create -f vsd.yml
conda activate vsd-0.2
```

You need to provide an index file for the pipeline specifying: sample name, the full path to the filtered fastq files and the size range to select for performing *de novo* assembly on each line. This index file needs to be located in the same location where you will run the nextflow run command (see below) or you can specify its location by giving the full path to the nextflow run command.

```
sampleid,samplepath,minlen,maxlen
MT222,/work/hia_mt18005/diagnostics/2021/6_AGRF_CAGRF21046430_H5VVMDRXY/fastp/MT222_H5VVMDR
XY_ATCACG_L001_R1_trimmed.fastq,21,22,K9-21
MT223,/work/hia_mt18005/diagnostics/2021/6_AGRF_CAGRF21046430_H5VVMDRXY/fastp/MT223_H5VVMDR
XY_CGATGT_L001_R1_trimmed.fastq,21,22,K9-21
MT224,/work/hia_mt18005/diagnostics/2021/6_AGRF_CAGRF21046430_H5VVMDRXY/fastp/MT224_H5VVMDR
XY_TTAGGC_L001_R1_trimmed.fastq,21,22,K9-21
```

This is a typical pbs script example to run VSD on an HPC server:

```
#!/bin/bash -l
#PBS -N NextflowVSD
#PBS -l select=1:ncpus=2:mem=6gb
#PBS -l walltime=3:30:00
#PBS -j oe

cd $PBS_O_WORKDIR
module load java
NXF_OPTS='-Xms1g -Xmx4g'
nextflow run vsd/main.nf --blast_nt_db /work/hia_mt18005/blastDB/ -resume
```

This script can be run using:

```
qsub nextflow.pbs
```

The Nextflow pipeline creates a folder with multiple subfolders and under each of them, a separate folder with the sample name will include the results of each step tht is collected by the pipeline. The subfolders match each pipeline step:

- 01_readprocessing
- 02_velvet
- 03_cap3
- 04_getord

- 05_blastn
- 06_blastp
- 07_blastdbcmd
- 08_BlastTools

The blastn summary results located under 08_BlastTools called summary_*SampleName*_velvet_21-22nt_megablast_vs_NT_top5Hits_virus_viroids_final.txt (see **Table 2**) can then be filtered to retain the top virus and viroid hit and derive additional coverage statistics. A final script can be run on the filtered outputs to flag potential contaminants at the run-level.

5 CONTACTS FOR FURTHER INFORMATION

A/Professor Roberto Barrero | Senior Bioinformatics Solutions Architect | eResearch Office, DRI | QUT
P Block, Gardens Point Campus, 2 George St, Brisbane, QLD 4000 | GPO Box 2434, Brisbane, QLD 4001
t: +617 3138 4027 | f: +617 3138 4097 | email: roberto.barrero@qut.edu.au

6 ACKNOWLEDGEMENTS

We acknowledge our colleagues from the Department of Agriculture, Water and Environment (DAWE) for their collaboration to advance HTS diagnosis for viruses and viroids.

7 REFERENCES

7.1 Useful references for further information

1. Soliman T, Mourits MCM, Lansink AGJMO, van der Werf W. Quantitative economic impact assessment of an invasive plant disease under uncertainty - A case study for potato spindle tuber viroid (PSTVd) invasion into the European Union. Crop Prot. 2012;40:28-35.
2. Adams I, Fox A. Diagnosis of Plant Viruses Using Next-Generation Sequencing and Metagenomic Analysis. In: Wang A. ZX, editor. Current Research Topics in Plant Virology2016. p. 323-35.
3. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, et al. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol. 2009;10(4):537-45.
4. Al Rwahnih M, Daubert S, Golino D, Rowhani A. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology. 2009;387(2):395-401.
5. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, et al. 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. 2009;388(1):1-7.
6. Roossinck MJ, Martin DP, Roumagnac P. Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology. 2015;105(6):716-27.
7. Wu Q, Ding SW, Zhang Y, Zhu S. Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. Annu Rev Phytopathol. 2015;53:425-44.
8. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971;35(3):235-41.
9. Flores R, Gago-Zachert S, Serra P, Sanjuán R, Elena SF. Viroids: survivors from the RNA world? Annu Rev Microbiol. 2014;68:395-414.
10. Reuter JA, Spacek DV, Snyder MP. High-Throughput Sequencing Technologies. Mol Cell. 2015;58(4):586-97.
11. Miozzi L, Pantaleo V, Burgyan J, Accotto GP, Noris E. Analysis of small RNAs derived from tomato yellow leaf curl Sardinia virus reveals a cross reaction between the major viral hotspot and the plant host genome. Virus Res. 2013;178(2):287-96.
12. Zhu H, Guo HS. The role of virus-derived small interfering RNAs in RNA silencing in plants. Sci China Life Sci. 2012;55(2):119-25.
13. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344-9.
14. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009;324(5924):218-23.
15. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. P Natl Acad Sci USA. 2010;107(1):139-44.
16. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326(5950):289-93.
17. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. Mol Cell. 2011;44(4):667-78.
18. Meissner A, Mikkelsen TS, Gu HC, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766-U91.

19. Cokus SJ, Feng SH, Zhang XY, Chen ZG, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452(7184):215-9.
20. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, et al. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods. 2006;3(7):503-9.
21. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213-+.
22. Wan Y, Qu K, Ouyang ZQ, Kertesz M, Li J, Tibshirani R, et al. Genome-wide Measurement of RNA Folding Energies. Mol Cell. 2012;48(2):169-81.
23. Ding YL, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014;505(7485):696-+.
24. Buenrostro JD, Araya CL, Chircus LM, Layton CJ, Chang HY, Snyder MP, et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nat Biotechnol. 2014;32(6):562-+.
25. Sephton CF, Cenik C, Kucukural A, Dammer EB, Cenik B, Han YH, et al. Identification of Neuronal RNA Targets of TDP-43-containing Ribonucleoprotein Complexes. J Biol Chem. 2011;286(2):1204-15.
26. German MA, Pillay M, Jeong DH, Hetawal A, Luo SJ, Janardhanan P, et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol. 2008;26(8):941-6.
27. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30(3):265-+.
28. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics. 2015;14(2):130-42.
29. Barrero RA, Guerrero FD, Moolhuijzen P, Goolsby JA, Tidwell J, Bellgard SE, et al. Shoot transcriptome of the giant reed, Arundo donax. Data Brief. 2015;3:1-6.
30. Sun F, Fang P, Li J, Du LL, Lan Y, Zhou T, et al. RNA-seq-based digital gene expression analysis reveals modification of host defense responses by rice stripe virus during disease symptom development in Arabidopsis. Virol J. 2016;13.
31. Kohlhase DR, O'Rourke JA, Owen MDK, Graham MA. Using RNA-seq to characterize responses to 4-hydroxyphenylpyruvate dioxygenase (HPPD) inhibitor herbicide resistance in waterhemp (Amaranthus tuberculatus). Bmc Plant Biol. 2019;19.
32. Kabekkodu SP, Shukla V, Varghese VK, Adiga D, Vethil Jishnu P, Chakrabarty S, et al. Cluster miRNAs and cancer: Diagnostic, prognostic and therapeutic opportunities. Wires Rna. 2019.
33. Roossinck MJ. Deep sequencing for discovery and evolutionary analysis of plant viruses. Virus Res. 2017;239:82-6.
34. Barrero RA, Napier KR, Cunnington J, Liefting L, Keenan S, Frampton RA, et al. An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. Bmc Bioinformatics. 2017;18.
35. Massart S, Chiumenti M, Jonghe K, Glover R, Haegeman A, Koloniuk I, et al. Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies. Phytopathology. 2019;109(3):488-97.
36. Roossinck MJ. Plant Virus Metagenomics: Biodiversity and Ecology. Annu Rev Genet. 2012;46:359-69.
37. Seifert KA. Progress towards DNA barcoding of fungi. Mol Ecol Resour. 2009;9:83-9.
38. Stackebrandt E, Goebel BM. A Place for DNA-DNA Reassociation and 16s Ribosomal-Rna Sequence-Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol. 1994;44(4):846-9.
39. Fung E, Hill K, Hogendoorn K, Glatz RV, Napier KR, Bellgard MI, et al. De novo assembly of honey bee RNA viral genomes by tapping into the innate insect antiviral response pathway. J Invertebr Pathol. 2018;152:38-47.
40. Kraberger S, Schmidlin K, Fontenele RS, Walters M, Varsani A. Unravelling the Single-Stranded DNA Virome of the New Zealand Blackfly. Viruses-Basel. 2019;11(6).
41. Li LL, Victoria JG, Wang CL, Jones M, Fellers GM, Kunz TH, et al. Bat Guano Virome: Predominance of Dietary Viruses from Insects and Plants plus Novel Mammalian Viruses. J Virol. 2010;84(14):6955-65.
42. Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. Metagenomic Analysis of RNA Viruses in a Fresh Water Lake. Plos One. 2009;4(9).
43. Roossinck MJ, Martin DP, Roumagnac P. Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology. 2015;105(6):716-27.
44. Kutnjak D, Elena SF, Ravnikar M. Time-Sampled Population Sequencing Reveals the Interplay of Selection and Genetic Drift in Experimental Evolution of Potato Virus Y. J Virol. 2017;91(16).
45. Olmos A, Boonham N, Candresse T, Gentit P, Giovani B, Kutnjak D, et al. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. EPPO Bulletin. 2018;48(2):219-24.
46. Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, et al. Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. PLOS ONE. 2014;9(7):e102945.
47. Fox A. Reconsidering causal association in plant virology. Plant Pathology. 2020;69(6):956-61.
48. Hart SN, Therneau TM, Zhang YJ, Poland GA, Kocher JP. Calculating Sample Size Estimates for RNA Sequencing Data. J Comput Biol. 2013;20(12):970-8.
49. Guo Y, Zhao S, Li CI, Sheng Q, Shyr Y. RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. Cancer Inform. 2014;13(Suppl 6):1-5.

50. Zhao SL, Li CI, Guo Y, Sheng QH, Shyr Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. Bmc Bioinformatics. 2018;19.
51. Blainey P, Krzywinski M, Altman N. Points of significance: replication. Nat Methods. 2014;11(9):879-80.
52. Fang ZD, Cui XQ. Design and validation issues in RNA-seq experiments. Brief Bioinform. 2011;12(3):280-7.
53. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis (vol 17, 13, 2016). Genome Biol. 2016;17.
54. Liu YW, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3):301-4.
55. Ching T, Huang SJ, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. Rna. 2014;20(11):1684-96.
56. Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Rna. 2016;22(6):839-51.
57. Gu X. Statistical detection of differentially expressed genes based on RNA-seq: from biological to phylogenetic replicates. Brief Bioinform. 2016;17(2):243-8.
58. Gaafar YZA, Ziebell H. Comparative study on three viral enrichment approaches based on RNA extraction for plant virus/viroid detection using high-throughput sequencing. PLoS One. 2020;15(8):e0237951.
59. Maliogka VI, Minafra A, Saldarelli P, Ruiz-García AB, Glasa M, Katis N, et al. Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. Viruses. 2018;10(8).
60. Pecman A, Kutnjak D, Gutiérrez-Aguirre I, Adams I, Fox A, Boonham N, et al. Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches. Frontiers in Microbiology. 2017;8(1998).
61. Visser M, Bester R, Burger JT, Maree HJ. Next-generation sequencing for virus detection: covering all the bases. Virol J. 2016;13:85.
62. Al Rwahnih M, Dave A, Anderson MM, Rowhani A, Uyemoto JK, Sudarshana MR. Association of a DNA Virus with Grapevines Affected by Red Blotch Disease in California. Phytopathology. 2013;103(10):1069-76.
63. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. Bmc Mol Biol. 2006;7.
64. Johnson MTJ, Carpenter EJ, Tian ZJ, Bruskiewich R, Burris JN, Carrigan CT, et al. Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes. Plos One. 2012;7(11).
65. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, et al. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 2004;32(3).
66. Yi H, Cho YJ, Won S, Lee JE, Yu HJ, Kim S, et al. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. Nucleic Acids Res. 2011;39(20).
67. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10(7):623-+.
68. Johne R, Müller H, Rector A, van Ranst M, Stevens H. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends in Microbiology. 2009;17(5):205-11.
69. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. Plos One. 2011;6(11).
70. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12).
71. van Gurp TP, McIntyre LM, Verhoeven KJF. Consistent Errors in First Strand cDNA Due to Random Hexamer Mispriming. Plos One. 2013;8(12).
72. Gubler U, Hoffman BJ. A Simple and Very Efficient Method for Generating Cdna Libraries. Gene. 1983;25(2-3):263-9.
73. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: A SMART (TM) approach for full-length cDNA library construction. Biotechniques. 2001;30(4):892-7.
74. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012;40(1):e3.
75. Budowle B, Connell ND, Bielecka-Oder A, Colwell RR, Corbett CR, Fletcher J, et al. Validation of high throughput sequencing and microbial forensics applications. Investig Genet. 2014;5:9.
76. Massart S, Chiumenti M, De Jonghe K, Glover R, Haegeman A, Koloniuk I, et al. Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. Phytopathology. 2019;109(3):488-97.
77. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics. 2018;19(1):30.
78. Sena JA, Galotto G, Devitt NP, Connick MC, Jacobi JL, Umale PE, et al. Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. Scientific Reports. 2018;8(1):13121.
79. Wright ES, Vetsigian KH. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC genomics. 2016;17(1):876-.
80. Jiang LC, Schlesinger F, Davis CA, Zhang Y, Li RH, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21(9):1543-51.

81. Munro SA, Lund SP, Pine PS, Binder H, Clevert DA, Conesa A, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun. 2014;5.
82. Wong T, Deveson IW, Hardwick SA, Mercer TR. ANAQUIN: a software toolkit for the analysis of spike-in controls for next generation sequencing. Bioinformatics. 2017;33(11):1723-4.
83. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. Semin Thromb Hemost. 2019;45(7):661-73.
84. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. 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. 2020;221(Suppl 3):S292-s307.
85. Cho HH, Davis J, Li X, Smith KS, Battle A, Montgomery SB. High-Resolution Transcriptome Analysis with Long-Read RNA Sequencing. Plos One. 2014;9(9).
86. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 2015;16.
87. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015;33(7):736-42.
88. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009-+.
89. Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing. G3-Genes Genom Genet. 2013;3(3):387-97.
90. Chang Z, Wang ZJ, Li GJ. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. Plos One. 2014;9(4).
91. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science. 2010;327(5961):78-81.
92. Porreca GJ. Genome sequencing on nanoballs. Nat Biotechnol. 2010;28(1):43-4.
93. Lei R, Ye KX, Gu ZL, Sun XP. Diminishing returns in next-generation sequencing (NGS) transcriptome data. Gene. 2015;557(1):82-7.
94. Zheng Y, Gao S, Padmanabhan C, Li RG, Galvez M, Gutierrez D, et al. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology. 2017;500:130-8.
95. Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A. Evolution and ecology of plant viruses. Nat Rev Microbiol. 2019;17(10):632-44.
96. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.
97. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: A set of tools for quality control and analysis Of high-throughput sequence data. Methods. 2013;63(1):41-9.
98. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. Plos One. 2013;8(12).
99. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
100. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.
101. Lambert C, Braxton C, Charlebois RL, Deyati A, Duncan P, La Neve F, et al. Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection. Viruses. 2018;10(10).
102. Baizan-Edge A, Cock P, MacFarlane S, McGavin W, Torrance L, Jones S. Kodoja: A workflow for virus detection in plants using k-mer analysis of RNA-sequencing data. J Gen Virol. 2019;100(3):533-42.
103. Hohn T, Richert-Pöggeler KR, Staginnus C, Harper G, Schwarzacher T, Teo CH, et al. Evolution of Integrated Plant Viruses. In: Roossinck MJ, editor. Plant Virus Evolution. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 53-81.
104. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology. 2020;21(1):241.
105. Visser M, Bester R, Burger JT, Maree HJ. Next-generation sequencing for virus detection: covering all the bases. Virol J. 2016;13.
106. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15(3):R46.
107. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications. 2016;7(1):11257.
108. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn. 2018;20(1):4-27.
109. Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Brüggenwirth HT, Lekanne Deprez RH, et al. Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. Human Mutation. 2013;34(10):1313-21.
110. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19(5):455-77.

111. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-U130.
112. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-9.
113. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155.
114. Martin J, Bruno VM, Fang ZD, Meng XD, Blow M, Zhang T, et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. Bmc Genomics. 2010;11.
115. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909-U62.
116. Chouvarine P, Wiehlmann L, Moran Losada P, DeLuca DS, Tummler B. Filtration and Normalization of Sequencing Read Data in Whole-Metagenome Shotgun Samples. Plos One. 2016;11(10):e0165015.
117. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9(9):868-77.
118. Castro CJ, Ng TFF. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. Journal of computational biology : a journal of computational molecular cell biology. 2017;24 11:1071-80.
119. Ho T, Tzanetakis IE. Development of a virus detection and discovery pipeline using next generation sequencing. Virology. 2014;471:54-60.
120. Altschul S. Hot papers - Bioinformatics - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs by S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman - Comments. Scientist. 1999;13(8):15-.
121. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus : architecture and applications. Bmc Bioinformatics. 2009;10.
122. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60.
123. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24(16):1757-64.
124. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178-92.
125. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform. 2013;14(2):193-202.
126. Schmid-Burgk JL, Hornung V. BrowserGenome.org: web-based RNA-seq data analysis and visualization. Nat Methods. 2015;12(11):1001-.
127. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14(2):144-61.

Table 1 List of target pathogens and their host

African cassava mosaic begomovirus	Cassava (Manihot spp.)
African peanut clump pecluvirus	Clonal grasses
Agropyron mosaic potyvirus	Clonal grasses
Alfalfa mosaic virus	Potato (Solanum tuberosum), Hops (Humulus spp.)
Algerian navel orange virus	Citrus (Citrus spp.)
American hop latent virus (Hop strain)	Hops (Humulus spp.)
American plum line pattern ilarvirus	Stonefruit (Prunus spp.)
Andean potato latent virus	Potato (Solanum tuberosum)
Andean potato mottle virus	Potato (Solanum tuberosum)
Anthoxanthum latent blanching hordeivirus	Clonal grasses
Anthoxanthum mosaic potyvirus	Clonal grasses
Apricot latent foveavirus	Stonefruit (Prunus spp.)
Apricot latent ringspot nepovirus	Stonefruit (Prunus spp.)
Arabis mosaic nepovirus, Arabis mosaic virus (ArMV)	Olives (Olea spp.), Stonefruit (Prunus spp.), Strawberry (Fragaria spp.), Hops (Humulus spp.), Grapevines (Vitis spp.)
Arkansas fig virus 1	Figs (Ficus carica)
Arkansas fig virus 2	Figs (Ficus carica)
Arracacha virus B – Oca strain	Potato (Solanum tuberosum)
Artichoke Italian latent virus (AILV)	Grapevines (Vitis spp.)
Asian prunus foveavirus (1,2,3)	Stonefruit (Prunus spp.)
Bamboo mosaic potexvirus	Clonal grasses
Barley stripe mosaic (Barley stripe mosaic hordeivirus)	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
Barley yellow dwarf virus – rice infecting strain)-Giallume	Rice (Oryza spp.)
Bean common mosaic virus – PStV strain	Peanuts (Arachis spp.)
Beet curly top virus	Potato (Solanum tuberosum)
Bermuda grass etched-line marafivirus	Clonal grasses
Blackberry calico carlavirus	Raspberry and Blackberry (Rubus spp.)
Blackberry chlorotic ringspot ilarvirus	Raspberry and Blackberry (Rubus spp.)
Blackberry leaf mottle associated virus	Raspberry and Blackberry (Rubus spp.)
Blackberry virus E	Raspberry and Blackberry (Rubus spp.)
Blackberry virus S	Raspberry and Blackberry (Rubus spp.)
Blackberry virus Y potyvirus	Raspberry and Blackberry (Rubus spp.)
Blackberry yellow vein associated virus	Raspberry and Blackberry (Rubus spp.)
Blackcurrant reversion nepovirus (BRV)	Currants and Gooseberry (Ribes spp.)
Blackcurrant yellows	Currants and Gooseberry (Ribes spp.)
Blueberry latent spherical nepovirus (group C), Blueberry leaf mottle nepovirus,	Blueberry and Cranberry (Vaccinium spp.)
Blueberry leaf mottle nepovirus, Blueberry leaf mottle virus (BLMoV) New York strain	Blueberry and Cranberry (Vaccinium spp.), Grapevines (Vitis spp.)
Blueberry mosaic associated ophiovirus	Blueberry and Cranberry (Vaccinium spp.)
Blueberry necrotic ring blotch virus	Blueberry and Cranberry (Vaccinium spp.)
Blueberry red ringspot soymovirus	Blueberry and Cranberry (Vaccinium spp.)
Blueberry scorch carlavirus	Blueberry and Cranberry (Vaccinium spp.)

Blueberry shock ilarvirus, Blueberry shock ilarvirus (subgroup 3)	Blueberry and Cranberry (Vaccinium spp.)
Blueberry shoestring sobemovirus	Blueberry and Cranberry (Vaccinium spp.)
Blueberry virus A (unassigned)	Blueberry and Cranberry (Vaccinium spp.)
Broad bean mottle virus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Broad bean stain comovirus, Broad bean stain virus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Broad bean true mosaic comovirus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Brome mosaic bromovirus	Clonal grasses
Canary (reed) mosaic potyvirus	Clonal grasses
Carnation Italian ringspot tombusvirus	Stonefruit (Prunus spp.)
Cassava American latent nepovirus	Cassava (Manihot spp.)
Cassava brown streak ipomovirus	Cassava (Manihot spp.)
Cassava Colombian symptomless tentative potexvirus	Cassava (Manihot spp.)
Cassava common mosaic potexvirus	Cassava (Manihot spp.)
Cassava frogskin 'virus'	Cassava (Manihot spp.)
Cassava green mottle nepovirus	Cassava (Manihot spp.)
Cassava Ivorian bacilliform anulavirus	Cassava (Manihot spp.)
Cassava vein mosaic cavemovirus	Cassava (Manihot spp.)
Cassava virus X potexvirus	Cassava (Manihot spp.)
Cereal northern mosaic rhabdovirus	Clonal grasses
Cherry leaf roll nepovirus, Cherry leaf roll virus, Cherry leafroll virus (CLRV) – grape isolate	Blueberry and Cranberry (Vaccinium spp.), Currants and Gooseberry (Ribes spp.), Elms (Ulmus, Planera and Zelkova spp.), Hops (Humulus spp.), Kiwifruit (Actinidia spp.), Olives (Olea spp.), Pome (Malus, Pyrus and Cydonia spp.), Raspberry and Blackberry (Rubus spp.), Stonefruit (Prunus spp.), Walnuts (Juglans spp.), Hops (Humulus spp.), Grapevines (Vitis spp.),
Cherry mottle leaf trichovirus	Stonefruit (Prunus spp.)
Cherry rasp leaf nepovirus, Cherry rasp leaf virus, Cherry rasp leaf virus (Malus only)	Stonefruit (Prunus spp.), Raspberry and Blackberry (Rubus spp.), Pome (Malus, Pyrus and Cydonia spp.)
Cherry twisted leaf foveavirus	Stonefruit (Prunus spp.)
Cherry twisted leaf robigovirus	Stonefruit (Prunus spp.)
Citrus chlorotic dwarf associated virus	Citrus (Citrus spp.)
Citrus concave gum virus (Psorosis complex)	Citrus (Citrus spp.)
Citrus leaf blotch citrivirus	Citrus (Citrus spp.)
Citrus leaf rugose virus, Citrus leaf rugose ilarvirus (Psorosis complex)	Citrus (Citrus spp.)
Citrus leprosis rhabdovirus, Citrus leprosis virus	Citrus (Citrus spp.)
Citrus ringspot virus (Psorosis complex – Indian)	Citrus (Citrus spp.)
Citrus seed-borne virus	Citrus (Citrus spp.)
Citrus sudden death virus	Citrus (Citrus spp.)
Citrus tatter leaf capillovirus, Citrus tatter leaf virus, Citrus tatter leaf/stunt (Citrus tatter leaf capillovirus)	Citrus (Citrus spp.)
Citrus tristeza closterovirus, Citrus tristeza closterovirus – stem pitting strain, Citrus tristeza closterovirus - yellows strain, Citrus tristeza virus	Citrus (Citrus spp.)
Citrus variegation ilarvirus, Citrus variegation ilarvirus (Psorosis complex), Citrus variegation virus	Citrus (Citrus spp.)
Citrus vein enation virus	Citrus (Citrus spp.)
Citrus yellow mosaic badnavirus	Citrus (Citrus spp.)

Cocksfoot mild mosaic panicovirus	Clonal grasses
Cocksfoot streak potyvirus	Clonal grasses
Colocasia bobone disease-associated cytorhabdovirus	Taro (Colocasia spp.)
Cotton blue disease (Luteovirus [suspected])	Cotton (Gossypium spp.)
Cotton leaf curl virus	Hibiscus (Hibiscus spp.)
Crinkly leaf virus	Citrus (Citrus spp.)
Cynodon chlorotic streak rhabdovirus	Clonal grasses
Dasheen mosaic potyvirus (French Polynesian strain)	Taro (Colocasia spp.)
East African cassava mosaic begomovirus	Cassava (Manihot spp.)
East African cassava mosaic Cameroon begomovirus	Cassava (Manihot spp.)
East African cassava mosaic Malawi begomovirus	Cassava (Manihot spp.)
East African cassava mosaic Zanzibar begomovirus	Cassava (Manihot spp.)
East Asian passiflora potyvirus	Passionfruit (Passiflora spp.)
Eggplant mottled dwarf virus	Potato (Solanum tuberosum)
Elm mottle virus	Elms (Ulmus, Planera and Zelkova spp.)
Festuca leaf streak rhabdovirus	Clonal grasses
Festuca necrosis closterovirus	Clonal grasses
Fig fleck associated virus	Figs (Ficus carica)
Fig leaf mottle associated virus 1	Figs (Ficus carica)
Fig leaf mottle associated virus 2	Figs (Ficus carica)
Fig mild mottle associated virus	Figs (Ficus carica)
Foxtail mosaic (Foxtail mosaic potexvirus)	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat, Clonal grasses
Fragaria chiloensis latent virus	Strawberry (Fragaria spp.)
Geminiviruses	Potato (Solanum tuberosum)
Gooseberry vein banding associated badnavirus	Currants and Gooseberry (Ribes spp.)
Grapevine ajinashika virus (GAgV)	Grapevines (Vitis spp.)
Grapevine Anatolian ringspot virus (GARSV)	Grapevines (Vitis spp.)
Grapevine angular mosaic-associated virus (GAMaV)	Grapevines (Vitis spp.)
Grapevine asteroid mosaic associated virus (GAMV)	Grapevines (Vitis spp.)
Grapevine berry inner necrosis virus (GINV)	Grapevines (Vitis spp.)
Grapevine Bulgarian latent virus (GBLV)	Grapevines (Vitis spp.)
Grapevine chrome mosaic virus (GCMV)	Grapevines (Vitis spp.)
Grapevine deformation virus (GDefV)	Grapevines (Vitis spp.)
Grapevine fanleaf virus (GFLV)	Grapevines (Vitis spp.)
Grapevine leafroll associated virus (GLRaV – 6,7,10, 11)	Grapevines (Vitis spp.)
Grapevine line pattern virus (GLPV)	Grapevines (Vitis spp.)
Grapevine red blotch-associated virus (GRBaV)	Grapevines (Vitis spp.)
Grapevine red globe virus (GRGV)	Grapevines (Vitis spp.)
Grapevine Roditis leaf discolouration- associated virus (GRLDaV)	Grapevines (Vitis spp.)
Grapevine rupestris vein feathering virus (GRVFV)	Grapevines (Vitis spp.)
Grapevine Syrah virus 1, Grapevine syrah virus-I (GSyV-I)	Raspberry and Blackberry (Rubus spp.), Grapevines (Vitis spp.)

Grapevine Tunisian ringspot virus (GTRSV)	Grapevines (Vitis spp.)
Grapevine vein clearing virus	Grapevines (Vitis spp.)
Grapevine virus B (strains associated with corky bark) (GVB)	Grapevines (Vitis spp.)
Grapevine virus E (GVE)	Grapevines (Vitis spp.)
Grapevine virus F (GVF)	Grapevines (Vitis spp.)
Guinea grass mosaic potyvirus	Clonal grasses
High Plains tenuivirus	Clonal grasses
Holcus lanatus yellowing rhabdovirus	Clonal grasses
Holcus streak potyvirus	Clonal grasses
Humulus japonicus latent virus	Hops (Humulus spp.)
Impatiens necrotic spot virus	Potato (Solanum tuberosum)
Indian cassava mosaic begomovirus	Cassava (Manihot spp.)
Indian peanut clump potyvirus	Clonal grasses
Maize chlorotic mottle (Maize chlorotic mottle machlomovirus), Maize chlorotic mottle potyvirus	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
Maize dwarf mosaic (Maize dwarf mosaic potyvirus), Maize dwarf mosaic potyvirus	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
Maize mosaic rhabdovirus B	Clonal grasses
Maize rough dwarf fijivirus	Clonal grasses
Maize streak geminivirus	Clonal grasses
Maize white line mosaic (Maize white line mosaic virus)	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
Millet streak geminivirus	Clonal grasses
Miscanthus streak mastrevirus	Clonal grasses
Molinia streak sobemovirus	Clonal grasses
Moroccan watermelon mosaic potyvirus	Papaya (Carica papaya)
Mulberry green mosaic virus	Mulberry (Morus spp.)
Mulberry latent carlavirus	Mulberry (Morus spp.)
Mulberry mosaic leaf roll associated nepovirus	Mulberry (Morus spp.)
Mulberry mosaic virus	Mulberry (Morus spp.)
Mulberry ringspot nepovirus	Mulberry (Morus spp.)
Mulberry vein banding tospovirus	Mulberry (Morus spp.)
Mulberry yellow net vein virus	Mulberry (Morus spp.)
Myrobalan latent ringspot nepovirus	Mulberry (Morus spp.)
Oat necrotic mottle potyvirus	Stonefruit (Prunus spp.)
Olive latent 1 necrovirus, Olive latent 1 sobemovirus (only for olive latent virus1)	Clonal grasses
Olive latent 2 oleavirus, Olive latent 2 ourmiavirus (only for olive latent virus 2)	Olives (Olea spp.)
Olive latent 3 marafivirus	Olives (Olea spp.)
Olive latent ringspot nepovirus	Olives (Olea spp.)
Olive leaf yellowing associated closterovirus	Olives (Olea spp.)
Olive mild mosaic necrovirus	Olives (Olea spp.)
Olive semilatent virus	Olives (Olea spp.)
Olive vein yellowing associated potexvirus	Olives (Olea spp.)
Olive yellow mottling and decline associated virus	Olives (Olea spp.)

Panicum mosaic sobemovirus	Clonal grasses
Papaya apical necrosis rhabdovirus	Papaya (Carica papaya)
Papaya droopy necrosis rhabdovirus	Papaya (Carica papaya)
Papaya leaf curl geminivirus	Papaya (Carica papaya)
Papaya leaf distortion mosaic potyvirus	Papaya (Carica papaya)
Papaya lethal yellowing sobemovirus	Papaya (Carica papaya)
Papaya meleira virus	Papaya (Carica papaya)
Papaya mosaic potexvirus	Papaya (Carica papaya)
Papaya ringspot potyvirus, Papaya ringspot potyvirus – Type P	Papaya (Carica papaya)
Passiflora chlorosis potyvirus	Passionfruit (Passiflora spp.)
Passion fruit yellow mosaic tymoviridae	Passionfruit (Passiflora spp.)
Passionfruit crinkle potyvirus	Passionfruit (Passiflora spp.)
Passionfruit ringspot potyvirus	Passionfruit (Passiflora spp.)
Passionfruit severe leaf distortion begomovirus	Passionfruit (Passiflora spp.)
Passionfruit Sri Lankan mottle potyvirus	Passionfruit (Passiflora spp.)
Passionfruit vein clearing rhabdovirus	Passionfruit (Passiflora spp.)
Pea early browning tobravirus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Pea seed-borne mosaic virus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Peach mosaic trichovirus	Stonefruit (Prunus spp.)
Peach rosette mosaic nepovirus, Peach rosette mosaic nepovirus (group C), Peach rosette mosaic virus (PRMV),	Stonefruit (Prunus spp.), Blueberry and Cranberry (Vaccinium spp.), Grapevines (Vitis spp.)
Peanut bud necrosis virus (PBNV)	Peanuts (Arachis spp.)
Peanut clump (Peanut clump pecluvirus)	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
Peanut clump furovirus (PCV)	Peanuts (Arachis spp.)
Peanut stripe potyvirus (PStV)	Peanuts (Arachis spp.)
Peanut stunt cucumovirus (PSV)	Peanuts (Arachis spp.)
Pearl millet mosaic potyvirus	Clonal grasses
Pepino mosaic virus	Potato (Solanum tuberosum)
Petunia asteroid mosaic tombusvirus, Petunia asteroid mosaic virus, Petunia asteroid mosaic virus (PeAMV)	Stonefruit (Prunus spp.), Hops (Humulus spp.), Grapevines (Vitis spp.)
Plum pox potyvirus	Stonefruit (Prunus spp.)
Potato 14R virus	Potato (Solanum tuberosum)
Potato black ringspot nepovirus, Potato black ringspot virus	Potato (Solanum tuberosum)
Potato deforming mosaic virus	Potato (Solanum tuberosum)
Potato latent virus	Potato (Solanum tuberosum)
Potato mop-top virus	Potato (Solanum tuberosum)
Potato rough dwarf virus	Potato (Solanum tuberosum)
Potato virus M	Potato (Solanum tuberosum)
Potato virus S – Andean strain	Potato (Solanum tuberosum)
Potato virus T, Potato virus T tepovirus	Potato (Solanum tuberosum)
Potato virus U	Potato (Solanum tuberosum)
Potato virus V	Potato (Solanum tuberosum)
Potato virus X	Potato (Solanum tuberosum)

Potato virus Y	Potato (Solanum tuberosum)
Potato yellow dwarf virus	Potato (Solanum tuberosum)
Potato yellow vein virus	Potato (Solanum tuberosum)
Potato yellowing virus	Potato (Solanum tuberosum)
Psorosis complex	Citrus (Citrus spp.)
Raspberry latent virus	Raspberry and Blackberry (Rubus spp.)
Raspberry leaf blotch virus	Raspberry and Blackberry (Rubus spp.)
Raspberry leaf curl virus	Raspberry and Blackberry (Rubus spp.)
Raspberry ringspot nepovirus, Raspberry ringspot nepovirus – cherry strain, Raspberry ringspot nepovirus (RRSV), Raspberry ringspot virus, Raspberry ringspot virus (RpRSV) – grapevine strain	Raspberry and Blackberry (Rubus spp.), Stonefruit (Prunus spp.), Strawberry (Fragaria spp.), Strawberry (Fragaria spp.), Stonefruit (Prunus spp.), Currants and Gooseberry (Ribes spp.), Raspberry and Blackberry (Rubus spp.), Grapevines (Vitis spp.)
Raspberry vein chlorosis virus	Raspberry and Blackberry (Rubus spp.)
Red clover vein mosaic carlavirus	Legumes—chickpea, common bean, faba bean, field pea, lentil, lupin, mung bean and soybean
Rice black streaked dwarf fijivirus	Clonal grasses, Rice (Oryza spp.)
Rice dwarf virus	Clonal grasses, Rice (Oryza spp.)
Rice hoja blanca tenuivirus	Clonal grasses, Rice (Oryza spp.)
Rice stripe tenuivirus	Clonal grasses, Rice (Oryza spp.)
Rice tungro bacilliform badnavirus	Clonal grasses, Rice (Oryza spp.)
Rice tungro spherical waikavirus	Clonal grasses, Rice (Oryza spp.)
Rod-shaped viruses	Blueberry and Cranberry (Vaccinium spp.)
Rubus chlorotic mottle sobeomovirus	Raspberry and Blackberry (Rubus spp.)
Rubus yellow net virus	Raspberry and Blackberry (Rubus spp.)
Satsuma dwarf sadwavirus, Satsuma dwarf virus	Citrus (Citrus spp.)
Solanum apical leaf curling virus	Potato (Solanum tuberosum)
Sorghum mosaic potyvirus	Clonal grasses, Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat
South African cassava mosaic begomovirus	Cassava (Manihot spp.)
Sowbane mosaic virus (SoMV) – grape infecting strain	Grapevines (Vitis spp.)
Stocky prune cheravirus	Stonefruit (Prunus spp.)
Strawberry chlorotic fleck virus	Strawberry (Fragaria spp.)
Strawberry crinivirus 3	Strawberry (Fragaria spp.)
Strawberry crinivirus 4	Strawberry (Fragaria spp.)
Strawberry feather leaf	Strawberry (Fragaria spp.)
Strawberry latent C rhabdovirus	Strawberry (Fragaria spp.)
Strawberry latent ringpot nepovirus, Strawberry latent ringspot sadwavirus, Strawberry latent ringspot sadwavirus (SLRSV), Strawberry latent ringspot virus, Strawberry latent ringspot virus (exotic strain)	Strawberry (Fragaria spp.)
Strawberry leaf curl virus	Strawberry (Fragaria spp.)
Strawberry lethal decline	Strawberry (Fragaria spp.)
Strawberry mild yellow edge virus	Strawberry (Fragaria spp.)
strawberry mottle virus	Strawberry (Fragaria spp.)
Strawberry polerovirus 1	Strawberry (Fragaria spp.)
Strawberry pseudo mild edge carlavirus	Strawberry (Fragaria spp.)
Strawberry pseudo yellow mild edge virus	Strawberry (Fragaria spp.)

strawberry vein banding	Strawberry (Fragaria spp.)
Sunflower chlorotic mottle potyvirus	Sunflower (Helianthus spp.)
Sunflower mosaic potyvirus	Sunflower (Helianthus spp.)
Sweet potato caulimo-virus	Sweet potato (Ipomoea batatas)
Sweet potato chlorotic fleck virus	Sweet potato (Ipomoea batatas)
Sweet potato chlorotic stunt virus	Sweet potato (Ipomoea batatas)
Sweet potato latent virus	Sweet potato (Ipomoea batatas)
Sweet potato leaf curl virus	Sweet potato (Ipomoea batatas)
Sweet potato leaf speckling virus	Sweet potato (Ipomoea batatas)
Sweet potato mild speckling virus (C8)	Sweet potato (Ipomoea batatas)
Sweet potato ringspot virus	Sweet potato (Ipomoea batatas)
Sweet potato vein mosaic virus	Sweet potato (Ipomoea batatas)
Sweet potato virus 2	Sweet potato (Ipomoea batatas)
Sweet potato virus C6	Sweet potato (Ipomoea batatas)
Sweet potato yellow dwarf virus	Sweet potato (Ipomoea batatas)
Taro vein chlorosis rhabdovirus	Taro (Colocasia spp.)
Tobacco mosaic virus – potato strain	Potato (Solanum tuberosum)
Tobacco necrosis alphanecrovirus – plum strain, Tobacco necrosis virus (TNV) – grape strain, Tobacco necrosis virus (Hop strain),	Stonefruit (Prunus spp.), Potato (Solanum tuberosum), Raspberry and Blackberry (Rubus spp.), Grapevines (Vitis spp.), Hops (Humulus spp.), Hops (Humulus spp.)
Tobacco necrosis virus D	Strawberry (Fragaria spp.)
Tobacco rattle virus	Potato (Solanum tuberosum)
Tobacco streak virus – potato strain, Tobacco streak virus	Potato (Solanum tuberosum), Strawberry (Fragaria spp.)
Tomato black ring virus,Tomato black ring nepovirus (TBRV)	Potato (Solanum tuberosum), Raspberry and Blackberry (Rubus spp.), Stonefruit (Prunus spp.), Strawberry (Fragaria spp.), Currants and Gooseberry (Ribes spp.), Grapevines (Vitis spp.)
Tomato bushy stunt tombusvirus	Pome (Malus, Pyrus and Cydonia spp.), Strawberry (Fragaria spp.)
Tomato chlorosis virus	Potato (Solanum tuberosum)
Tomato infectious chlorosis virus	Potato (Solanum tuberosum)
Tomato leaf curl Albatinah begomovirus	Papaya (Carica papaya)
Tomato leaf curl New Delhi virus	Potato (Solanum tuberosum)
Tomato mottle Taino virus	Potato (Solanum tuberosum)
Tomato ringspot nepovirus,Tomato ringspot virus (ToRSV)	Blueberry and Cranberry (Vaccinium spp.),Pome (Malus, Pyrus and Cydonia spp.), Raspberry and Blackberry (Rubus spp.), Stonefruit (Prunus spp.), Strawberry (Fragaria spp.), Currants and Gooseberry (Ribes spp.), Pome (Malus, Pyrus and Cydonia spp.), Raspberry and Blackberry (Rubus spp.), Grapevines (Vitis spp.)
Tomato yellow mosaic virus	Potato (Solanum tuberosum)
Tomato zonate spot tospovirus	Taro (Colocasia spp.)
Tulare apple mosaic ilarvirus	Hazelnuts (Corylus spp.)
Wheat dwarf geminivirus A	Clonal grasses
Wheat soilbome mosaic furovirus	Clonal grasses
Wheat streak mosaic (Wheat streak mosaic virus) , Wheat streak mosaic potyvirus	Grains—barley, maize, millet, oats, rye, sorghum, triticale, wheat, Clonal grasses
Wheat yellow leaf closterovirus	Clonal grasses
Zoysia mosaic potyvirus	Clonal grasses

Table 2. Example of VSD output for homology searches against the non-redundant NCBI database:

sample-id	input-len	sacc	naccs	length	slen	cov	av-pident	stitle	qseqids
MT002-24-AGRF_CAGRF20031589_HVF77BGXF	24-24nt	MH427285	2	300	1929	0.156	98.387	Prunus_necrotic_ringspot_virus_isolate_PNRSV-RNA3-SA_movement_protein_and_coat_protein_genes,_complete_cds	MT002_24-24_K9-21_000030, MT002_24-24_K9-21_000028
MT002-24-AGRF_CAGRF20031589_HVF77BGXF	24-24nt	EU849602	1	176	852	0.207	98.864	Prunus_necrotic_ringspot_virus_isolate_PNRSV-AL1_movement_protein_mRNA,_complete_cds	MT002_24-24_K9-21_000028

Where:

- sample-id: ID of the sample sequenced;
- input-len: length of the quality-trimmed reads used to run VSD, here 21-22 nucloetides;
- sacc: accession number of the best match found on NCBI NR database;
- naacs: number of contigs assembled by the pipeline that return this match;
- length: length shared by the contig and match;
- slen: length of the matching subject;
- cov: coverage;
- av-pident: average percent identity calculated by the blast search;
- stitle: description of the matching sequence;
- qseqids: IDs of the contigs assembled by the pipeline.

Appendix 27: Hybridization probe evaluation and sensitivity II

27.1 Introduction

Currently, unculturable plant pathogens are detected in a post entry quarantine (PEQ) environment using molecular techniques specific to each target pathogen such as polymerase chain reaction (PCR). When samples require screening for multiple pathogens, this process becomes labour intensive and expensive. High throughput sequencing (HTS) is an essential tool for metagenomic studies, the study of genetic material from environmental samples (Ghosh et al., 2019). Metagenomic sequencing, where all DNA from a sample is sequenced, is an increasingly popular tool in diagnostics of unculturable or unknown pathogens as it has the capability to perform unbiased identification of all pathogens within a sample. However, one of the main limitations of metagenomic sequencing for pathogen detection is that the titre of a target pathogen's DNA can be very low compared to that of its hosts DNA. This means that to detect the pathogen, very deep sequencing may be required, which can be expensive, and in some cases the pathogen still may not be detected.

Targeted HTS using hybridization capture is more sensitive than metagenomics for pathogen detection. It has been shown to be as sensitive as polymerase chain reaction (PCR) assays and can detect multiple targets in one sample. The use of hybridization capture prior to HTS enables capture of target nucleic acids from a broader pool, enriches the sample for a specific target, reducing the amount of sequencing required. Hybridization capture has been used previously for a range of applications including diversity studies, molecular diagnostics and breed/cultivar improvement (Donaldson et al., 2017; Mascher et al., 2013; Muraya et al., 2015; Pankin et al., 2014; Whiston et al., 2017).

In Milestone 105, we designed and synthesized a hybridization capture assay for the detection of pathogenic bacteria within the National Priority Plant Pests (NPPP) list (2019), many of which are unculturable or fastidious, can be present in low titre in their plant hosts and difficult to accurately identify with routine diagnostics. In the design of the hybridization capture assay, multiple genomic targets were included for each bacterial pathogen to facilitate accurate identification, but the design strategy varied depending on the genomic sequence data available for the target pathogen and closely related species, and the level of taxonomic identification required for the pathogen (e.g. species, sub-species or sequence type (ST)). Consequently, a total of 5,901 dsDNA hybridization probes were designed for *Xylella fastidiosa* (Xf), *X. taiwanensis* (Xt), *Xanthomonas citri* subsp. *citri* (Xcc), *Erwinia amylovora* (Ea), '*Candidatus* Liberibacter asiaticus' (CLas), '*Ca*. Liberibacter americanus' (CLam), '*Ca*. Liberibacter africanus' (CLaf) and '*Ca*. Liberibacter solanacearum' (CLso), and 16SrI group phytoplasmas genomes (Table 27.1) and were used in a single hybridization capture assay for multiplex detection of the target pathogens.

Results from the first hybridization capture proof-of-concept experiment (Milestone 106) were very positive. The probes were applied to metagenomic libraries of plants naturally infected with CLaf, CLso, *Candidatus* Phytoplasma stylosanthis (StLL; phytoplasma group 16SrXXXVII), and *Ca.* Phytoplasma australasia (strain WFB, phytoplasma group 16SrII) and plant DNA spiked with serially diluted DNA of *Xylella fastidiosa* subsp. *fastidiosa* (Xff), Xcc and Ea to determine the ability of the probes to detect the specific pathogens and also the sensitivity of this detection technique compared to other tests. Results showed that the probes were able to detect all the pathogens to varying levels, and that the sensitivity of Xf, Xcc and Ea detection using the hybridization capture assay increased by 1,000 – 10,000 X compared to metagenomic sequencing and achieved equivalent sensitivity to PCR-based tests. Sensitivity of detection of unculturable CLaf and CLso were also significantly improved using the hybridization capture assay compared to metagenomic sequencing. Probes designed for detection of exotic 16SrI phytoplasmas did not cross react with *Ca.* Phytoplasma stylosanthis or *Ca.* Phytoplasma Australasia from phytoplasma group 16SrXXXVII and 16SrII respectively, indicating they are likely to be relatively specific, and different probes need to be designed to detect a broader range of phytoplasma 16Sr groups. DNA of phytoplasmas in the 16SrI group is still being sourced to evaluate these probes against a positive control.

In this milestone, additional improvements to the hybridization probe targets were attempted and the hybridization capture assay was further evaluated. Specifically:

- An additional gene was added to the *Xylella* spp. detection panel to enable ST determination.
- Additional phytoplasma genomes were included in probe design to evaluate detection of a broader range of phytoplasma 16Sr groups (16SrII, III, V, X, XI, XII and 16SrXXXVII).
- Comparison of the hybridization capture assay with metagenomic sequencing (non-probed) and current laboratory approaches for samples provided by the PEQ, Northern Australia Quarantine Strategy (NAQS) and Agriculture Victoria Research (AVR). Additionally, the level of taxonomic identification (e.g. species, subspecies, haplotype and ST) that can be determined from the HTS hybridization capture data was also assessed.

- Comparison of the hybridization capture assay with metagenomic sequencing (non-probed) on spiked samples with multiple pathogens. Additionally, the level of taxonomic identification (e.g. species, subspecies, haplotype and ST) that can be determined from the HTS hybridization capture data was also assessed.
- Modifications to probe protocols to reduce costs was assessed.

27.2 Material and Methods

27.2.1 Probe design

Probe set 1 from Milestone 106 was updated to create probe set 2. These updates include the addition of *petC* to the *Xylella* spp. detection targets and additional phytoplasma genomes to enable detection of a broader range of phytoplasma 16Sr groups (Table 27.1). The complete list of phytoplasma 16Sr groups used for developing the probes for target enrichment were *Ca.* Phytoplasma asteris strain aster yellows witches-broom (16SrI; Genbank ID NC_007716.1), *Ca.* Phytoplasma asteris strain Chrysanthemum coronarium phytoplasma (16SrI; Genbank ID BBIY00000000.1), Echinacea purpurea witches' broom phytoplasma (16SrII; Genbank ID CP040925.1), Milkweed yellows phytoplasma (16srIII; Genbank ID AKIL00000000.1), *Ca*. Phytoplasma ziziphi (16SrV; Genbank ID NZ_CP025121.1), *Ca*. Phytoplasma mali (16SrX; Genbank ID NC_011047.1), *Ca*. Phytoplasma oryzae (16SrXI; Genbank ID JHUK00000000.1), *Ca*. Phytoplasma australiense (16SrXII, Genbank ID AM422018.1), *Ca*. Phytoplasma solani (16SrXII; Genbank ID FO393428.1) and *Ca.* Phytoplasma stylosanthis (16SrXXXVII; Rodrigues Jardim et al., 2021).

Similar to probe set 1, probe set 2 were designed with 4 X tiling to increase the capture rate of low titre pathogens. Phytoplasma genomes are known to have AT-rich and low complexity regions and to achieve near full coverage of the phytoplasma target genes present across 16Sr groups, stringency for low complexity regions during the bioinformatic pipeline for hybridization probe design needed to be reduced. Each probe consists of a 120 mer dsDNA oligonucleotide with a biotin molecule at one end to capture targeted genes using Streptavidin bound to magnetic beads.

27.2.2 Samples used for hybridization capture assay evaluation

27.2.2.1 Pathogen DNA

Pathogen DNA extracted from pure culture or from host plant tissue previously diagnosed to contain fastidious and/or unculturable phytopathogenic bacteria were provided by collaborators for hybridization capture assay evaluation (Table 27.2).

27.2.2.2 DNA extraction from plant material

Nucleic acids were extracted from desiccated plant samples infected with phytopathogenic bacteria provided by collaborators (Table 27.2) and from uninfected fresh leaf tissue from olive, lemon, apple and potato. Desiccated tissues were macerated in CTAB buffer using pestles in 1.5 mL Eppendorf tubes and extracted using a Qiagen DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Fresh tissue was extracted as described in Milestone 106 using CTAB and Qiagen DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. These uninfected tissues were used as host background for diluting pathogens for the mock mixed infections.

Table 27.1. The genes that were chosen as hybridization capture probe targets for detection and identification of*, Xylella fastidiosa, Xylella taiwanensis*, *Candidatus* Liberibacter species and *Candidatus* Phytoplasma species in groups 16SrI, II, III, V, X, XI, XII and XXXVII.

Unique regions		**MLST genes+ unique regions**		**MLST genes**			
Xanthomonas citri* subsp. *citri		***Xylella fastidiosa* [1]/ *Xylella taiwanensis***		***Candidatus* Liberibacter sp.[2]**		***Candidatus* Phytoplasma[3]**	
Gene annotation	Gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)	Gene annotation	Average gene length (bp)
AMDFNCPH_00526	1518	group_806	2257	*rplL*	396	rbfA	342
AMDFNCPH_00529	1185	group_1256	5311	*rplJ*	519	nusA	1074
AMDFNCPH_01859	2988	*mshA*	2568	*metG*	1536	rpsS-rplV-rpsC	1434
AMDFNCPH_02036	2109	group_7658 (subsp. *pauca* only)	2093	*adk*	606	rplF	555
AMDFNCPH_02047	2118	*copB*	933	*mutS*	2703	secY	1242
AMDFNCPH_02358	10173	group 1350	1378	*grpE*	636	map	744
Total length	27810	*dacC*	1446	*dnaG*	1935	secA	2508
		group_852	2217	*glnA*	1404	Total length	7899
		gltT	1313	Total length	9735		
		lacF	881				
		nuoL	2142				
		holC	425				
		leuA	1559				
		cysG	1429				
		petC [3]	737				
		Total length	25951				

[1] Includes *Xylella taiwanensis*, *X. fastidiosa* subsp. *fastidiosa*, *X. fastidiosa* subsp. *multiplex*, *X. fastidiosa* subsp. *pauca* and *X. fastidiosa* subsp. *sandyi*

[2] Includes *Candidatus* Liberibacter solanacearum, *Ca.* Liberibacter africanus, *Ca.* Liberibacter americanus and *Ca.* Liberibacter asiaticus

1 Update for probe set 2: Addition of *petC* and *Candidatus* Phytoplasma group 16SrII, III, V, X, XI, XII and 16SrXXXVII

Table 27.2 The samples used for evaluation of targeted high throughput sequencing using the hybridization capture assay.

Sample ID	Pathogen identified	Plant host species	Identification Method	Origin	Sample type	Source	DNA concentration (ng/µl)
PEQ-1	*Ca*. Liberibacter asiaticus	*Citrus* sp.	NDP 25 V1.1 PCR using A2 J5 Primers. Gel Positive	Timor Leste	Desiccate[1]	NAQS	40.8
PEQ-2	*Ca*. Liberibacter asiaticus	*Citrus* sp.	NDP 25 V1.1 PCR using A2 J5 Primers. Gel Positive	Timor Leste	Desiccate[1]	NAQS	66.2
PEQ-3	*Ca*. Liberibacter asiaticus	*Citrus* sp.	NDP 25 V1.1 PCR using A2 J5 Primers. Gel Positive	Timor Leste	Nucleic acid	NAQS	44.8
PEQ-4	*Ca*. Phytoplasma sp.	*Jatropha gossypifolla*	*Unknown*	Western Australia	Desiccate[1]	NAQS	5.78
PEQ-5	*Ca*. Phytoplasma sp.	*Carica papaya*	*Unknown*	Western Australia	Desiccate[1]	NAQS	11
PEQ-6	*Ca*. Phytoplasma sp.	*Carica papaya*	*Unknown*	Northern Territory	Desiccate[1]	NAQS	11.7
PEQ-7	*Ca*. Phytoplasma aurantifolia	*Melochia corchorifolia*	Christensen et al. (2004) qPCR & Nested PCR P1/P7, R16F2/R2, Sanger seq	Northern Territory	Desiccate[1]	NAQS	4.48
PEQ-8	*Ca*. Phytoplasma aurantifolia	*Ipomea batatas*	Nested PCR P1/P7, R16F2/R2, Sanger seq	Timor Leste	Desiccate[1]	NAQS	12.3
PEQ-9	*Ca*. Phytoplasma aurantifolia	*Ipomea batatas*	Nested PCR P1/P7, R16F2/R2, Sanger seq	Timor Leste	Desiccate[1]	NAQS	4.24
PEQ-10	*Ca*. Phytoplasma aurantifolia	*Carica papaya*	Christensen et al. (2004) qPCR & Nested PCR P1/P7, R16F2/R2, Sanger seq	Northern Territory	Nucleic acid	NAQS	20.2
PEQ-11	*Ca*. Phytoplasma aurantifolia	*Luffa acutangula*	Christensen et al. (2004) qPCR & Nested PCR P1/P7, R16F2/R2, Sanger seq	Northern Territory	Nucleic acid	NAQS	8.3
PEQ-12	*Ca*. Phytoplasma aurantifolia	*Phaseolus vulgaris*	Hodgetts qPCR & Nested PCR P1/P7, R16F2/R2, Sanger seq	Northern Territory	Nucleic acid	NAQS	19.5
PEQ-13	*Ca*. Phytoplasma 16SrII-A	*Carica papaya*	Christensen et al. (2004) qPCR & Nested PCR P1/P7, R16F2/R2, Sanger seq	Western Australia	Nucleic acid	NAQS	10.2

PEQ-14	*Ca.* Phytoplasma sp.	*Solanaceae lycopersicum*	*Unknown*	Queensland	Desiccate[1]	PEQ	14.5
PEQ-15	*Ca.* Phytoplasma sp.	*Solanaceae annuum*	*Unknown*	Queensland	Desiccate[1]	PEQ	22
PEQ-16	*Ca.* Phytoplasma australasia	Solanum tuberosum	HTS Sequencing	Australia	Nucleic acid	AVR	7.48
PEQ-17[2]	*Ca.* Phytoplasma australiense (Australian grapevine yellows (AGY))	*Vitis vinifera*	HTS Sequencing	Australia	Nucleic acid	AVR	4.26
PEQ-18[2]	*Ca.* Phytoplasma stylosanthis	*Solanum tuberosum*	HTS Sequencing	Australia	Nucleic acid	AVR	6.14
ICMP 15971 (XFF)	*Xylella fastidiosa* subsp. *fastidiosa*	Pure culture	HTS Sequencing	USA	Nucleic acid	MPI	2.4
ICMP 8739 (XFM)	*Xylella fastidiosa* subsp. *multiplex*	Pure culture	HTS Sequencing	USA	Nucleic acid	MPI	4.1
NCPPB 4612 (XT)	*Xylella taiwanensis*	Pure culture	HTS Sequencing	Taiwan	Nucleic acid	MPI	2
CLbr A1 (CLbr)	*Ca.* Liberibacter brunswickensis	*Solanum melongena*	HTS Sequencing	Australia	Nucleic acid	AVR	1.1
1 Lso N (CLso)	*Ca.* Liberibacter solanacearum	*Solanum tuberosum*	HTS Sequencing	New Zealand	Nucleic acid	PFR	1.42
BRIP 70839 (CLaf)	*Ca.* Liberibacter africanus	*Citrus sinensis* (L.) Osbeck	HTS Sequencing	South Africa	Nucleic acid	DAF	0.49
Xcc 83 (Xcc)	*Xanthomonas citri* pv *citri*	Pure culture	HTS Sequencing	Australia	Nucleic acid	AVR	4.5

NAQS = Northern Australia Quarantine Strategy, PEQ = Post Entry Quarantine, AVR = Agriculture Victoria Research, MPI = Ministry for Primary Industries, DAF = Department of Agriculture and Fisheries, PFR = Plant and Food Research, New Zealand.

[1] Desiccated plant tissues (mostly leaf)

2 Samples were also used for the modified pooling protocol

27.2.2.3 DNA quality assessment and confirmation of pathogen presence

The concentration of each DNA sample was measured using the Quantus™ fluorometer following manufacturer's instructions (Table 27.2).

DNA of pathogens in host background were tested using a plant housekeeping qPCR targeting the *cox* gene (Weller et al., 2000) to confirm DNA presence. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR (Promega), 1 µL cox primer mixture (Weller et al., 2000), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water) and a positive plant control. The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

The identity of pure cultures of Xcc and Xf were confirmed with specific pathogen PCR and qPCR in Milestone 106.

To confirm the identity of specific species of *Ca.* Liberibacter in desiccated DNA samples qPCR was used (Li et al., 2006; 2009). Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL, 150 nM target probe, 300 nM (each) internal control primers, 150 nM internal control probe (Li et al., 2006; 2009), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water). The cycling conditions were 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 10 seconds and 58°C for 30 seconds on a QuantStudio 3 (Thermo Fisher).

To confirm the presence of *Ca.* Phytoplasma in desiccated DNA samples the *Ca.* Phytoplasma generic qPCR of Christensen et al., (2004) was used. Each qPCR reaction was performed in a final volume of 20 µL. The reaction mixture contained 10 µL GoTaq® Probe qPCR Master Mix 2X + CXR, 1 µL 300 µM forward primer, 900 µM reverse primer, 200 µM probe (Christensen et al., 2004), 4 µL molecular biology grade water and 5 µL of template DNA. Each test included a negative template control (water). The cycling conditions were 50°C for 2 minutes, 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 1 min on a QuantStudio 3 (Thermo Fisher).

27.2.2.4 Samples for sensitivity evaluation

Hybridization probe set 2 includes the gene *petC* needed for ST determination of Xf. The Xf dilution series from Milestone 106 (Table 27.3) was used to evaluate the use of hybridization capture using probe set 2 to determine ST of Xf and the sensitivity of the technique compared to metagenomic sequencing, qPCR, LAMP and the hybridization capture assay using probe Set 1. Preparation of the Xf dilution series, qPCR, LAMP metagenomic sequencing and probe detection using probe set 1 are described in Milestone 106.

Table 27.3 The *Xylella fastisiosa* subsp. *fastidiosa* (ICMP 15971) dilution series used to assess the sensitity of targeted high throughput sequencing using the hybridization capture assay.

Sample[1]	Concentration of *Xylella* DNA (ng/µL)
XF-2-0	0.24
XF-2-1	2.4×10^{-02}
XF-2-2	2.4×10^{-03}
XF-2-3	2.4×10^{-04}
XF-2-4	2.4×10^{-05}
XF-2-5	2.4×10^{-06}
XF-2-6	2.4×10^{-07}
XF-2-7	2.4×10^{-08}

1 Samples were also used for the modified pooling protocol

27.2.2.5 Preparation of mock mixed infection samples

DNA of plant samples naturally infected with *Candidatus* Liberibacter or phytoplasma, and DNA extracted from pure cultures of *Xylella* sp and Xcc (Table 27.2) were combined to produce mock mixed infections to evaluate the ability of the probes to

detect multiple target pathogens in the one sample. Naturally infected plants and DNA from pure culture were combined in host background DNA according to Table 27.4 to generate mock multiple titre pathogens. Details of pathogens can be found in Table 27.2.

Table 27.4 Combinations of bacterial pathogens used to evaluate the specificity and sensitivity of targeted high throughput sequencing using hybridization capture using probe set 2.

Sample	Species	Isolate/natural infection
Combo-1[1]	Clso	Natural infection
	Clas (PEQ-2)	Natural infection
Combo-2	Claf	Natural infection
	CLbr	Natural infection
Combo-3[1]	StLL (PEQ-18)	Natural infection
	AGY(PEQ-17)	Natural infection
	WFB (PEQ-16)	Natural infection
Combo-4[1]	XFF	Isolate
	XCC	Isolate
Combo-5[1]	XFF	Isolate
	XFM	Isolate
	XT	Isolate
Combo-6[1]	Clso	Natural infection
	Clas (PEQ-2)	Natural infection
	StLL (PEQ-18)	Natural infection
	AGY(PEQ-17)	Natural infection
	WFB (PEQ-16)	Natural infection
	XFF	Isolate
	XFM	Isolate
	XT	Isolate

1 Samples were also used for the modified pooling protocol

27.2.3 Target enrichment: library preparation, hybridization, and sequencing

DNA concentrations of all samples of single and mock mixed infections were measured using a Quantus™ fluorometer. Approximately 20 ng of DNA was used for library preparation except combo-1 and combo-2 (Table 27.4) where 15 ng was used due to lower concentrations of DNA. Libraries were generated using half reaction volumes of the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit and PerkinElmer NEXTFLEX® Unique Dual Index (UDI) Barcodes (384 barcodes) (PerkinElmer). A total of 105 libraries were prepared, which included two libraries for each pathogen sample and the olive, citrus and apple DNA (one library for hybridization capture and the other for metagenomic sequencing), four libraries per *Xylella* dilution so both HTS approaches (hybridization capture and metagenomics) could be done in duplicate and one no template control library.

The DNA concentration of each library was quantified using a Quantus™ fluorometer and the 2200 TapeStation system (Agilent technologies) was used to assess library quality and fragment size distribution.

For hybridization capture, eight libraries were pooled into one reaction to make a total of 1,500 ng of DNA (187.5 ng per samples) according to the manufacture's instruction. To reach 187.5 ng, each sample needed a minimum concentration of 4 ng/µL. For libraries with concentrations below 4 ng/µL, an aliquot of 25 µL of library was used for hybridization capture reaction. A total of seven hybridization capture reaction pools were prepared for samples listed in Tables 27.2, 27.3 and 27.4. The probed PEQ-13 (PEQ-P-13) library was only 0.156 ng/µL and it was not detected on the TapeStation. However, the library was still used for sequencing using 25 µL for hybridization capture. Unprobed libraries were sequenced as metagenomic samples at even ratios.

To assess whether the number of libraries pooled in a hybridization capture reaction could be increased and the amount of library added to the pool could be decreased, the total number of libraries included in single reaction pool was increased to 18 libraries (compared to eight in the manufacturer's instructions) to make a total of 1,500 ng of DNA (50-100 ng per sample). Pooled libraries were subjected to hybridization capture using the Twist Target Enrichment Protocol (Twist Bioscience, USA) according to manufacturer's instructions.

Hybridization capture and metagenomic libraries were paired end sequenced (2 x 250 bp) using the Illumina NovaSeq on an SP flow cell.

27.2.4 Bioinformatic analysis

A bioinformatic protocol to enable high throughput analysis of samples is proposed and evaluated in this milestone. This includes an initial triage step for target read detection to exclude negative samples from further analysis, multiple identification methods to enable identification of pathogens to varying taxonomic classifications and a step to identify what organism/s off-target species reads were captured from (Figure 27.1). Below are the methods evaluated for incorporation into the pipeline.

Raw Illumina sequence reads were adaptor trimmed and quality filtered using fastp (Chen et al., 2018). The detection of target reads was then assessed using Kraken 2 (Wood et al., 2019) and read mapping to determine the likelihood of a target pathogens presence in a sample. Trimmed reads were analysed using Kraken 2 (Wood et al., 2019) to assess the percentage of reads classified into pathogenic genera. Sequences of all alleles of genes used for probe design, including an extra 1 kbp sequence of the 5' and 3' end which includes 70 phytoplasma regions (seven MLSA genes from 10 representative phytoplasma species), 28 *Ca.* Liberibacter regions (seven MLSA genes from four *Ca.* Liberibacter species) and 35 *Xylella* regions (seven MLSA/MLST genes from Xt and four subspecies of Xf) were concatenated into one file and used as a reference database for mapping trimmed reads (database is labelled as the hybridization probe reference (HPR) database) using the "--fast" option of the mapping tool, bowtie2 (Langmead et al., 2012) which typically generates lower stringency (faster but less sensitive and less accurate) mapping than the default parameters. Reads were extracted from BAM files using the program bam2fastq (https://github.com/jts/bam2fastq).

For the identification step, mapped reads were stringently mapped using the "--sensitive" option in bowtie2 to databases of the MLSA/MLST genes of individual *Ca.* Liberibacter species, *Ca.* Phytoplasma species and *Xylella* subspecies (Xf database 1 - 5, Phy database 1 - 10 and Lib database 1 - 4; Figure 27.1). Identification through mapping (Method 1) is indicated by mapping of reads to all probed MLSA/MLST genes from a specific species/subspecies database.

Reads from stringent mapping to individual databases were also assembled using SPAdes v3.6.0 (Bankevich et al., 2012). Contigs generated from the same database mapping were concatenated and used for phylogenetic analysis using MEGA (version 7.0.26) (Kumar et al., 2016). In this report the concatenated genes *rplL/rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA* were analysed for the *Ca*. Liberibacter species, *secA* for the *Ca.* Phytoplasma species and *gltT*, *lacF*, *nuoL*, *holC*, *leuA*, *cysG* and *petC* for *Xylella* species/subspecies. Identification through phylogenetic analysis (Method 2) is based on clustering of the concatenated output with known pathogens within the phylogenetic tree.

Assembled contigs were also subjected to BLASTn using the NCBI non-redundant database for identification of complete and incomplete genes (Method 3).

Identification of off-target reads (Method 4) associated with non-specific binding of probes required de novo assembly of low stringency mapped reads using SPAdes v3.6.0 (Bankevich et al., 2012) and BLASTn of the assembled contigs to the NCBI non-redundant database.

A k-mer based mapping tool, StringMLST v. 0.6.3 (Gupta et al., 2017), was assessed to assign the Xf ST using mapped reads extracted from hybridization capture data from samples in the Xf dilution series.

Metrics around mapping efficiency can potentially be used to indicate hybridization probe capture efficiency/quality. To compare the mapping efficiency of reads captured using probe set 1 and probe set 2, BAM files of mapped Xf dilution series samples generated using both probe sets were subjected to picard CollectHsMetrics (version 2.26) to calculate AT and GC dropout metrics. The AT/GC dropout metrics measured the reduction of mapping efficiencies due to the increased numbers of errors associated with excessive or deficient numbers of GC bases in a sequencing region. The output is a percentage of misaligned reads that correlate with low (%-GC is < 50%) or high (%-GC is > 50%) GC content respectively.

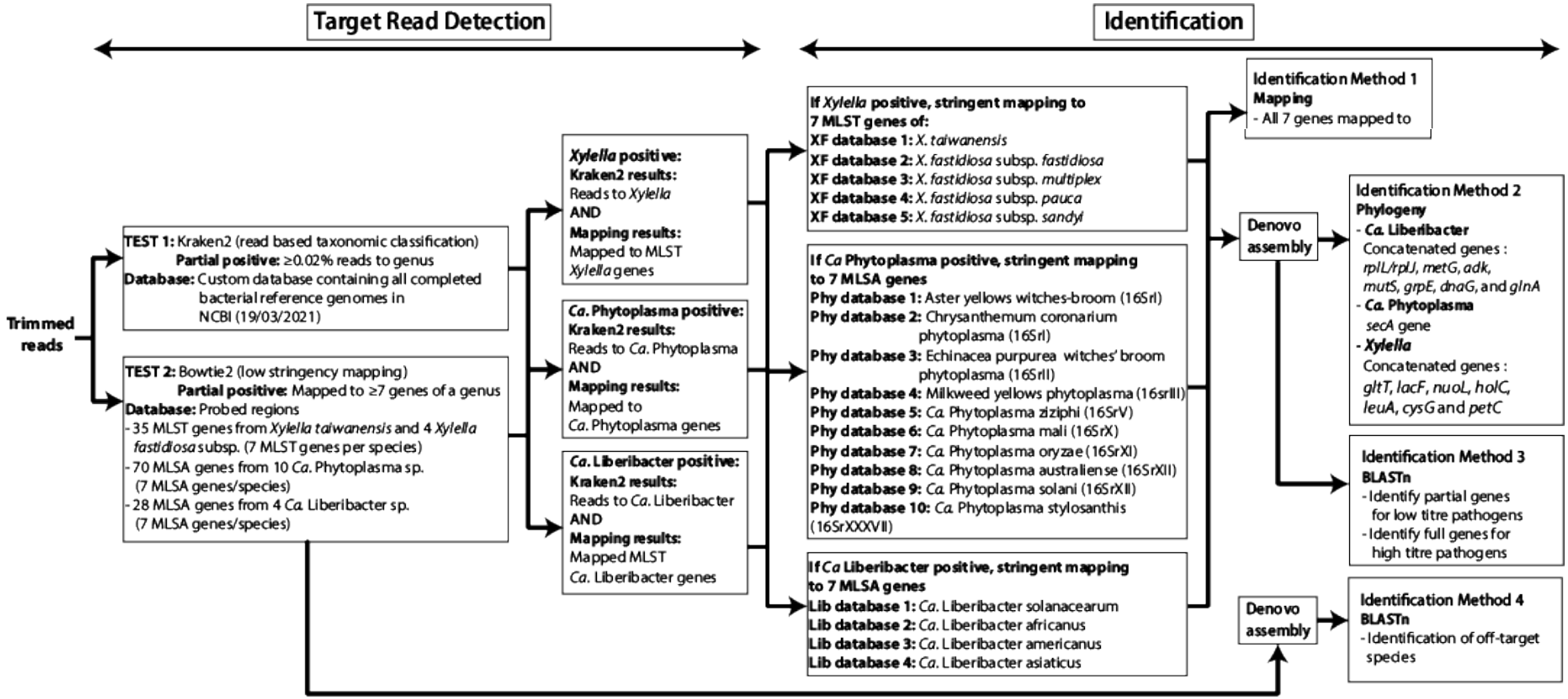


Figure 27.1 Hybridization probe bioinformatic pipeline flowchart for pathogen detection.

27.3 Results and discussion

27.3.1 DNA quality assessment and confirmation of pathogen presence

The Quantus™ fluorometer and the *cox* qPCR assay were used to confirm the presence and quality of DNA in the acquired DNA samples and the nucleic acids extractions from desiccated samples. All samples from plant hosts had quantifiable DNA (concentrations listed in Table 27.2) and were amplifiable using the *cox* PCR (Weller et al. 2000).

DNA supplied by NAQS and AVR had been previously tested and *Ca*. Phytoplasma and *Ca.* Liberibacter species were detected. The presence of *Ca*. Phytoplasma species in the suspected positive desiccated samples (PEQ-P-4 to 9, 15 and 16) was confirmed by qPCR with cycle threshold (Ct) values ranging from 12.64 to 22.91, except in PEQ-P-6 which gave no Ct value below 40. PEQ-P-6 has been noted to be a weak positive by NAQS. The presence of CLas species in the suspected positive desiccated samples (PEQ-P-1 and 2) was confirmed by qPCR (Ct 19.84 and 21.82, respectively).

27.3.2 Sequencing output

The data outputs for each library after quality trimming are presented in Table 27.5.

Table 27.5: The volume of sequence data (megabyte; Mb) and the total number of reads generated by metagenomic sequencing and targeted sequencing using hybridization capture for libraries generated using uninfected plant controls, *Xylella* serial dilutions, pathogen and the mock mixed infection DNA. The libraries generated from the *Xylella* DNA serial dilutions were prepared in duplicate, and the average amount of data and number of reads is presented.

Uninfected plant				
	Metagenomic		Probed set 2	
Sample	Data generated (Mb)	Reads	Data generated (Mb)	Reads
Lemon	1154	9537154	451	4299638
Olive	1287	10904952	2400	23974124
Apple	2025	17039172	878	8014996
No template control	37	317112		

***Xylella* serial dilutions**						
	Metagenomic		Probed set 2		Probe set 1 (Milestone 106)	
Sample	Data generated (Mb)	Reads	Data generated (Mb)	Reads	Data generated (Mb)	Reads
XF-0	1083	9251507	1805.5	17582508	6,281	60447357
XF-1	1853.5	16382275	1604.5	16122794	2,194	20865534
XF-2	1154	9761320	1339	13292467	998	9298002
XF-3	1068.5	9085304	1296	13086849	736	7181871
XF-4	1203	10346505	1388	13821331	690	6276356
XF-5	1133	9571244	1651.5	16582184	693	6664122
XF-6	1247	10658723	1501	14884823	575	5377575
XF-7	1310	11170495	1521.5	15231881	457	4458065

Pathogen				
	Metagenomic		Probed set 2	
Sample	Data generated (Mb)	Reads	Data generated (Mb)	Reads

PEQ-1	909	7555580	634	6470874
PEQ-2	848	6996888	7.7	77970
PEQ-3	1340	11557028	390	3982598
PEQ-4	970	8298854	3200	31216248
PEQ-5	904	7466838	6400	63758630
PEQ-6	1986	17094014	77	662322
PEQ-7	727	6181556	2007	19184226
PEQ-8	1021	8873840	615	5792610
PEQ-9	874	8222990	312	3282076
PEQ-10	1134	9444882	1801	16566796
PEQ-11	1190	9778770	1669	15125096
PEQ-12	970	8396652	2400	22034140
PEQ-13	890	7414292	2.4	20566
PEQ-14	1150	9521816	2400	21884398
PEQ-15	1047	8647312	1169	10529602
PEQ-16	884	7844796	3400	34635266
PEQ-17	1355	12008038	1005	9856766
PEQ-18	1130	9339094	981	8572350

Mock mixed infections

	Metagenomic		Probed set 2	
Sample	Data generated (Mb)	Reads	Data generated (Mb)	Reads
Combo-1	829	7052544	1889	18806850
Combo-2	1179	9695452	603	5464438
Combo-3	1429	12473348	1750	16831372
Combo-4	1309	10999188	2400	22000990
Combo-5	1163	9887416	1954	19252926
Combo-6	1322	11209860	2600	24886750

18 library pool

	Probed set 2	
Sample	Data generated (Mb)	Reads
R-P-Blank-apple	275	2347242
R-P-Blank-lemon	367	3234132
R-P-Blank-olive	484	4442450
R-P-PEQ-P-17	445	3959222
R-P-PEQ-P-18	1161	10377296
R-P-Combo-1	662	5992982
R-P-Combo-3	8400	83831832
R-P-Combo-4	688	6185076
R-P-Combo-5	1008	9458114
R-P-Combo-6	1221	11007144
R-P-XF-0	3200	29748170
R-P-XF-1	704	6548972

R-P-XF-2	600	5606848
R-P-XF-3	680	6340606
R-P-XF-4	818	7764534
R-P-XF-5	779	7321088
R-P-XF-6	876	8291884
R-P-XF-7	624	5857106

27.3.3 Detection and identification of pathogens in PEQ samples

The computational resources required, and the time taken to analyse HTS data can vary greatly depending on the programs used. For example, *de novo* assembly of reads into contigs and comparison of those contigs to large databases using BLASTn uses considerably more computational resources and time (hours) compared to fast (secs) k-mer based read classification tools like Kraken2. In Milestone 106, we used a number of different tools/programs to analyse HTS data from metagenomic and hybridization capture sequencing for the detection and identification of target pathogens. In this milestone we have proposed a draft analysis pipeline that incorporates different analysis tools in a practical, resource and time efficient manner (Figure 27.1). We are trialling an initial target read detection step that includes two methods for read classification, Kraken2 and low stringency mapping to the HPR database, that are very fast and do not require large computation resources. It is proposed that we can utilise the combined results from these analyses to determine whether reads from our target pathogens have been detected. This step will quickly triage which datasets require further, more in-depth analysis for identification and eliminate negative results without wasting time and resources.

Identification of pathogens using the reads detected in the target read detection step, has been evaluated using three different methods as presented in Figure 27.1. These include: high stringency mapping to the pathogen MLST databases (XF, Phy and Lib databases) and quantification of the number of genes mapped to and the breadth of coverage (Method 1) and *de novo* assembly of the stringently mapped reads followed by phylogenetic analysis (Method 2) and sequence comparison of the contigs to the non-redundant database at NCBI using BLASTn (Method 3). Additionally, to determine what off-target reads were being mapped to the target regions in metagenomic data, and what off-target reads were being captured in the hybridization capture assay, reads mapped in the low stringency mapping (at the target read detection step) were also *de novo* assembled and compared to sequences in the non-redundant database at NCBI using BLASTn. The off target capture data is presented and discussed in section 3.7.

27.3.3.1 Candidatus Phytoplasma sp.

27.3.3.1.1 Target read detection

Kraken 2 classified the hybridization capture reads as *Ca.* Phytoplasma in 13 out of 15 of the samples previously identified as being infected by *Ca.* Phytoplasma. The percentage of reads classified as *Ca.* Phytoplasma in these samples ranged from 0.08 to 11.17% (Table 27.6). In samples PEQ-P-6 and PEQ-P-13, 0% of reads were assigned to *Ca.* Phytoplasma using this method. Analysis of the metagenomic samples with Kraken2 classified reads as *Ca.* Phytoplasma in 12 samples out of the 15 samples not detecting *Ca*. Phytoplasma reads in samples PEQ-P-6, PEQ-P-13 and PEQ-P-17 (Table 27.6). The percentage of reads classified as *Ca*. Phytoplasma was significantly lower in metagenomic data compared to the hybridization capture data, ranging from 0.02 - 0.63% of the reads. Kraken2 results indicate that there was a 17.73 - 116.5 fold increase in the number of *Ca*. Phytoplasma reads detected using hybridization capture compared to metagenomic sequencing of the same samples. *Ca.* Phytoplasma reads were only detected in the hybridization capture data of PEQ-P-17 and not in the metagenomic data indicating that the titre of the pathogen in this sample is too low for detection using metagenomic sequencing and highlighting the increased sensitivity of hybridization capture.

When reads from hybridization capture were mapped with low stringency to the HPR database, using bowtie2, all seven *Ca.* Phytoplasma MLSA genes had reads mapped to them in 13 out of 15 samples (Table 27.6). A small number of reads from the PEQ-P-6 hybridization capture data were mapped to one *Ca.* Phytoplasma MLSA gene and no reads from PEQ-P-13 data mapped to the *Ca.* Phytoplasma MLSA genes. These results correlated with the Kraken2 results indicating that there is no *Ca.* Phytoplasma DNA at a detectable concentration in these samples. The reads from PEQ-P-6 that mapped to one *Ca.* Phytoplasma MLSA gene were assembled and compared to the non-redundant database NCBI using BLASTn (Method 4, Figure 27.1). The contig generated was only 134 bp and had no significant sequence similarity to anything in the non-redundant database further confirming the absence of detectable *Ca.* Phytoplasma DNA in this sample.

Low stringency mapping of metagenomic reads to the HPR database resulted in reads mapping to all seven *Ca.* Phytoplasma MLSA genes in 12 out of 15 samples which was consistent with Kraken2 results. No reads from metagenomic data from PEQ-

6 and PEQ-17 mapped to any of the *Ca.* Phytoplasma MLSA genes. One *Ca.* Phytoplasma MLSA gene had reads mapped to it from PEQ-13 metagenomic data, however, those reads were insufficient for any contigs to be assembled.

It was not surprising that no Phytoplasma reads were detected in samples PEQ-P-6 and PEQ-P-13. The DNA extracted from the desiccated PEQ-P-6 sample, tested negative using a *Ca.* Phytoplasma qPCR (section 3.1) and the output libraries for hybridization capture for both samples were low concentration (PEQ-P-6: 1.21 ng/µL and PEQ-P-13: 0.156 ng/µL). The library for PEQ-P-13 was exceptionally low concentration and it is unclear as to whether the library preparation was not successful, the DNA was not good enough quality, or the pathogen was below detectable limits. We are currently exploring the use of an internal endogenous controls to help identify whether this negative result was due to a problem with library preparation or degradation of the supplied DNA.

As no pathogen reads were detected for PEQ-P-6 and PEQ-P-13 using Kraken2 and bowtie2, they were not included in further bioinformatic analyses.

27.3.3.1.2 Identification

Overall, the results of this proof-of-concept study showed that targeted hybridization capture HTS was able to identify the expected phytoplasma group or species in the sample on which it was tested. In most cases it was able provide a good level of demarcation and differentiation between phytoplasma groups.

Stringent mapping of reads from hybridization capture data to MLSA genes in the HPR database of individual *Ca.* Phytoplasma species (Method 1; Phy database 1-10) resulted in complete coverage of the seven genes of 16SrII group Echinacea purpurea witches' broom phytoplasma for samples PEQ-P-4 to 16. When *de novo* assembled contigs from stringent mapping were compared to the non-redundant database at NCBI using BLASTn (Method 3), Echinacea purpurea witches' broom phytoplasma was the top hit for all seven genes from samples PEQ-P-4 to 16 (≥99% sequence identity). This confirms the original detection of a 16SrII *Ca.* Phytoplasma species in samples PEQ-P-7-13 and 16. For samples PEQ-P-4-6,14 and 15, which were recorded only as *Ca.* Phytoplasma sp., comparison to the HRP database and by BLASTn provided a more specific level of taxonomic classification to the 16SrII phytoplasma group level. Echinacea purpurea witches' broom phytoplasma is most closely related to *Ca.* Phytoplasma australasia (Chang et al., 2015; Tseng et al., 2012) however previous identification, based on Sanger sequencing of 16S rRNA genes identified samples PEQ-P-7-12 *Ca*. Phytoplasma aurantifolia. Further analysis will be done to compare the similarity of all MLSA genes used in this study, between a range of 16SrII group phytoplasmas, and determine if in a finer taxonomic resolution of the different 16SrII isolates used in this study can be achieved, to match identification based on the 16SrRNA gene.

For PEQ-P-17, stringent mapping to MLSA genes of individual phytoplasma species (Method 1; Phy database 1-10) resulted in the majority of reads mapping to phytoplasma 16SrXII group, *Ca.* Phytoplasma australiense, confirming its previous diagnosis based metagenomic sequencing. Similarly, when *de novo* assembled contigs from stringent mapping were compared to the non-redundant database at NCBI using BLASTn the sequence similarities indicated that the phytoplasma present in this sample was more closely related to another 16SrXII phytoplasma, the *Ca.* Phytoplasma australiense strain strawberry lethal yellows (≥99% sequence identity). The sequence identity of several other related 16SrXII groups species in different subgroups, such as *Ca.* Phytoplasma solani was less than 90%. This suggests that when ample gene sequence data is available the method will provide a correct identification at the species and/or subgroup level.

Stringent mapping of PEQ-P-18 reads to MLSA genes of individual phytoplasma species (Phy database 1-10) resulted in complete coverage of the seven MLSA genes of *Ca.* Phytoplasma stylosanthis (phytoplasma group 16SrXXXVII) which correlated with the AVR identification. When *de novo* assembled contigs from stringent mapping were compared to the non-redundant database at NCBI using BLASTn, the top hits for six out of the seven MLSA genes were *Ca*. Phytoplasma luffae (phytoplasma group 16SrVIII) but they only had 79-85% sequence identity (Table 27.9). There was only one top hit to *Ca.* Phytoplasma stylosanthis, which was for the gene region *rpsS-rplV-rpsC* (99.96% sequence identity). The low sequence identity for six genes is due to the underrepresentation of genes for *Ca.* Phytoplasma stylosanthis in the NCBI GenBank database and the BLASTn algorithm prioritised the matches to the more complete set of genes from *Ca*. Phytoplasma luffae.

The only gene available for analysis across all phytoplasma groups and species in sequence reference databases for phylogenetic analysis was *secA.* Phylogenetic analysis (Method 2) using *secA* clustered PEQ-P-4 -16 with 16SrII phytoplasma species, which was consistent with results using methods 1 and 3 and most samples clustered with and confirmed their identification as strains of *Ca.* Phytoplasma australasia (subgroup 16SrII-D) (Hodgetts et al 2008; Tran-Nguyen et al., 2012) (Figure 27.2).

Sample PEQ-P-17 clustered with *Ca.* P. australiense (subgorup 16SrXII-B) and PEQ-P-18 clustered with *Ca*. Phytoplasma stylosanthis (group 16SrXXXVII) which confirmed the previous identification (Figure 27.2).

Phylogenetic results showed that the *secA* gene was sufficient in identifying the 16Sr *Ca.* Phytoplasma 16Sr groups of all

samples and some samples were identified to the species level. The *secA* gene is an important genetic marker and is widely used studying phytoplasma diversity and ecology (Hodgetts et al., 2009; Martini et al., 2019). Although it is well represented in the GenBank database amongst various phytoplasma groups, species and strains, compared to the other MLSA genes used in this study, it is not as extensively represented as the 16S gene which is used for primary identification and classification. Unlike the 16S rRNA gene secA is much more diverse within phytoplasma species, which could aid the identification of phytoplasmas, however distinct nucleotide percentage identities need to be well defined to determine cut offs between groups, sub-groups species and strains. The cut-off value to delineate groups, sub-groups, species and strains for this gene will be explored in future analyses.

It is noted that some of the phytoplasma MLSA genes used in this study are poorly represented on the NCBI GenBank database, which could impact the finer taxonomic demarcation within and between some *Ca.* Phytoplasma 16Sr groups, sub-groups, species, and strains during mapping to the HPR database and during BLASTn analysis of *de novo* assembles sequences. This highlights the importance of generating key gene sequences for all genes used for detection and identification of key targets using the hybridization probes, to support the development of a verified database that enables fine resolution of taxonomic assignments. Other phytoplasma genes, such as the 16S rRNA and elongation factor Tu (tuf) genes, that are commonly used for taxonomy and to study diversity and ecology, are better represented across many more phytoplasma groups, sub-groups species and strains. However, these genes are also highly conserved across a broad range of taxa. Probes designed to these regions would be prone to significant off-target capture in conserved genic regions, something that could significantly impact the sensitivity of detection. The development of probes to these genes will be considered and assessed bioinformatically with the aim of better delineating between *Ca.* Phytoplasma species. Nevertheless, this study has provided a demonstration of the benefits MLSA approach using targeted hybridization probes *Ca.* Phytoplasma species detection and identification during PEQ. Further proof will be gained when the outputs of the hybridization probes are compared to metagenomic sequencing for phytoplasma detection and identification.

27.3.3.2 Candidatus Liberibacter sp.

27.3.3.2.1 Target read detection

Kraken 2 analysis classified hybridization capture reads as *Ca.* Liberibacter for all three samples (PEQ-1 to 3), previously identified as CLas. The percentage of reads classified as *Ca.* Liberibacter in these samples ranged from 0.08 to 23.17% (Table 27.7). Analysis of the metagenomic samples with Kraken2 classified reads as *Ca.* Liberibacter in all three samples but at significantly lower read percentages ranging from 0.1 to 0.33% (Table 27.7). The percentage of reads classified as Ca. Liberibacter using hybridization capture increased by 8 – 82 fold when to compared to metagenomic sequencing.

When reads from hybridization capture were mapped with low stringency to the HPR database using bowtie2, all seven *Ca.* Liberibacter MLSA genes in all 3 samples were mapped to (Table 27.7). Low stringency mapping of metagenomic reads to the HPR database resulted in reads mapping to all seven *Ca.* Liberibacter MLSA genes for samples PEQ-P-1 and PEQ-P-2. Only one *Ca.* Liberibacter MLSA gene had reads mapped for sample PEQ-P-3 indicating that the titre of Ca. Liberibacter in this sample was too low to be detected using the metagenomic dataset.

27.3.3.2.2 Identification

Stringent mapping of reads from hybridization capture data to MLSA genes of individual *Ca.* Liberibacter species (Method 1; Lib databases 1-4) resulted in complete coverage of the seven genes of CLas. Comparison of the assembled contigs from the stringent mapped reads to the non-redundant database at NCBI using BLASTn (Table 27.8) revealed contigs with high read coverage had the highest percentage sequence identity to CLas. Similarly, phylogenetic analysis showed concatenated contigs from the stringently mapped reads from samples PEQ-P-1 to 3 clustering with CLas (Figure 27.3).

Table 27.6: Detection of *Xylella*, *Ca.* Phytoplasma and *Ca.* Liberibacter reads in PEQ samples 1-18. Kraken2 (K) results are presented as the percentage of total reads identified for each genera and bowtie2 (M) results are presented as the number of target genera MLSA/MLST genes mapped with the number of contigs in the HPR database mapped to in brackets. Fold enrichment = fold increase of Kraken2 percentage from metagenomic to probed samples for each genus.

		Metagenomic						Probed							
Sample	PEQ diagnosis	Xylella		Phytoplasma		Liberibacter		Xylella		Phytoplasma			Liberibacter		
		M	K	M	K	M	K	M	K	M	K	Fold enrichment	M	K	Fold enrichment
PEQ-1	*Ca.* Liberibacter asiaticus	0	0	0	0.01	7 (8)	0.33	4	0	2	0	-	7 (23)	23.17	70.21
PEQ-2	*Ca.* Liberibacter asiaticus	0	0	0	0.01	7	0.06	0	0	0	0	-	7 (9)	4.93	82.17
PEQ-3	*Ca.* Liberibacter asiaticus	0	0	0	0.01	1	0.01	1	0	2	0	-	7	0.08	8
PEQ-4	*Ca.* Phytoplasma sp.	0	0	7 (8)	0.55	0	0.01	0	0	7 (32)	10.64	19.35	0	0	0
PEQ-5	*Ca.* Phytoplasma sp.	0	0	7 (11)	0.63	0	0.01	1	0	7 (34)	11.17	17.73	2	0	0
PEQ-6	*Ca.* Phytoplasma sp.	0	0	0	0.01	0	0.01	1	0	1	0	0	0	0	0
PEQ-7	*Ca.* Phytoplasma aurantifolia	1	0	7	0.18	0	0	4	0	7 (26)	9.92	55.11	0	0	-
PEQ-8	*Ca.* Phytoplasma aurantifolia	1	0	7	0.11	0	0.01	3	0	7 (22)	8.02	72.91	0	0	0
PEQ-9	*Ca.* Phytoplasma aurantifolia	1	0	7 (8)	0.02	0	0.01	2	0	7 (18)	2.33	116.5	0	0	0
PEQ-10	*Ca.* Phytoplasma aurantifolia	0	0	7 (8)	0.06	0	0.01	1	0	7 (24)	5.84	97.33	0	0	0
PEQ-11	*Ca.* Phytoplasma aurantifolia	0	0	7 (8)	0.07	0	0.01	1	0	7 (23)	5.84	83.43	0	0	0
PEQ-12	*Ca.* Phytoplasma aurantifolia	0	0	7 (9)	0.15	0	0.01	1	0	7 (25)	7.72	51.47	0	0	0
PEQ-13	*Ca.* Phytoplasma 16SrII-A	0	0	1	0.01	0	0.01	0	0	0	0	-	0	0	0
PEQ-14	*Ca.* Phytoplasma sp.	6	0	7 (8)	0.28	0	0.01	15	0	7 (26)	9.6	34.29	0	0	0
PEQ-15	*Ca.* Phytoplasma sp.	0	0	7	0.14	0	0.01	1	0	7 (22)	8.9	63.57	0	0	0
PEQ-16	*Ca.* Phytoplasma australasia	0	0	7 (9)	0.17	0	0.01	1	0	7 (28)	8.86	52.12	0	0	0
PEQ-17	*Ca.* Phytoplasma australiense	0	0	0	0	0	0.01	1	0	7 (11)	0.08	-	0	0.01	1
PEQ-18	*Ca.* Phytoplasma stylosanthis	0	0	7 (11)	0.04	0	0	1	0	7 (35)	4.4	110	0	0	-

Table 27.7: Identification of *Ca.* Phytoplasma in PEQ samples in comparison to identification with traditional methods using mapping identification (all 7 phytoplasma MLSA genes mapped to with ≥85% gene coverage), BLASTn identification (top hits with % identity of ≥90% to isolates from non-redundant database from NCBI), percentage of identified pathogen contigs and taxonomic identification based on phylogenetic analysis.

Sample	Traditional identification	Mapping identification (Method 1)		BLASTn identification (Method 3)		Method 4 Contig % (contig no.)[1]	Phylogeny identification (Method 2)	
		16Sr group	Identification	16Sr group	Identification		16Sr group	Identification
PEQ-P-4	*Ca.* Phytoplasma sp.	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	43.98 (3973/9034)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-5	*Ca.* Phytoplasma sp.	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	76.68 (7146/9319)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-7	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	74.13 (1229/1658)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-8	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	83.81 (207/247)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-9	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	87.5 (7/8)	16SrII	Ipomoea phytoplasma
PEQ-P-10	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	91.91 (432/470)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-11	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	93.68 (489/522)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-12	*Ca.* Phytoplasma aurantifolia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	91.65 (999/1090)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-14	*Ca.* Phytoplasma sp.	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	84.9 (832/980)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-15	*Ca.* Phytoplasma sp.	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	90.34 (458/507)	16SrII	Australian big bud phytoplasma
PEQ-P-16	*Ca.* Phytoplasma australasia	16SrII	Echinacea purpurea witches' broom phytoplasma	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	94.9 (1675/1765)	16SrII	Sweet potato little leaf phytoplasma
PEQ-P-17	*Ca.* Phytoplasma australiense (AGY)	16SrXII	*Ca.* Phytoplasma australiense	16SrXII	*Ca.* Phytoplasma australiense (Strawberry lethal yellows phytoplasma (CPA) str. NZSb11)	63.64 (7/11)	16SrXII	*Ca.* Phytoplasma australiense
PEQ-P-18	*Ca.* Phytoplasma stylosanthis	16SrXXXVII	*Ca.* Phytoplasma stylosanthis	16SrVIII	*Ca*. Phytoplasma luffae isolate NCHU2019[2]	47.37 (9/19)	16SrXXVII	*Ca.* Phytoplasma stylosanthis

[1] Contigs based on Method 3 BLASTn identification

[2] BLASTn sequence identity ≥75% on 6 genes.

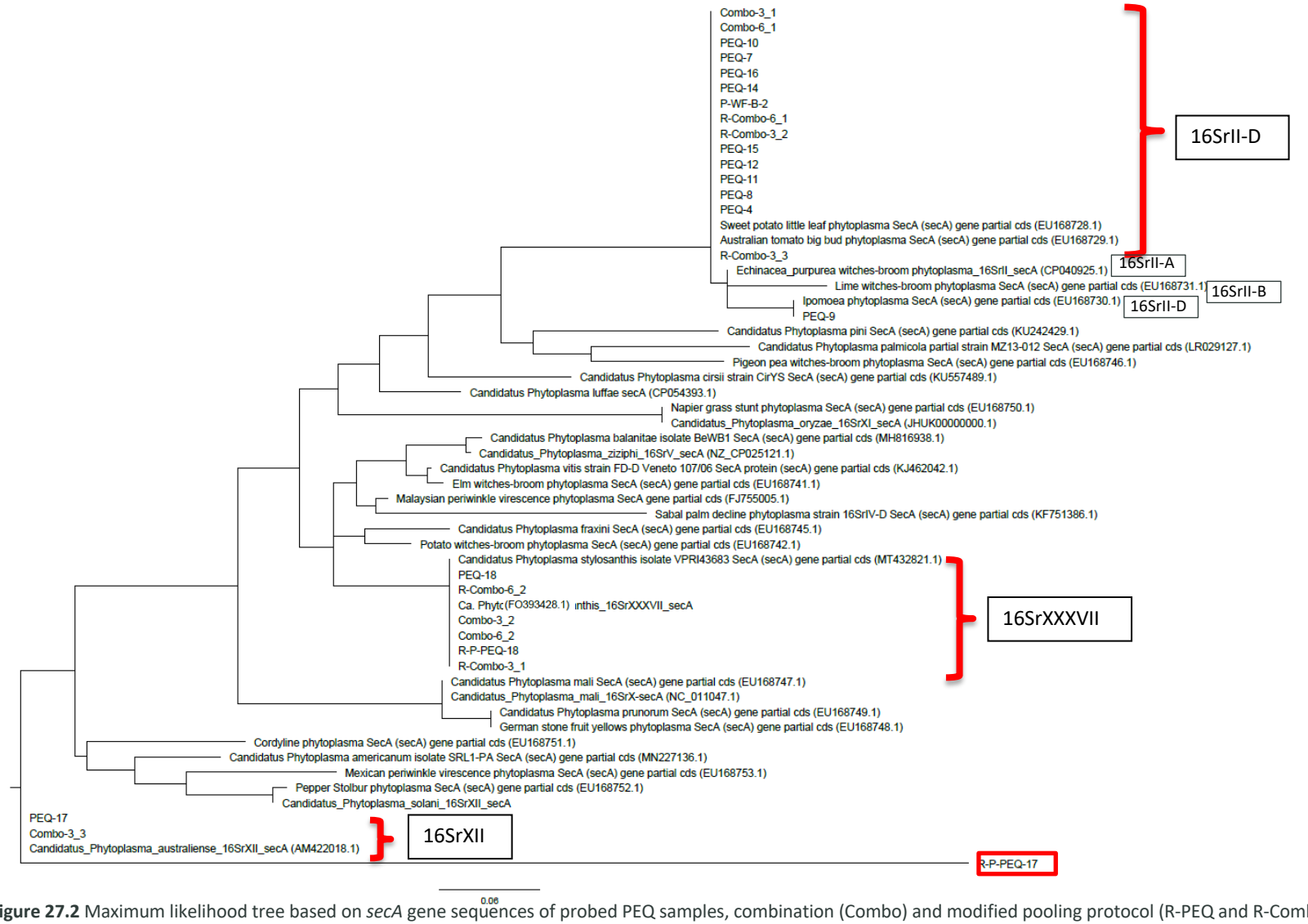


Figure 27.2 Maximum likelihood tree based on *secA* gene sequences of probed PEQ samples, combination (Combo) and modified pooling protocol (R-PEQ and R-Combo) samples and validated *secA* gene sequences of Phytoplasma species from NCBI.

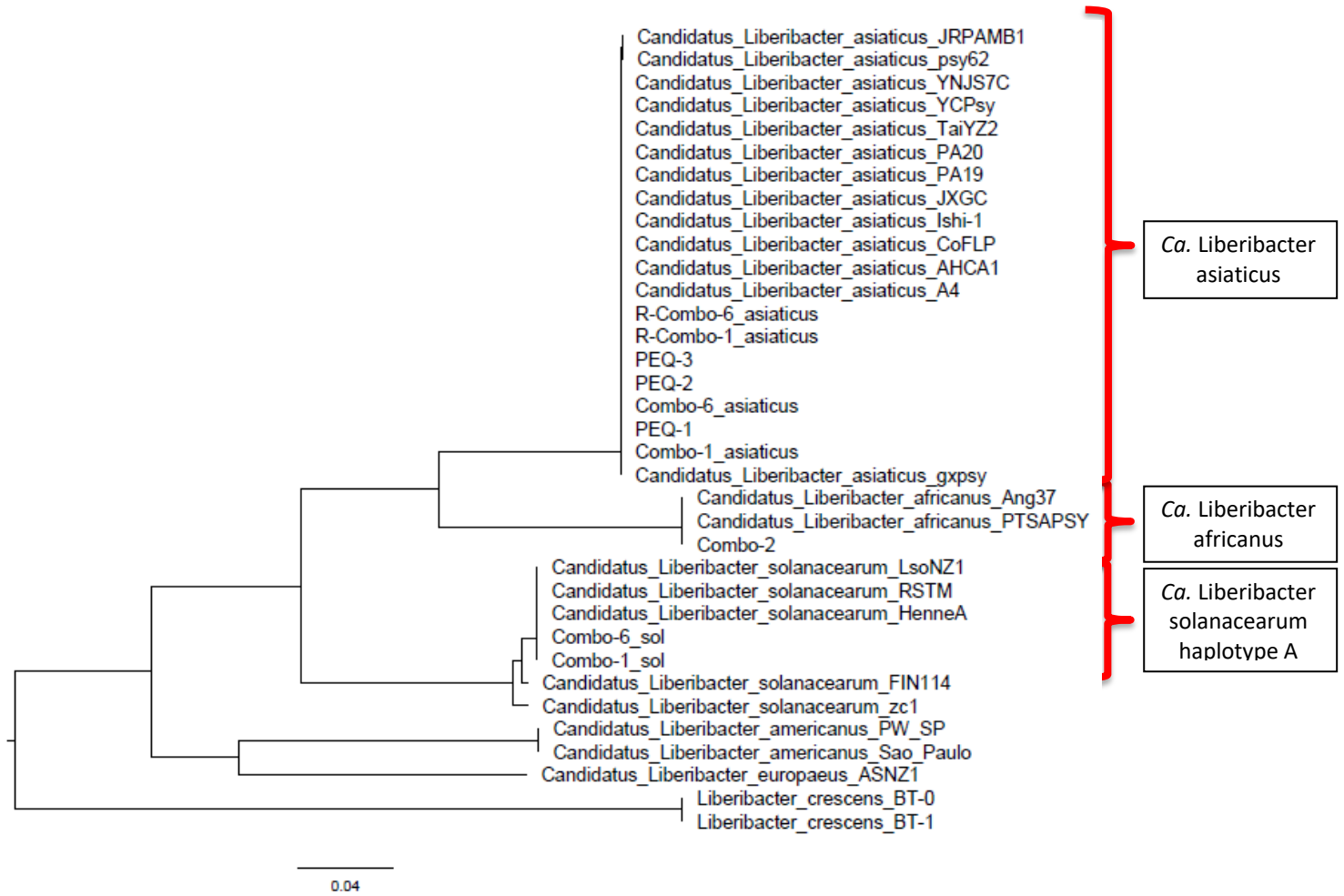


Figure 27.3 Maximum likelihood tree based on *rplL*, *rplJ*, *metG*, *adk*, *mutS*, *grpE*, *dnaG*, and *glnA* gene sequences of probed PEQ, combination (Combo) and modified pooling protocol (R-PEQ and R-Combo) samples and their closest phylogenetic neighbours.

Table 27.8 Identification of *Ca.* Liberibacter in PEQ samples using traditional methods compared to identification using mapping identification (all 7 Liberibacter MLSA genes mapped to with ≥85% gene coverage), BLASTn identification (top hits with % identity of ≤90% to isolates from non-redundant database from NCBI), percentage of identified pathogen contigs and taxonomic identification based on phylogenetic analysis.

Sample	Traditional identification	Mapping identification (Method 1)	BLASTn identification (Method 3)	Method 4 Contig % (contig no.)[1]	Phylogeny identification (Method 2)
PEQ-P-1	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus strain A4	53.06 (26/49)	*Ca.* Liberibacter asiaticus
PEQ-P-2	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus strain A4	90.34 (458/507)	*Ca.* Liberibacter asiaticus
PEQ-P-3	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus strain A4	100 (7/7)	*Ca.* Liberibacter asiaticus

[1] Contigs based on Method 3 BLASTn identification

27.3.4 Detection and molecular identification of pathogens in mixed samples

27.3.4.1 Target read detection

Kraken 2 analysis classified reads as *Ca.* Liberibacter in Combo-1, 2 and 6, *Ca.* Phytoplasma in Combo-3 and 6, and *Xylella* in combo-5 and 6 which was expected based on spiking of the samples. The percentage of reads classified as each pathogen genera in these samples ranged from 0.24 to 7.24% (Table 27.9). The only discrepancy with expected results was that 0% of reads were classified as *Xylella* in Combo-4 which had Xff spiked in at low concentration.

Analysis of the metagenomic samples with Kraken2 classified reads as *Ca.* Liberibacter in Combo-1, *Ca.* Phytoplasma in Combo-3 and *Xylella* in Combo-5 and 6, detecting three fewer target pathogen genera than with hybridization capture. The percentage of reads classified as each pathogen in these samples were also significantly lower, ranging from 0.04% to 0.08% (Table 27.10). This is reflected by the fold enrichment of Kraken2 results from 24 to 687 fold more when using hybridization probes in comparison to metagenomic samples.

When reads from hybridization capture were mapped with low stringency to the HPR database using bowtie2, all seven *Ca.* Liberibacter MLSA genes had reads mapped to them in Combo-1, 2 and 6; all seven *Ca.* Phytoplasma MLSA genes had reads mapped to them in Combo-3 and 6; and all seven *Xylella* MLST genes had reads mapped to them in Combo-5 and 6 (Table 27.10). Read data for Combo-4 mapped to four *Xylella* and one *Ca.* Phytoplasma target regions.

Metagenomic reads were mapped with low stringency to all seven *Ca.* Liberibacter MLSA genes in Combo-1; all seven *Ca.* Phytoplasma MLSA genes in Combo-3 and 6; and all seven *Xylella* MLST genes in Combo-5 and 6 respectively detecting two fewer pathogens using metagenomic sequencing compared to hybridization capture.

Due to the lack of read data available for Xf in Combo 4, this data has not been progressed to the identification step. A short analysis did identify about 100 reads were mapped to the *leuA* gene of Xff, However, this was not sufficient to assemble a contig for identification.

27.3.4.2 Identification

Stringent mapping of reads from hybridization capture data to target MLSA genes of phytoplasma species (Phy database 1-10) (Method 1) resulted in complete coverage of the seven MLSA genes of Echinacea purpurea witches' broom phytoplasma (phytoplasma group 16SrII) in Combo-3 and 6, *Ca.* Phytoplasma australiense (phytoplasma group 16SrXII) in combo 3 and *Ca.* Phytoplasma stylosanthis (Phytoplasma group 16SrXXXVII) in combo 3 and 6 (Table 27.10). These were expected results based on the spiking of samples except that *Ca.* Phytoplasma australiense was not detected in combo 6.

Sequence comparison of contigs assembled from stringent mapping to the non-redundant database using BLASTn (Method 3) correctly identified spiked phytoplasma species to the 16Sr groups in Combo-3 and Combo-6. The top hits were Echinacea purpurea witches'-broom phytoplasma strain NCHU2014 and Strawberry lethal yellows phytoplasma (CPA) str. NZSb11 which are phytoplasma group 16SrII and 16SrXII respectively. Similar to Section 3.3.1, *Ca.* Phytoplasma stylosanthis (Phytoplasma group 16SrXXXVII) spiked into Combo-3 and Combo-6 were identified as *Ca*. Phytoplasma luffae (Phytoplasma group 16SrVIII) with reduced sequence identity (79-85% sequence identity) and only three contigs were identified as *Ca*. Phytoplasma luffae in comparison to the one contig identified as *Ca.* Phytoplasma stylosanthis in Combo-6. These species share high *in silico* RFLP and nucleotide identity (Rodrigues Jardim et al., 2021).

Candidatus Liberibacter species, Clso, CLas and Claf, spiked into Combo-1, 2 and 6 were all correctly identified using methods 1, 2 and 3 (Table 27.12). Off-target detection and impact of spiking CLbr into Combo 2 with Claf is discussed in 3.8.3.

Stringent mapping of reads from hybridization capture data to target MLST genes of *Xylella* species (Xf database 1-5) (Method 1) resulted in complete coverage of the seven MLST genes of Xff, Xfm, *Xylella fastidiosa* subsp. *sandyi* and Xt, in Combo-5 and Combo-6 and also *Xylella fastidiosa* subsp. *pauca* (Xfp) for combo 6. This result was not unexpected given that there is limited genetic diversity amongst the MLST target genes, with a small number of base pair differences between species and subspecies across the gene subset. When reads from stringent mapping were *de novo* assembled and the contigs compared to the non-redundant database (Method 27.3) and phylogenetically analysed (Method 2) both Xt and Xfm were identified in Combo-5 and 6 (Table 27.10 and Figure 27.4), both of which were spiked into the sample. Xff was also spiked into Combo-5 and 6 at a very low concentration. However, based on the reduced sensitivity of probe set 2 (described in section 27.3.5) it is likely that the concentration at which it was spiked would be within the limit of detection of this probe experiment.

Table 27.10 Detection of pathogen reads of *Xylella*, *Ca.* Phytoplasma and *Ca.* Liberibacter in mixed samples using Kraken2 and low stringency mapping to seven MLSA/MLST genes from each genus. Kraken2 results (K) are presented as the percentage of reads classified as the genera and mapping results (M) are presented as the number of MLST/MLSA genes with reads mapped.

		Metagenomic						Probed								
Sample	Species	Xylella		Phytoplasma		Liberibacter		Xylella			Phytoplasma			Liberibacter		
		M	K	M	K	M	K	M	K	Fold enrichment	M	K	Fold enrichment	M	K	Fold enrichment
Blank-apple	-	2	0	0	0	0	0	1	0	-	4	0	-	0	0	-
Blank-lemon	-	1	0	0	0.01	0	0	1	0	-	0	0	-	0	0	-
Blank-olive	-	1	0	0	0	0	0	1	0	-	1	0	-	0	0	-
Combo-1	Clso Clas	0	0	0	0.01	7 (9)	0.05	2	0	-	0	0	-	7 (21)	2.35	47
Combo-2	Claf CLbr	0	0	1	0.01	5	0.01	1	0	-	0.01	0	-	7 (19)	6.87	687
Combo-3	StLL AGY WFB	0	0	7 (15)	0.08	0	0.01	1	0	-	7 (35)	7.24	90.5	0	0	-
Combo-4	XFF XCC	1	0	2	0.01	0	0	4	0	-	1	0	-	0	0	-
Combo-5	XFF XFM XT	7 (24)	0.05	0	0	0	0	7 (34)	2.7	54	2	0	-	0	0	-
Combo-6	Clso Clas StLL AGY WFB XFF XFM XT	7 (23)	0.04	10	0.01	3	0.01	7 (34)	2.47	61.75	7 (27)	0.62	62	7 (19)	0.24	24

Table 27.11 Identification of mixed pathogen samples in comparison to original identification using mapping (all 7 MLSA/MLST genes mapped to with ≥85% gene coverage), BLASTn identification (top hits with % identity of ≥90% to isolates from non-redundant database from NCBI), percentage of identified pathogen contigs, and taxonomic identification based on phylogenetic analysis.

Sample	Species	Mapping identification (Method 1)	BLASTn Identification (Method 3)	BLASTn off-target identification % (Contigs) (Method 4)	Phylogeny identification (Method 2)
Combo-1	Clso	*Ca.* Liberibacter asiaticus	*Candidatus* Liberibacter asiaticus strain A4	*Ca.* Liberibacter asiaticus strain A4 57.89 (11/19)	*Ca.* Liberibacter asiaticus
	Clas	*Ca.* Liberibacter solanacearum	*Candidatus* Liberibacter solanacearum CLso-ZC1	*Ca.* Liberibacter solanacearum CLso-ZC1 42.11 (8/19)	*Ca.* Liberibacter solanacearum haplotype A
Combo-2	Claf	*Ca.* Liberibacter africanus	*Candidatus* Liberibacter africanus PTSAPSY	*Ca.* Liberibacter africanus PTSAPSY 87.5 (7/8)	*Ca.* Liberibacter africanus
	CLbr	*Not detected*	*Not detected*	*Not detected*	*Not detected*
Combo-3	WFB	16SrII-Echinacea purpurea witches' broom phytoplasma	16SrII - Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	Echinacea purpurea witches'-broom phytoplasma 92.16 (576/625)	16SrII-D- Sweet potato little leaf phytoplasma
	AGY	16SrXII - *Ca.* Phytoplasma australiense	16SrXII - Strawberry lethal yellows phytoplasma (CPA) str. NZSb11	Strawberry lethal yellows phytoplasma (CPA) str. NZSb11 1.6 (10/625)	16SrXII- *Ca.* Phytoplasma australiense
	StLL	16SrXXXVII - *Ca. Phytoplasma stylosanthis*	*Not detected*	*Ca.*Phytoplasma stylosanthis isolate 935[1] 0.16 (1/625)	16SrXXXVII- *Ca. Phytoplasma stylosanthis*
				Ca. Phytoplasma luffae isolate NCHU2019 0.48 (3/625)	
Combo-4	XFF	*Not detected*	*Not detected*	-	*Not detected*
	XCC	N/A	N/A	*Xanthomonas citri* pv *citri* 100 (10/10)	N/A
Combo-5	XFM	*Xylella fastidiosa* subsp. *multiplex*	*Xylella fastidiosa* subsp. *multiplex* (multiple strains)	*Xylella fastidiosa* subsp. *multiplex* (multiple strains) 58.97 (23/39)	*Xylella fastidiosa* subsp. *multiplex*
	XT	*Xylella taiwanensis*	*Xylella taiwanensis* strain PLS229	*Xylella taiwanensis* strain PLS229 35.9 (14/39)	*Xylella taiwanensis*

	XFF	*Xylella fastidiosa* subsp. *fastidiosa*	*Not detected*	*Not detected*	*Not detected*
		Xylella fastidiosa subsp. *sandyi*			
Combo-6	XT	*Xylella taiwanensis*	*Xylella taiwanensis* strain PLS229	*Xylella taiwanensis* strain PLS229 19.18 (14/73)	*Xylella taiwanensis*
	XFM	*Xylella fastidiosa* subsp. *multiplex*	*Xylella fastidiosa* subsp. *multiplex* (multiple strains)	*Xylella fastidiosa* subsp. *multiplex* (multiple strains) 28.77 (21/73)	*Xylella fastidiosa* subsp. *multiplex*
	XFF	*Xylella fastidiosa* subsp. *fastidiosa*	*Not detected*	*Not detected*	*Not detected*
		Xylella fastidiosa subsp. *pauca*	*Not detected*	*Ca.* Phytoplasma luffae isolate NCHU2019 4.11 (3/73)	
		Xylella fastidiosa subsp. *sandyi*	*Not detected*	*Not detected*	
	StLL	16SrXXXVII - *Ca. Phytoplasma stylosanthis*	*Not detected*	*Ca.*Phytoplasma stylosanthis isolate 935[1] 1.37 (1/73)	16SrXXXVII - *Ca. Phytoplasma stylosanthis*
	AGY	*Not detected*	*Not detected*	-	*Not detected*
	WFB	16SrII-A -Echinacea purpurea witches' broom phytoplasma	16SrII-A - Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	16SrII-A - Echinacea purpurea witches'-broom phytoplasma 13.7 (10/73)	16SrII-D - Sweet potato little leaf phytoplasma
	Clas	*Ca.* Liberibacter asiaticus	*Ca.* Liberibacter asiaticus strain A4	*Ca.* Liberibacter asiaticus strain A4 9.59 (7/73)	*Ca.* Liberibacter asiaticus
	Clso	*Ca.* Liberibacter solanacearum	*Ca.* Liberibacter solanacearum CLso-ZC1	*Ca.* Liberibacter solanacearum CLso-ZC1 15.07 (14/73)	*Ca.* Liberibacter solanacearum haplotype A

[2] BLASTn sequence identity ≥75%.

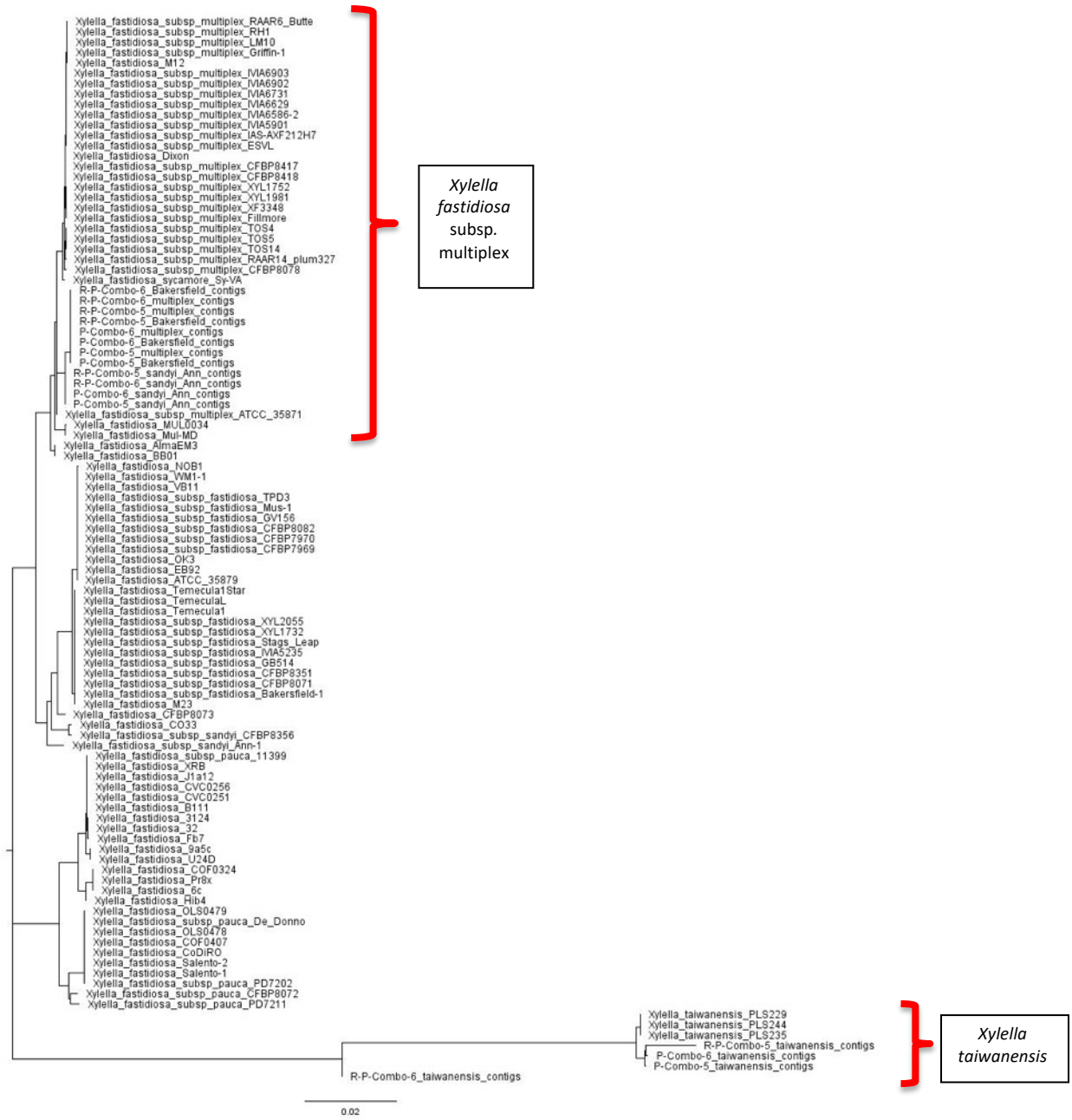


Figure 27.4 Maximum likelihood tree based on sequences of *Xylella* MLST genes *gltT, lacF, nuoL, holC, leuA, cysG* and *petC* from probed PEQ, combination (Combo) and modified pooling protocol (R-PEQ and R-Combo) samples and their closest phylogenetic neighbours.

27.3.5 Hybridization probe sensitivity

A serial dilution of Xf in DNA from olive tissue was used to assess the sensitivity of the use of hybridization capture probe sets 1 and 2 compared to metagenomic sequencing (non-probed), LAMP and qPCR.

Using both Kraken2 and the low stringency mapping approach for target read detection, the presence of DNA from the *Xylella* genus was detected in metagenomic samples as low as 2.4 x 10^{-02} ng/µL and in hybridization capture samples (Probe set 2) as low as 2.4 x 10^{-04} ng/µL (Table 27.12).

As expected, the percentage of Xf reads in metagenomic data was consistent across experiments assessing probe set 1 and probe set 2 (Table 27.12).

Using probe set 2, the percentage of total Xf reads was higher for probed samples in comparison to metagenomic samples. This trend was also seen for probe set 1. Using Kraken2, Xf was detected confidently down to 2.4 x 10^{-02} ng/µL for metagenomic Xf samples, 1,000 X and 10,000 X less sensitive than LAMP and qPCR respectively. Using probe set 1, Xf was detected confidently down to 2.4 x 10^{-06} ng/µL, 10 X more sensitive than LAMP and equally as sensitive as qPCR. However, using the probe set 2, Xf was detected confidently down to 2.4 x 10^{-04} ng/µL, 100X less sensitive than the probe set 1.

Using StringMLST on trimmed reads, the ST was unable to be identified in the first dilution in metagenomic Xf samples using probe set 2. This is because in the first dilution, 0.24 ng/µL, one replicate was identified as ST1, while the other was identified as ST2. At Xf concentration 2.4 x 10^{-02} ng/µL, only four out of seven MLST genes were identified, which was insufficient to allocate an ST (Table 27.12).

Probe set 1 did not contain the probes for the *petC* gene so ST, which requires seven genes to identify, was not determined using these probes. Nevertheless, StringMLST was still able to detect the rest of the six MLST genes up to Xf concentration 2.4 x 10^{-05} ng/µL which could at least be used to assign subspecies. StringMLST was also able to identify some MLST genes at Xf concentration 2.4 x 10^{-06} ng/µL using probe set 1, however, not all six genes were detected.

Using probe set 2, which included *petC,* ST was detected down to Xf concentration 2.4 x 10^{-03} ng/µL. Whilst it is not a direct comparison, this is 100X less sensitive than the detection of all six Xf genes included in probe set 1 and is consistent with the trend of reduced sensitivity using probe set 2 observed in the kraken and mapping results. StringMLST was also able to identify four out of seven MLST genes at Xf concentration 2.4 x 10^{-04} ng/µL, however, it was insufficient to allocate an ST (Table 27.14).

The reduced sensitivity of probe set 2 to detect *Xylella* indicates that a concentration of *Xylella* below 2.4 x 10^{-04} would not be detected. In section 27.3.4, the concentration of Xff spiked into combo 5 was 2.4 x 10^{-05} ng/µL (CT 34.2) which was below the detection limit for this new set of probes. Xff was not detected in Combo-6 either as it was a mix of Combo-1, Combo-3 and Combo-5, hence, diluting the Xff concentration even further.

The reduced sensitivity achieved by hybridization capture Probe set 2 is explored further in section 27.3.7.3.

Table 27.12 LAMP and qPCR results from Milestone 105, and results for metagenomic (meta) and hybridization capture sequencing of serially diluted *Xylella fastidiosa* (Xf) DNA into olive DNA using probe set 2 (this report) and probe set 1 (Milestone 106). Kraken2 (K) was used to assign *Xylella* reads and results are presented as a percentage of the total reads. Low stringency mapping results (M) are presented as the number of MLST gene that had full read coverage. StringMLST results (S) are presented either the sequence type (ST) of *Xylella* identified or Xf if sufficient information was retrieved to identify the species. UD = Undetermined, Ct = threshold cycle of amplification in qPCR.

				Probe set 2						Probe set 1			
				M		K		S		K		S	
Sample	**Concentration of *Xylella* DNA (ng/µL)**	**qPCR (Ct)**	**LAMP (amp time)**	**Meta**	**Probed**	**Meta**	**Probed**	**Meta**	**Probed**	**Meta**	**Probed**	**Meta**	**Probed[2]**
XF-0	0.24	21.09	9.9	7(16)	7(33)	0.37	31.6	ST2/ST1	ST2	0.4	96.49	ST2	XF
XF-1	2.4×10^{-02}	24.52	11.2	7(5)	7(26)	0.02	2.29	XF[1]	ST2	0.02	72.37	XF[1]	XF
XF-2	2.4×10^{-03}	27.87	12.9	2	7(19)	0	0.25	-	ST2	0	20.29	-	XF
XF-3	2.4×10^{-04}	31.17	14.7	0	7(8)	0	0.02	-	XF[1]	0	2.78	-	XF
XF-4	2.4×10^{-05}	34.21	20.4	0	1	0	0	-	-	0	0.34	-	XF
XF-5	2.4×10^{-06}	37.45	UD	0	0	0	0	-	-	0	0.05	-	XF[1]
XF-6	2.4×10^{-07}	UD	UD	0	0	0	0	-	-	0	0	-	-
XF-7	2.4×10^{-08}	UD	UD	0	1(2)	0	0	-	-	0	0	-	-
Negative (olive)	0	UD	UD	0	0	0	0	-	-	0	0	-	-

[1] Incomplete gene typing

[2] *petC* was not included in the probe set

27.3.6 Evaluation of modified library pooling protocol

A modified library pooling protocol was trialled to reduce costs by increasing the number of libraries per hybridization capture reaction. Instead of eight libraries pooled with 187.5 ng library per sample, 18 libraries were pooled with 50-100 ng library per sample. The 18 libraries used to assess the modified pooling protocol were a subsample of the libraries assessed in this report using the manufacturers pooling protocol (Table 27.2, 27.3, and 27.4).

The results for target read detection using Kraken2 analysis and low stringency mapping using data generated with the modified pooling protocol were consistent with the results generated with the same libraries using the manufacturers protocol (data not shown).

Pathogen identification in samples using assembled contigs from stringent mapping for phylogenetic analysis (Method 1) using the modified pooling protocol data, generated results consistent with those generated with the manufacturers protocol with most samples (Figure 27.2 and 27.4). Pooled protocol results that were not consistent with those generated using the manufactures protocol were that *Ca.* Phytoplasma 16SrXII was not detected in R-Combo-3, CLso was not detected in R-Combo-1 and R-Combo-6), contigs from R-PEQ-P-17 did not cluster with *Ca.* Phytoplasma 16SrXII (Figure 27.2) and contigs from R-Combo-6 clustered just outside the other *X. taiwanenesis* contigs (Figure 27.4).

These discrepancies in the results were mostly due to a lack of read depth to generate full contigs for further analysis. Even though MLSA genes had ≥90% gene coverage, when assembled using *de novo* assembly, the *secA* gene from R-Combo-3 only covered 60% of the reference gene while the *Xylella taiwanensis* from R-Combo-6 only covered 87%. Though coverage was almost complete for all genes, the depth was too low to assemble full genes using *de novo* assembly. When looking at output data, there was a disproportionate amount of data (8.5 GB) was generated for R-Combo-3, which reduced the output of other samples. This disproportionate allocation of data within a probe reaction was not observed for the normal probe protocol indicating that reducing the amount of DNA per library per probe reaction may affect the output or an error occurred during the pooling step. It has been noted by the manufacturer that reducing amount of DNA per library may affect the output as it would lead to more off-target binding. In this case, even though there was a high proportion of reads generated for R-Combo-3 (83,831,832 reads), we were unable to detect *Ca.* Phytoplasma australiense (phytoplasma 16SrXII) whereas using the manufacture's protocol we detected it from the same sample library from 16,831,372 reads. Based on these results we will not be pursuing the use of the modified pooling protocol.

3.7 Off-target capture

3.7.1 Off-target detection

Contigs assembled using reads from low stringency mapping were compared to non-redundant sequence database at NCBI using BLASTn. The percentage of contigs that had high sequence similarity (>90% identify) to the target pathogens was calculated for each sample and these are listed in Table 7 and 8.

In samples known to be infected with a *Ca.* Phytoplasma the percentage of contigs that had highest sequence similarity to the target pathogen ranged between 43.98% and 94.9%. In these samples, the majority of the non-target contigs had high sequence similarity to plant DNA. However, in PEQ-P-4 and PEQ-P-18, which had lower percentages of target contigs identified than other samples, 43.89% and 47.37% respectively, additional contigs had higher sequence identity to other *Ca.* Phytoplasma species (Table 9). For PEQ-P-4, 1,067 contigs (11.81%) shared higher sequence similarity to *Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma and Sesame phyllody phytoplasma. For PEQ-P-18, six contigs (31.58%) shared sequence similarity to *Ca.* Phytoplasma stylosanthis, *Ca.* Phytoplasma ziziphi and 'Vitis vinifera' yellows phytoplasma.

No significant off target detection was observed for *Ca.* phytoplasma spp. Not all MLSA genes were able to be used for phylogenetic analysis due to the lack of sequence data available for the species studied. As described earlier the taxonomic demarcation criteria for genes other than the 16S is not clearly defined and this could lead to incorrect assignment of a species. There may also be biases due to the incomplete phytoplasma genes and genomes in the non-redundant database at NCBI. This may lead to multiple *Ca.* Phytoplasma identification as smaller contigs may hit to the partial genes of the correct or incorrect phytoplasma species especially when they align to more conserved regions. Therefore, it is critical to ensure good breadth of gene coverage, to ensure accurate identification. The minimum breadth and depth of coverage needs to be determined. Since all samples were paired end sequenced (2 x 250 bp), small contig size could also indicate poor quality reads that could affect identification as a few SNPs could easily misidentify species within the same group, therefore the quality of reads used for analysis also need to be defined. We cannot, however, rule out the possibility that mismatched could represent a mixed infection of two or more *Ca*. Phytoplasma species.

In samples known to be infected with a *Ca.* Liberibacter the percentage of contigs that had highest sequence similarity to the target pathogen ranged between 53.06% and 100%. Unlike in the *Ca.* Phytoplasma analysis, the non-target contigs

assembled in these samples only shared high sequence similarity to plant DNA.

Probed reads from sample PEQ-P-14, which was infected with Echinacea purpurea witches'-broom phytoplasma, not only mapped to the target pathogen using low stringency mapping, but also mapped to 15 *Xylella* contigs in the HPR database representing four genes (*holC*, *nuoL*, *leuA* and *lacF*). However, no reads were classified as *Xylella* using Kraken2. When the assembled contigs were compared to the non-redundant database at NCBI using BLASTn, 84.9% of contigs had top hits to the target pathogen Echinacea purpurea witches'-broom phytoplasma strain NCHU2014. BLASTn sequence similarity results also indicated that *Xanthomonas arboricola* is present with six contigs (0.6%) sharing high sequence identity to the *leuA* gene of *Xylella fastidiosa* subsp *pauca* (31.6% gene coverage and 78% sequence identity), the *nuoL* gene of *Xylella fastidiosa* subsp *sandyi Ann* (16.5% gene coverage and 80.6% sequence identity) and the *lacF* gene of *Xylella fastidiosa* subsp *pauca* (21.7% gene coverage and 79.2% sequence identity).

These results indicate that there were varying amounts of off-target hybridization capture in each sample. Much of this could be the result of the lower stringency of probe set 2 (discussed further in Section 3.8) as compared to probe set 1 and is likely to explain some of the off-target capture of plant DNA. However, off-target capture of closely related species or genera, as seen with the capture of *Xanthomonas* reads with the *Xylella* probes is of interest. Whilst this assay is not designed to identify unknown pathogens, it is an added value that closely related species may also be detected depending on how you choose to interrogate the data. In Milestone 106 we demonstrated that when hybridization probe set 1 was used on a sample infected with CLbr, a *Ca.* Liberibacter species most closely related to CLaf, all seven *Ca.* Liberibacter brunswickensis housekeeping genes had reads partially mapped to them. However, when probed CLbr reads were mapped to CLaf MLSA genes, only 158 reads were mapped in comparison to the 538,619 reads mapped to CLbr MLSA genes. This indicates that the bioinformatic pipeline may be excluding some off-target reads. The off-target capture of CLbr is further explored in section 3.7.2, where it has been spiked into a sample with CLaf.

Further work needs to be done to assess how much off-target capture occurs with bacteria closely related to the target species of the current probe set, and it is envisaged that off-target capture will vary for each target region and will need to be assessed for each new target region added.

Table 27.13 Identification of target *Ca.* Phytoplasma and off-target species in *Ca.* Phytoplasma PEQ samples using BLASTn identification (top hits with % identity of ≥75% to isolates from non-redundant database from NCBI), and percentage of identified species contigs.

Sample	16Sr group	Identification/ target pathogen	Contig (%) target pathogen	Off-target contig identification	Contig (%) off-target
PEQ-P-4	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	43.98 (3973/9034)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, Sesame phyllody phytoplasma	11.81 (1,067/9034)
PEQ-P-5	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	76.68 (7146/9319)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, Sesame phyllody phytoplasma	4.96 (462/9319)
PEQ-P-7	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	74.13 (1229/1658)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma, Sesame phyllody phytoplasma	12.85 (213/1658)
PEQ-P-8	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	83.81 (207/247)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma	9.3 (23/247)
PEQ-P-9	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	87.5 (7/8)	-	-
PEQ-P-10	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	91.91 (432/470)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma	5.53 (26/470)
PEQ-P-11	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	93.68 (489/522)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma, 'Gendarussa vulgaris' witches'-broom phytoplasma, 'Tephrosia purpurea' witches'-broom phytoplasma	5.94 (31/522)
PEQ-P-12	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	91.65 (999/1090)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma, 'Gendarussa vulgaris' witches'-broom phytoplasma	8.17 (89/1090)
PEQ-P-14	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	84.9 (832/980)	*Xanthomonas arboricola* pv. juglandis, *Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus'	8.67 (85/980)

				phyllody phytoplasma	
PEQ-P-15	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	90.34 (458/507)	'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma, 'Gendarussa vulgaris' witches'-broom phytoplasma	8.48 (43/507)
PEQ-P-16	16SrII	Echinacea purpurea witches'-broom phytoplasma strain NCHU2014	94.9 (1675/1765)	*Ca*. Phytoplasma aurantifolia, 'Cicer arietinum' phyllody phytoplasma, 'Parthenium hysterophorus' phyllody phytoplasma, 'Gendarussa vulgaris' witches'-broom phytoplasma	3.8 (67/1765)
PEQ-P-17	16SrXII	*Ca.* Phytoplasma australiense (Strawberry lethal yellows phytoplasma (CPA) str. NZSb11)	63.64 (7/11)	*Ca*. Phytoplasma australiense	9.09 (1/11)
PEQ-P-18	16SrVIII	*Ca*. Phytoplasma luffae isolate NCHU2019	47.37 (9/19)	*Ca.* Phytoplasma stylosanthis, *Ca.* Phytoplasma ziziphi, and, 'Vitis vinifera' yellows phytoplasma	31.586/19)

27.3.7.2 Off-target capture of closely related Ca. Liberibacter sp

To assess off-target hybridization capture of sequence from closely related species, data sets generated with probe set 1 from one sample infected with CLbr and another infected CLaf were analysed. The two *Ca.* Liberibacter species are closely related, and probes were only designed for CLaf, not CLbr.

As expected, visualisation of the hybridization capture data from the CLaf infected sample and Combo-2 (CLaf and CLbr) mapped to the CLaf MLSA genes shows complete breadth of coverage of all genes of the target pathogen (Figure 27.7). When reads generated from hybridization capture of the CLbr infected sample were mapped to CLaf genes, only 158 reads mapped providing insufficient reads for identification analysis (data not shown). However, when the same data set were mapped to the CLbr genome, 545,599 total reads mapped across all seven genes even though the hybridization probes were not designed to detect this species (Figure 27.7). Only partial coverage of the CLbr MLSA genes was observed and *de novo* assembly of mapped reads only generated three contigs large enough for comparison. These three contigs were compared to the CLaf genes using the BLASTn algorithm. Results indicated that requirements for off-target capture could be as low as 81.53% sequence identity with an e-value of 2e-69 (Figure 27.7).

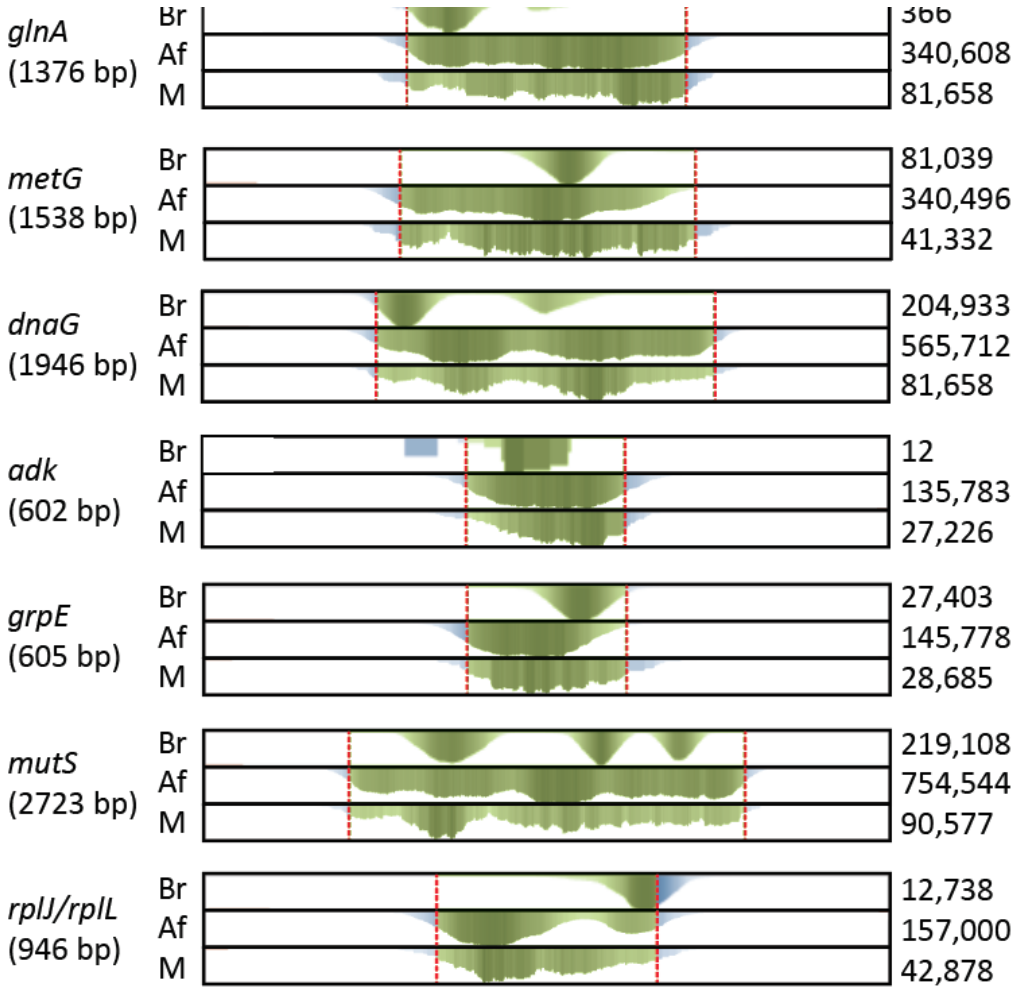


Figure 27.5 Mapping of hybridization data from a *Ca.* Liberibacter brunswickensis infected sample (Br) using *Ca.* Liberibacter brunswickensis MLSA genes as reference, and a *Ca.* Liberibacter africanus infected sample (Af), and a sample infected with both *Ca.* Liberibacter aricanus and *Ca.* Liberibacter brunswickensis (Combo 2) (M) using *Ca.* Liberibacter africanus MLSA genes as reference. Green regions within the red dotted lines indicate gene location while blue regions are 1 kb sequences 5' and 3' of the gene.

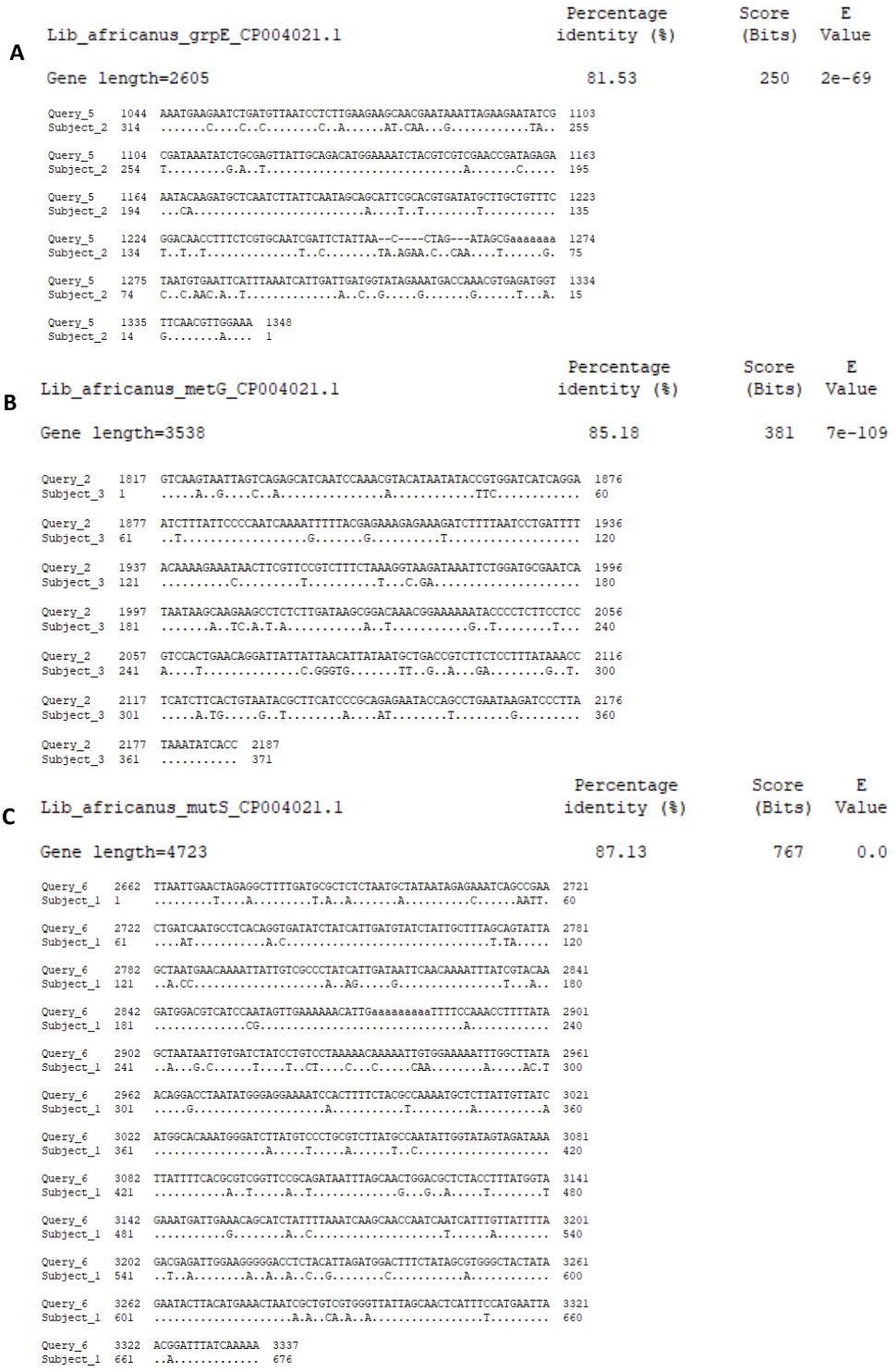


Figure 27.6 BLASTn nucleic acid sequence alignment, percentage identity and e-value of *Ca.* Liberibacter brunswickensis contigs (Subject 1-3) compared to *Ca.* Liberibacter africanus genes (A) *grpE*, (B) *metG* and (C) *mutS* genes.

27.3.7.3 AT/GC dropout

Picard CollectHsMetrics collects hybrid-selection (HS) metrics for a mapping output file, SAM or BAM. One of the Picard CollectHsMetrics outputs of interest is the AT/GC dropout. The AT/GC dropout metrics indicate the degree of reduced mapping efficiencies of a particular region based on its AT or GC content. This is because AT rich regions are known to reduce binding stability of the hybridization capture probes leading to more off-target binding, and GC rich regions could lead to difficulty in capture and sequencing bias (Bodi et al., 2013; Dannemann et al., 2009). It has been identified that when using hybridization capture, off-target binding increases in the absence of target DNA. For this analysis, we used the low stringency mapped BAM files of the Xf dilution series from this milestone and Milestone 106 to initially assess the use of the AT/GC dropout metric as the concentration of target DNA reduced. as expected, with the reduction in concentration of Xf, the AT/GC dropout percentage increased indicating reduced mapping efficiency (Table 15).

When picard CollectHsMetrics AT/GC dropout results for probe set 1 were compared to AT/GC dropout results for probe set 2 it was observed that probe set 1 maintained AT/GC dropouts below 50% until *Xylella* concentration was as low as 2.4 x 10^{-05} ng/µL whilst AT dropouts of above 50% were recorded for probe set 2 from the first dilution (0.24 ng/µL) (Table 18). These results indicate that the reduction in stringency of low complexity regions during the design of probe set 2 have impacted the overall sensitivity of this probe set for pathogen detection.

Table 27.14 Picard CollectHsMetrics AT/GC dropout results for the serially diluted *Xylella fastidiosa* (XF) samples using the probe set 1 (Milestone 106) and probe set 2.

Sample	Concentration of *Xylella* DNA (ng/µL)	AT/GC dropouts	Probe set 2	Probe set 1
XF-0	0.24	AT	55.7970	41.6918
		GC	17.9252	25.2187
XF-1	2.4 x 10^{-02}	AT	55.9052	41.7198
		GC	18.0802	26.2214
XF-2	2.4 x 10^{-03}	AT	56.2446	41.5625
		GC	20.5648	25.2056
XF-3	2.4 x 10^{-04}	AT	55.0192	43.1824
		GC	33.2805	29.3585
XF-4	2.4 x 10^{-05}	AT	54.0154	47.8220
		GC	39.1110	49.2757
XF-5	2.4 x 10^{-06}	AT	55.9508	46.7115
		GC	42.0411	53.3166
XF-6	2.4 x 10^{-07}	AT	54.2761	45.3531
		GC	42.0411	56.8207
XF-7	2.4 x 10^{-08}	AT	58.1343	45.6010
		GC	40.7951	56.8207

27.3.8 Bioinformatic pipeline

27.3.8.1 Rapid target read detection

The proposed initial step in the bioinformatic pipeline for analysis of hybridization capture data involves the use of the K-mer based tool (Kraken2) and the mapping program (bowtie2) as a means to detect target reads in samples. This first assessment of the data is intended to provide a swift method to exclude samples from further analysis; assessment of 9 million reads takes approximately 10 mins using Kraken2 and approximately 4 minutes mapping with bowtie2. Data of this analysis can be seen in Table 6, 9 and 11, and would act as a checkpoint to exclude negative samples from further processing. In this Milestone, we assessed the cut-off points to identify 'positive' samples. When analysing Kraken2 data from Milestone 106 where pathogen reads and percentages were assessed in detail, ≥0.02% of pathogen genus reads was the minimum read number associated with positive detection. This minimum value could also accurately be used in this report as a cut off for a positive sample and will be further assessed with validation on closely related non-target species in subsequent experiments.

The second analysis used in the rapid detection of target reads was a low stringency mapping to MLSA/MLST genes

from multiple species/sub-species from a pathogenic genus.

Though results from Kraken2 were straightforward (e.g. a ratio of reads matching the target genus), there are multiple considerations for the low stringency mapping step. Firstly, since the database used for mapping includes representative MLSA/MLST genes from multiple species/subspecies per genera (e.g. MLST genes from four representative strains of Xf, therefore a total of 28 representative genes were included in the target reference database for *Xylella*) more than seven contigs were mapped to for many probed samples (Table 6, 9 and 11). This result could be due to multiple species/subspecies being present within a sample or due to off-target mapping to conserved regions of the MLSA/MLST genes. An example of this was observed for sample PEQ-P-14 where 15 *Xylella* contigs in the HPR database from multiple subspecies had reads mapped to them but no/low percentage of reads were classified as *Xylella* using Kraken2. When further analysed using BLASTn of assembled contigs to compare to the non-redundant database in NCBI, the results indicated that the reads that mapped to the *Xylella* MLST genes were from *Xanthomonas arboricola*. *Xanthomonas* is the most closely related genera to *Xylella* and it is not surprising that reads would map to the conserved regions of the reference genes. To overcome this problem, we initially propose that mapping to all target seven MLSA/MLST genes (regardless of the number of contigs with mapped reads) is required for samples to move forward for further analysis. Additionally, we believe that adding a cut-off parameter for the breadth of read coverage of target regions from the mapping output would exclude samples with off-target mapping to conserved regions. These additional steps will be further assessed in subsequent experiments to give confidence in pathogen detection in this triage step.

It is worth noting that these cut-off parameters are proposed for analysis of reads from probed samples and may be too stringent for use with metagenomic samples. For example, in Combo-P-6 which was spiked with three phytoplasma species, reads mapped to 10 *Ca.* Phytoplasma genes and only 0.01% of reads were *Ca.* Phytoplasma. Due to the low number of reads, no contigs could be assembled.

In summary for the initial data assessment step:

- Using K-mer based read classification (Kraken2) and low stringency mapping combined as a rapid tool for initial assessment of hybridization capture data was successful. We propose that using parameters ≥0.02% of pathogen reads classified as the target genus (Kraken2) and mapping to all MLSA/MLST genes was sufficient to indicate the presence of the target pathogens in this experiment.
- The addition of another cut-off parameter for breadth of coverage for low stringency mapping will be assessed to determine whether this rapid data assessment could more definitively define a positive result.
- Both Kraken2 and low stringency mapping results were needed in combination for a confident indication of target pathogen presence in the initial assessment step. Neither method provided accurate results independently.
- The parameters proposed here are for analysis of hybridization capture data and our analysis indicates that they will not always provide accurate results for metagenomic data.

27.3.8.2 Identification

The initial analysis assessed for pathogen identification used reads from low stringency mapping for *de novo* assembly (Method 4). However, when multiple closely related pathogens were present within a sample, differentiation of genes and subsequent identification of pathogens proved problematic. To circumvent this issue, stringent mapping prior to *de novo* assembly (Method 1) was also trialled in comparison to low stringency mapping. Noting that *de novo* assembly using SPAdes can take up to 8 hours for 9 million reads, this step can be a bottle neck of this bioinformatic pipeline and it is desirable that the input is as accurate as possible and the step only needs to be run once. High stringency mapping of reads to a target database (Method 1) enabled accurate identification of closely related subspecies present within the same sample that could not be separated using contigs assembled from low stringency mapping (Method 4) but also provided accurate data sets for *de novo* assembly of contigs for use in methods 2 and 3.

We found that the most informative means of pathogen identification was phylogenetic analysis of the contigs assembled from mapped reads (Method 2) as it enabled the relatedness of an organism to closely related species to be estimated. However, for low titre infections there were sometimes insufficient reads to assemble full length genes for phylogenetic analysis, but the assembled contigs (partial genes) could still enable pathogen identification by sequence comparison to a reference database using BLASTn (Method 3). Hence, moving forward with this analysis in the future we will continue to output the results of the three identification methods outlined in this report. For BLASTn of contigs to reference databases and stringent mapping, good databases are required for confident identification.

In summary:

- *De novo* assembly of low stringency mapped genes (Method 4) was time and compute-intensive, and did not always enable sufficient differentiation of closely related bacteria (e.g. subspecies of Xf) for identification. For this reason, this method will not be considered as an identification tool in future analysis, instead, only incorporated to assess off target hybridisation where required.
- The addition of a stringent mapping step to a specific target database (Method 1) not only provided a preliminary identification but also aided in streamlining *de novo* read assembly for subsequent analysis. In particular, the addition of this step aided identification of multiple pathogen infections in the one sample, specifically closely related species/subspecies.
- Phylogenetic analysis of assembled contigs was the most informative and accurate identification method providing information of taxonomy and relatedness to other species or strains. However, for some low titre infections there were insufficient reads to assemble full genes for analysis.
- Comparison of assembled contigs (full or partial) to vouchered reference databases using Method 3 was a valuable analysis that complemented the output of Method 2. When results from Methods 2 and 3 were the same it provided confidence in the results. If only partial contigs were assembled for low titre pathogens, Method 3 could still provide results with a level of confidence that the target pathogen had been detected.
- For identification of target pathogens from hybridization data it is recommended that outputs from Methods 1,2 and 3 are assessed together to provide an accurate result.

27.4 Conclusion

In conclusion, even though probe set 2 had reduced sensitivity this protocol shows further value for the use of targeted HTS hybridization capture for unculturable and cultural pathogen detection and identification from plant samples with single or multiple pathogens infections.
The major findings include:

- Initial target pathogen read detection was successfully achieved using a rapid triage step that conserved computational resources.
- The inclusion of all Xf MLST genes enabled sensitive sequence type assignment of Xf.
- Identification required a number of analyses that successfully enabled the taxonomic classification of target pathogens to a high resolution (to subspecies, sequence type or haplotype for some pathogens). However, accurate identification was reliant on high quality databases of genomic data. For example, taxonomic classification for *Ca*. Phytoplasma spp. was impacted by the availability of reference sequence material.
- We determined that increasing the number of libraries pooled into a single capture reaction decreased the sensitivity of the assay.
- The reduction of stringency for low complexity regions in the design of new probes for detection of *Ca.* Phytoplasmas decreased the sensitivity of hybridization capture and increased the amount of off-target capture. This has been a valuable discovery and will aid in future probe design.

Acknowledgments

Multiple people and organizations have aided this project in providing samples for the hybridization probe evaluation along with their expertise. This included Bianca Rorigues Jardim and the “VG16086 Area wide management of vegetable diseases: viruses and bacteria” team for providing phytoplasma samples; the NAQS team for providing *Ca.* Liberibacter asiaticus and phytoplasma DNA and desiccates; and the PEQ team for phytoplasma plant material. We would also like to thank the Ministry for Primary Industries for providing *Xylella* DNA. Bianca and the VG16086 Area wide management of vegetable diseases: viruses and bacteria also provided full genome sequences of *Ca.* Phytoplasma stylosanthis used in this study.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., ... & Zianni, M. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. *Journal of biomolecular techniques: JBT*, *24*(2), 73.

Bronzato Badial, A., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., & King, J. (2018). Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. *Plant Dis, 102*(8), 1648-1652. doi:10.1094/pdis-04-17-0488-re

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., . . . Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology, 68*(2), 229-238. doi:10.1111/ppa.12957

Chang, S. H., Cho, S. T., Chen, C. L., Yang, J. Y., & Kuo, C. H. (2015). Draft Genome Sequence of a 16SrII-A Subgroup Phytoplasma Associated with Purple Coneflower (Echinacea purpurea) Witches' Broom Disease in Taiwan. *Genome announcements*, *3*(6), e01398-15.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Christensen, N. M., Nicolaisen, M., Hansen, M., & Schulz, A. (2004). Distribution of phytoplasmas in infected plants as revealed by real-time PCR and bioimaging. *Molecular Plant-Microbe Interactions*, *17*(11), 1175-1184.

Dannemann, M., Lorenc, A., Hellmann, I., Khaitovich, P., & Lachmann, M. (2009). The effects of probe binding affinity differences on gene expression measurements and how to deal with them. *Bioinformatics*, *25*(21), 2772-2779.

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics, 6*(348). doi:10.3389/fgene.2015.00348

Ghosh, A., Mehta, A., & Khan, A. M. (2019). Metagenomic Analysis and its Applications. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 184-193). Oxford: Academic Press.

Gupta, A., Jordan, I. K., & Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics, 33*(1), 119-121. doi:10.1093/bioinformatics/btw586

Harper, S. J., Ward, L. I., & Clover, G. R. G. (2010). Development of LAMP and Real-Time PCR Methods for the Rapid Detection of Xylella fastidiosa for Quarantine and Field Applications. *Phytopathology®, 100*(12), 1282-1288. doi:10.1094/PHYTO-06-10-0168

Hodgetts, J., & Dickinson, M. (2010). Phytoplasma phylogeny and detection based on genes other than 16S rRNA. *Phytoplasmas: Genomes, Plant Hosts, and Vectors. Wallingford, UK: CABI Publishing*, 93-113.

Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., & Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. *Phytobiomes Journal, 3*(2), 92-101. doi:10.1094/pbiomes-01-19-0004-r

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, *35*(6), 1547.

Li, W., Hartung, J. S., & Levy, L. (2006). Quantitative real-time PCR for detection and identification of Candidatus Liberibacter species associated with citrus huanglongbing. *Journal of microbiological methods*, *66*(1), 104-115.

Martini, M., Quaglino, F., & Bertaccini, A. (2019). Multilocus genetic characterization of phytoplasmas. In *Phytoplasmas: plant pathogenic bacteria-III* (pp. 161-200). Springer, Singapore.

Nelson, W. R., Sengoda, V. G., Alfaro-Fernandez, A. O., Font, M. I., Crosslin, J. M., & Munyaneza, J. E. (2013). A new haplotype of “Candidatus Liberibacter solanacearum” identified in the Mediterranean region. *European Journal of Plant Pathology*, *135*(4), 633-639

Jardim, B. R., Kinoti, W. M., Tran-Nguyen, L. T., Gambley, C., Rodoni, B., & Constable, F. E. (2021). ‘Candidatus Phytoplasma stylosanthis’, a novel taxon with a diverse host range in Australia, characterised using multilocus sequence analysis of 16S rRNA, secA, tuf, and rp genes. *International Journal of Systematic and Evolutionary Microbiology*, *71*(1)

Thompson, S. M., Johnson, C. P., Lu, A. Y., Frampton, R. A., Sullivan, K. L., Fiers, M. W., ... & Smith, G. R. (2015). Genomes of ‘Candidatus Liberibacter solanacearum’haplotype A from New Zealand and the United States suggest significant genome plasticity in the species. *Phytopathology*, *105*(7), 863-871

Tran-Nguyen, Lucy & Smith, S. & Liberato, José. (2012). Sweet potato little leaf strain V4 phytoplasma associated with snake bean in the Northern Territory, Australia. Australasian Plant Disease Notes. 7. 10.1007/s13314-012-0071-9

Tseng, Y. W., Deng, W. L., Chang, C. J., Su, C. C., Chen, C. L., & Jan, F. J. (2012). First Report of a 16SrII-A Subgroup Phytoplasma Associated with Purple Coneflower (Echinacea purpurea) Witches'-Broom Disease in Taiwan. *Plant disease*, *96*(4), 582-582.

Weller, S. A., Elphinstone, J. G., Smith, N. C., Boonham, N., & Stead, D. E. (2000). Detection of Ralstonia solanacearum strains with a quantitative, multiplex, real-time, fluorogenic PCR (TaqMan) assay. *Applied and Environmental Microbiology, 66*(7), 2853-2858. doi:10.1128/aem.66.7.2853-2858.2000

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:10.1186/s13059-019-1891-0

Zhao, Y., Wei, W., Lee, M., Shao, J., Suo, X., & Davis, R. E. (2009). Construction of an interactive online phytoplasma classification tool, iPhyClassifier, and its application in analysis of the peach X-disease phytoplasma group (16SrIII). *International Journal of Systematic and Evolutionary Microbiology*, *59*(Pt 10), 2582.

Appendix 28 – Improvement of RNA extraction and assessing small RNA-seq controls.

Improvement to nucleic acids extraction protocol

Degradation of RNA can result in the over representation of non-informative genetic sequences in the HTS datasets including ribosomal RNA, chloroplast RNAs, mitochondrial RNA and other host RNAs. The SSG team performed several assays and identified that modifying the timing of the homogeinisation buffer incorporation (Maxwell Total RNA extraction kit – Promega) during RNA extraction yielded improved results:

- less RNA degradation, recovery of higher RNA quality
- increase proportion of usable/informative read content (e.g., ribosomal RNA content drops to < 20% compared to > 55%) (**Figure 28.1**)
- cleaner small RNA read size distribution profile (**Figure 28.2**)

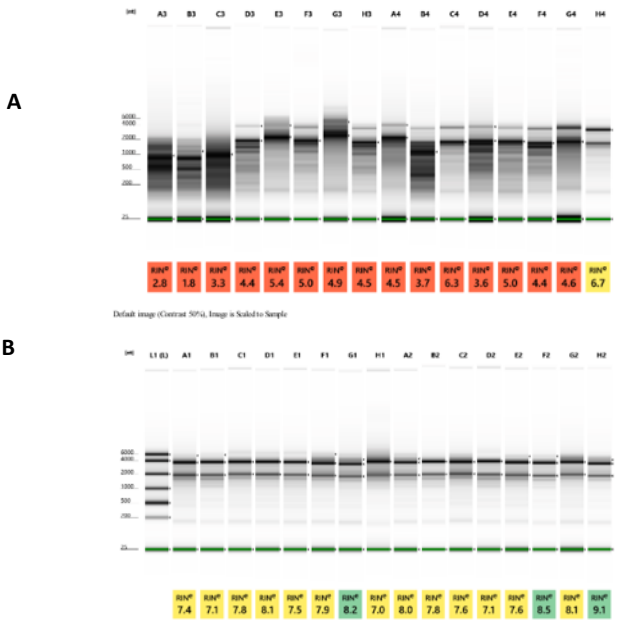


Figure 28.1: Example of electrophoresis gels via TapeStation from RNA extractions derived before (A) and after (B) modifying the incorporation of the homogeinisation buffer.

A significant improvement of the small RNA profiles after modification of the extraction protocol was achieved (Figure 28.2), particularly, small RNA peaks at 21, 22 and 24 nt in length are clearly defined showing that there was no RNA degradation as compared to the initial approach that yielded additional peaks typically corresponding to degraded ribosomal RNAs and other host plant RNAs.

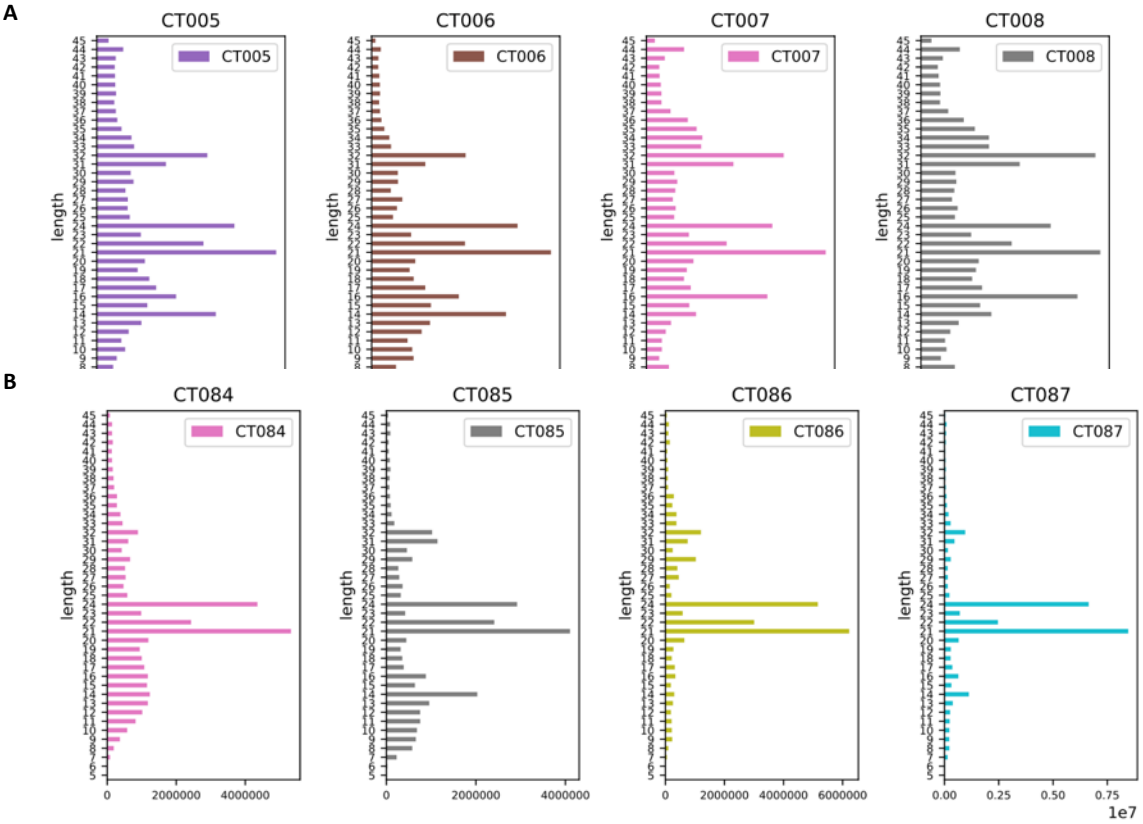


Figure 28.2: Example of sRNA read distribution profiles derived from RNA extractions performed before (A) and after (B) modifying the incorporation of the homogeinisation buffer.

Implementation of wet lab controls

Three wet lab controls were introduced to assess the quality of sample collection, cross-sample contamination and sensitivity of detection of the small RNA-seq HTS protocol.

Negative control

One Maxwell cartridge with just homegenisation and lysis buffer was processed along with cartridges with sample tissue lysates (**Figure 28.3**). Negative control shows no nucleic acids fragments (**Figure 28.4**).

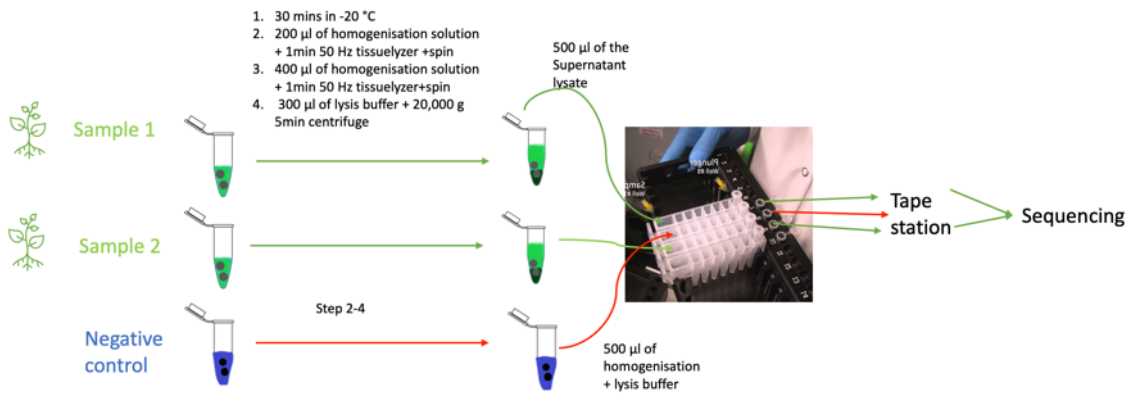


Figure 28.3 Processing of negative control along with plant sample lysates.

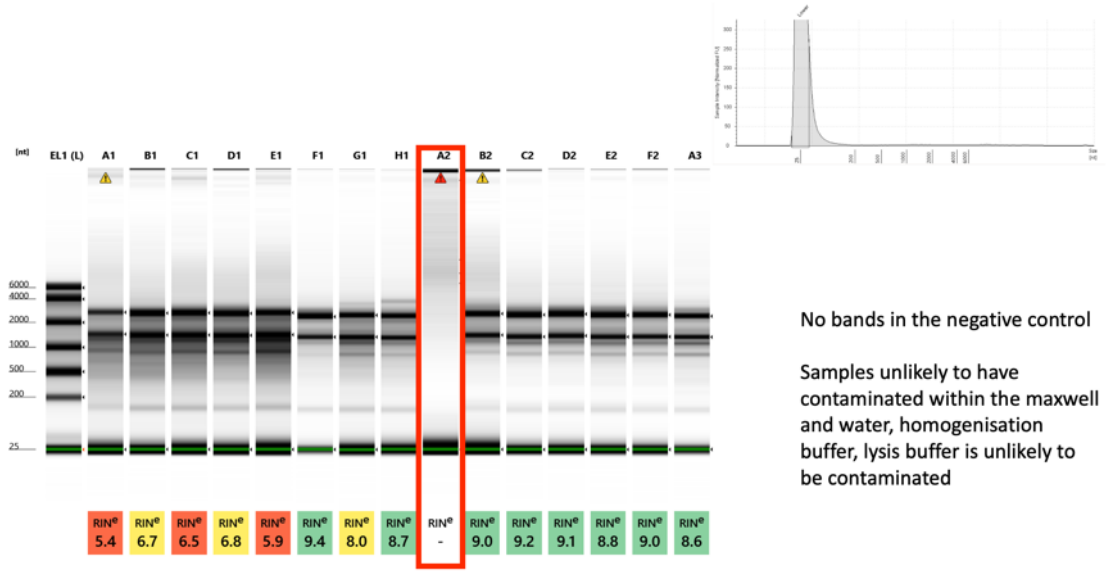


Figure 28.4 RNA profiles of plant lysates along with a negative control sample (red box). Source: SSG, DAFF team.

Alien control

A control plant infected with a known viral pathogen was selected as an alien control to monitor possible cross-sample contamination events in multiplex HTS experiments. We selected a *Mischantus sinensis* plant infected with a novel *Miscanthus sinensis* mosaic virus (MsiMV) that we reported in quarantined plant held at PEQ (LeBlanc et al., 2022). The infected plant shows a high MsiMV titre accounting for up to 50% of the total small RNAs sequenced, which is ideal to assess possible index hoping events in a range of plant commodities, except the *M. sinensis* ornamental grass.

Spike-in oligo control

A synthetic microRNA oligonucleotide control from the worm *Caenorhabditis elegans* (cel-miR-39) that has no sequence identity to any plant gene has been tested and spiked-in during the initial steps of the RNA extraction protocol to:

- account for sample heterogeneity (different tissue types, plant species and technical variability)
- evaluate detection limit for a given sample
- validate negatives (demonstrate that a virus-free sample is not due to RNA extraction issues)
- assess technology reproducibility/precision using replicates
- define the recovery thresholds of spike-in control for specific plant commodities

Figure 2.5 depicts the spiked-in of the synthetic miRNA during the sample homogenation step by adding 4 ml of 33 fmol/ml or 3.3 fmol/ ml solution into Fragaria and Mischantus samples. Real time PCR amplification of the spiked-in synthetic oligo detected ~1/100 and ~1/1000 dilution of the synthetic oligo in Mischantus and Fragaria, respectively in the final RNA extract sample (**Figure 2.6**).

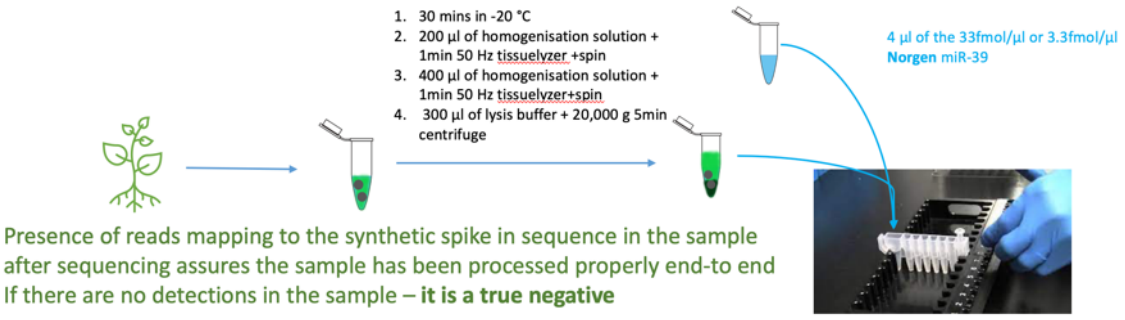
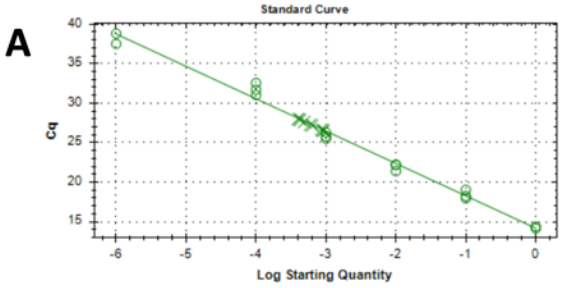
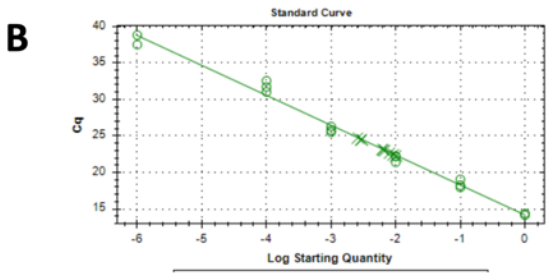


Figure 28.5 Spike-in of synthetic cel-miR39 oligo to plant lysates.



Norgen celmiR39 extraction efficiency in Fragaria – 0.000637



Norgen celmiR39 extraction efficiency in Miscanthus – 0.006199

Figure 28.6 Real time PCR of spiked in samples shows distinct recovery efficiencies for (**A**) Fragaria and (**B**) Mischantus samples showing Ct values between 26 to 28 and 22 to 25, respectively.

In agreement with real time PCR measurements, HTS sequencing of spiked samples showed a larger titre of cel-miR39 in Mischantus sinensis than in Fragaria yield in average 248K and 35K normalised reads per kilobase per million (RPKM) reads, respectively (**Tables 28.1 and 28.2**).

Sample ID	MT number	Species +Tissue	Type of spike + quantity (fmol)	Total RNA concentration (ng/µl)	RIN (Ramaciotti)	Deduplicated read count	Total quality filtered reads	RPKM	Cq	SQ fmol/µl)
Fn1	MT512	Fragaria +Petiole	Norgen+132	33.1	8.4	19,489	24,489,896	36,173	28.86	3.47E-04
Fn2	MT513	Fragaria +Petiole	Norgen+132	38.7	8.5	27,641	30,082,141	41,766	28.36	4.58E-04
Fn3	MT514	Fragaria +Petiole	Norgen+132	41	8.0	13,426	21,878,540	27,894	29.47	2.47E-04
							Avg	35,277	29	2.50E-04

Table 28.1: Detection of spike-in synthetic oligo in Fragaria samples.

Sample ID	MT number	Species	Type of spike + quantity (fmol)	Total RNA concentration (ng/µl)	RIN (PEQ+ Ramaciotti)	Deduplicated read count	Total quality filtered reads	RPKM	Cq	SQ fmol/µl)
Cn1	MT515	Miscanthus+ leaf/midrib	Norgen+132	45.2	6.4	366,271	77,069,672	216,021	24.86	3.35E-03
Cn2	MT516	Miscanthus+ leaf/midrib	Norgen+132	39.1	6.8	300,840	51,662,796	264,688	25.89	1.84E-03
Cn3	MT517	Miscanthus+ leaf/midrib	Norgen+132	61.3	7.3	197,341	33,872,306	264,819	24.52	3.98E-03
							Avg	248,510	25.09	3.06E-03

Table 28.2: Detection of spike-in synthetic oligo in Fragaria samples.

Subsequent processing of larger batch of Citrus, Fragaria and Stenotaphrum plants shows that ornamental grasses tend to show higher tither of synthetic spiked in oligos compared to other Dicotyledonous plants species (**Table 3**). These results show the importance to define recovery rates of synthetic oligos per commodity and tissue sample used to collect RNA.

Table 28.3: Normalised Fragment per kilobase per million (FPKM) counts of synthetic oligo cel-mir-39 (Norgen) across tested plant commodities.

Sample	Mean FPKM	Min FPKM	Max FPKM
Strawberry (n = 6)	64,915	34,740	92,257
Citrus woody indicator plant cambium (n = 15)	172,855	77,218	263,772
Citrus woody indicator plant leaf (n = 16)	214,035	89,085	373,092
Citrus STG leaf (n = 7)	210,503	103,542	258,701
Clonal grass Miscanthus (n =3)	195,736	152,173	227,921
Clonal grass Stenotaphrum (n=18)	516,051	243,417	1,034,058

For spike-in Fragaria samples, we also measure the level of detection of the spiked-in oligo by performing a real time PCR using an aliquot of the RNA extract sent to HTS data generation. Results showed Fragaria samples spiked with 4ul of 33 fmol/ml oligo recovered ~1/100 the added oligo concentration at the completion of the RNA extraction protocol, which is equivalent to ~13.2 amol/ml or to Ct20 as measured by real time PCR. Thus, *the sensitivity of detection of the small RNA-seq HTS protocol is comparable to that of current molecular assays used at PEQ.*

Appendix 29: Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass)

Archives of Virology (2022) 167:1701–1705
https://doi.org/10.1007/s00705-022-05445-3

ANNOTATED SEQUENCE RECORD

Complete genome sequence of a novel potyvirus infecting *Miscanthus sinensis* (silver grass)

Zacharie Leblanc[1] · Marie-Emilie Gauthier[1] · Ruvini Lelwala[1] · Candace Elliott[2] · Cassie McMaster[2] · Robin Eichner[2] · Kevin Davis[4] · Lia Liefting[5] · Jeremy Thompson[5] · Adrian Dinsdale[3] · Mark Whattam[3] · Julie Pattemore[2] · Roberto A. Barrero[1]

Received: 9 November 2021 / Accepted: 7 March 2022 / Published online: 17 May 2022
© The Author(s) 2022

Abstract
Here, we describe the full-length genome sequence of a novel potyvirus, tentatively named "Miscanthus sinensis mosaic virus" (MsiMV), isolated from *Miscanthus sinensis* (silver grass) held in a post-entry quarantine facility after being imported into Western Australia, Australia. The MsiMV genome is 9604 nucleotides (nt) in length, encoding a 3071-amino-acid (aa) polyprotein with conserved sequence motifs. The MsiMV genome is most closely related to that of sorghum mosaic virus (SrMV), with 74% nt and 78.5% aa sequence identity to the SrMV polyprotein region. Phylogenetic analysis based on the polyprotein grouped MsiMV with SrMV, sugarcane mosaic virus (SCMV), and maize dwarf mosaic virus (MDMV). This is the first report of a novel monopartite ssRNA virus in *Miscanthus sinensis* related to members of the genus *Potyvirus* in the family *Potyviridae*.

Miscanthus sinensis, commonly known as silver grass, is a member of the family Poaceae and is native to eastern Asia, where it is a keystone species in grasslands. In addition to ornamental applications, *M. sinensis* has been investigated for use as a biofuel crop due to its high biomass yield. Miscanthus has been found to be an alternate host for viruses infecting plants of the family Poaceae, including switchgrass mosaic virus and barley yellow dwarf virus [1].

Potyviruses represent the largest and most economically damaging genus of known plant RNA viruses [2]. Potyviruses are aphid-borne, infecting both monocotyledonous and eudicotyledonous angiosperms [3]. In response to viral infection, plants can disrupt viral transcripts via RNA interference using Dicer-like endonucleases [4]. Infections by potyviruses typically result in silencing responses, predominantly using 21- and 22-nt small RNAs [5].

All potyviruses have a monopartite genome containing a long open reading frame (ORF) [6], which encodes a polyprotein that is proteolytically processed by three viral proteases into 10 functional proteins: P1, HC-Pro, P3, 6K1, cylindrical inclusion (CI) protein, 6K2, viral genome-linked protein (VPg), NIa-Pro, nuclear inclusion b (NIb; also known as RNA-dependent RNA polymerase), and capsid protein (CP) [7, 8].

Handling Editor: Massimo Turina.

✉ Roberto A. Barrero
roberto.barrero@qut.edu.au

[1] eResearch Office, Queensland University of Technology, Brisbane, QLD 4000, Australia

[2] Science and Surveillance Group, Department of Agriculture, Water and the Environment, Post Entry Quarantine, Mickleham, VIC 3064, Australia

[3] Department of Agriculture, Water and the Environment, Plant Innovation Centre @ Post Entry Quarantine, Mickleham, VIC 3064, Australia

[4] Department of Agriculture, Water and the Environment, Australian Government, Canberra City, ACT 2601, Australia

[5] Plant Health and Environment Laboratory, Ministry for Primary Industries, P.O. Box 2095, Auckland 1140, New Zealand

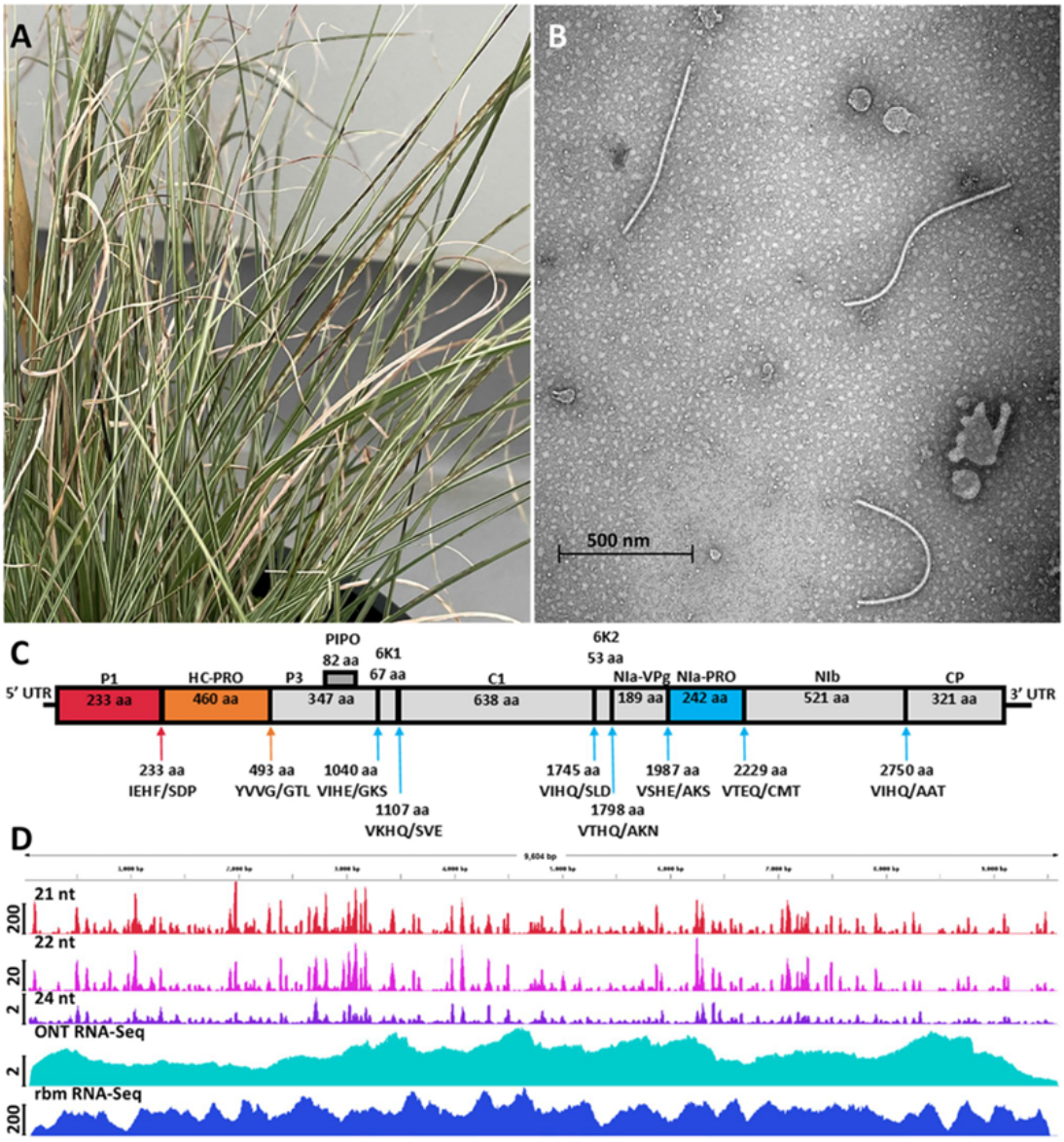


Fig. 1 (A) A *Miscanthus* plant infected with MsiMV, showing leaf mottling symptoms. (B) Morphology of MsiMV virions viewed by transmission electron microscopy. Samples were prepared using a modified 'quick dip' method [14]. Freshly cut leaf pieces (2 × 2 mm) were placed into 2% (w/v) phosphotungstic acid stain (pH 7.07) and adsorbed onto a carbon-coated mesh grid (formvar carbon film 400 mesh, copper). Grids were dried at 37 °C for at least 30 min then examined at 26,000x magnification using a Tecnai G2 Spirit transmission electron microscope. (C) Schematic representation of the genome organization of MsiMV. The 5′ (140 nt) and 3′ (251 nt) untranslated regions are represented by solid horizontal bars, while ORFs are depicted as open boxes. Putative sites of polyprotein cleavage mediated by P1, HC-Pro, and NIa-Pro are indicated by red, orange, and blue arrows, respectively. Protein domain sizes in aa are indicated below the protein name. (D) Alignment of small RNAs, ONT, and Illumina ribominus RNA-seq sequences with the MsiMV genome sequence. Scale bars on the *y*-axis show the total number of mapped reads (x1000).

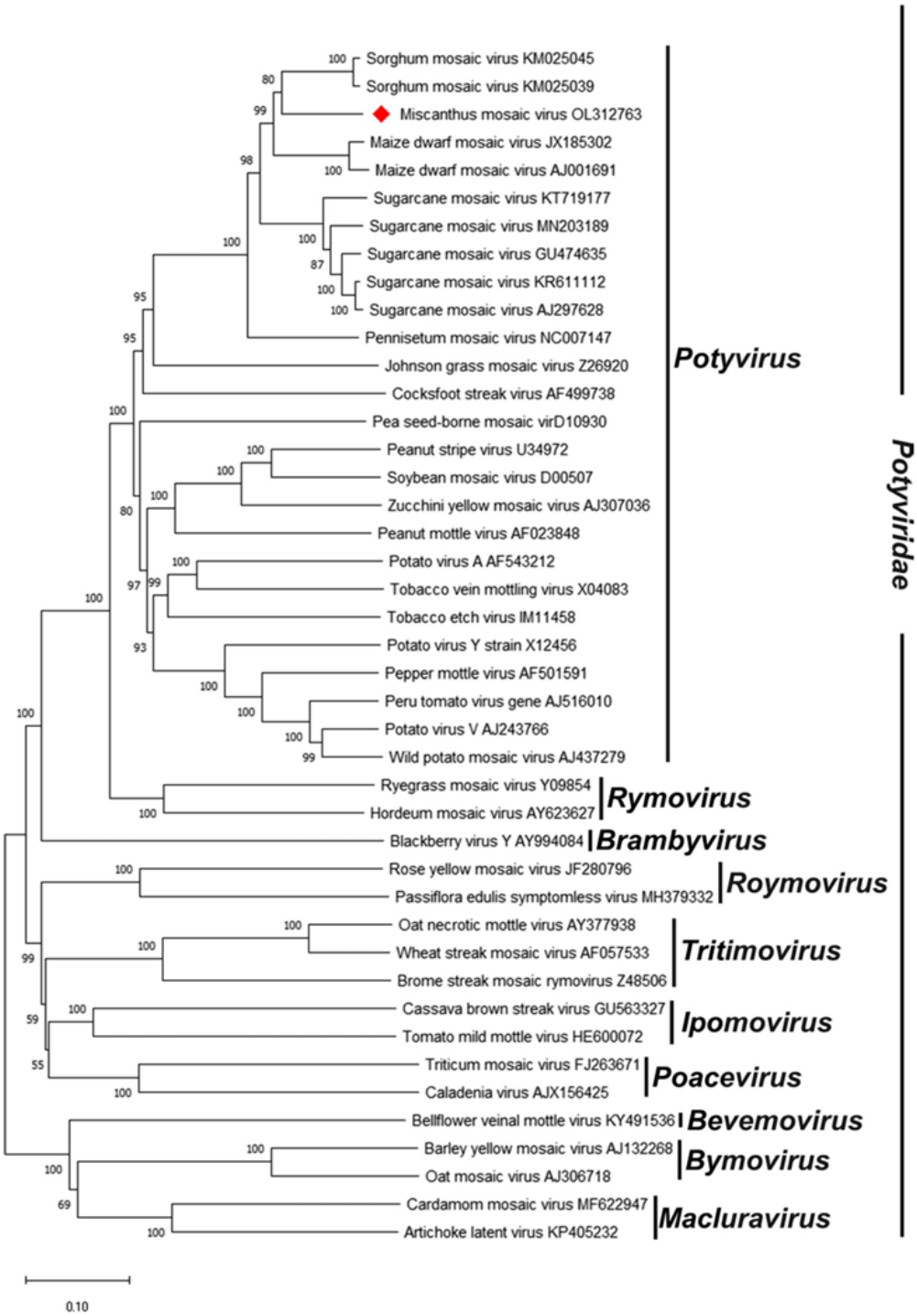
A *Miscanthus sinensis* cultivar "Morning Light" plant imported into Western Australia from the USA in 1985 was transferred to the Plant Quarantine Station, Rydalmere, New South Wales, for virus screening, as per new import conditions for clonal grasses at the time [9]. Double-stranded RNA extracted from 38 g of mature leaf tissue electrophoresed in a 0.75% agarose gel revealed bands suspected to be of viral RNA, and subsequent mechanical inoculation of ground grass tissue onto plants of *Zea mays* cv. Supagold resulted in mosaic symptoms [10]. The infected Morning Light cultivar was never released from post-entry quarantine and was maintained as a positive control for biological indexing (Fig. 1A). Sap from the infected plant contained filamentous particles approximately 600 nm in length (Fig. 1B), further confirming the presence of a potyviral pathogen.

Collection of RNA samples and *in silico* assembly of 21- to 22-nt small RNA sequences were performed as described previously [11], yielding a single 9604-nt sequence containing a single long ORF from nt 141 to 9353 and flanking noncoding sequences, including a polyA tail at the 3' end, which is typical of potyviruses. Mapping of small RNA reads onto the MsiMV genome sequence, allowing no mismatches, resulted in 11,370,140 aligned reads (~25Kx coverage) of which 89.2%, 7.6%, and 0.3% were 21 nt, 22 nt, and 24 nt long, respectively. The high proportion of mapped 21-nt reads is consistent with them being derived from the antiviral silencing response [5].

The completeness of the small-RNA-based genome assembly was confirmed by mapping of both Illumina and long-read Oxford Nanopore Technology (ONT reads > 500 bp) ribosomal-depleted RNA-seq data (Fig.1D). Overall, 7,154,030 Illumina (~74.5Kx coverage) and 21,495 ONT (2140x coverage) reads were mapped onto the MsiMV genome using bowtie [12] and minimap2 [13], respectively. Identical assembled MsiMV genome sequences were obtained using all three technologies. Raw data are available under BioProject PRJNA752836.

The qualifiers accepted by the International Committee on Taxonomy of Viruses for a new species of potyvirus are < 76% nt and < 82% aa ORF sequence identity [8]. Comparison of the MsiMV polyprotein and coat protein sequences against those of SrMV (KM025045) showed that the new virus met these criteria. Additionally, the polyprotein is predicted to be cleaved into 10 subdomains corresponding to those of other potyviruses (Fig. 1C).

Phylogenetic analysis based on polyprotein sequences placed MsiMV within a clade with SrMV and MDMV, with SCMV as a sister clade (Fig. 2). The clade formation of these viruses may be reflective of the similarity of their plant hosts, which are all in the subfamily Panicoideae of the family Poaceae. The sequence similarity of these viruses could also explain overlaps in host susceptibility seen with SrMV, which can infect *Miscanthus,* corn, and sugarcane. The phylogenetic position of MsiMV, separate from SrMV, SCMV, and MDMV, further supports the classification of MsiMV as a member of a novel species in the genus *Potyvirus*.


Sorghum mosaic virus KM025045
Sorghum mosaic virus KM025039
Miscanthus mosaic virus OL312763
Maize dwarf mosaic virus JX185302
Maize dwarf mosaic virus AJ001691
Sugarcane mosaic virus KT719177
Sugarcane mosaic virus MN203189
Sugarcane mosaic virus GU474635
Sugarcane mosaic virus KR611112
Sugarcane mosaic virus AJ297628
Pennisetum mosaic virus NC007147
Johnson grass mosaic virus Z26920
Cocksfoot streak virus AF499738
Pea seed-borne mosaic virD10930
Peanut stripe virus U34972
Soybean mosaic virus D00507
Zucchini yellow mosaic virus AJ307036
Peanut mottle virus AF023848
Potato virus A AF543212
Tobacco vein mottling virus X04083
Tobacco etch virus IM11458
Potato virus Y strain X12456
Pepper mottle virus AF501591
Peru tomato virus gene AJ516010
Potato virus V AJ243766
Wild potato mosaic virus AJ437279
Ryegrass mosaic virus Y09854
Hordeum mosaic virus AY623627
Blackberry virus Y AY994084
Rose yellow mosaic virus JF280796
Passiflora edulis symptomless virus MH379332
Oat necrotic mottle virus AY377938
Wheat streak mosaic virus AF057533
Brome streak mosaic rymovirus Z48506
Cassava brown streak virus GU563327
Tomato mild mottle virus HE600072
Triticum mosaic virus FJ263671
Caladenia virus AJX156425
Bellflower veinal mottle virus KY491536
Barley yellow mosaic virus AJ132268
Oat mosaic virus AJ306718
Cardamom mosaic virus MF622947
Artichoke latent virus KP405232
Potyvirus
Rymovirus
Brambyvirus
Roymovirus
Tritimovirus
Ipomovirus
Poacevirus
Bevemovirus
Bymovirus
Macluravirus
Potyviridae
0.10

◂**Fig. 2** Neighbor-joining tree of polyprotein aa sequences of MsiMV and selected members of the family *Potyviridae*. Bootstrap analysis was applied, using 1000 bootstrap replicates in MEGA X [15]. Percent bootstrap values are shown at each node.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Project Improving plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing. MT18005 is funded by Hort Innovation, using the Hort Innovation Citrus, Grape Tables, Rubus, Potato and Nursery research and development levy, co-investment from Queensland University of Technology and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bolus S, Malapi-Wight M, Grinstead SC, Fuentes-Bueno I, Hendrickson L, Hammond RW, Mollov D (2020) Identification and characterization of Miscanthus yellow fleck virus, a new polerovirus infecting Miscanthus sinensis. PLoS ONE 15:e0239199
2. Braidwood L, Muller SY, Baulcombe D (2019) Extensive recombination challenges the utility of Sugarcane mosaic virus phylogeny and strain typing. Sci Rep 9:20067
3. Gibbs AJ, Hajizadeh M, Ohshima K, Jones RA (2020) The potyviruses: an evolutionary synthesis is emerging. Viruses 12:132
4. Blevins T, Rajeswaran R, Shivaprasad PV, Beknazariants D, Si-Ammour A, Park H-S, Vazquez F, Robertson D, Meins F Jr, Hohn T (2006) Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res 34:6233–6246
5. Ho T, Wang H, Pallett D, Dalmay T (2007) Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins. FEBS Lett 581:3267–3272
6. Cui H, Wang A (2019) The biological impact of the hypervariable N-terminal region of potyviral genomes. Annu Rev Virol 6:255–274
7. Shen W, Shi Y, Dai Z, Wang A (2020) The RNA-dependent RNA polymerase NIb of potyviruses plays multifunctional, contrasting roles during viral infection. Viruses 12
8. Wylie SJ, Adams M, Chalam C, Kreuze J, Lopez-Moya JJ, Ohshima K, Praveen S, Rabenstein F, Stenger D, Wang A, Zerbini FM, Ictv Report C (2017) ICTV virus taxonomy profile: Potyviridae. J Gen Virol 98:352–354
9. Service AQaI (1986) Quarantine Circular Memorandum GM:PG L85/1386. Canberra, Australian Capital Territory.
10. Davis K, Gillings M (1996) Detection of latent viruses and double-stranded RNAs in clonal grasses during post-entry quarantine. In: ACIAR PROCEEDINGS. Australian Centre for International Agricultural Research, pp 95–99
11. Barrero RA, Napier KR, Cunnington J, Liefting L, Keenan S, Frampton RA, Szabo T, Bulman S, Hunter A, Ward L (2017) An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. BMC Bioinformatics 18:1–12
12. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
13. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100
14. Dijkstra J, de Jager CP (1998) Practical plant virology: protocols and exercises. Springer-Verlag, Berlin Heidelberg New York, pp 291–292
15. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 30: Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants

viruses

MDPI

Article

Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants

Ruvini V. Lelwala [1,2], Zacharie LeBlanc [1], Marie-Emilie A. Gauthier [1], Candace E. Elliott [2], Fiona E. Constable [3], Greg Murphy [4], Callum Tyle [4], Adrian Dinsdale [5], Mark Whattam [5], Julie Pattemore [2] and Roberto A. Barrero [1,*]

1 eResearch, Research Infrastructure, Academic Division, Queensland University of Technology, Brisbane, QLD 4001, Australia; ruvini.lelwala@aff.gov.au (R.V.L.); z.leblanc@qut.edu.au (Z.L.); marieemilie.gauthier@qut.edu.au (M.-E.A.G.)
2 Science and Surveillance Group, Post Entry Quarantine, Department of Agriculture, Fisheries and Forestry, Mickleham, VIC 3064, Australia; candace.elliott@aff.gov.au (C.E.E.); julie.pattemore@aff.gov.au (J.P.)
3 Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia; fiona.constable@agriculture.vic.gov.au
4 Technology Infrastructure Branch, Information Services Division, Department of Agriculture, Fisheries and Forestry, Canberra, ACT 2601, Australia; greg.murphy@aff.gov.au (G.M.); callum.tyle@aff.gov.au (C.T.)
5 Plant Innovation Centre, Post Entry Quarantine, Department of Agriculture, Fisheries and Forestry, Mickleham, VIC 3064, Australia; adrian.dinsdale@aff.gov.au (A.D.); mark.whattam@aff.gov.au (M.W.)
* Correspondence: roberto.barrero@qut.edu.au

Abstract: High-throughput sequencing (HTS) of host plant small RNA (sRNA) is a popular approach for plant virus and viroid detection. The major bottlenecks for implementing this approach in routine virus screening of plants in quarantine include lack of computational resources and/or expertise in command-line environments and limited availability of curated plant virus and viroid databases. We developed: (1) virus and viroid report web-based bioinformatics workflows on Galaxy Australia called GA-VirReport and GA-VirReport-Stats for detecting viruses and viroids from host plant sRNA extracts and (2) a curated higher plant virus and viroid database (PVirDB). We implemented sRNA sequencing with unique dual indexing on a set of plants with known viruses. Sequencing data were analyzed using GA-VirReport and PVirDB to validate these resources. We detected all known viruses in this pilot study with no cross-sample contamination. We then conducted a large-scale diagnosis of 105 imported plants processed at the post-entry quarantine facility (PEQ), Australia. We detected various pathogens in 14 imported plants and discovered that de novo assembly using 21–22 nt sRNA fraction and the megablast algorithm yielded better sensitivity and specificity. This study reports the successful, large-scale implementation of HTS and a user-friendly bioinformatics workflow for virus and viroid screening of imported plants at the PEQ.

Keywords: high-throughput sequencing; small RNA; plant virus and viroid detection; post-entry quarantine; VirReport

check for updates

Citation: Lelwala, R.V.; LeBlanc, Z.; Gauthier, M.-E.A.; Elliott, C.E.; Constable, F.E.; Murphy, G.; Tyle, C.; Dinsdale, A.; Whattam, M.; Pattemore, J.; et al. Implementation of GA-VirReport, a Web-Based Bioinformatics Toolkit for Post-Entry Quarantine Screening of Virus and Viroids in Plants. *Viruses* **2022**, *14*, 1480. https://doi.org/10.3390/v14071480

Academic Editor: Jozef Nissimov

Received: 8 June 2022
Accepted: 29 June 2022
Published: 5 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Plant pathogenic viruses and viroids are an increasing threat to global food security [1]. Apart from the economic impact on primary production by reducing yield and quality, plant viruses and viroids can also impair market access and cause severe ecological consequences to native biodiversity. With the ever-growing global movement of plant material and the discovery of novel viruses, improvements in the detection and surveillance of viruses and viroids have become critical to biosecurity and border protection authorities [2–4]. Plant quarantine is a biosecurity measure implemented to prevent the entry of economically important plant pests to an area free of those pests or where such pests are present at

manageable levels and are officially controlled [5]. Regardless of the point of quarantine enforcement, nationally or regionally, at pre-entry or post-entry, using reliable diagnostic solutions is paramount for making accurate regulatory decisions.

A variety of methods are available to diagnose viruses and viroids in plants. Among these, more conventional methods such as visual inspection and biological assays can result in false positives/negatives due to the absence of disease symptoms, temporal manifestation of symptoms, similarity to abiotic stress phenotypes and symptoms caused by a co-infection of multiple viruses [6]. Moreover, biological assays require a specialist skillset that nowadays is in decline and are resource-intensive, requiring special glasshouse facilities to maintain quarantined plants over long periods (up to 2 years). These challenges add to the cost of biosecurity clearance and delay the release of new genetic material, impacting the ability of Australia's agricultural sector to maintain its competitiveness [3]. While electron microscopy detection of viruses using particle morphology enables the discovery of infecting viruses, narrowing it down to the species taxonomic level and its low intrinsic sensitivity remain a challenge [7]. Serological and molecular methods are sensitive and specific, allowing pathogen identification up to genus and species levels, but their application focuses on known and well-characterized pests [8]. Furthermore, both serological and molecular assays usually target a single or few closely related pathogens, and they may not readily detect genetically divergent viruses, leading to false-negative detections. Additionally, the increase in the number of viruses of biosecurity concern over time requires implementing a large set of molecular assays, which is increasingly becoming an unsustainable approach to large-scale quarantine testing [3,7].

The advent of high-throughput sequencing (HTS) has revolutionized plant virus diagnostics by facilitating the detection and identification of pathogens without a priori knowledge of infectious status [8]. The large number of data generated through HTS enables detection of all viral agents present, and classification into the lowest taxonomy level is feasible [4,9]. Therefore, HTS can reduce the time and resources required for quarantine screening by eliminating the need to test for a large panel of individual pathogens over a lengthy quarantine period [3,10]. HTS can assist in optimizing the allocation of diagnostic resources during PEQ testing, effectively informing what molecular assays to run to verify HTS-detected pests. Moreover, HTS sequencing can lead to the discovery of new viruses or divergent strains that can be emerging pests [11,12]. With rapid advancements in sequencing technologies and the plummeting cost per base, HTS is becoming an increasingly popular and economical tool for high-throughput virus and viroid diagnostics [9].

Analysis of RNA sequences from host plants is an effective virus and viroid detection method considering most plant viruses have RNA as their genetic material and DNA viruses produce RNA transcripts [12]. Total RNA (commonly ribosomal-depleted), double-stranded RNA and small RNA (sRNA) have been explored as the targeted nucleic acids to be sequenced [13]. The sRNA method explicitly targets the products of the host antiviral response mediated by plant Dicer-like enzymes (DCLs) that cleave viral RNA and replication intermediates into 21–24 nt-long fragments (reviewed in [14]). Moreover, the reduced costs of sRNA-seq and its ability to flag actively replicating viral pathogens compared to RNA-seq have made it a widely adopted approach for viral pathogen diagnostics [10,15–17].

Bioinformatics analysis for virus and viroid diagnostics using sRNA-seq data include: (1) quality processing of reads, (2) identifying and removing host reads, (3) de novo and/or reference-guided genome assembly, (4) sequence similarity searches of the contigs against public or custom viral databases and (5) the summarizing and curation of the results. Although theoretically simple, these analyses are often a complex task requiring specialist knowledge in bioinformatics [17]. Several bioinformatics toolkits were recently developed for streamlining the detection of viruses and viroids from host plant sRNA-seq data. Some of these tools include the viral surveillance and diagnosis (VSD) toolkit [10], VirusDetect [18], VirFind [19] and VirReport [20]. In plant quarantine testing settings, the bottleneck for using some of the existing automated toolkits is the requirement for either local installation or knowledge of the Linux operating system and IT security barriers relating to the

potentially sensitive nature of the data being generated. Although the tools VirFind [19] and VirusDetect [18] provide public web servers, the individual steps remain hidden from the users. Commercial software such as CLC Genomics Workbench [21] or Geneious Prime [22] enable the building of all-in-one customized workflows with visualization tools. However, these tools incur ongoing licensing fees and require appropriate computational resources at the user's end to analyze large datasets. As Jones et al. [12] emphasized, the availability of a simple web-based analytical environment is of better use to the research community and diagnosticians because the users can modify processing parameters.

Another challenge associated with the bioinformatics analysis of sequencing data for diagnostic purposes, in general, is the dependency on sequence databases for taxonomic assignment of detections [12,17]. Although public databases such as the GenBank database maintained at National Center for Biotechnology Information (NCBI) are the most up to date, the enormous nature of these databases limits their portability across different computing resources. Moreover, sequencing data submitted by users to these public databases may not be accurate, and annotations provided in the GenBank record, such as the taxonomy assignment, which is crucial for accurate diagnosis, may be inaccurate and/or out of date [23]. Additionally, the descriptors of the sequences in the public databases are not harmonized and lack taxonomic information, posing an additional challenge to validating sequence-homology-based pathogen detections.

Here, we introduce a web-based counterpart of VirReport [20] that is executable on the Galaxy Australia (GA) platform [23], which we called GA-VirReport, and a harmonized plant virus and viroid database (PVirDB). A secondary GA-VirReport-Stats pipeline generates confidence statistics for detections and includes reference-guided consensus sequence generation. This work describes the validation and subsequent implementation of the above resources in screening for viruses and viroids in plants quarantined at the PEQ.

2. Materials and Methods

2.1. Plant Material

2.1.1. Control Plants

A subset of eight plants from the Gauthier et al. [20] study were selected for comparative pilot implementation and testing of the GA-VirReport workflow. These plants belonging to *Citrus* ($n = 3$), *Prunus* ($n = 1$), *Rubus* ($n = 1$), *Miscanthus* ($n = 1$), *Iris* ($n = 1$) and *Ipomoea* ($n = 1$) genera were maintained in the PEQ glasshouses, Mickleham, Australia, and were re-sampled in March 2021 for the current study.

2.1.2. Imported Plants

A total of 105 imported plants held in quarantine at PEQ were selected for a large-scale study using HTS and GA-VirReport for quarantine testing (Supplementary Table S7). These plants were from a range of plant genera, namely *Solanum* ($n = 9$), *Fragaria* ($n = 14$), *Vitis* ($n = 4$), *Malus* ($n = 3$), *Pyrus* ($n = 6$) and *Prunus* ($n = 69$). All plants were sampled between September and November 2021 according to [20].

2.2. Nucleic Acid Extraction

For all except *Rubus* and *Fragaria*, 50 mg of the leaf/midrib was subsampled for nucleic acid extraction. For *Rubus* and *Fragaria*, 50 mg of the petioles were used for nucleic acid isolation as per PEQ sampling procedures. Total RNA extractions were performed as previously described [20]. The quality and concentration of total RNA were measured using a NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer (ThermoFisher Scientific, Waltham, MA, USA) following manufacturer's protocols.

2.3. Library Construction and Sequencing

Aliquots of 14–15 µL of the total RNA extract were shipped on dry ice to the Ramaciotti Centre for Genomics, University of New South Wales, Australia. The concentration and the purity of the samples were checked using the Epoch Microplate Spectrophotometer

(Agilent, Santa Clara, CA, USA), and the RNA integrity was evaluated using the 4200 TapeStation System (Agilent) using the RNA Screen Tape device (Agilent). A total of 100 ng of the quality-checked RNA was used in the library construction using the QIAseq miRNA Library Kit (Qiagen, Hilden, Germany). QIAseq miRNA UDI 96 index plate with unique dual-index adapters was used in the library preparation according to the manufacturer's protocol. With each library prep batch, Qiagen XpressRef Universal Total RNA (Qiagen) and nuclease-free water were used as the positive and negative controls, respectively. Invitrogen™ Quant-iT™ PicoGreen™ dsDNA Assay Kit (ThermoFisher Scientific) and the LabChip DNA High Sensitivity Reagent kit (PerkinElmer, Waltham, MA, USA) were used to perform quality checks on individual sequencing libraries. Libraries were then normalized, equimolarly pooled and cleaned up with Qiagen QMN beads (Qiagen). Quality check on the cleaned pooled libraries was performed using the Qubit dsDNA HS assay kit on a Qubit Fluorometer (ThermoFisher Scientific) and a TapeStation System (Agilent) High Sensitivity D1000 ScreenTape device (Agilent). Sequencing was performed on the Illumina NovaSeq 6000 instrument with SP 100c flowcell for the control samples and the S1 100c or S2 100c flowcell for the imported samples, generating 75 bp single-end reads with unique molecular identifiers (UMIs). PhiX control (Illumina, San Diego, CA, USA) was used as a control for sequencing runs.

2.4. Bioinformatics Analyses

The sequencing provider performed demultiplexing of the raw reads using the bcl2fastq software [24]. All raw FASTQ files from the control dataset were deposited in the Short Read Archive (SRA) database under the BioProject PRJNA752836. We deposited the GA-VirReport workflow for multiple input files (https://doi.org/10.5281/zenodo.6387492) and GA-VirReport-Stats workflow (https://doi.org/10.5281/zenodo.6387504) (Figure 1) at Zenodo. We used these workflows to process raw sequencing data and custom bash scripts for the downstream bioinformatic processing. GA tools used in individual tasks are detailed below.

2.4.1. Preparation of Input Files for the GA-VirReport Workflow

The upload of raw files into the GA platform [23] was carried out via the 'Upload Data' tool in the GA toolshed [25] or remotely via an API gateway (unpublish). For each sample, the demultiplexed fastq files were uploaded into individual histories in GA, and a 'Dataset list' was created listing all the sequencing files associated with a sample. The multiFASTA file for the PVirDB database (see below) was also uploaded using the 'Upload Data' tool. A dataset named 'non-informative database' consisting of rRNA, mitochondria and chloroplast sequences derived from plant species held at PEQ was also uploaded to the same history and made publicly available at Zenodo's data repository (https://doi.org/10.5281/zenodo.6387507). The GA-VirReport workflow for multiple input files was selected from the 'Workflows' tab in the GA interface and linked to the respective raw fastq inputs before running the workflow for each sample.

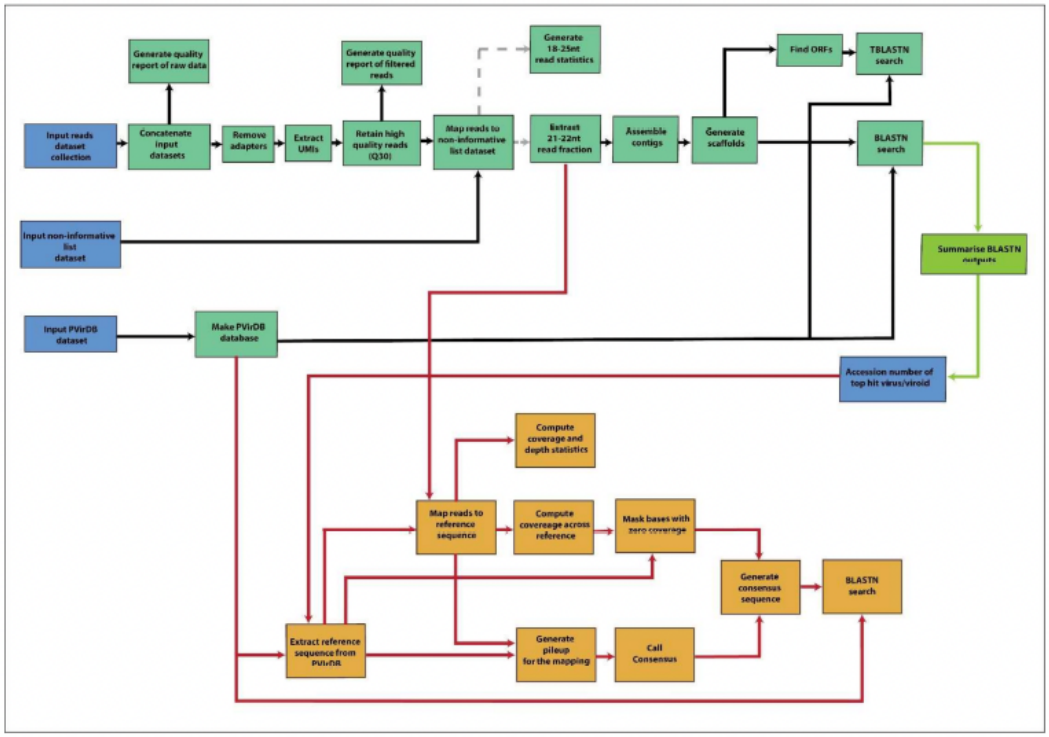


Figure 1. GA-VirReport and GA-VirReport-Stats workflows. GA-VirReport workflow is shown in green, and the GA-VirReport-Stats workflow is shown in orange.

2.4.2. GA-VirReport Workflow

Multiple sequencing read files in the dataset list corresponding to the same sample were merged using the 'Concatenate multiple datasets tail-to-head (Galaxy Version 0.2)' tool. Sequencing adapters were then trimmed from reads using the 'Cutadapt (Remove adapter sequences from FASTQ/FASTA (Galaxy Version 3.5+galaxy1))' tool, which uses cutadapt (Version 3.5) [26]. The adapters specified were the Illumina TruSeq dual-indexing adapters 5′-AGATCGGAAGAGCACACGTCTGAACTCCAGTCA-3′ as the 3′ adapter and '3′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-5′ as the 5′ adapter. Adapter filtered reads were filtered for UMIs using the 'UMI-tools extract (Extract UMI from fastq files (Galaxy Version 1.1.2+galaxy2))' tool, which utilizes UMI-tools (Version 1.1.2) [27]. UMI-extracted reads were then subjected to quality filtering with a Q score threshold of 30 and an unknown base limit of zero using 'fastp (fast all-in-one pre-processing for FASTQ files (Galaxy Version 0.20.1+galaxy0))' tool with fastp (Version 0.20.1) [28]. Fastp and 'FastQC (Read Quality reports (Galaxy Version 0.72+galaxy1))' tool with FastQC (Version 0.11.8) [29] were used to generate quality reports. The quality-filtered reads were filtered for plant RNA by mapping trimmed reads onto non-informative host plant sequences using the 'Map with Bowtie for Illumina (Galaxy Version 1.2.0)' tool with bowtie (Version 1.2.0) [30]. Filtered reads were subjected to size selection to select the 21–22 nt read fraction using the 'fastp (fast all-in-one pre-processing for FASTQ files (Galaxy Version 0.20.1+galaxy0))' tool. For the control dataset, the 24 nt read fraction was also derived, and the downstream workflow was separately conducted on this read fraction. De novo genome assembly was performed with the 'velveth (Prepare a dataset for the Velvet velvetg Assembler (Galaxy Version 1.2.10.3))' tool with a kmer length of 15 and the 'velvetg (Velvet sequence assembler

for very short reads (Galaxy Version 1.2.10.2))' tool, which uses velvet (Version 1.2.10) [31]. The 'cap3 (Sequence Assembly tool (Galaxy Version 2.0.0))' with cap3 (Version 10.2011) [32] was used for scaffolding. The open reading frames were predicted using the 'getorf (Finds and extracts open reading frames (ORFs) (Galaxy Version 5.0.0.1))' from the EMBOSS suite (Version 5.0.0) [33]. The custom plant virus and viroid database PVirDB was generated within GA using the multiFASTA file as an input for the 'NCBI BLAST+ makeblastdb (Make BLAST database (Galaxy Version 2.10.1+galaxy0))' tool with the BLAST (Version 2.10.1) [34]. The scaffolds from cap3 were used as queries for all-vs-all sequence similarity search against the PVirDB database using 'NCBI BLAST+ blastn (Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy0))' tool using BLAST (Version 2.10.1) [34]. The predicted ORFs were used as queries against an all-vs-all sequence similarity search of the translated PVirDB using the 'NCBI BLAST+ tblastn (Search translated nucleotide database with protein query sequence(s) (Galaxy Version 2.10.1+galaxy0))' with BLAST (Version 2.10.1) [34]. The scaffolds from the cap3 and ORFs from getorf in FASTA format, and the sequence similarity blastn and tblastn outputs were downloaded for further processing. Summarizing and curation of the blastn outputs were performed using BlastTools java script [35] to obtain a summary output with statistics of top hits for each sample.

2.4.3. GA-VirReport-Stats Workflow

All positive detections were subjected to a secondary complementary workflow, the GA-VirReport-Stats. The 21–22 nt read fraction and the PVirDB database were selected as input files. The accession number of the top hit for an individual detection, which was sourced from the blastn summary output, was also entered manually before executing the workflow. The following section describes the tools used in the GA-VirReport-Stats workflow.

Reference sequences were extracted from the PVirDB database with the 'NCBI BLAST+ blastdbcmd entry(s) Extract sequence(s) from BLAST database (Galaxy Version 2.10.1+galaxy0)' tool, which uses BLAST (Version 2.10.1) [34]. The 21–22 nt reads were mapped onto this reference sequence with Map with Bowtie for Illumina (Galaxy Version 1.2.0), which uses bowtie (Version 1.2.0) [30]. 'SAM-to-BAM convert SAM to BAM (Galaxy Version 2.1.1)' tool, which applies samtools (Version 1.9) [36], was used to convert the obtained the sequence alignment map to BAM format. Coverage and depth statistics were generated for the mapped data with the 'QualiMap BamQC (Galaxy Version 2.2.2d+galaxy3)' tool that uses Qualimap (Version 2.2.2d) [37]. The 'bcftools mpileup Generate VCF or BCF containing genotype likelihoods for one or multiple alignment (BAM or CRAM) files (Galaxy Version 1.10)' tool was used to generate VCF/BCF files, and variant calling was performed with the 'bcftools call SNP/indel variant calling from VCF/BCF (Galaxy Version 1.10)', which runs bcftools (Version 1.10) [36]. The coverage of the reference sequence/genome was derived with the 'bedtools Genome Coverage compute the coverage over an entire genome (Galaxy Version 2.30.0)' [38], which uses BEDTools (Version 2.30.0) [39]. Regions of the reference sequence with no coverage were extracted with the 'Text reformatting with awk (Galaxy Version 1.1.2)' tool [38] that uses gawk (Version 4.2.0) and masked with Ns using 'bedtools MaskFastaBed use intervals to mask sequences from a FASTA file (Galaxy Version 2.30.0)' tool, which uses BEDTools (Version 2.30.0) [39]. A reference-guided consensus sequence was derived from the VCF variants with 'bcftools consensus Create consensus sequence by applying VCF variants to a reference FASTA file (Galaxy Version 1.10+galaxy1)', which uses bcftools (Version 1.10). A final sequence similarity search of the consensus sequence against the PVirDB was performed using 'NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy0)' tool as described above [34]. The reads per kilobase of transcript per million mapped reads (RPKM) in Equation (1) was computed as follows for each virus and viroid in a sample:

$$\text{RPKM} = \frac{\text{Number of reads mapped to the reference} \times 10^9}{\text{Total number of quality filtered reads} \times \text{reference length in bp}} \quad (1)$$

2.4.4. Curated Plant Virus–Viroid Database: PVirDB

Figure 2 depicts the workflow used to generate PVirDB. All available nucleotide sequences (n = 17,153) of regulated viruses and viroids at PEQ and endemic viruses of interest to Australian plant industries were retrieved in GenBank format using an Entrez text search on the NCBI nucleotide database [40]. The obtained list was checked manually, and any non-specific sequences removed. From here on, the above datasets will be referenced as 'Regulated' and 'Endemic', respectively, and 'PEQ dataset' collectively. To prevent off-target detections, we also included in the custom database other reference viral sequences derived for higher plant species and abundant host plant sequences (i.e., chloroplast genomes). A nucleotide accession list of 120,945 viruses from higher plants, Embryophyta (taxid 3193), was obtained from the NCBI Virus [41] (October 2021), and their nucleotide sequences were downloaded in GenBank format using NCBI Batch Entrez [42]. From here on the above dataset will be addressed as 'Higher plant dataset'. Additionally, complete chloroplast genome sequences from the same genera of host plants held at PEQ were retrieved from NCBI nucleotide database [40] in GenBank format and will be referred to as 'Chloroplast dataset' from here on. The GenBank-formatted sequences were converted into FASTA format with annotated taxonomic information using a custom script by Gutiérrez et al. [43]. In the regulated dataset, the unknown base percentages of each sequence were determined using the Seqtk version 1.3 [44], and sequences with >5% N content were removed along with host plant sequences and sequences >20 Kb, except for complete chloroplast genomes that ranged from 100 to 200 kbp. To avoid ambiguous detection and reporting of regulated and endemic pests, we clustered highly similar sequences (≥99.0%) using CD-HIT version 4.8.1 [45] to select a single representative sequence per cluster. Similarly, we clustered the higher plant dataset and retained a single representative sequence for each cluster. Next, duplicated accessions in PEQ and higher plant datasets were removed from the higher plant dataset. Furthermore, highly similar sequences in higher plant dataset to the PEQ dataset (identity > 97% and coverage > 70%) were also removed. The representative sequences for 'regulated', 'endemic' and 'higher plant' datasets were merged into a single FASTA file with chloroplast genomes. To make the database globally compatible, we removed the information on biosecurity nature from the headers of the sequences in the public version of the database.

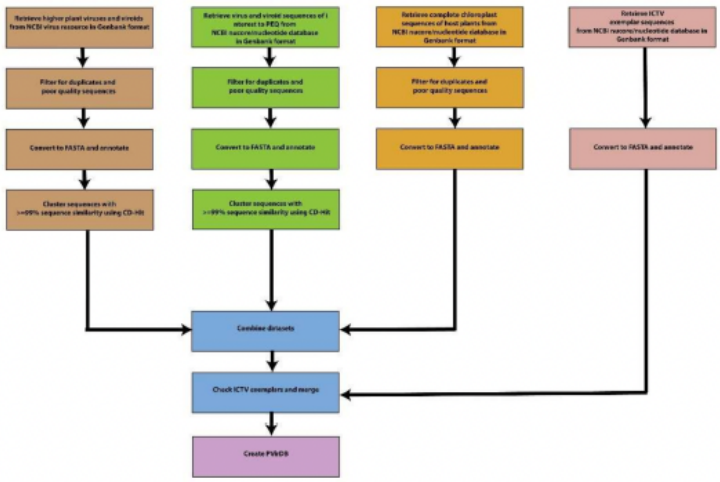


Figure 2. Workflow for generation of PVirDB.

We compared their sequence similarity to exemplar sequences from the International Committee for Taxonomy of Viruses (ICTV) to validate the selected representative se-

quences. Where appropriate, we ingested these ICTV exemplar sequences into PVirDB. To this end, we initially obtained virus sequences from the Virus Metadata Repository file available on the ICTV website [46,47]. This file was then filtered, removing all non-exemplar sequences and non-plant host viruses, generating an ICTV plant virus exemplar list. This list was then used as a guide to verify and correct genus and family taxonomic data in the PVirDB fasta headers. ICTV exemplar sequences that were not already represented in the PVirDB within a 97% sequence identity and 95% sequence coverage threshold were appended to the PVirDB FASTA file as new entries. A copy of the dataset for PVirDB was made publicly available at Zenodo's data repository (https://doi.org/10.5281/zenodo.6609576).

3. Results

3.1. Pilot Study Implementing and Testing the GA-VirReport Bioinformatics Workflow

In this study, we re-sampled host plant small RNAs of eight virus-positive reference plants held at PEQ (Table 1) from the same pool used in Gauthier et al. [20] sequenced using unique dual indexing adapters on a NovaSeq 6000 sequencing instrument. These known positive control plants from diverse genera, including *Citrus*, *Prunus*, *Iris*, *Ipomoea*, *Miscanthus* and *Rubus*, were used to test the implementation of the GA-VirReport bioinformatics workflow in the GA platform and compare its performance against VirReport and routine PEQ diagnostic assays.

Table 1. PEQ-positive control plants used in the pilot study infected with known viruses and viroids.

Sample	Commodity	Species	Known PEQ Detections [1]	Detections from VirReport [2]	Detections from GA-VirReport [3]
MT001	Citrus	Citrus Troyer × Frost-Lisbon	CEVd	CEVd	CEVd
MT002	Stonefruit	*Prunus persica* (Nectarine)	PNRSV	PNRSV	PNRSV
MT003	Citrus	*Citrus aurantiifolia* (Christm.) Swingle	CTV, CVEV, CDVd, HSVd	CTV, CVEV, CDVd, HSVd	CTV, CVEV, CDVd, HSVd
MT005	Raspberry	*Rubus idaeus*	RBDV	RBDV, RYNV	RBDV, RYNV
MT008	Citrus	*Citrus sinensis*	CDVd, HSVd	CDVd, HSVd, CiVA	CDVd, HSVd, CiVA
MT010	Ornamental grass	*Miscanthus sinensis* 'Morning light'	Novel *Potyvirus* (MsiMV)	Novel *Potyvirus* (MsiMV)	Novel *Potyvirus* (MsiMV)
MT012	Iris	*Iris* sp. 'Crimson colossus'	ISMV, TRSV	ISMV, TRSV, CRLV	ISMV, TRSV, CRLV
MT016	Sweet Potato	*Ipomoea batatas* 'GRF0334'	SPFMV	SPFMV, SPSMV-1, SPBV	SPFMV, SPSMV-1, SPBV

[1] Presence of regulated viruses confirmed using molecular (PCR and ELISA) and bioassays (biological and woody indexing): [2] Viruses and viroids reported through the VirReport pipeline according to Gauthier et al. [20]: [3] Viruses and viroids reported through GA-VirReport pipeline presented in this study; CEVd = citrus exocortis viroid, PNRSV = prunus necrotic ringspot virus, CTV = citrus tristeza virus, CVEV = citrus vein enation virus, CDVd = citrus dwarfing viroid, HSVd = hop stunt viroid, RBDV = raspberry bushy dwarf virus, RYNV = rubus yellow net virus, CiVA = citrus virus A, ISMV = iris severe mosaic virus, TRSV = tobacco ringspot virus, CRLV = cherry rasp leaf virus, SPFMV = sweet potato feathery mottle virus, SPSMV-1 = sweet potato symptomless virus 1, SPBV = sweet potato badnavirus.

3.1.1. Sequencing and Assembly Statistics

We generated, on average, 48.79 million reads per sample derived from the sequencing of eight positive control plants. Upon removing adapters, deconvolution of UMIs, and filtering of poor-quality bases and reads, we retained an average of 42.49 million (87.11%) high-quality reads per sample (Supplementary Table S1). Next, we removed non-informative sequences that mapped onto plant rRNAs, mitochondria and chloroplasts, leaving, on average, 12.39 million high-quality 18–25 nt-long reads per sample. For de-

tecting viruses and viroids, we used the 21–22 nt- or the 24 nt-long read fractions, which were on average 7.36 million (15.09% of the raw reads) and 3.32 million (6.80% of the raw reads), respectively. Only 1.52 million reads were preserved in the 21–22 nt fraction for sample MT002.

3.1.2. GA-VirReport Detected All PEQ-positive Viruses in Control Plants Using the 21–22 nt Read Fraction

All viruses and viroids detected by PEQ assays in the eight control plants were also detected by the GA-VirReport pipeline (Table 1) when using the 21–22 nt read fraction for de novo assembly and utilizing BLASTN to search for sequence similarity to PVirDB and NCBI NT databases (Supplementary Table S2). BLASTN searches against both PVirDB and NT generated the same virus/viroid detection for all samples, therefore validating the use of the PVirDB database.

For all cases except MT010, where a novel potyvirus tentatively named *Miscanthus sinensis* mosaic virus (MsiMV) was recently reported [48], the average percentage identities of the contigs to the top hit in the PVirDB database were above 92%. In MT010, the contigs showed on average 76.1% and 85.1% sequence similarity to the maize dwarf mosaic virus (MDMV) and sorghum mosaic virus (SrMV), respectively. The 1× read coverage of mapping 21–22 nt reads onto the respective reference sequence was over 80% in most cases except for the novel *Potyvirus* in MT010 and for citrus vein enation virus (CVEV) in MT003. The number of 21–22 nt reads mapping onto the CVEV genome (MN187043.1) was 835, and the percentage of bases covered at 1× depth was 59.1% (Figure 3 and Supplementary Table S2). However, the percentage of bases of the reference covered at 10× depth with 21–22 nt reads was over 70% for most cases, except the prunus necrotic ringspot virus (PNRSV) in MT002 at 35.44%, the novel *Potyvirus* in MT010, ranging from 20.85% to 31.12%, and CVEV in MT003, decreasing to just 7.87%. The percentage coverage of the reference genome by the aligned de novo-assembled contigs, defined as 'contig coverage', was also lowest in CVEV (0.82%). Compared to the other viruses in the MT003, with reads per kilobase million (RPKM) values ranging between 2366 and 5804, the lower value 7.40 in CVEV suggested the titer of the virus was significantly low in MT003 (Supplementary Table S2). These results suggest that a low number of 21–22 nt reads due to the low titer of the CVEV have impaired de novo assembly, generating only one short contig. The 10× read coverage showed a significant positive correlation ($p < 0.05$, $R^2 = 0.22$) to the contig coverage. No significant correlation was observed between the 1× read coverage and contig coverage from 21 to 22 nt reads, suggesting the depth of the reads mapping to the reference is crucial for generating read overlaps necessary for successful de novo assembly.



Figure 3. Read coverage of 21–22 nt and 24 nt reads mapping to the citrus vein enation virus genome (NC_021564.1) computed through the GA-VirReport-Stats pipeline. The scales of the mapped read depth for 21–22 nt reads and 24 nt reads are [0–48] and [0–10], respectively.

Identification of predicted ORFs showing homology to viral protein sequences in NCBI NT database provided additional evidence for the presence of infecting viruses detected via BLASTN (Supplementary Table S3). Among the ORFs predicted from the contigs assembled using the 21–22 nt fraction, nine ORFS showed above 98.6% sequence identity to citrus tristeza virus (CTV) in MT003, and one ORF showed 99% sequence homology to the raspberry bushy dwarf virus (RBDV) replicase protein (AGH55579.1) in MT005. In MT016, nine ORFs showed sequence homology above 98% to sweet potato feathery mottle virus (SPFMV) polyproteins, and one ORF showed homology to a putative disease resistance protein RGA4 [49].

3.1.3. The Sparser Number of Reads Recovered in the 24 nt Read Fraction Can Impair Viral Detection Relying on De Novo Assembly

The mapping read count and the percentage reference bases covered at 10× were significantly lower ($p < 0.01$) for the 24 nt read fraction than the 21–22 nt read fraction. Although the overall number of contigs generated by the 24 nt pipeline was significantly higher than that generated by the 21–22 nt pipeline ($p < 0.05$) (Supplementary Table S4), the number of contigs showing homology to viruses/viroids was significantly higher in the 21–22 nt pipeline ($p < 0.05$). We also observed an increase in the contig coverage of the 21–22 nt read fraction compared to the 24 nt read fraction ($p < 0.01$). The minimum scaffold/contig length returned from both pipelines was 29 bp, whereas the maximum scaffold lengths for the 21–22 nt and 24 nt pipelines were on average 366 bp and 262 bp, respectively. Interestingly, the 24 nt pipeline scaffolds' mean length was significantly shorter ($p < 0.01$) than that of their 21–22 nt counterparts. These results suggest that the 24 nt pipeline generates a larger number of short contigs during de novo assembly, which cannot be merged into scaffolds due to insufficient read overlap.

The PNRSV in MT002 and CVEV in MT003 were not detected through the de novo assembly of the 24 nt read fraction. Direct mapping onto the tripartite reference genomes of PNRSV aligned significantly more reads from the 21–22 nt read fraction than the 24 nt read fraction (1243 vs. 18 reads onto the PNRSV RNA1 (NC_004362.1); 1443 vs. 8 reads aligned onto the PNRSV RNA2 (NC_004363.1), and 13,437 vs. 107 reads aligned to the PNRSV RNA3 (NC_004364.1) for the 21–22 nt and 24 nt read fractions, respectively (Figure 4). Both read fractions showed the highest coverage across the PNRSV RNA3 with 84.47% of bases of reference covered by 21–22 nt reads and only 2.5% of bases covered by 24 nt reads at 10× read depth. Although the read coverage of the 24 nt reads for the PNRSV RNA3 was significantly higher than the read coverage for the PNRSV RNA1 and PNRSV RNA2, the overlap between reads were not significant to assemble contigs (Figure 4).

Similarly, for sample MT003, alignment of reads onto the reference CVEV genome (MN187043) yielded 835, and 10 mapped reads were derived from the 21–22 nt and 24 nt read fractions, respectively (Figure 3). The genome coverage of CVEV at 10× read depth represented 7.8% and 0% when mapping the 21–22 nt or 24 nt fractions, respectively. A lower mapping depth of 24 nt reads resulted in fewer read overlaps, impairing de novo assembly of both CVEV and PNRSV genomes.

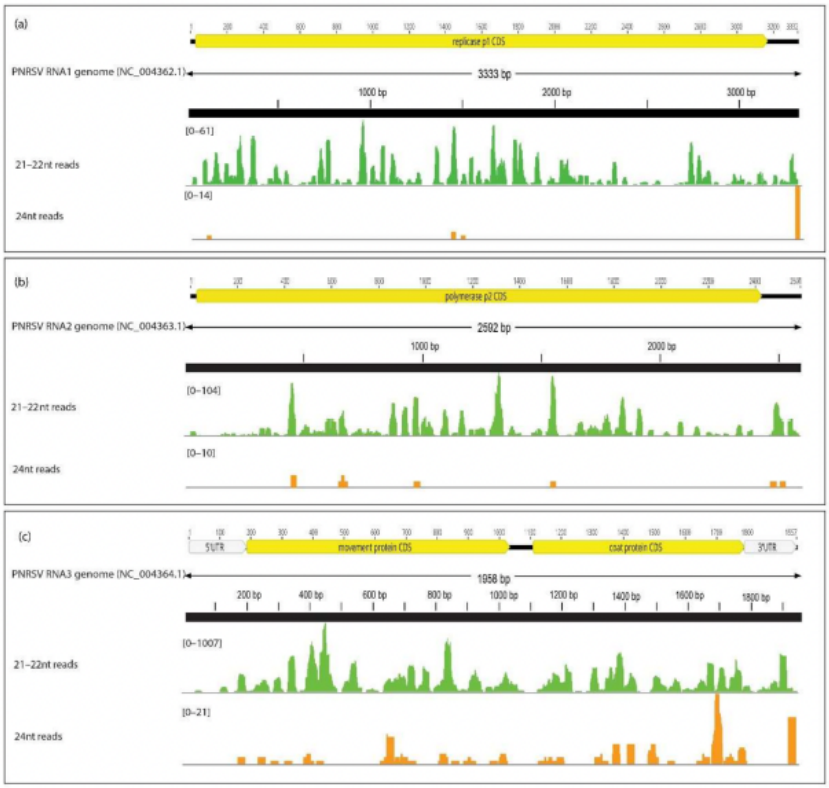


Figure 4. Mapping of filtered 21–22 nt reads and 24 nt reads of MT002 to the tripartite reference genomes of prunus necrotic ringspot virus (**a**) RNA1, (**b**) RNA2 and (**c**) RNA3.

3.1.4. GA-VirReport Detected Additional Viruses in Control Plants

All viruses and viroids detected by PEQ assays and HTS using VirReport [20] were also detected by the GA-VirReport pipeline using both 21–22 nt and 24 nt read fractions (Table 1). A rubus yellow net virus (RYNV), citrus virus A (CiVA) and cherry rasp leaf virus (CRLV) were detected in *Prunus*, *Citrus* and *Iris* control plants, respectively. Two viruses, the sweet potato symptomless virus 1 (SPSMV-1), and sweet potato badnavirus (SPBV), were also found in the sweet potato control plant (Table S2).

Overall, the average percentage identity of the contigs mapping to the top hit accessions of these viruses was above 91.98%. The genome coverage at 10× was above 49.05%, with the lowest for both statistics reported for the CRLV in MT012. The RPKM value of the CRLV (673.31) was significantly lower than that of the iris severe mosaic virus (ISMV; 1712.50) and tobacco ringspot virus (TRSV; 3786.39) reported from the same sample. Predicted ORFs with homology to known proteins provided additional support for these detections. In MT008, an ORF showed homology (with 96% identity) to the nucleocapsid protein of CiVA (QEE94613.1), and in MT012, three ORFs showing homology to CRLV with 91–97% identities were recovered.

3.1.5. Cross-Sample Contamination Is Not Detected

Gauthier et al. [20] reported the detection of viruses and viroids identified in control plants in other multiplexed samples in the same HTS experiment. In this study, to avoid index-hopping events found in HTS data, we used a different library preparation kit with

unique dual indexes. We did not detect any cross-sample contamination across any of the control samples tested. Moreover, none of the flagged false positives previously reported in a subset of control samples used in the current study were detected, thus supporting the cross-sample contamination cases proposed by Gauthier et al. [20].

3.1.6. Short Contigs of Plant Material Result in Misassignment in Sequence Similarity Searches against the PVirDB

We found four non-specific detections from the blastn searches of 21–22 nt (n = 1) and 24 nt (n = 3) read fractions due to the presence of short contigs (20–44 bp) showing high sequence similarity (>87.5%) to accessions in the PVirDB over a short alignment length (<57 bp). Furthermore, those non-specific cases showed low genome coverage at 1× (<11.28%) (Table S5). To assess if the assembled short contigs originated from host plant sequences [50], we subjected the short contigs to sequence similarity against the NCBI NT database. The top hits were non-coding RNAs (ncRNAs), including mitochondria and U1 spliceosomal ncRNA sequences showing higher nucleotide identities (≥98.21%) than top hits in PVirDB (Table S6). The exception was a short 30 nt-long contig that showed 100% sequence similarity to both a mitochondrion (LC697740.1) and a citrus exocortis viroid isolate (CEVd; MT561434.1). The use of megablast as the default blastn algorithm prevented the detection of non-specific hits in the 21–22 nt fraction. However, CEVd non-specific hits remained in the 24 nt fraction. Further investigation revealed that the CEVd sequence MT561434.1 was showing top similarity to FJ751930.1, FJ662762.1, DQ318794.1 and DQ318790.1, which are known host-derived sequences erroneously annotated in GenBank as CEVd [51,52]. Therefore, the CEVd sequence set was revised, and sequences with high sequence similarity to plant sequences were removed from PVirDB.

3.1.7. GA-VirReport-Stats Pipeline Produces Longer Consensus Sequences

We implemented a complementary GA-VirReport-Stats pipeline that uses the top BLASTN hits reported by the GA-VirReport pipeline to estimate read mapping statistics for each detected virus and viroid including genome coverage at different read depths, mean genome coverage and normalized read counts (RPKMs). Additionally, the GA-VirReport-Stats pipeline creates a reference-guided consensus viral genome sequence for each positive detection. The initial de novo assembly strategy minimizes false positives, and the subsequent reference-guided approach aims to extend the genome coverage of the detected viruses and viroids. We found, on average, a 52.43% increase in genome coverage by consensus sequences compared to de novo-assembled contigs (Supplementary Table S6). Top hits with sequence similarity to de novo-assembled contigs below 90% did not yield complete consensus sequences due to mapping small RNAs with more than three mismatches onto a reference genome, which allows for detecting sequences with up to 10–12% sequence divergence. The exception of a hit with high sequence similarity was CVEV; despite showing 100% sequence similarity to MN187043, the consensus genomic sequence only resolved 59.1% of the reference genome (Supplementary Table S6). The low expression titer of CVEV with an RPKM value of 7.40 and only 7.87% base coverage at ≥10× is likely associated with the presence of a large fraction of Ns in the predicted consensus sequence.

3.2. Large-Scale Testing and Implementation of GA-VirReport Bioinformatics Workflow on Imported Plants under Quarantine

To assess the detection of plant viruses and viroids in quarantined plants held at the PEQ, we selected 105 plants, including *Prunus spp.* (n = 69), *Fragaria ananassa* (n = 14), *Pyrus communis* (n = 5), *Malus domestica* (n = 4), *Solanum tuberosum* (n = 9) and *Vitis vinifera* (n = 4) (Table S7). The following sections focus on the results of this large-scale study.

3.2.1. Sequencing Statistics

Small RNA-seq HTS was performed in two batches, batch 1 with 59 plant samples (9 *S. tuberosum* and 50 *Prunus* spp.) and batch 2 with 46 samples (19 *Prunus* spp., 5 *P. communis*, 4 *M. domestica*, 4 *V. vinifera* and 14 *F. ananassa*). Batch 1 and 2 sequencing

produced an average of 57 million and 39.2 million raw reads per sample, respectively (Supplementary Table S7). Upon quality filtering, the read fraction of 21–22 nt was higher than 3 million for all 17 plant samples with positive detections (Supplementary Table S8).

3.2.2. GA-VirReport Detected a Variety of Pests in Plants Quarantined at PEQ

Based on our results from the pilot study on positive control plants, we only used the 21–22 nt read fraction and megablast with the GA-VirReport pipeline to detect viruses and viroids in the subsequent large-scale study of imported plants. Out of 105 plants tested, 14 showed the presence of a total of 23 viruses/viroids. These included 12 *Prunus* spp. (4 nectarines, 2 Japanese plums, 3 sweet cherries, 1 peach, 2 hybrids of apricot and plum), 1 *M. domestica* and 1 *V. vinifera* (Table 2 and Supplementary Table S9).

Table 2. Detections of viral sequences in the large-scale high-throughput sequencing testing of quarantined plants using GA-VriReport.

Sample ID	Commodity	Species	[1] GA-VirReport Detection	[2] Subject Accession	[3] Average Percent Identity	[4] Percentage Contig Coverage	[5] Percentage Bases 10× Read Depth
MT338	Stone fruit	*Prunus salicina*	PrLV	MF510412	100	41.26	85.54%
MT341	Stone fruit	*Prunus salicina*	PrLV	MF510412	100	25.95	66.43%
MT350	Stone fruit	*Prunus avium*	CVA	LC422952	99.41	32.41	79.78%
			PDV	MT013233	98.8	64.01	96.00%
MT357	Stone fruit	*Prunus salicina* × armeniaca	NeVM	KT273412	99.48	47.71	95.01%
MT362	Stone fruit	*Prunus persica* var. nucipersica	NSPaV	KT273410	99.15	76.86	98.24%
MT363	Stone fruit	*Prunus persica* var. nucipersica	NSPaV	KT273410	99.06	76.32	97.44%
MT365	Stone fruit	*Prunus persica* var. nucipersica	NeVM	KT273412	99.16	60.24	96.51%
			NSPaV	KT273410	99.29	77.84	98.16%
MT368	Stone fruit	*Prunus persica* var. nucipersica	NeVM	KT273412	98.61	50.19	96.83%
MT370	Stone fruit	*Prunus persica*	NSPaV	KT273410	99.34	68.48	96.31%
MT378	Stone fruit	*Prunus salicinia* × *avium*	NeVM	KT273412	99.03	56.28	97.82%
MT383	Stone fruit	*Prunus avium*	CVA	KX370827	98.19	38.27	95.67%
			PDV	MT013233	98.81	74.61	97.96%
MT385	Stone fruit	*Prunus avium*	CVA	LC523004	92.44	14.1	98.07%
			PDV	MT013233	98.75	72.99	79.58%
MT410	Apple	*Malus domestica*	CCGaV	MK940543	99.79	33.56	76.70%
			ASPV	KF321967	96.87	24.55	59.02%
			ACLSV	KR606325	98.04	47.2	97.31%
MT411	Grapevine	*Vitis vinifera*	HSVd	MF979532	100	51.34	100.00%
			GRSPaV	MG938345	98.19	43.42	90.52%
			GYSVd	DQ377130	100	34.9	96.12%
			AGVd	GU327604	100	14.68	79.82%

[1] Viruses and viroids reported through GA-VirReport pipeline presented in this study: PrLV = prunus latent virus, CVA = cherry virus A, PDV = prune dwarf virus, NeVM = nectarine virus M, NSPaV = nectarine-stem-pitting-associated virus, CCGaV = citrus-concave-gum-associated virus, ASPV = apple stem pitting virus, ACLSV = apple chlorotic leaf spot virus, HSVd = hop stunt viroid, GRSPaV = grapevine-rupestris-stem-pitting-associated virus, GYSVd = grapevine yellow speckle viroid, AGVd = Australian grapevine viroid; [2] subject accession from PVirDB; [3] Average percentage identity of the cumulative alignment and the [4] coverage of the subject achieved by the alignment of de novo contig was computed through BlasTools java script [35] summarizing BLASTN outputs of GA-VirReport. [5] Percentage bases of the reference covered by 10× mapping depth of filtered 21–22 nt reads.

The average nucleotide percentage identity of the contigs to the respective virus/viroid sequence showing the best homology was over 98% for all detections, except for CVA in MT385, which was 92.4% identical to LC523004 (Table 2). The percentage of the reference sequence covered by the alignment of de novo-assembled contigs was highest (77.84%) in the nectarine-stem-pitting-associated virus (NSPaV) detection in MT365. The 1× genome

coverage for most viruses was ≥87.67%. Additionally, in MT350, one ORF showed 99% sequence identity to the PDV coat protein (ACY25757.1), and another predicted ORF showed 95.18% identity to replicase P1 of PDV (QIQ53097.1).

3.3. Curation of a Custom Plant Virus and Viroid Blast Database (PVirDB)

The PVirDB blast database comprises 49,206 individual entries. We verified the taxonomic information against the ICTV plant virus exemplar list, evaluating the latest available taxonomic classification. As ICTV is the accepted authority on virus taxonomic classification, the taxonomic information collected from GenBank records that were misassigned, outdated, or missing was replaced with ICTV taxonomic information. Among the FASTA entries in the PEQ dataset derived from GenBank data, 10.2% had missing genus data, and 20.0% had missing family data. Most entries for viroid species (7.8% of the PEQ dataset) did not contain family or genus information. We also found 222 incorrect genera and 291 erroneous family information data in GenBank records compared to ICTV taxonomic information. Among these entries, there was outdated taxonomic information that was no longer recognized by the ICTV but remained in GenBank files, such as: the *Mandarivirus*, which has since been demoted to a subgenus of *Potexvirus* [53], *Nucleorhabdovirus*, which has been split into three genera, namely *Alphanucleorhabdovirus*, *Betanucleorhabdovirus* and *Gammanucleorhabdovirus* [54], and the family *Luteoviridae*, which has since been abolished [55]. Specific taxonomy changes affecting 22.6% of the PEQ detections are noted in Supplementary Table S10. After incorporating the current curations, the PVirDB database contains representative sequences for 47,312 viruses and 1894 viroids belonging to 43 families and 185 genera (Supplementary Table S11).

4. Discussion

HTS technologies have transformed many fields of life sciences, including pest diagnostics. The ability to detect one or more infecting plant viruses or viroids without a priori knowledge enables reporting of all pests that may be present in a sample using a single assay. However, a critical factor that can increase the complexity of detections is index hopping events between multiplexed samples in the same HTS run, potentially leading to false positives [20]. In this study, we adopted an HTS library preparation protocol that prevents cross-sample contamination, ensures diagnostic accuracy and mitigates the need for post-processing procedures to identify false positives. We also present GA-VirReport, a user-friendly web-based analytical workflow for detecting plant viruses and viroids using small RNA-seq data derived from quarantined plants held at a PEQ.

The automated GA-VirReport workflow runs from processing raw data to detecting viruses and viroids against a custom blast database PVirDB. A custom script runs the BlastTools java script [28], which parses GA-VirReport outputs and summarizes detected pests in each sample. A subsequent GA-VirReport-Stats workflow uses the best hits identified by BlastTools to generate reference-guided consensus genomic sequences and mapping statistics for plants with positive detections. Both GA-VirReport and GA-VirReport-Stats run on the GA platform, an open, web-based analytical environment that enables plant pathologists and molecular biologists with limited bioinformatics expertise to execute diagnostic pipelines. The automated workflows are publicly available as sharable workflows that GA users can import into their profile and use with either the GA-integrated NT database or the PVirDB introduced through this study. The user-friendly one-click solution nature of the workflows, the descriptive statistics and the graphical quality reports reduce the bottleneck for bioinformatic data analysis and interpretation when using HTS in quarantine testing of plant viruses and viroids. Moreover, the reference-guided consensus sequence generated through the GA-VirReport-Stats pipeline enables primer design for independent validation for pathogens where PCR assays are unavailable. The mapping statistics generated through this post-processing pipeline provide an additional level of confidence supporting positive detections.

In this study, we also developed a curated plant virus and viroid database: PVirDB. Identifying viruses and viroids to the lowest taxonomy levels is crucial for regulatory decisions regarding the high-risk plants processed at quarantine facilities [2,3]. Usage of public databases for diagnostics can be challenging due to a lack of consistency among descriptors, incorrect user annotations and the large scale of public databases [56,57]. PVirDB contains a curated subset of representative sequences of viruses reported from higher plants in NCBI. With harmonized sequence descriptors with taxonomy (Genus, Family) information curated with the latest ICTV taxonomy release, PVirDB provides accurate taxonomic assignment of viruses and viroids detected through the GA-VirReport workflows. The portability of the database makes it compatible with other wet-lab protocols and bioinformatics pipelines for plant virus and viroid detection as well as other applications such as virus discovery and phylogenetics. Sequences in public databases can have erroneously assigned species names. For example, a plant-derived sequence erroneously annotated as CEVd caused false positives in the analysis of control plants. Additional checkpoints, including scrutinizing the sequence similarity of representative virus/ viroids sequences against ICTV exemplar sequences based on species demarcation criteria, will ensure the identification of other misannotated entries.

We detected viruses and viroids from 22 of the 113 plants tested through the GA-VirReport pipeline. We recovered all true-positive viruses and viroids in the plants diagnosed by PEQ assays [20]. The fact that the true-positive detections were the same across different sequencing platforms, commercial kits, sequencing providers and different sampling times indicates the robustness of the HTS as an approach for the diagnosis of plant viruses and viroids.

The abundance of pests varied among samples, with mapping read counts ranging from hundreds to thousands. However, in the control samples MT003 and MT008, where more than one virus was present, the viral titer did not necessarily match the abundance reported previously. In MT003, Gauthier et al. [20] reported CDVd as the lowest abundant pathogen, but in this study, it was CVEV. In the same study, HSVd was the lowest abundant pathogen in MT008. However, in this study, it was CDVd. The changes observed in titer among different viruses in the same plant sampled a year apart may highlight temporal differences. The viral titer within a host is primarily affected by seasonality [58]. An increase in environmental temperature alters the competition among viral strains within the plant hosts [59]. However, Gauthier et al. [20] and this study maintained the tested plants in PEQ glasshouses under the same growing conditions. Thus, changes in environmental factors are unlikely to explain the observed virus titer differences. Another possibility is the change in tissue sampling location between studies. Many viral determinants facilitate cell-to-cell and long-distance movement of the virus, eventually causing the infection to become systemic [60,61]. However, it is also well-documented that different viruses vary in their potential to cause systemic disease, and viral titer can be affected by the change in sampling location. Additionally, using different cDNA library kits can yield different biases [46], potentially affecting the small RNA content in samples. Finally, the interaction between co-infecting viruses and the plant immune response could impact viral abundance over time. These observations emphasize the importance of cDNA library kit selection, sampling time, growing conditions and collecting multiple tissue types for viral diagnostics.

We did not observe cross-sample contamination among samples tested in the pilot and large-scale studies presented here. Although this can occur at various stages during the wet-lab procedure, contamination due to index misassignment is an inherent issue of HTS using Illumina's ExAmp patterned flow cells [62]. Index misassignment occurs in pooled libraries due to the residual excessive primers and adapters used in library preparation, causing spurious extension of fragments with the incorrect sample index. When single or combinatorial dual indexing is used in pooled libraries, swapped indices can cause misassignment of samples [63]. This phenomenon is especially problematic in quarantine testing, as this can lead to false positives and potential destruction of a healthy plant if no confirmatory independent molecular assays are performed. Processing a single

sample in an HTS run is not a cost-effective strategy to mitigate index swapping events when processing a large volume of samples. Gauthier at al. [20] reported cross-sample contamination events in two different sRNA-seq approaches that used DNA nanoball (DNB)-based HTS and the Illumina NextSeq 500 platform, which are reported to have low levels of index misassignment compared to patterned flow cells [64]. Usage of unique dual indexing, blocking adapters and pooling libraries together just before sequencing has been proposed to mitigate cross-sample contamination [62,63]. When unique dual indices (UDIs) are incorporated during library preparation, misassigned reads can be identified via demultiplexing using both i7 and i5 indexes [63]. The use of unique molecular identifiers (UMIs) in addition to the i7 and i5 adapters can further increase sensitivity towards low-frequency variant detection and reduce PCR-induced errors in library preparation [64]. In the current study, we incorporated UDIs together with UMIs through the QIAseq miRNA UDI 96 index plate, which yielded accurate assignment of sequencing reads across all pooled libraries. Moreover, libraries were cleansed using a bead-based approach to remove residual primers and adapters [62,64]. These findings reinforce the necessity of using UDIs and optionally UMIs for small RNA-seq datasets as standard best practices to ensure the accuracy of HTS-based quarantine testing, especially with sequencing instruments that use patterned flow cells such as NovaSeq.

The host antiviral response against viruses is driven by Dicer 4 (DCL4) and Dicer 2 (DCL2) enzymes that lead to the cleavage of viral dsRNA molecules into small RNAs 21nt and 22 nt in length, respectively [65,66]. Both DCL4 and DCL2 are essential for host plant viral defense by RNA silencing [66]. Additionally, 24 nt-long sRNAs are recognized and cleaved by Dicer 3 (DCL3), an enzyme involved in the silencing of transposons and repetitive elements [67] that has also been reported to be involved in immunity against DNA viruses by inducing methylation in viral cytosines [68–70]. Interestingly, DCL4 can act cooperatively with DCL3 in triggering the antiviral response of DNA viruses [70]. Barrero et al. [10] reported that the 21–22 nt sRNA fraction detected most viruses, except for a citrus endogenous pararetrovirus that was better assembled with the 24 nt read fraction. Therefore, we initially implemented two separate assembly pipelines for both size fractions in the pilot study, the 21–22 nt pipeline to detect active viral infections and the 24 nt pipeline to flag possible host-genome-integrated viruses. Our results suggest that using the 21–22 nt read fraction improves and is sufficient for the reliable detection of viruses and viroids by generating longer contigs and reducing the risk of spurious detections. Comparatively shorter contigs generated through the 24 nt read fraction caused false positives due to random hits of assembled host RNA sequences to viral genomes. We detected four DNA viruses, namely, PGVA and SPSMV-1, which are single-stranded, and RYNV and SPBV, which are double-stranded DNA viruses using the 21–22 nt fraction. In all four DNA viruses, we found a greater number of 21–22 nt mapped reads than 24 nt reads, further supporting using the 21–22 nt read fraction for downstream de novo assembly and detection of viral pathogens. Therefore, we have made the 24 nt pipeline optional in the GA-VirReport workflow as the focus of routine diagnostics is to detect active de novo viral infections triggering the host plant RNAi response.

Another important finding was that the megablast algorithm outperformed the blast algorithm in blastn sequence similarity searches performed in this study. Blastn resulted in a higher number of spurious hits to the sequences in the PVirDB compared to the megablast in the pilot study. Compared to the blast algorithm, megablast is more optimized for finding closely related sequences with percentage identities >95%. Megablast can handle longer DNA sequences, and due to the word size being larger, it is also faster than blastn [71]. In the context of screening for regulated pests where we are interested in finding the best match to known pests rather than new-to-science pests, megablast was therefore the favored option.

While our bioinformatics toolkit provided the same level of sensitivity compared to existing PEQ molecular assays and reported additional pathogens that are not routinely tested at PEQ, an integrated method for distinguishing false negatives/positives and true

negatives/positives could provide an additional layer of confidence. This is of paramount importance in HTS approaches, including in this study, where the number of total reads generated by the sequencing instrument did not necessarily correlate to the 'number of usable reads', defined in the current study as the number of remaining reads after filtering out plant ribosomal, mitochondrial and chloroplast RNAs. Usable reads are 21–22 nt-long and can be used for de novo assembly. Inadequate recovery of 'usable reads' can cause false negatives if the viral titer is low. The number of usable reads is affected by the sampling methods, efficiency of the RNA extraction from different plant commodities and the success of library preparation, which may be prone to variability. Although sensitivity can be improved by increasing the sequencing depth [17], we have observed that attempts at deeper sequencing as high as 75 million reads to compensate for the low titer in the virus or poor quality of the RNA sample tends to introduce false positives. Massart et al. [17] proposed a minimum sequencing depth of 2.5 million raw reads for sRNA viral diagnostics, and Gauthier et al. [20] reported that at least 2.5 million quality-filtered reads were required to obtain 100% sensitivity. In the current study, the minimum number of 21–22 nt reads that was used to detect a pathogen was 1.51 million reads. However, we have observed that defining a universal threshold for the 'usable reads' is challenging due to variability in the quality and composition of RNA from different plants and in the genome size and titer of the pathogens present.

The use of appropriate controls at different sample preparation and sequencing steps is important to generate high-quality HTS data. Although controls were used to assess the library preparation and sequencing success by the sequencing service provider, additional controls are recommended to test for efficiency of the entire wet-lab process [72] and to account for variability among commodities. The introduction of an external positive control for each targeted virus introduces a risk of plant contamination at the PEQ [9]. Therefore, usage of known nucleic acids as an internal control is recommended. Kesanakurti et al. spiked plant tissue with leaf discs from a plant with two known endogenous viruses as an alien internal control for the nucleic acid extraction [73]. Allocating additional resources for maintenance of plants may not be cost-effective in post-entry quarantine facilities. Therefore, we propose the use of synthetic oligonucleotides, for example, invertebrate miRNA genes that are not present in plant genomes, as spiked in controls added during the RNA extraction as an internal control for future HTS testing at the PEQ. Detection of a known concentration of internal controls in the sequencing data will enable benchmarking the analytical sensitivity of the HTS approach and enable normalizing the read counts of the pathogens to eliminate noise due to contamination in addition to assuring the success of the wet-lab procedure.

5. Conclusions

Small RNA-Seq HTS using UDIs and subsequent processing of 21–22 nt-long read fractions through the GA-VirReport pipeline detected all viruses and viroids reported by PEQ assays and additional viruses reported by the VirReport pipeline. In this study, using a 21–22 nt read fraction that reflects the host antiviral RNAi response was better-suited for detecting plant viruses and viroids. Furthermore, the screening of 105 quarantined plants to assess their phytosanitary status represents a case study for adopting HTS for quarantine testing. HTS, as a first-pass screening method, enables targeted confirmatory testing using molecular assays in HTS-positive detections, thereby accelerating the quarantine testing process.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v14071480/s1. Table S1: sequencing statistics of the pilot study; Table S2: detections of the GA-VirReport pipeline from the pilot study from the 21–22 nt- and 24 nt-long filtered read fractions; Table S3: ORFs predicted in control samples; Table S4: assembly statistics of the 21–22 nt and 24 nt pipelines; Table S5: false positives resulting from 21–22 nt and 24 nt pipelines with the blastn algorithm; Table S6: comparison of de novo assembly versus reference-guided consensus assembly; Table S7: metadata of the large-scale study; Table S8: sequencing statistics of the

plants with positives detections in the large-scale study; Table S9: positive detections in large-scale study; Table S10: curations in the taxonomy assignments performed on the GenBank accessions based on ICTV; Table S11: family and genus composition of PVirDB.

Author Contributions: Conceptualization, R.A.B., J.P. and M.W.; methodology, R.V.L., M.-E.A.G. and R.A.B.; software, R.V.L., Z.L., M.-E.A.G., R.A.B., G.M. and C.T.; validation, R.V.L. and C.E.E.; formal analysis, R.V.L. and Z.L.; investigation, R.V.L., M.-E.A.G. and C.E.E.; resources, J.P., C.E.E., M.W. and A.D.; writing—original draft preparation, R.V.L.; writing—review and editing; R.V.L., Z.L., M.-E.A.G., R.A.B., J.P., A.D., C.E.E. and F.E.C.; supervision, R.A.B. and J.P.; project administration, R.A.B.; funding acquisition, R.A.B., J.P. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Biosecurity Innovation Program, project number 202048, and in part by the Hort Innovation, project number MT18005, 'Improving plant industry access to new genetics through faster and more accurate diagnostics using Next Generation Sequencing (NGS)' using the Hort Innovation Citrus, Grape Tables, Rubus, Potato and Nursery research and development levy, co-investment from Queensland University of Technology and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this pilot study are openly available in the Short Read Archive (SRA) database under the BioProject PRJNA752836 (https://dataview.ncbi.nlm.nih.gov/BioProject/PRJNA752836, accessed on 15 February 2022). The data presented in the large-scale quarantine testing study are not publicly available due to being commercial in confidence client data. Access to viruses' data can be provided upon request. GA-VirReport workflow for single-input files (https://doi.org/10.5281/zenodo.6387498), multiple-input files (https://doi.org/10.5281/zenodo.6387492), GA-VirReport-Stats workflow (https://doi.org/10.5281/zenodo.6387504) and PVirDB dataset (https://doi.org/10.5281/zenodo.6545171) are available through the Zenodo data repository (https://zenodo.org/, accessed on 27 June 2022).

Acknowledgments: We acknowledge the Science and Surveillance Group at PEQ for their assistance in growing and collecting plant material and the Galaxy Australia platform and support team for their support in providing access to infrastructure. We are also thankful to QUT's eResearch team for their support in accessing local compute infrastructure to optimize the bioinformatics pipeline. This work is supported by Galaxy Australia, a service provided by the Australian Biocommons and its partners. The service receives NCRIS funding through Bioplatforms Australia and the Australian Research Data Commons (https://doi.org/10.47486/PL105), as well as The University of Melbourne and Queensland Government RICF funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, S.; Krainer, K.M.C. Pandemics of People and Plants: Which Is the Greater Threat to Food Security? *Mol. Plant* **2020**, *13*, 933–934. [CrossRef] [PubMed]
2. Rodoni, B. The role of plant biosecurity in preventing and controlling emerging plant virus disease epidemics. *Virus Res.* **2009**, *141*, 150–157. [CrossRef] [PubMed]
3. Whattam, M.; Dinsdale, A.; Elliott, C. Evolution of Plant Virus Diagnostics Used in Australian Post Entry Quarantine. *Plants* **2021**, *10*, 1430. [CrossRef] [PubMed]
4. Martin, R.R.; Constable, F.; Tzanetakis, I.E. Quarantine regulations and the impact of modern detection methods. *Ann. Rev. Phytopathol.* **2016**, *54*, 189–205. [CrossRef]
5. FAO. Framework for pest risk analysis. *Int. Stand. Phytosanitary Meas.* **2007**, *1*, 13–14.
6. Gucek, T.; Trdan, S.; Jakse, J.; Javornik, B.; Matousek, J.; Radisek, S. Diagnostic techniques for viroids. *Plant Pathol.* **2017**, *66*, 339–358. [CrossRef]
7. Varma, A.; Singh, M.K. Chapter 6-Diagnosis of plant virus diseases. In *Applied Plant Virology*; Awasthi, L.P., Ed.; Academic Press: Cambridge, MA, USA, 2020; pp. 79–92.
8. Massart, S.; Olmos, A.; Jijakli, H.; Candresse, T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* **2014**, *188*, 90–96. [CrossRef]

9. Maree, H.J.; Fox, A.; Al Rwahnih, M.; Boonham, N.; Candresse, T. Application of HTS for Routine Plant Virus Diagnostics: State of the Art and Challenges. *Front. Plant Sci.* **2018**, *9*, 1082. [CrossRef]
10. Barrero, R.A.; Napier, K.R.; Cunnington, J.; Liefting, L.; Keenan, S.; Frampton, R.A.; Szabó, T.O.; Bulman, S.; Hunter, A.; Ward, L.; et al. An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. *BMC Bioinform.* **2017**, *18*, 26. [CrossRef]
11. Rubio, L.; Galipienso, L.; Ferriol, I. Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution. *Front. Plant Sci.* **2020**, *11*, 1092. [CrossRef]
12. Jones, S.; Baizan-Edge, A.; MacFarlane, S.; Torrance, L. Viral Diagnostics in Plants Using Next Generation Sequencing: Computational Analysis in Practice. *Front. Plant Sci.* **2017**, *8*, 1770. [CrossRef] [PubMed]
13. Bester, R.; Cook, G.; Breytenbach, J.H.J.; Steyn, C.; De Bruyn, R.; Maree, H.J. Towards the validation of high-throughput sequencing (HTS) for routine plant virus diagnostics: Measurement of variation linked to HTS detection of citrus viruses and viroids. *Virol. J.* **2021**, *18*, 61. [CrossRef] [PubMed]
14. Vivek, A.; Zahra, S.; Kumar, S. From current knowledge to best practice: A primer on viral diagnostics using deep sequencing of virus-derived small interfering RNAs (vsiRNAs) in infected plants. *Methods* **2020**, *183*, 30–37. [CrossRef] [PubMed]
15. Velasco, L.; Padilla, C.V. High-Throughput Sequencing of Small RNAs for the Sanitary Certification of Viruses in Grapevine. *Front. Plant Sci.* **2021**, *12*, 682879. [CrossRef] [PubMed]
16. Santala, J.; Valkonen, J.P.T. Sensitivity of Small RNA-Based Detection of Plant Viruses. *Front. Microbiol.* **2018**, *9*, 939. [CrossRef] [PubMed]
17. Massart, S.; Chiumenti, M.; De Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Komínek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L.; et al. Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies. *Phytopathology* **2019**, *109*, 488–497. [CrossRef]
18. Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Ling, K.-S.; Kreuze, J.; Fei, Z. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **2017**, *500*, 130–138. [CrossRef]
19. Ho, T.; Tzanetakis, I.E. Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **2014**, *471–473*, 54–60. [CrossRef]
20. Gauthier, M.-E.A.; Lelwala, R.V.; Elliott, C.E.; Windell, C.; Fiorito, S.; Dinsdale, A.; Whattam, M.; Pattemore, J.; Barrero, R.A. Side-by-Side Comparison of Post-Entry Quarantine and High Throughput Sequencing Methods for Virus and Viroid Diagnosis. *Biology* **2022**, *11*, 263. [CrossRef]
21. Qiagen CLC Genomics Workbench. Available online: https://digitalinsights.qiagen.com (accessed on 1 July 2022).
22. Geneious. Available online: https://www.geneious.com/ (accessed on 8 April 2022).
23. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef]
24. bcl2fastq2 Conversion Software v2.20. Available online: https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html (accessed on 8 April 2022).
25. Blankenberg, D.; Von Kuster, G.; Bouvier, D.; Baker, D.; Afgan, E.; Stoler, N.; Taylor, J.; Nekrutenko, A.; Galaxy the Galaxy Team. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* **2014**, *15*, 403. [CrossRef] [PubMed]
26. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**, *17*, 3. [CrossRef]
27. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **2017**, *27*, 491–499. [CrossRef] [PubMed]
28. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
29. FastQC. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 27 March 2022).
30. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef]
31. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef]
32. Huang, X.; Madan, A. CAP3: A DNA Sequence Assembly Program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef]
33. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
34. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
35. Blast-Tools. Available online: https://github.com/schmidda/blast-tools (accessed on 27 March 2022).
36. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
37. Okonechnikov, K.; Conesa, A.; García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **2015**, *32*, 292–294. [CrossRef] [PubMed]
38. Galaxy-Tools. Available online: https://github.com/bgruening/galaxytools (accessed on 27 March 2022).
39. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]

40. Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2013**, *42*, D7–D17.
41. Hatcher, E.L.; Zhdanov, S.A.; Bao, Y.; Blinkova, O.; Nawrocki, E.P.; Ostapchuck, Y.; Schäffer, A.A.; Brister, J.R. Virus Variation Resource–improved response to emergent viral outbreaks. *Nucleic Acids Res.* **2017**, *45*, D482–D490. [CrossRef]
42. Ganten, D.; Ruckpaul, K. Batch Entrez. In *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*; Springer: Berlin, Germany, 2006; p. 131.
43. Gutiérrez, P.; Rivillas, A.; Tejada, D.; Giraldo, S.; Restrepo, A.; Ospina, M.; Marín, M. PVDP: A portable open source pipeline for detection of plant viruses in RNAseq data. A case study on potato viruses in Antioquia (Colombia). *Physiol. Mol. Plant Pathol.* **2021**, *113*, 101604. [CrossRef]
44. Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **2016**, *11*, e0163962. [CrossRef]
45. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
46. Wright, C.; Rajpurohit, A.; Burke, E.E.; Williams, C.; Collado-Torres, L.; Kimos, M.; Brandon, N.J.; Cross, A.J.; Jaffe, A.E.; Weinberger, D.R.; et al. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genom.* **2019**, *20*, 513. [CrossRef]
47. International Committee of Taxonomy of Viruses (ICTV). VMR File Repository—VMR. Available online: https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/ (accessed on 30 June 2022).
48. Leblanc, Z.; Gauthier, M.-E.; Lelwala, R.; Elliott, C.; McMaster, C.; Eichner, R.; Davis, K.; Liefting, L.; Thompson, J.; Dinsdale, A.; et al. Complete genome sequence of a novel potyvirus infecting Miscanthus sinensis (silver grass). *Arch. Virol.* **2022**, *167*, 1–5. [CrossRef]
49. Césari, S.; Kanzaki, H.; Fujiwara, T.; Bernoux, M.; Chalvon, V.; Kawano, Y.; Shimamoto, K.; Dodds, P.; Terauchi, R.; Kroj, T. The NB-LRR proteins RGA 4 and RGA 5 interact functionally and physically to confer disease resistance. *EMBO J.* **2014**, *33*, 1941–1959. [CrossRef]
50. Wei, L.; Gu, L.; Song, X.; Cui, X.; Lu, Z.; Zhou, M.; Wang, L.; Hu, F.; Zhai, J.; Meyers, B.C.; et al. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3877–3882. [CrossRef] [PubMed]
51. Di Serio, F.; Flores, R.; Verhoeven, J.T.J.; Li, S.F.; Pallás, V.; Randles, J.W.; Owens, R.A. Current status of viroid taxonomy. *Arch. Virol.* **2014**, *159*, 3467–3478. [CrossRef] [PubMed]
52. Poojari, S.; Alabi, O.J.; Fofanov, V.Y.; Naidu, R.A. Correction: A Leafhopper-Transmissible DNA Virus with Novel Evolutionary Lineage in the Family Geminiviridae Implicated in Grapevine Redleaf Disease by Next-Generation Sequencing. *PLoS ONE* **2016**, *11*, e0147510. [CrossRef]
53. International Committee of Taxonomy of Viruses (ICTV). Proposal 2020.031P—Abolish the Genus Mandarivirus, Create a New Subgenus (Mandarivirus) in the Genus Potexvirus and One New Subgenus (Acarillexivirus) in the Genus Allexivirus (Tymovorales: Alphaflexiviridae). Available online: https://talk.ictvonline.org (accessed on 30 May 2022).
54. Walker, P.J.; Siddell, S.G.; Lefkowitz, E.J.; Mushegian, A.R.; Adriaenssens, E.M.; Alfenas-Zerbini, P.; Davison, A.J.; Dempsey, D.M.; Dutilh, B.E.; García, M.L.; et al. Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch. Virol.* **2021**, *166*, 2633–2648. [CrossRef] [PubMed]
55. Scheets, W.M.K.; Somera, M. (2020, Abolish the Family Luteoviridae (Tolivirales) and Move Its Genera to the Families Tombusviridae (Tolivirales) and Solemoviridae (Sobelivirales). Available online: https://ictv.global/ictv/proposals/2020.026P.R.Abolish_Luteoviridae.zip (accessed on 30 May 2022).
56. Adams, M.J.; Antoniw, J.F. DPVweb: A comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res.* **2006**, *34*, D382–D385. [CrossRef]
57. Bagheri, H.; Severin, A.J.; Rajan, H. Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics* **2020**, *36*, 4699–4705. [CrossRef]
58. Honjo, M.N.; Emura, N.; Kawagoe, T.; Sugisaka, J.; Kamitani, M.; Nagano, A.J.; Kudoh, H. Seasonality of interactions between a plant virus and its host during persistent infection in a natural environment. *ISME J.* **2020**, *14*, 506–518. [CrossRef]
59. Alcaide, C.; Sardanyés, J.; Elena, S.F.; Gómez, P. Increasing temperature alters the within-host competition of viral strains and influences virus genetic variability. *Virus Evol.* **2021**, *7*, veab017. [CrossRef]
60. Hipper, C.; Brault, V.; Ziegler-Graff, V.; Revers, F. Viral and Cellular Factors Involved in Phloem Transport of Plant Viruses. *Front. Plant Sci.* **2013**, *4*, 154. [CrossRef]
61. Rodrigo, G.; Zwart, M.P.; Elena, S.F. Onset of virus systemic infection in plants is determined by speed of cell-to-cell movement and number of primary infection foci. *J. R. Soc. Interface* **2014**, *11*, 20140555. [CrossRef]
62. Illumina, I. Effects of Index Misassignment on Multiplexing and Downstream Analysis. Available online: www.illumina.com (accessed on 30 May 2022).
63. Costello, M.; Fleharty, M.; Abreu, J.; Farjoun, Y.; Ferriera, S.; Holmes, L.; Gabriel, S. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genom.* **2018**, *19*, 332. [CrossRef] [PubMed]

64. MacConaill, L.E.; Burns, R.T.; Nag, A.; Coleman, H.A.; Slevin, M.K.; Giorda, K.; Thorner, A.R. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genom.* **2018**, *19*, 30. [CrossRef] [PubMed]
65. Bouché, N.; Lauressergues, D.; Gasciolli, V.; Vaucheret, H. An antagonistic function for Arabidopsis DCL2 in development and a new function for DCL4 in generating viral siRNAs. *EMBO J.* **2006**, *25*, 3347–3356. [CrossRef] [PubMed]
66. Deleris, A.; Gallego-Bartolome, J.; Bao, J.; Kasschau, K.D.; Carrington, J.C.; Voinnet, O. Hierarchical Action and Inhibition of Plant Dicer-Like Proteins in Antiviral Defense. *Science* **2006**, *313*, 68–71. [CrossRef]
67. Xie, Z.; Johansen, L.K.; Gustafson, A.M.; Kasschau, K.D.; Lellis, A.D.; Zilberman, D.; Weigel, D. Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLOS Biol.* **2004**, *2*, e104. [CrossRef]
68. Wang, Q.; Xue, Y.; Zhang, L.; Zhong, Z.; Feng, S.; Wang, C.; Xiao, L.; Yang, Z.; Harris, C.J.; Wu, Z.; et al. Mechanism of siRNA production by a plant Dicer-RNA complex in dicing-competent conformation. *Science* **2021**, *374*, 1152–1157. [CrossRef]
69. Deng, Z.; Ma, L.; Zhang, P.; Zhu, H. Small RNAs Participate in Plant–Virus Interaction and Their Application in Plant Viral Defense. *Int. J. Mol. Sci.* **2022**, *23*, 696. [CrossRef]
70. Dunoyer, P. La bataille du silence. *Med. Sci.* **2009**, *25*, 505–512. [CrossRef]
71. Zhang, Z.; Schwartz, S.; Wagner, L.; Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* **2000**, *7*, 203–214. [CrossRef]
72. Parker, V.L.; Cushen, B.F.; Gavriil, E.; Marshall, B.; Waite, S.; Pacey, A.; Heath, P.R. Comparison and optimisation of microRNA extraction from the plasma of healthy pregnant women. *Mol. Med. Rep.* **2021**, *23*, 258. [CrossRef]
73. Kesanakurti, P.; Belton, M.; Saeed, H.; Rast, H.; Boyes, I.; Rott, M. Screening for plant viruses by next generation sequencing using a modified double strand RNA extraction protocol with an internal amplification control. *J. Virol. Methods* **2016**, *236*, 35–40. [CrossRef] [PubMed]

Appendix 31: HTS testing for viruses and viroids across diverse plant commodities undergoing quarantine testing at PEQ.

Background

A total 206 plant consignments including Citrus, Rubus, Prunus, Grapevines and ornamental grasses were sequenced using small RNA-seq HTS and screened for the presence of quarantinable pests.

Raw reads for each sample were analysed using the VirReport pipeline. VirReport specifically reported quarantinable and endemic viruses and viroids of interest from screened plants with some exceptions: clonal grasses, border enforcement samples and special requests from clients (e.g., garlic samples). Therefore, we summarised in separate tables plants that were: i) only screened for quarantinable and endemics of interest (see **Table 31.1** and **31.2**) and ii) screened for all viruses/viroids detected on the samples (**Table 31.3**).

Major Findings

Samples screened specifically for quarantinable/endemics of interest (total = 206)

16 imported plants tested positive for the presence of quarantinable pests (15 viruses and one viroid). Three imported plants tested positive for the presence of endemic virus of interest (6 viruses). These detections were all in agreement with molecular testing (ELISA or Sanger of PCR product) that was performed at SSG either on the same or a different extract from the same plant (**Table 31.1 and 31.2**).

Contigs matching to Nectarine marafivirus M were detected in 5 prunus samples. This virus was not initially part of the prescribed testing at SSG and after review by PIO, its status was changed to quarantinable and independently validated by PCR and Sanger sequencing (**Table 31.1 and 31.2**).

NGS failed to detect a cherry virus F (CVF) present at low titer in a prunus sample (MT258) sequenced using AGRF but this virus was detected by another bioinformatics tool, VirusDetect, which uses reference-guided assembly prior to de novo assembly to identify viruses. MT258 is a re-sample of plant MT094 and MT098 and CVF presence was molecularly validated in all extracts using Nepovirus B PCR followed by Sanger sequencing. This suggests that the read abundance was too low in this sample so that our de novo assembly step could successfully assemble them into a contig (**Table 31.1 and 31.2**).

NGS failed to detect the endemic virus, cherry virus A (CVA) present at low titer in a prunus sample (MT399). This was molecularly validated at SSG using a nested generic Betaflexiridae PCR as well as CVA-specific PCR followed by sander sequencing. The fact that a nested PCR was required to get a PCR product with the generic Betaflexiridae method and that only a faint band was obtained with the CVA-specific PCR method confirm that virus particles were present at low concentration, which might explain why this virus was not detected by NGS (we only recovered one read matching this virus in the sRNA-Seq data) (**Table 31.1 and 31.2**).

Short contigs matched at low depth to a candidate novel virus with closest match to grapevine red globe virus were also detected in three prunus persica var. nectarina and one prunus avium samples. This virus is not known to infect prunus plants. Sanger sequencing returned chimeric sequences matching both plants and virus. Subsequent sequencing using MinION nanopore sequencing, performed in collaboration with PIC@PEQ with Shamila Abeynayake, was able to resolve around 5000 bp using nine overlapping amplicons (**Table 31.1 and 31.2**).

Fourteen rubus plants tested positive for rubus yellow net virus (RYNV) (MT306-MT317, MT320 and MT321). In our previous report, we already described how raspberries are known to harbour up to six endogenous RYNV integrated sites in their genome and thus most RYNV detection in rubus plants are likely to be matches to these integrated sequences. By mapping our contigs to a local database incorporating endogenous RYNVs, which have been provided by Dr Thien Ho, we found that the contigs showed highest homology to these endogenous RYNVs. Thus, we interpret these detections as endogenous RYNVs integrated in the genome of the rubus plant sampled (**Table 31.2**). Therefore, these do not contribute to our episomal viral counts in **Table 31.1**.

Table 31.1: Comparison of HTS and PEQ detections of regulated and endemic viruses.

	Number of plants		Number of viruses/viroids	
Method	**Regulated**	**Endemic**	**Regulated**	**Endemic**
PEQ +ve / HTS +ve	16	3	16	6
PEQ -ve / HTS +ve	5	0	5 (NeVM)	0
PEQ +ve / HTS -ve	1	1	1 (CVF)	1 (CVA)

Table 31.2. Top BLASTN results of assembled quarantinable and targeted endemic virus and viroid sequences, coverage statistics. Confirmed false negatives are represented in orange. sacc = subject (reference) accession sequence, slen = subject (reference) length, av-pident = average sequence identity percent of assembled sequence(s) to reference, PCT_10X = fraction of bases that attained at least 1X sequence coverage. Category = viral detection classification.

Sample	Batch	Species	PEQ	Virus/Viroid	Type	sacc	slen	av-pident	% bases at 10X	Category	Note
MT242	3 (A)	Vitis	Y	Grapevine virus B	Quarantinable	EF583906	7599	89.19	0.15	KNOWN FRAGMENT	
MT246	3 (A,R)	Citrus	Y	Citrus tristeza virus	Quarantinable	AB046398	19302	99.58	0.93	KNOWN	
MT247	3 (A,R)	Citrus	Y	Citrus tristeza virus	Quarantinable	MK779711	19260	98.22	0.88	KNOWN	
MT247	3 (A,R)	Citrus	Y on different extraction	Citrus vein enation virus	Quarantinable	MN187044	5982	99.46	0.95	KNOWN	
MT249	3 (A,R)	Citrus	Y	Citrus tristeza virus	Quarantinable	AB046398	19302	99.26	0.94	KNOWN	
MT249	3 (A,R)	Citrus	Y on different extraction	Hop stunt viroid	Quarantinable	MT917186	302	100	0.99	KNOWN	
MT258	3 (A)	Stonefruit	Y	ND [missed CVF, detected with VirusDetect]	Quarantinable	NA	NA	NA	NA	NA	
MT306	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	92.42	0.58	KNOWN FRAGMENT	
MT307	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	91.81	0.6	KNOWN FRAGMENT	
MT308	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	97.77	0.54	KNOWN FRAGMENT	
MT309	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	98.22	0.56	KNOWN FRAGMENT	
MT310	5	Rubus	NT	RYNV (endogenous)	Endogenous	MZ358192	7816	85.67	0.42	KNOWN FRAGMENT	
MT311	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	89.64	0.43	KNOWN FRAGMENT	
MT312	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	96.57	0.79	KNOWN	
MT313	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	95.76	0.77	KNOWN	
MT314	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	95.74	0.79	KNOWN	
MT315	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	96.94	0.79	KNOWN	
MT316	5	Rubus	NT	RYNV (endogenous)	Endogenous	MZ358192	7816	85.08	0.19	KNOWN FRAGMENT	
MT317	5	Rubus	NT	RYNV (endogenous)	Endogenous	MZ358192	7816	83.1	0.18	CANDIDATE NOVEL	
MT320	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	97.65	0.47	KNOWN FRAGMENT	
MT321	5	Rubus	NT	RYNV (endogenous)	Endogenous	MN245240	7844	99.51	0.51	KNOWN FRAGMENT	
MT331	5	Prunus	N	Grapevine red globe virus	Quarantinable in grapevine	MZ451069	6850	84.02	0.07	EXCLUDE	Novel virus
MT332	5	Prunus	N	Grapevine red globe virus	Quarantinable in grapevine	MZ451071	6847	86.01	0.07	EXCLUDE	Novel virus
MT338	5	Stonefruit	Y	Prunus geminivirus A	Quarantinable	MF504038	3174	99.86	0.86	KNOWN	
MT341	5	Stonefruit	Y	Prunus geminivirus A	Quarantinable	MF504038	3174	99.93	0.67	KNOWN FRAGMENT	
MT349	5	Prunus	N	Grapevine red globe virus	Quarantinable in grapevine	MZ451070	6850	86.83	0.11	EXCLUDE	Novel virus
MT350	5	Prunus	Y	Cherry virus A	Endemic	KY510908	7431	99.55	0.91	KNOWN	
MT350	5	Stonefruit	Y	Prune dwarf virus RNA1	Endemic	MZ220991	3345	99.5	0.96	KNOWN	
MT350	5	Stonefruit		Prune dwarf virus RNA2	Endemic	MZ220875	2518	99.61	0.9	KNOWN	
MT350	5	Stonefruit		Prune dwarf virus RNA3	Endemic	MZ221014	2039	98.8	0.99	KNOWN	
MT357	5	Stonefruit	Y	Nectarine marafivirus M	Quarantinable	KT273411	6421	99.38	0.97	KNOWN	
MT365	5	Stonefruit	Y	Nectarine marafivirus M	Quarantinable	KT273411	6421	99.12	0.99	KNOWN	
MT368	5	Stonefruit	Y	Nectarine marafivirus M	Quarantinable	KT273411	6421	98.61	0.99	KNOWN	
MT378	5	Stonefruit	Y	Nectarine marafivirus M	Quarantinable	KT273411	6421	98.94	0.99	KNOWN	
MT383	6	Prunus	Y	Cherry virus A	Endemic	KY510909	7418	98.95	0.96	KNOWN	
MT383	6	Stonefruit	Y	Prune dwarf virus RNA1	Endemic	MZ291947	3350	99.56	0.98	KNOWN	
MT383	6	Stonefruit		Prune dwarf virus RNA2	Endemic	MZ220884	2529	99.38	0.93	KNOWN	
MT383	6	Stonefruit		Prune dwarf virus RNA3	Endemic	MZ221007	2081	99.05	0.99	KNOWN	
MT384	6	Prunus	N	Grapevine red globe virus	Quarantinable in grapevine	KX109927	6863	91.18	0.02	EXCLUDE	
MT385	6	Prunus	Y	Cherry virus A	Endemic	KY510909	7418	94.05	0.83	KNOWN	
MT385	6	Stonefruit	Y	Prune dwarf virus RNA1	Endemic	MZ220991	3345	99.49	0.98	KNOWN	
MT385	6	Stonefruit		Prune dwarf virus RNA2	Endemic	MZ220875	2518	99.25	0.94	KNOWN	
MT385	6	Stonefruit		Prune dwarf virus RNA3	Endemic	MZ221007	2081	99.32	0.99	KNOWN	
MT399	6	Prunus	Y	ND [missed CVA, only recovered one read]	Endemic	NA	NA	NA	NA	NA	
MT429	7	Citrus	Y	Citrus vein enation virus	Quarantinable	LC433634	5982	99.54	0.98	KNOWN	
MT447	7	Stonefruit	Y	Prunus latent virus/Prunus geminivirus A	Quarantinable	MF504040	3174	99.71	0.96	KNOWN	
MT448	7	Stonefruit	Y	Prunus latent virus/Prunus geminivirus A	Quarantinable	MF504040	3174	99.71	0.93	KNOWN	
MT449	7	Stonefruit	Y	Nectarine marafivirus M	Quarantinable	KT273413	6701	97.33	0.96	KNOWN	
MT449	7	Stonefruit	Y	Prunus geminivirus A	Quarantinable	MF504038	3174	99.95	0.92	KNOWN	
MT451	7	Stonefruit	Y	Prunus geminivirus A	Quarantinable	MF504038	3174	99.77	0.98	KNOWN	
MT452	7	Stonefruit	Y	Prunus geminivirus A	Quarantinable	MF504038	3174	99.85	0.98	KNOWN	
MT453	7	Vitis	Y	Grapevine virus B	Quarantinable	EF583906	7599	91.04	0.94	KNOWN	
MT454	7	Vitis	Y	Grapevine virus B	Quarantinable	EF583906	7599	90.59	0.54	KNOWN FRAGMENT	

Samples screened for all viruses/viroids (total = 61)

Sixty-one plants (clonal grasses, bushes, trees and garlics) were screened for the presence of any viruses/viroids (**Table 31.3**). For these samples, only pests deemed of biosecurity concern were independently validated.

Four apple hammerhead viroids were detected in clonal grasses in samples sequenced at AGRF (MT259, MT260, MT262 and MT263) but these detections were flagged as false positives originating from apple samples present in the same sequencing batch (**Table 31.3**). This viroid is not of biosecurity concern.

Three clonal grasses (MT260, MT262 and MT263) harboured sequences that showed 85-95% homology to several badnaviruses of the family Caulimoviridae which were present at low titer (% bases covered at 10X < 10). Endogenous badnaviruses from this family occur in banana (Banana streak virus) and rice (Rice tungro bacilliform virus). Based on the low depth recovered for reads that match to these viruses and the fragmented nature of the read alignment to these references, we infer that our detections are likely to be also resulting from viral integrations present in the clonal grass genomes; and that it is unlikely that full ORFs are transcribed and subsequently translated, to cause disease (**Table 31.3**).

Eight individual plants representative of a consignment of symptomatic *Adenium obesum* were sent for assessment at SSG. These plants had already been screened by the client using ImmunoStrip and were known to be infected with cucumber mosaic virus and either a potex or a potyvirus. SSG confirmed via TEM that the samples were infected with a potexvirus. Sanger sequencing using generic potexvirus primers amplified the same 700 bp PCR product in all samples. This product spanned the RNA-dependent RNA polymerase (RdRP) with the best match to the predicted protein sharing only 72% identity with the most closely related potexvirus, which was a pepino mosaic virus. NGS data independently detected blastn matches showing 82-88% to potexviruses in sample MT481, MT482, MT485, MT487 and MT488 (**Table 31.3**). Based on the Sanger sequencing results, we combined the reads from the eight samples and were able to assemble ~7000 bp of the viral genome. The reconstituted genome is composed of 5 ORFs and encodes conserved domains present in potexviruses: an RNA-dependent RNA polymerase (RdRP), a coat protein and a complete triple gene block. It showed 60% homology via blastx to the RdRP of nerine virus X. A phylogenetic analysis on the predicted RNA-dependent RNA polymerase grouped the novel potexvirus with Nerine virus X and Cassava virus X (**Figure 31.1**). Given that the ICTV species demarcation criteria for the potexvirus genus state that viral genomes sharing <80% amino acid identity over the RdRP are distinct species, we conclude that the virus detected here represent a novel virus that has not been previously described. Further consultation with DAFF policy regulators deemed this virus as a pest of regulatory concern.

Finally, 14 garlic samples were infected with several viruses none of which were deemed to be of biosecurity concern (**Table 31.3**). The matches in grey presented in the table were confirmed to be spurious short hits that can be excluded.

Table 31.3. Top BLASTN results of assembled virus and viroid sequences, coverage statistics and contamination flag to identify false positives for clonal grasses, garlic samples and border enforcement samples originating from *Adenium obesum*. Confirmed false postitives are represented in red. Contigs matching to a confirmed novel candidate potexvirus are represented in blue. sacc = subject (reference) accession sequence, slen = subject (reference) length, av-pident = average sequence identity percent of assembled sequence(s) to reference, PCT_10X = fraction of bases that attained at least 1X sequence coverage, Category = viral detection classification. CF contamination flag, FP = False positive.

Sample	Batch	Species	PEQ	Virus/Viroid	Type	sacc	slen	av-pident	% bases at 10X	CATEGORY	CF
MT259	3 (A)	Clonal grasses	NT	Apple hammerhead viroid	Higher plant viruses	MK188696	435	100	0.39	KNOWN FRAGMENT	FP
MT260	3 (A)	Clonal grasses	NT	Banana streak UL virus	Higher plant viruses	KF545135	1239	85.96	0.15	EXCLUDE	
MT260	3 (A)	Clonal grasses	NT	Banana streak UM virus	Higher plant viruses	HQ593110	7532	81.34	0.02	EXCLUDE	
MT260	3 (A)	Clonal grasses	NT	Banana_streak_VN_virus	Higher plant viruses	KF318350	575	83.33	0.11	EXCLUDE	
MT260	3 (A)	Clonal grasses	NT	Dioscorea bacilliform virus	Higher plant viruses	KY555511	528	88.64	0.12	KNOWN FRAGMENT	
MT260	3 (A)	Clonal grasses	NT	Sugarcane bacilliform virus	Higher plant viruses	JN377535	7613	88.89	0.01	EXCLUDE	
MT261	3 (A)	Clonal grasses	NT	Apple hammerhead viroid	Higher plant viruses	MH049330	433	97.92	0.27	KNOWN FRAGMENT	FP
MT262	3 (A)	Clonal grasses	NT	Apple hammerhead viroid	Higher plant viruses	MK690630	374	99.29	0.41	KNOWN FRAGMENT	FP
MT262	3 (A)	Clonal grasses	NT	Banana streak UM virus	Higher plant viruses	KF545127	1239	80.14	0.03	EXCLUDE	
MT262	3 (A)	Clonal grasses	NT	Sugarcane bacilliform virus	Higher plant viruses	KM214357	7764	82.85	0.02	EXCLUDE	
MT263	3 (A)	Clonal grasses	NT	Apple hammerhead viroid	Higher plant viruses	MK188697	435	99.32	0.37	KNOWN FRAGMENT	FP
MT263	3 (A)	Clonal grasses	NT	Banana streak UL virus	Higher plant viruses	HQ593109	7401	82.76	0.13	EXCLUDE	
MT263	3 (A)	Clonal grasses	NT	Banana streak UM virus	Higher plant viruses	KF545119	1239	83.33	0.13	EXCLUDE	
MT263	3 (A)	Clonal grasses	NT	Cacao red vein banding virus	Higher plant viruses	MH785300	6846	95.92	0.01	EXCLUDE	
MT481	8	Evergreen shrub	Y	Potato virus X	Higher plant viruses	MK116552	6423	88.23	0.02	EXCLUDE	
MT481	8	Evergreen shrub	Y	Cassava common mosaic virus	Quarantinable	KT002433	609	81.94	0.11	CANDIDATE NOVEL	
MT481	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	KX525728	3356	97.46	0.97	KNOWN	
MT481	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	KX525732	3044	97.7	0.99	KNOWN	
MT481	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	MK330012	2220	100	0.99	KNOWN	
MT481	8	Evergreen shrub	NT	Hydrocharis dubia cucumovirus	Higher plant viruses	MN841296	2130	100	0.94	KNOWN	
MT481	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA3	Higher plant viruses	MT067620	2521	99.66	0.96	KNOWN	
MT481	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	100	0.92	KNOWN	
MT482	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	KX525728	3356	96.58	0.97	KNOWN	
MT482	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	MH709143	3046	97.58	0.99	KNOWN	
MT482	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	MK330012	2220	99.14	0.99	KNOWN	
MT482	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA3	Higher plant viruses	MT067620	2521	100	0.96	KNOWN	
MT482	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	100	0.9	KNOWN	
MT482	8	Evergreen shrub	NT	Potato virus X	Higher plant viruses	MK116552	6423	88.23	0.02	EXCLUDE	
MT483	8	Evergreen shrub	Y	Cassava common mosaic virus	Quarantinable	KT002433	609	83.33	0.11	CANDIDATE NOVEL	
MT483	8	Evergreen shrub	NT	Potato virus X	Higher plant viruses	MK116552	6423	83	0.02	EXCLUDE	
MT483	8	Evergreen shrub	NT	Hydrocharis dubia cucumovirus	Higher plant viruses	MN841294	3249	98.34	0.93	KNOWN	
MT483	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	100	0.9	KNOWN	
MT484	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	KX525728	3356	96.28	0.98	KNOWN	
MT484	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	MN422334	3051	98.32	0.99	KNOWN	
MT484	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	MK330012	2220	100	0.98	KNOWN	
MT484	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA1	Higher plant viruses	MT067619	3195	97.92	0.92	KNOWN	
MT484	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841298	3222	99.11	0.95	KNOWN	
MT485	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	KX525728	3356	96.61	0.97	KNOWN	
MT485	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	KT921315	3046	100	0.94	KNOWN	
MT485	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	MK330012	2220	99.77	0.99	KNOWN	
MT485	8	Evergreen shrub	NT	Lettuce chlorosis virus RNA1	Higher plant viruses	KX685958	8615	93.71	0.33	KNOWN FRAGMENT	
MT485	8	Evergreen shrub	NT	Lettuce chlorosis virus RNA2	Higher plant viruses	KY430286	8456	96.72	0.38	KNOWN FRAGMENT	
MT485	8	Evergreen shrub	Y	Potato aucuba mosaic virus	Higher plant viruses	MG356507	7125	82.58	0.04	EXCLUDE	
MT485	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	99.71	0.91	KNOWN	
MT485	8	Evergreen shrub	Y	White clover mosaic virus	Higher plant viruses	MT151614	5793	87.14	0.04	EXCLUDE	
MT486	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	KX525728	3356	96.62	0.97	KNOWN	
MT486	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	KT921315	3046	100	0.95	KNOWN	
MT486	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	MK330012	2220	100	0.99	KNOWN	
MT486	8	Evergreen shrub	NT	Hydrocharis dubia cucumovirus	Higher plant viruses	MN841294	3249	100	0.94	KNOWN	
MT486	8	Evergreen shrub	NT	Lettuce chlorosis virus RNA1	Higher plant viruses	KY430285	8602	100	0.3	KNOWN FRAGMENT	
MT486	8	Evergreen shrub	NT	Lettuce chlorosis virus RNA2	Higher plant viruses	KY430286	8456	95.53	0.37	KNOWN FRAGMENT	
MT486	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA1	Higher plant viruses	MT067619	3195	100	0.92	KNOWN	
MT486	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA3	Higher plant viruses	MT067620	2521	100	0.96	KNOWN	
MT486	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	100	0.92	KNOWN	
MT487	8	Evergreen shrub	Y	Cassava common mosaic virus	Quarantinable	KT002433	609	82.86	0.11	EXCLUDE	
MT487	8	Evergreen shrub	Y	Strawberry mild yellow edge virus	Higher plant viruses	MZ326681	2934	100	0.08	EXCLUDE	
MT488	8	Evergreen shrub	NT	Cucumber mosaic virus RNA1	Higher plant viruses	MK421105	3363	99.4	0.97	KNOWN	
MT488	8	Evergreen shrub	NT	Cucumber mosaic virus RNA2	Higher plant viruses	LC510820	3040	99.44	0.99	KNOWN	
MT488	8	Evergreen shrub	NT	Cucumber mosaic virus RNA3	Higher plant viruses	KY794710	2215	100	0.99	KNOWN	
MT488	8	Evergreen shrub	NT	Hydrocharis dubia cucumovirus	Higher plant viruses	MN841294	3249	99.28	0.97	KNOWN	
MT488	8	Evergreen shrub	Y	Nerine virus X	Higher plant viruses	AB219105	6582	97.22	0.06	EXCLUDE	
MT488	8	Evergreen shrub	NT	Populus tomentosa cucumovirus RNA1	Higher plant viruses	MT067619	3195	99.76	0.95	KNOWN	
MT488	8	Evergreen shrub	Y	Strawberry mild yellow edge virus	Higher plant viruses	MZ326681	2934	97.22	0.08	EXCLUDE	
MT488	8	Evergreen shrub	NT	Toona sinensis cucumovirus	Higher plant viruses	MN841299	2505	100	0.92	KNOWN	

MT498	9	Garlic	NT	Garlic latent virus/ Shallot latent virus	Higher plant viruses	MN059207	8364	98.36	1	KNOWN
MT498	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059504	10112	99.55	1	KNOWN
MT498	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059565	10531	99.52	0.99	KNOWN
MT499	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059314	8726	87.83	0.72	KNOWN
MT499	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	91.64	0.82	KNOWN
MT499	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059147	8419	89.89	0.87	KNOWN
MT499	9	Garlic	NT	Garlic virus D	Higher plant viruses	AB010303	4497	91.47	0.78	KNOWN
MT499	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	KP258216	10341	85.11	0.75	KNOWN
MT499	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059636	10202	87.71	0.63	KNOWN FRAGMENT
MT500	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059388	8420	89.67	0.86	KNOWN
MT500	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059632	10527	86.86	0.6	KNOWN FRAGMENT
MT501	9	Garlic	NT	Garlic virus B	Higher plant viruses	MN059179	8342	98.17	0.99	KNOWN
MT501	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059368	8074	88.61	0.86	KNOWN
MT501	9	Garlic	NT	Narcissus yellow stripe virus	Higher plant viruses	LC314392	9637	79.03	0.04	EXCLUDE
MT501	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059630	10514	89.63	0.83	KNOWN
MT502	9	Garlic	NT	Garlic latent virus/ Shallot latent virus	Higher plant viruses	MN059236	8357	97.68	1	KNOWN
MT502	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059516	10115	99.47	0.99	KNOWN
MT502	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059603	10528	98.96	0.97	KNOWN
MT503	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059314	8726	89.38	0.71	KNOWN
MT503	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	90.09	0.82	KNOWN
MT503	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059147	8419	91.05	0.86	KNOWN
MT503	9	Garlic	NT	Garlic virus D	Higher plant viruses	AB010303	4497	92.02	0.78	KNOWN
MT503	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	KP258216	10341	85.83	0.74	KNOWN
MT503	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059631	10520	88.27	0.6	KNOWN FRAGMENT
MT503	9	Garlic	NT	Zucchini tigre mosaic virus	Higher plant viruses	MK988416	10331	94.87	0.05	EXCLUDE
MT504	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059299	8716	98	0.44	KNOWN FRAGMENT
MT504	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059368	8074	90.58	0.86	KNOWN
MT504	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059630	10514	88.97	0.83	KNOWN
MT505	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059367	8422	88.96	0.86	KNOWN
MT505	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059630	10514	91.17	0.82	KNOWN
MT505	9	Garlic	NT	Yam mosaic virus	Higher plant viruses	OK239700	9608	97.22	0.04	EXCLUDE
MT506	9	Garlic	NT	Garlic common latent virus	Higher plant viruses	KX255694	8604	95.41	0.98	KNOWN
MT506	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059313	8721	97.92	0.95	KNOWN
MT506	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	92.98	0.83	KNOWN
MT506	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059142	8413	96.58	0.98	KNOWN
MT506	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059368	8074	91.56	0.84	KNOWN
MT506	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059548	10364	92.69	0.93	KNOWN
MT506	9	Garlic	NT	Lettuce mosaic virus	Higher plant viruses	KJ161172	10079	90.7	0.07	EXCLUDE
MT506	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059601	10447	98.2	0.99	KNOWN
MT506	9	Garlic	NT	Telosma mosaic virus	Higher plant viruses	DQ851493	9689	93.18	0.06	EXCLUDE
MT507	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059322	8471	88.17	0.68	KNOWN FRAGMENT
MT507	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	92.5	0.86	KNOWN
MT507	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059147	8419	90.56	0.88	KNOWN
MT507	9	Garlic	NT	Garlic virus D	Higher plant viruses	AB010303	4497	92.58	0.79	KNOWN
MT507	9	Garlic	NT	Iranian johnsongrass mosaic virus	Higher plant viruses	KT899778	9544	89.47	0.08	EXCLUDE
MT507	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	KP258216	10341	87.16	0.74	KNOWN
MT507	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059638	10522	96.33	0.98	KNOWN
MT507	9	Garlic	NT	Zucchini tigre mosaic virus	Higher plant viruses	MK988416	10331	76.47	0.06	EXCLUDE
MT508	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059142	8413	96.57	0.99	KNOWN
MT508	9	Garlic	NT	Garlic virus D	Higher plant viruses	AB010303	4497	92.84	0.78	KNOWN
MT508	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059516	10115	99.22	1	KNOWN
MT508	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059630	10514	98.65	1	KNOWN
MT509	9	Garlic	NT	Bean yellow mosaic virus	Higher plant viruses	MG878448	275	88.14	0.27	KNOWN FRAGMENT
MT509	9	Garlic	NT	Cyrtanthus elatus virus A	Higher plant viruses	JQ723475	9908	85	0.05	EXCLUDE
MT509	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059321	8761	90.48	0.71	KNOWN
MT509	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059147	8419	90.83	0.87	KNOWN
MT509	9	Garlic	NT	Garlic virus D	Higher plant viruses	AB010303	4497	91.5	0.79	KNOWN
MT509	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	KP258216	10341	87.66	0.75	KNOWN
MT509	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059633	10523	87.66	0.64	KNOWN FRAGMENT
MT510	9	Garlic	NT	Garlic virus A	Higher plant viruses	MN059322	8471	91.39	0.79	KNOWN
MT510	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	90	0.73	KNOWN
MT510	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059142	8413	98.49	0.99	KNOWN
MT510	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059391	8425	98.25	0.97	KNOWN
MT510	9	Garlic	NT	Garlic virus X	Higher plant viruses	MN059426	8262	95.4	0.96	KNOWN
MT510	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059494	10366	94.57	0.98	KNOWN
MT510	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059597	10525	95.79	0.99	KNOWN
MT510	9	Garlic	NT	Shallot virus X	Higher plant viruses	MH389250	8907	83.33	0.29	CANDIDATE NOVEL
MT510	9	Garlic	NT	Wheat umbra-like virus	Higher plant viruses	OK573479	3532	97.56	0.11	EXCLUDE
MT511	9	Garlic	NT	Garlic common latent virus	Higher plant viruses	MN059138	8612	92.13	0.82	KNOWN
MT511	9	Garlic	NT	Garlic virus B	Higher plant viruses	AB010301	5077	89.03	0.84	KNOWN
MT511	9	Garlic	NT	Garlic virus C	Higher plant viruses	MN059146	8446	91.46	0.86	KNOWN
MT511	9	Garlic	NT	Garlic virus D	Higher plant viruses	MN059388	8420	93.48	0.95	KNOWN
MT511	9	Garlic	NT	Leek yellow stripe virus	Higher plant viruses	MN059494	10366	92.81	0.95	KNOWN
MT511	9	Garlic	NT	Onion yellow dwarf virus	Higher plant viruses	MN059638	10522	95.97	0.99	KNOWN
MT511	9	Garlic	NT	Potato virus M	Higher plant viruses	MH558035	8524	84.75	0.02	EXCLUDE

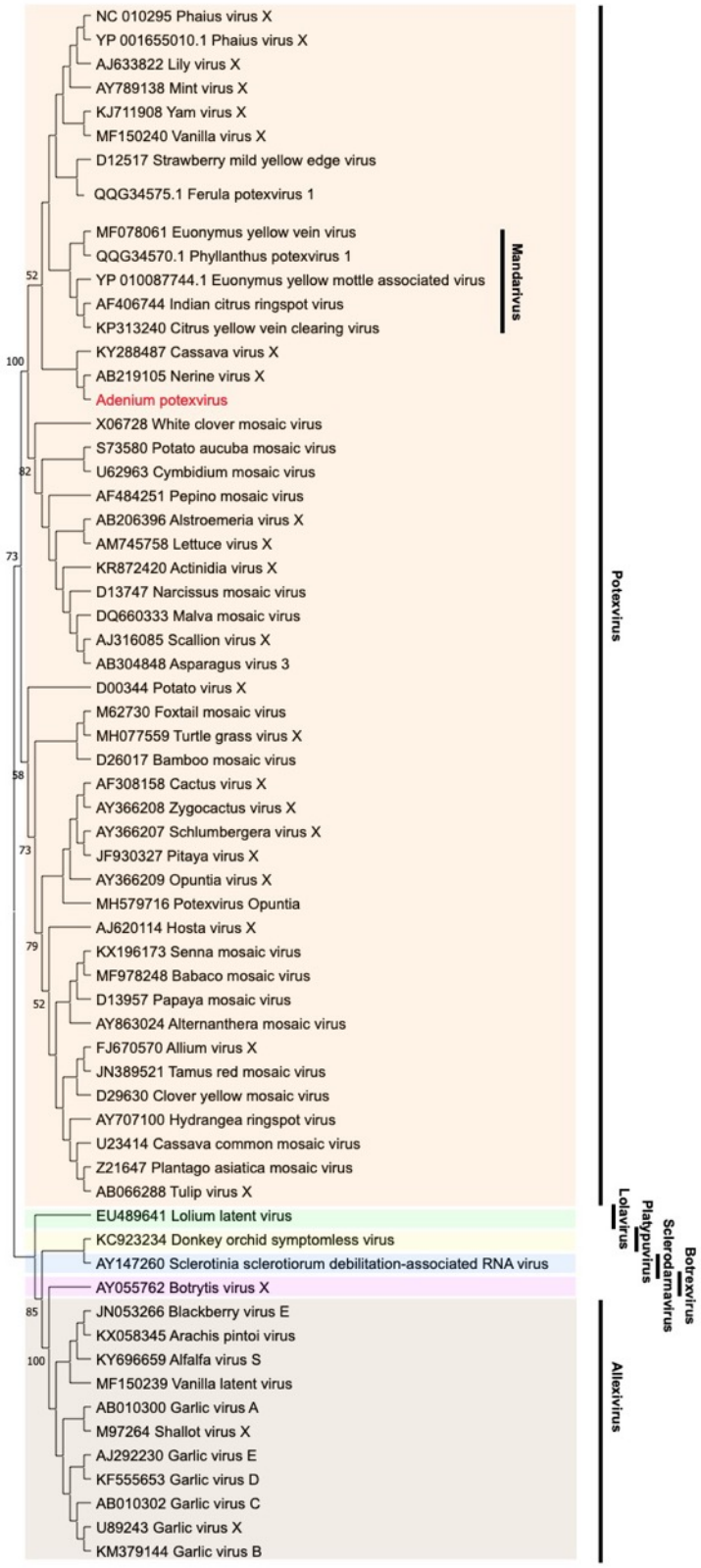


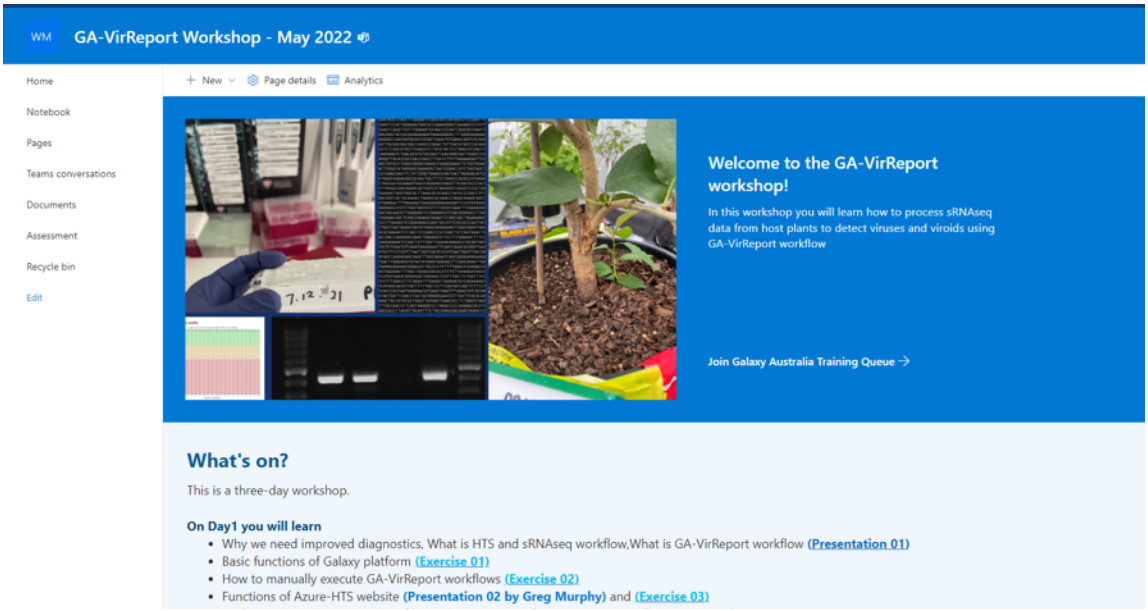
Figure 31.1. Phylogenetic analysis by maximum likelihood on the predicted RNA-dependent RNA polymerase.

Appendix 32: End-user hands-on training workshops

To enable training, access and dissemination of resources implemented in the project, we three in-person hands-on end-user workshops in Melbourne two specifically delivered to the SSG team at DAFF who are operationalizing the use of HTS for routine testing and another as part of an Australasian conference to engage diagnosticians from Australia States and Territories, New Zealand and Fiji. Additionally, a virtual training of researchers from NZ MPI, DAFF and AgVic to be delivered January 2023.

In person hands-on training workshops: Three face-to-face hands-on training workshops have been organised and delivered to train the SSG team and other groups across Australia and from overseas. These workshops include:

- **SSG hands-on training workshop (23-25 May 2022):** This workshop up-skilled SSG staff on the use of GA-VirReport pipeline that was developed based on the VirReport pipeline implementation in collaboration with Dr. Ruvini Lelwala. Training material and resources were shared with the SSG team via GovTeams' Share Point:





- **Hands-on training workshop at the 11th Annual Diagnosticians' Workshop (ADW) and the 5th Annual Surveillance Workshop (ASW) held at AgriBio on 30 August 2022**: A half-day end-user hands-on training workshop entitle "Diagnosis of Plant Viruses and Viroids using a Bioinformatics toolkit" was organised and delivered to diagnosticians from various organisations across Australia and overseas including DAFF, NZ MPI, Parks and Horticulture Authority NT, Biosecurity authority Fiji, Sugar Research Australia and

Department of Natural Resources and Environment Tasmania. The 23 participants were introduced to both VirReport and its counterpart GA-VirReport.

Workshop programme

Time	Topic
8.00 - 8.30 (recommended)	Setting up computers
8.30 - 8.35	Greetings and Welcome
8.35 - 9.10	**Introduction to sRNAseq and GA-VirReport workflow for processing sRNAseq data (Presentation 01)**
9.10 - 10.30	**Execution of GA-VirReport workflow in Galaxy Australia (Exercise 01)**
	Summarising GA-VirReport blastn outputs (Presentation 02) and (Demonstration 01)
	Execution of GA-VirReport-Stats workflow in Galaxy Australia (Presentation 03) and (Exercise 02)
10.30 - 10.45	Short break
10.45 - 11.00	**Processing of sRNAseq data with the prototype Azure-HTS website (Demonstration 02)**
11.00 - 11.30	**Introduction to VirReport pipeline (Presentation) and execution (Demonstration 03)**
11.30 - 12.00	**Interpretation of results and QnA (Interactive discussion)** **Post workshop survey**

- **SSG hands-on training workshop (14-16 May 2022):** In this workshop the participants learned about the changes to the GA processes, functions of PEQ-HiTS website, the VirReport pipeline and reporting and interpretation of outputs. As part of the workshop, 7 power point presentations were developed to train the SSG staff on how to use VirReport. Two exercises were also performed to check that the staff had assimilated the content of the lectures. All training material has been made available to all participants through GovTeams' Share Point space.

What's on?

This is a three-day workshop.

Day1

- Changes to the GA-VirReport workflows, summary process and PVirDB since workshop in May **(Presentation 01)**
- Understanding concepts of HTS controls and deduplication
- Demonstration and optional (Exercise 01) on running GA-VirReport with the new inputs
- Understanding the new outputs and intermediate files of GA workflows
- New functions of the PEQ-HiTS website (previously Azure-HTS website) **(Presentation 02 - Greg Murphy)**
- Understanding automated GA process, folder structure on SharePoint and content **(Presentation 03)**
- Demonstration and Hands-on (Exercise 02) on triggering entire workflow for assigned workshop samples and triggering individual steps for mock sample through the **PEQ-HiTS website**
- Providing feedback on **SOP5**

Day 2

- Introduction to VirReport, Nextflow, inputs required to run the pipeline and structure of output folder **(Presentation 04)**
- Running VirReport and introduction to NGS reporting template **(Presentation 05)**
- Understanding processes and outputs for the quality filtering step **(Presentation 06)**
- Hands on: populate the QC tabs of the report template using pre-computed results generated by VirReport for workshop batch using **VirReport SOP6** (Exercise 03)
- Q&A. Discussion about current UMI incorporation issue
- Understanding processes and outputs for the de novo read assembly step **(Presentation 07)**
- Understanding processes and outputs for the blast homology search and reference-guided assembly (viral database mode) **(Presentation 08)**
- Understanding processes and outputs for the blast homology search and reference-guided assembly (NCBI mode) **(Presentation 09)**
- Q&A. Get feedback on **SOP5**, **VirReport SOP6** and PEQ-HiTS website functionalities

Day3

- Understanding the **HTS metadata sheet** and **GA reporting template** **(Presentation 10)**
- Demonstration and Hands-on (Exercise 04) on populating GA-reporting template for workshop samples that you analysed on Day1 using GA-VirReport SOP6
- Hands on: populate the report template using pre-computed results generated by VirReport for workshop batch and review results obtained with VirReport and GA-VirReport using **VirReport SOP6** (Exercise 05)
- Interpretation of results and evidence categories **(Presentation 11)**
- Introducing the post-HTS decision making workflow and examples **(Presentation 12- Candace Elliot)**
- **Providing Feedback**

Virtual online VirReport and GA-VirReport workshops

A half-day online demonstration will be held on the 24th January 2023 to introduce both VirReport and GA-VirReport to end-users from Northern Australia Quarantine Strategy, NZ MPI, AgriBio and the Plant Innovation Centre at Post-Entry Quanrantine (PIC@PEQ, DAFF).

PEQ-HiTS website portal for routine HTS diagnosis

A PEQ-HiTS portal (**Figure 1**) developed by DAFF's Technology Infrastructure Branch has automated the execution of our developed VirReport and GA-VirReport end-to-end analysis pipelines enabling automated HTS data download from the accredited service provider, long term archival of data meeting data life cycle guidelines, and data processing of HTS data from genome assembly to reporting of pathogen detections. The implemented end-to-end automated solution will enable routine quarantine testing of imported plants at PEQ.

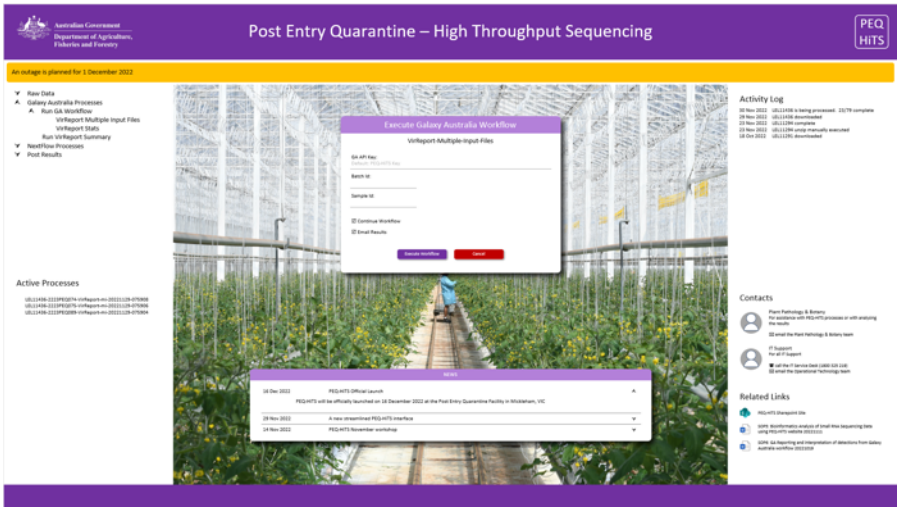


Figure 1: PEQ-HiTS portal – Post Entry Quarantine HTS website.

In partnership with DAFF's SSG, PIC@PEC, PIO, PSaRA and ISD teams the launch of PEQ-HiTS will take place on 16th December 2022 to enable the routine screening of viruses and viroids in imported plants via PEQ.

Date: Friday 16 December 2022
Location: Teams (virtual); PEQ, Mickleham, Vic
Duration: 2:30 to 3:15 pm.

Time	Overview	Presenter
2:30 pm	Acknowledgment of country	tba
2:32pm	Departmental overview outlining the importance of adding HTS to the diagnostic toolkit.	Biosecurity Plant & Science Services Division Executive (tba) Wayne See Kee, Assistant Secretary, Science and Surveillance Group Ben Wilson, Principle Director, Post Entry Quarantine
2:45pm	The start of the HTS journey at PEQ	Mark Whattam, Director Plant Innovation Centre
2:50pm	Our collaborating project lead	Associate Professor Roberto Barrero, Queensland University of Technology
2:55pm	Implementing HTS at PEQ	Julie Pattemore, Assistant Director SSG
3:00pm	PEQ-HiTS – a bespoke solution to managing big data	Greg Murphy, Director Operational Technology
3:10pm	Toasting the achievement	Virtual 'popping of the cork' – sparkling non-alcoholic beverages on site at PEQ

Appendix 33: End-User VirReport manual

33.1. Introduction

eresearchqut/VirReport is a bioinformatics pipeline based upon the scientific workflow manager Nextflow. It has been designed to help phytosanitary diagnostics of viruses and viroid pathogens in quarantine facilities. It takes small RNA-Seq fastq files as input. These can either be in raw format (currently only samples specifically prepared with the QIAGEN QIAseq miRNA library kit can be processed this way) or quality filtered.

For target identification, VirReport uses a hybrid *de novo* assembly approach to build contigs that are then annotated using blast homology searches against either a virus database or/and a local copy of NCBI nr and nt databases.

Nextflow is a workflow management software which enables the writing of scalable and reproducible scientific workflows. It can integrate various software package and environment management systems such as Docker, Singularity, and Conda. It allows for existing pipelines written in common scripting languages, such as Python and R, to be seamlessly coupled together. It implements a Domain Specific Language (DSL) that simplifies the implementation and running of workflows on cloud or high-performance computing (HPC) infrastructures. For a good introduction to Nextflow please refer to the following training materials:

- https://www.nextflow.io/docs/latest/getstarted.html
- https://training.seqera.io/
- https://carpentries-incubator.github.io/workflows-nextflow/

33.2. Pipeline summary

The VirReport workflow (Figure 33.1) will perform the following steps by default:

- Retain reads of a given length (21-22 nt long by default) from fastq file(s) provided in the index.csv file (**READPROCESSING**)
- *De novo* assembly using both Velvet and SPAdes. The contigs obtained are collapsed into scaffolds using cap3. By default, only contigs > 40 bp will be retained (**DENOVO_ASSEMBLY**)
- Run megablast homology search against either a local virus database or NCBI NT/NR databases:
- Searches against a local virus database:
    - Run megablast homology searches on *de novo* assembly against local virus and viroid database. Homology searches against blastn are also run in parallel for comparison with the megablast algorithm (**BLASTN_VIRAL_DB_CAP3**)
    - Retain top megablast hit and restrict results to virus and viroid matches. Summarise results by grouping all the *de novo* contigs matching to the same viral hit and deriving the cumulative blast coverage and percent identity for each viral hit (**FILTER_BLASTN_VIRAL_DB_CAP3**)
    - Align reads to top hit, derive coverage statistics and consensus sequence matching to top blast hit (**FILTER_BLAST_NT_VIRAL_DB_CAP3, COVSTATS_VIRAL_DB**)
    - Run tblastn homolgy search on predicted ORF >= 90 bp derived using getORF (**TBLASTN_VIRAL_DB**)
- Searches against local NCBI NT and NR databases:
    - Retain top 5 megablast hits and restrict results to virus and viroid matches. Summarise results by grouping all the de novo contigs matching to the same viral hit and deriving the cumulative blast coverage and percent ID for each viral hit (**BLASTN_NT_CAP3**)
    - Align reads to top hit, derive coverage statistics, consensus sequence and VCF matching to top blast hits (**COVSTATS_NT**)
    - Run blastx homolgy search on contigs >= 90 bp long for which no match was obtained in the megablast search. Summarise the blastx results and restrict to virus and viroid matches (**BLASTX**)

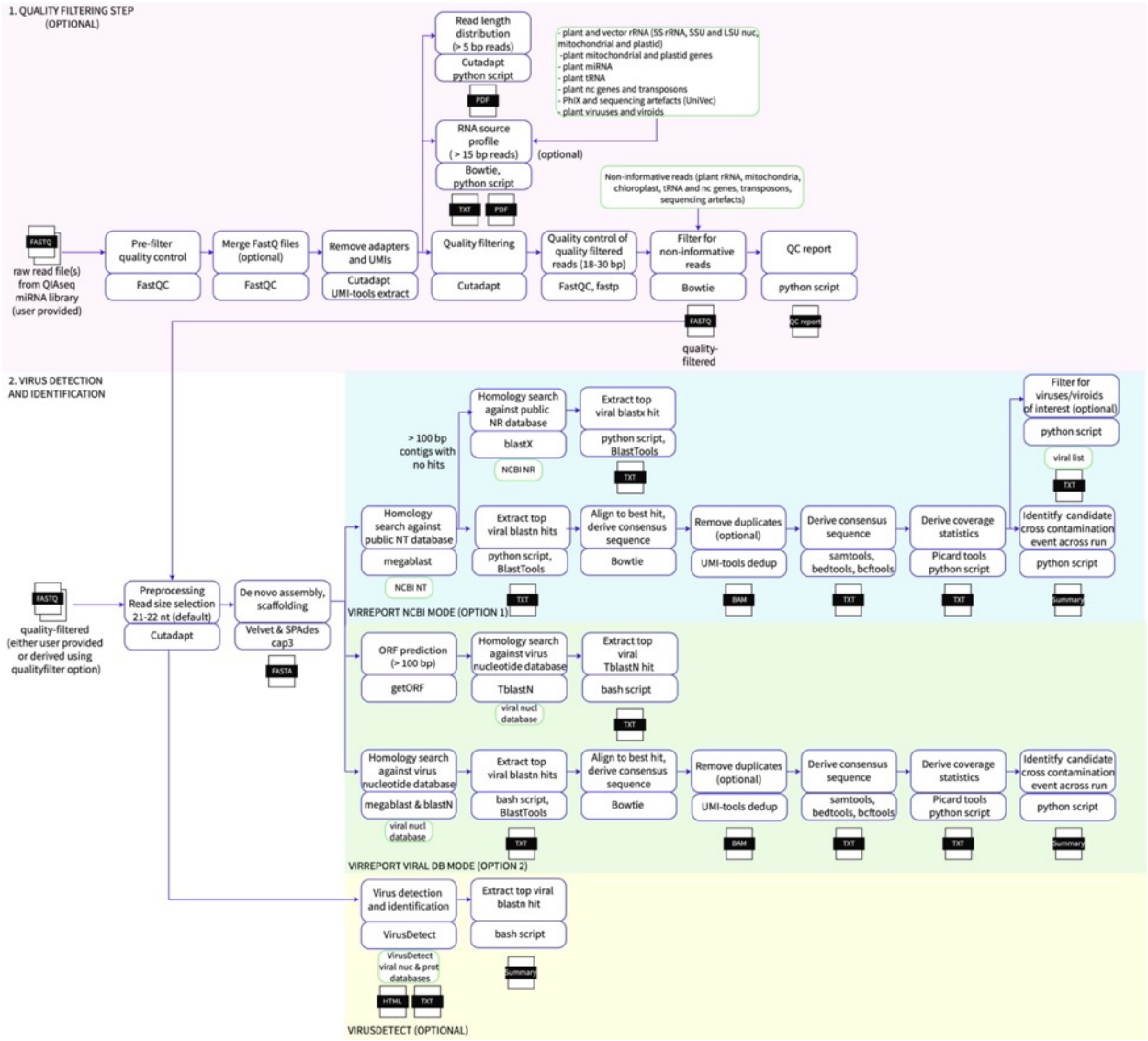


Figure 33.1. Overview of the VirReport end-to-end workflow (Gauthier et al., Biology, 2022).

- The pipeline can perform additional optional steps, which include:
    - A quality filtering step on raw fastq files (currently the workflow only processes samples prepared using QIAGEN QIAseq miRNA library kit). After performing quality filtering (**FASTQC_RAW, ADAPTER_TRIMMING, QUAL_TRIMING_AND_QC, DERIVE_USABLE_READS**). The pipeline will also derive a qc report (QCREPORT). An RNA souce profile can also be included as part of the quality filtering step (**RNA_SOURCE_PROFILE, RNA_SOURCE_PROFILE_REPORT**)
    - VirusDetect version 1.8 can also be run in parallel. A summary of the top virus/viroid blastn hits will be separately output (**VIRUS_DETECT, VIRUS_IDENTIFY, VIRUS_DETECT_BLASTN_SUMMARY, VIRUS_DETECT_BLASTN_SUMMARY_FILTERED**)

33.3 Pipeline prerequisites

- Basic unix command line knowledge (https://researchcomputing.princeton.edu/education/external-online-resources/linux; https://swcarpentry.github.io/shell-novice/

- Familiarity with one unix text editors (e.g. VIM (https://bioinformatics.uconn.edu/vim-guide/; https://missing.csail.mit.edu/2020/editors/or Nano (https://engineering.purdue.edu/ECN/Support/KB/Docs/BasictutorialforNanouhttps://www.howtogeek.com/howto/42980/the-beginners-guide-to-nano-the-linux-command-line-text-editor/)
- Java 11 or later, Nextflow, and Docker/Singularity/Conda to suit your environment.

Installing Java and Nextflow

You can follow the steps outlined in the Nexflow documentation to install Java and Nextflow on your local machine or server:https://www.nextflow.io/docs/latest/getstarted.html
This link specifically describes the steps to take to load Java and install Nextflow on our local HPC at QUT (Lyra): Nextflow

Installing a suitable environment management system

To run the VirReport pipeline, you will need to install a suitable environment management system such as Docker, Singularity or Conda to suit your environment.
We recommnend using Singularity.

Installing VirReport

The open-source VirReport code is available at https://github.com/eresearchqut/VirReport

Run the following command to get a copy of the source code:

```
git clone https://github.com/eresearchqut/VirReport.git
```

All the required modules are included in the environment.yml file in the pipeline base directory and shows all the tools used in the pipeline.

33.4 Running the pipeline

You can either invoke the pipeline by pointing to the location of main.nf in the version of VirReport you cloned, for example:

```
nextflow run ~/path_to_local_VirReport_copy/main.nf
```

or run directly `eresearchqut/VirReport`

```
nextflow run eresearchqut/VirReport
```

You will have to specify a profile to use to run the pipeline

```
nextflow run eresearchqut/VirReport -profile {docker, singularity or conda}
```

We recommend using Singularity as the profile.

Cached environment will be built in your home directory under the cached singularity directory. This step will take some time the first time you run the pipeline.

33.4A. Testing the pipeline on minimal test dataset:

Running these test datasets requires 2 cpus and 8 Gb mem and should take less than 5 mins to complete.

Make sure you have your nextflow config file set to “local" mode to run these tests:

```
process {
  executor = 'local'
  beforeScript = {
    """
    source $HOME/.bashrc
    source $HOME/.profile
    """
    }
```

This first command will test your installation using a single quality filtered fastq file (called test.fastq.gz) derived from a sample infected with citrus exocortis viroid and will run VirReport using a mock ncbi database (we recommend to use singularity on our local Lyra server at QUT):

```
nextflow -c conf/test.config run eresearchqut/VirReport -profile
test,{docker, singularity or conda} -latest -r main
```

```
N E X T F L O W  ~  version 20.10.0
Launching `eresearchqut/VirReport` [nostalgic_morse] - revision: 2a1fe2d8e3 [main]
NOTE: Your local project version looks outdated - a different revision is available in the remote repository [f9534bd5ed]
executor >  local (5)
[fd/771896] process > READPROCESSING (test)   [100%] 1 of 1 ✔
[85/c781f5] process > DENOVO_ASSEMBLY (test)  [100%] 1 of 1 ✔
[48/613bb9] process > BLASTN_NT_CAP3 (test)   [100%] 1 of 1 ✔
[c4/eef934] process > COVSTATS_NT (test)      [100%] 1 of 1 ✔
[74/34d6d4] process > CONTAMINATION_DETECTION [100%] 1 of 1 ✔
Completed at: 26-Aug-2022 15:10:58
Duration    : 1m 45s
CPU hours   : (a few seconds)
Succeeded   : 5
```

This second command will test your installation using a pair of raw fastq files (called test_pair_1.fastq.gz and test_pair_2.fastq.gz) derived from a sample infected with citrus tristeza virus and will run VirReport using a mock viral database (we recommend to use singularity on our local Lyra server at QUT):

```
nextflow -c conf/test2.config run eresearchqut/VirReport -profile
test2,{docker, singularity or conda} -latest -r main
```

```
(base) gauthiem@lyra02:~/code/github/clean/VirReport> nextflow -c conf/test2.config run eresearchqut/VirReport -profile test2,singularity
N E X T F L O W  ~  version 20.10.0
Launching `eresearchqut/VirReport` [small_baekeland] - revision: db9c5b4c90 [main]
executor >  local (13)
[ca/0f444e] process > FASTQC_RAW (test_pair)                  [100%] 2 of 2 ✔
[23/b4cd5d] process > MERGE_LANES (test_pair)                 [100%] 1 of 1 ✔
[74/1e7a6f] process > ADAPTER_AND_QUAL_TRIMMING (test_pair)   [100%] 1 of 1 ✔
[dc/5a0a84] process > QC_POST_QUAL_TRIMMING (test_pair)       [100%] 1 of 1 ✔
[93/b27f99] process > DERIVE_USABLE_READS (test_pair)         [100%] 1 of 1 ✔
[eb/bcc7f1] process > QCREPORT                                [100%] 1 of 1 ✔
[64/410699] process > DENOVO_ASSEMBLY (test_pair)             [100%] 1 of 1 ✔
[f6/baf266] process > BLAST_NT_VIRAL_DB_CAP3 (test_pair)      [100%] 1 of 1 ✔
[ca/e31461] process > FILTER_BLAST_NT_VIRAL_DB_CAP3 (test_pair) [100%] 1 of 1 ✔
[ce/e55245] process > COVSTATS_VIRAL_DB (test_pair)           [100%] 1 of 1 ✔
[e8/851660] process > CONTAMINATION_DETECTION_VIRAL_DB        [100%] 1 of 1 ✔
[70/48078b] process > TBLASTN_VIRAL_DB (test_pair)            [100%] 1 of 1 ✔
Completed at: 03-Aug-2022 13:22:38
Duration    : 2m 57s
CPU hours   : 0.1
Succeeded   : 13
```

- If both of these tests finish successfully, this means that the pipeline was set up properly.
- You can have a look at the files that have been created under the results folder.
- You are now all set to analyse your own samples.

33.4B. Running the pipeline with your own data

Provide an index.csv file

Create a TAB delimited text file that will be the input for the workflow to run. By default the pipeline will look for a file called “index.csv” in the base directory but you can specify any file name using the --indexfile [filename] in the nextflow run command. This text file requires the following columns (which needs to be included as a header):

```
sampleid,samplepath
```

- **sampleid** will be the sample name that will be given to the files created by the pipeline
- **samplepath** is the full path to the fastq files that the pipeline requires as starting input

An index_example.csv is included in the base directory:

```
sampleid,samplepath
MT212,/work/diagnostics/2021/MT212_21-22bp.fastq
MT213,/work/diagnostics/2021/MT213_21-22bp.fastq
```

If you need to set additional parameters, you can either include these in your nextflow run command:

```
nextflow run eresearchqut/VirReport -profile {singularity, docker or conda}
--indexfile index_example.csv --contamination_detection
```

or set them to **true** in the nextflow.config file.

```
params {
  contamination_detection = true
}
```

Provide a database

- By default, the pipeline is set to run homology blast searches against a local plant virus/viroid database (this is set in the nextflow.config file with parameter `--virreport_viral_db = true`. You will need to provide this database to run the pipeline. A curated database is provided at https://github.com/maelyg/PVirDB.git. Ensure you use NCBI BLAST+ makeblastdb to create the database. For instance, to set up this database, you would take the following steps:

  ```
  git clone https://github.com/maelyg/PVirDB.git
  cd PVirDB
  gunzip PVirDB_v1.fasta.gz
  makeblastdb -in PVirDB_v1.fasta -parse_seqids -dbtype nucl
  ```

  Then specify the full path to the database files including the prefix in the nextflow.config file. For example:

  ```
  params {
    blast_local_db_path = '/path_to_viral_DB/viral_DB_name'
  }
  ```

- If you also want to run homology searches against public NCBI databases, you need to set the parameter `virreport_ncbi` in the nextflow.config file to `true`:

  ```
  params {
    virreport_ncbi = true
  }
  ```

  or add it in your nextflow command:

  ```
  nextflow run eresearchqut/VirReport -profile {docker, singularity or
  conda} --virreport_ncbi
  ```

  Download these locally, following the detailed steps available at https://www.ncbi.nlm.nih.gov/books/NBK569850/ . Create a folder where you will store your NCBI databases. It is good practice to include the date of download. For instance:

  ```
  mkdir blastDB/30112021
  ```

You will need to use the update_blastdb.pl script from the blast+ version used with the pipeline.

For example:

```
perl update_blastdb.pl --decompress nt [*]
perl update_blastdb.pl --decompress nr [*]
perl update_blastdb.pl taxdb
tar -xzf taxdb.tar.gz
```

Make sure the taxdb.btd and the taxdb.bti files are present in the same directory as your blast databases.
Specify the path of your local NCBI blast nt and nr directories in the nextflow.config file.

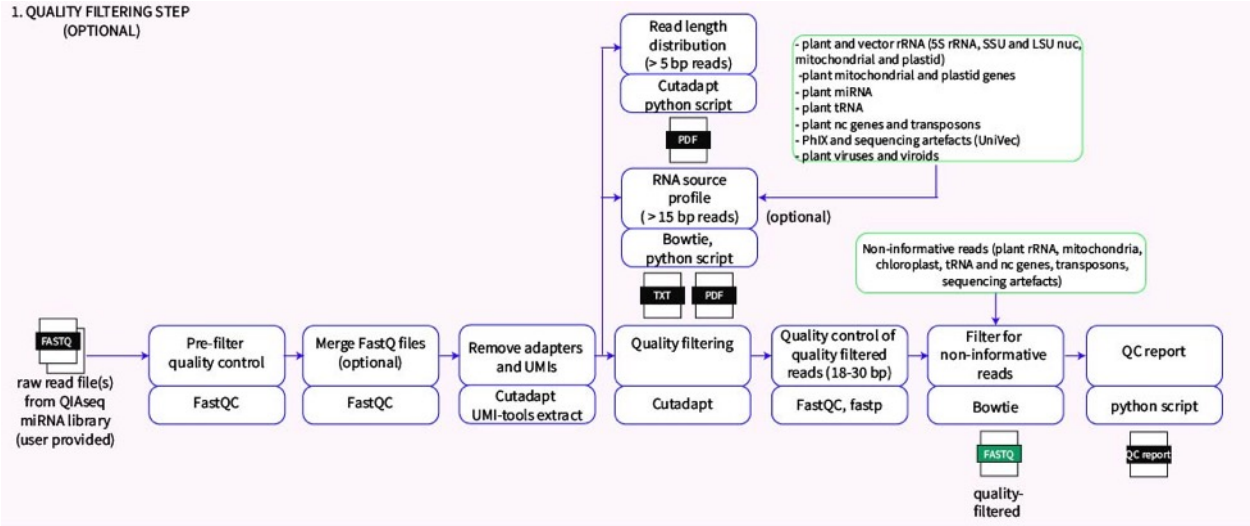
For instance:

```
params {
blast_db_dir = '/work/hia_mt18005_db/blastDB/20220408'
}
```

Fastq files

You can either provide raw or pre-filtered fastq files to the pipeline.

- By default the pipeline expects a single quality-filtered fastq file per sample.
- If you want to provide raw fastq files, samples have to be specifically prepared with the QIAGEN QIAseq miRNA library kit. If you want to run the initial quality filtering step on your raw fastq files, you will need to set the `--qualityfilter` paramater to `true` in the config.file and specify the path to the directory which holds the required bowtie indices (using the `--bowtie_db_dir` parameter) to: 1) filter non-informative reads (using the **blacklist** bowtie indices for the DERIVE_USABLE_READS process) and 2) optionally derive the origin of the filtered reads obtained (RNA_SOURCE_PROFILE process).



The required fasta files are available at https://github.com/maelyg/bowtie_indices.git and bowtie indices can be built from these using the command:

```
git clone https://github.com/maelyg/bowtie_indices.git
gunzip blacklist_v2.fasta.gz
#you might need to activate your environment cached in either your
conda or singularity environment in order to run bowtie
#for example
```

```
conda activate /path_to_cached_environment/virreport-
77d02f3abe1d8ba5f8dfdff194142de9
#then run the bowtie command
bowtie-build -f blacklist_v2.fasta blasklist
```

The directory in which the bowtie indices are located will need to be specified in the nextflow.config file:

```
params {
  bowtie_db_dir = '/path_to_bowtie_idx_directory'
}
```

If you are interested to derive an RNA source profile of your fastq files you will need to specify:

```
params {
  rna_source_profile = true
}
```

And build the other indices from the fasta files included in https://github.com/maelyg/bowtie_indices.git (i.e. rRNA, plant_tRNA, plant_noncoding, plant_pt_mt_other_genes, artefacts, miRNA, virus).
The quality filtering step will create the 00_quality_filtering folder under the results folder:

```
results/
├── 00_quality_filtering
    └── sample_name
    │   ├── sample_name_18-25nt_cutadapt.log
    │   ├── sample_name_fastqc.html
    │   ├── sample_name_fastqc.zip
    │   ├── sample_name_21-22nt_cutadapt.log
    │   ├── sample_name_21-22nt.fastq.gz
    │   ├── sample_name_24nt_cutadapt.log
    │   ├── sample_name_blacklist_filter.log
    │   ├── sample_name_fastp.html
    │   ├── sample_name_fastp.json
    │   ├── sample_name_qual_filtering_cutadapt.log
    │   ├── sample_name_quality_trimmed_fastqc.html
    │   ├── sample_name_quality_trimmed_fastqc.zip
    │   ├── sample_name_quality_trimmed.fastq.gz
    │   ├── sample_name_read_length_dist.pdf
    │   ├── sample_name_read_length_dist.txt
    │   ├── sample_name_truseq_adapter_cutadapt.log
    │   └── sample_name_umi_tools.log
    └── qc_report
        ├── read_origin_counts.txt
        ├── read_origin_detailed_pc.txt
        ├── read_origin_pc_summary.txt
        ├── run_qc_report.txt
        └── run_read_size_distribution.pdf
```

If your sequencing run was split on multiple lanes, you might have several raw fastq files per sample, and you can directly feed these to the pipeline and specify the `--merge-lane` parameter. The fastq files will be collapsed to one fastq file before performing downstream analysis. The sample name used will be the `sampleid` provided in the index.csv file. In the example below 2 fastq files were generated for 1 sample named CT103:

```
sampleid,samplepath
CT103,/path_to_fastq_files_directory/CT_103_S10_L001_R1_001.fastq.gz
CT103,/path_to_fastq_files_directory/CT_103_S10_L002_R1_001.fastq.gz
```

Running homology searches against viral database PVirDB

```
nextflow run eresearchqut/VirReport -profile singularity -resume --
indexfile index.csv \
                                    --merge_lane --qualityfilter --
rna_source_profile \
                                    --bowtie_db_dir /path_to_bowtie_indices
\
                                    --virreport_viral_db --
blast_viral_db_path /path_to_local_viral_database
```

Running homology searches against NCBI NT/NR database

```
nextflow run eresearchqut/VirReport -profile singularity -resume --
indexfile index.csv \
                                    --merge_lane --qualityfilter --
rna_source_profile \
                                    --bowtie_db_dir /path_to_bowtie_indices
\
                                    --virreport_ncbi --blast_db_dir
/path_to_ncbi_databases
```

Deduplicate reads using unique molecular identifiers and mapping coordinates

- If you used the QIAGEN QIAseq miRNA library kit for cDNA library preparation, your reads will include unique molecular identifiers (UMIs) at their 5' end. These are automatically extracted from reads and incorporated at the end of the read name when the pipeline is run using the `--qualityfilter` parameter.

- If you want to deduplicate reads from the BAM files that are derived at the alignment step, you will need to specify the `--dedup` parameter. UMI extraction and deduplication are performed with the tool umitools.

Summarising detections and flagging potential contamination events

- If you want to derive a summary of detections for all the samples included in the index file, specify the `--detecion_reporting_viral_db` or the `--detection_reporting_nt` option. This will create a summary text file under the Summary tab with a column called **contamination_flag**

- With the contamination flag, the assumption is that if a pest is present at high titer in a given sample and detection of reads matching to this pathogen in other samples occur at a significantly lower abundance, there is a risk that this lower signal is due to contamination (e.g. index hopping from high-titer sample). We first calculate the maximum FPKM value recorded for each virus and viroid identified on a run. If for a given virus, the FPKM value reported for a sample represented less than a percentage of this maximum FPKM value, it is then flagged as a contamination event. We apply a 1% threshold value as default. This is just indicative and method cannot discriminate between false positives and viruses present at very low titer in a plant. It is then recommended to compare the sequences obtained, check the SNPs and validate through independent method.

Running in diagnostic mode (SSG team only internal use)

- If you want to run VirReport in diagnostics mode (`--diagno`), the pipeline will also add an evidence category (ie KNOWN, KNOWN_FRAGMENT and CANDIDATE_NOVEL) to each detection based on **av-pident** and **% bases 10X**.
- If you are running homology searches against the NCBI NT database, you will also need to provide a list of pests of interest in the Targetted_Virus_Viroid.txt file located in the bin folder. If some of the detections match to this pest list, they will be categorised as **Quarantinable** versus **Higher_plant_viruses** in the final summary.
- Finally, if sample information is provided (**--sampleinfo --sampleinfo_path /path/to/sampleinfo.txt and --samplesheet_path /path/to/Sample_Sheet.csv**), this will be added to the final summary.

Sampleinfo.txt file example:

```
Sample  PEQ_index_number       LIMS_ID_RAMACIOTTI     Host_species
        Host_common_name       Plant_tissue_collected
MT498   P30     ELL110002A1    Allium sativum Garlic  50 mg leaf
```

Running VirusDetect

VirusDetect version 1.8 can also be run in parallel.
See http://virusdetect.feilab.net/cgi-bin/virusdetect/index.cgi for details about this separate pipeline.

Example of PBS script to run on an HPC with torque batch system

Make sure to either specify the full path to your index.csv file in the PBS script or place a copy of the index.csv file in the folder you will run the PBS script in.
Example 1:

- The PBS script example below (VirReport_nextflow.sh) will run on raw fastq files that will need to be merged and then quality filtered.
- We are also asking to run a process that will derive an RNA source profile for each samples during the quality filtering step.
- Blastn (using the megablast algorithm) and tblastn homology searches will be run against the PVirDB.
- Finally we will want the reads to be de-duplicated after mapping.

```
#!/bin/bash -l
#PBS -N VirReport
#PBS -l select=1:ncpus=2:mem=8gb
#PBS -l walltime=05:00:00


cd $PBS_O_WORKDIR
module load java
NXF_OPTS='-Xms1g -Xmx4g'

nextflow run eresearchqut/VirReport -profile singularity -resume --
indexfile index.csv \
                                    --merge_lane --qualityfilter --
rna_source_profile \
                                    --bowtie_db_dir /path_to_bowtie_indices
\
                                    --dedup \
```

```
                                        --virreport_viral_db --
blast_viral_db_path /path_to_local_viral_database \
                                        --detecion_reporting_viral_db
```

Example 2:
In the PBS job below, blastn (using the megablast algorithm) and blastx homology searches will be run against NCBI NR and NT respectively.

```
#!/bin/bash -l
#PBS -N VirReport
#PBS -l select=1:ncpus=2:mem=8gb
#PBS -l walltime=05:00:00

cd $PBS_O_WORKDIR
module load java
NXF_OPTS='-Xms1g -Xmx4g'

nextflow run eresearchqut/VirReport -profile singularity \
        -resume --indexfile index.csv \
        --merge_lane --qualityfilter \
        --rna_source_profile \
        --bowtie_db_dir /path_to_bowtie_indices \
        --dedup \
        --virreport_ncbi
        --blast_viral_db_path /path_to_ncbi_databases \
        --detecion_reporting_nt
```

Example 3:
In the PBS job below, homology searches will be run against NCBI and the PVirDB. The pipeline will also run VirusDetect in parallel.

```
#!/bin/bash -l
#PBS -N VirReport
#PBS -l select=1:ncpus=2:mem=8gb
#PBS -l walltime=05:00:00

cd $PBS_O_WORKDIR
module load java
NXF_OPTS='-Xms1g -Xmx4g'

nextflow run eresearchqut/VirReport \
        -profile singularity -resume \
        --indexfile index.csv \
        --merge_lane \
        --qualityfilter \
        --rna_source_profile \
        --bowtie_db_dir /path_to_bowtie_indices \
        --dedup \
        --virreport_ncbi \
        --blast_viral_db_path /path_to_ncbi_databases \
        --detecion_reporting_viral_nt \
        --virreport_viral_db \
        --blast_viral_db_path /path_to_local_viral_database \
        --detecion_reporting_viral_db \
        --virusdetect \
        --virusdetect_db_path
```

Example of samplesheet.csv (index) file:

```
sampleid,samplepath
MT447,/work/hia_mt18005/raw_data/20220218_RAMACIOTTI_LEL10024/MT447_S40_L001_R1_001.fastq.gz
MT447,/work/hia_mt18005/raw_data/20220218_RAMACIOTTI_LEL10024/MT447_S40_L002_R1_001.fastq.gz
MT449,/work/hia_mt18005/raw_data/20220218_RAMACIOTTI_LEL10024/MT449_S33_L001_R1_001.fastq.gz
MT449,/work/hia_mt18005/raw_data/20220218_RAMACIOTTI_LEL10024/MT449_S33_L002_R1_001.fastq.gz
```

Submit your job using the qsub command:

```
qsub VirReport_nextflow.sh
```

You can monitor your jobs using the command:

```
qstat -u $USER
```

Alternatively use the following command to check on the jobs you are running:

```
qjobs
```

You can also check the ".nextflow.log" file for details on progress.
Finally, if you have configured the connection to the Nextflow Tower (NFTower) you can logon and check your run.

33.5. Outputs

33.5A. Nextflow folder structure:

-**Work** folder where the intermediate results are kept so you can easily resume execution from the last successfully executed step. This can be deleted once analysis is finalised.

-**Results** folder where all the generated data files to be kept are saved. The pipeline will populate outputs under separate folders for each step. These will be stored in subfolders for each sample.

33.5B. VirReport results folder structure

Under the **Results** folder, the folders are structured as follows:

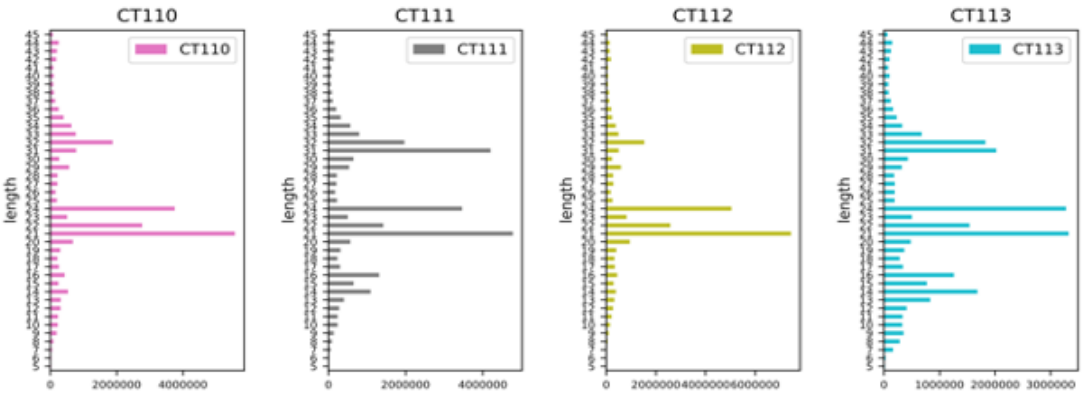
```
results/
├── 00_quality_filtering
│   └── sample_name
│       ├── sample_name_18-25nt_cutadapt.log
│       ├── sample_name_fastqc.html
│       ├── sample_name_fastqc.zip
│       ├── sample_name_21-22nt_cutadapt.log
│       ├── sample_name_21-22nt.fastq.gz
│       ├── sample_name_24nt_cutadapt.log
│       ├── sample_name_blacklist_filter.log
│       ├── sample_name_fastp.html
│       ├── sample_name_fastp.json
│       ├── sample_name_qual_filtering_cutadapt.log
│       ├── sample_name_quality_trimmed_fastqc.html
│       ├── sample_name_quality_trimmed_fastqc.zip
│       ├── sample_name_quality_trimmed.fastq.gz
│       ├── sample_name_read_length_dist.pdf
│       ├── sample_name_read_length_dist.txt
│       ├── sample_name_truseq_adapter_cutadapt.log
│       └── sample_name_umi_tools.log
│   └── qc_report
│       ├── read_origin_counts.txt
│       ├── read_origin_detailed_pc.txt
│       ├── read_origin_pc_summary.txt
│       ├── read_origin_pc_summary.txt
│       ├── run_qc_report.txt
│       └── run_read_size_distribution.pdf
├── 01_VirReport
│   └── sample_name
│       └── alignments
│           └──NT
│               ├── sample_name_21-22nt_all_targets_with_scores.txt
```

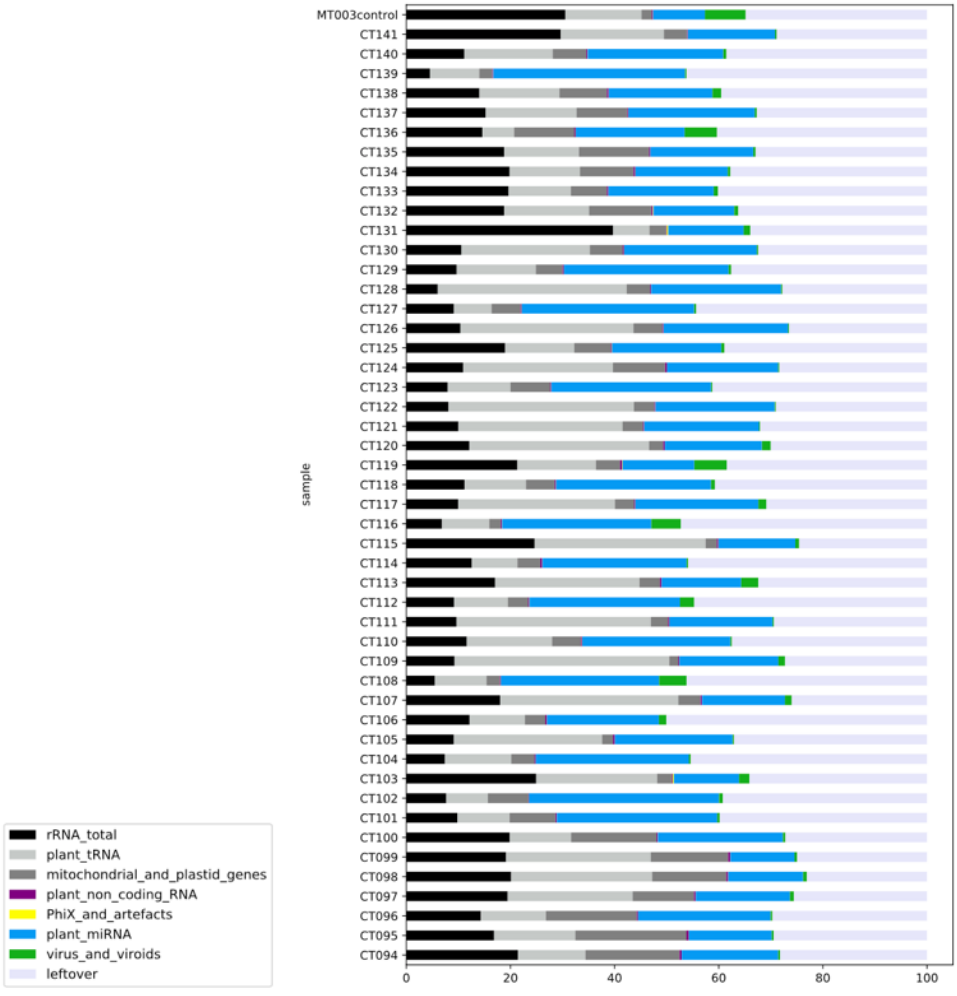
```
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_bowtie_log.txt
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name.consensus.fasta
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.dedup.bam
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.dedup.bam.bai
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.fa
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.fa.fai
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_norm.bcf
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_norm.bcf.csi
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_norm_flt_indels.bcf
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_norm_flt_indels.bcf.csi
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_picard_metrics.txt
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_sequence_variants.vcf.gz
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_sequence_variants.vcf.gz.csi
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_umi_tools.log
│   │   │   │   ├── sample_name_21-22nt_top_scoring_targets.txt
│   │   │   │   └── sample_name_21-
22nt_top_scoring_targets_with_cov_stats.txt
│   │   │   └──viral_db
│   │   │   │   ├── sample_name_21-22nt_all_targets_with_scores.txt
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_bowtie_log.txt
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name.consensus.fasta
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.dedup.bam
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.dedup.bam.bai
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.fa
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name.fa.fai
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_norm.bcf
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_norm.bcf.csi
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_norm_flt_indels.bcf
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_norm_flt_indels.bcf.csi
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_picard_metrics.txt
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_sequence_variants.vcf.gz
│   │   │   │   ├── sample_name_21-
22nt_GenBankID_virus_name_sequence_variants.vcf.gz.csi
│   │   │   │   ├── sample_name_21-22nt_GenBankID_virus_name_umi_tools.log
│   │   │   │   └── sample_name_21-
22nt_top_scoring_targets_with_cov_stats_viraldb.txt
│   │   └── assembly
│   │   │   ├── sample_name_cap3_21-22nt.fasta
│   │   │   ├── sample_name_spades_assembly_21-22nt.fasta
│   │   │   ├── sample_name_spades_log
│   │   │   ├── sample_name_velvet_assembly_21-22nt.fasta
│   │   │   └── sample_name_velvet_log
│   │   └──blastn
│   │   │   └── NT
│   │   │   │   ├── sample_name_cap3_21-22nt_blastn_vs_NT.bls
│   │   │   │   ├── sample_name_cap3_21-22nt_blastn_vs_NT_top5Hits.txt
│   │   │   │   ├── sample_name_cap3_21-
22nt_blastn_vs_NT_top5Hits_virus_viroids_final.txt
│   │   │   │   ├── sample_name_cap3_21-
22nt_blastn_vs_NT_top5Hits_virus_viroids_seq_ids_taxonomy.txt
```

```
│   │   │   │   └── summary_sample_name_cap3_21-
22nt_blastn_vs_NT_top5Hits_virus_viroids_final.txt
│   │   │   └── viral_db
│   │   │       ├── sample_name_cap3_21-22nt_blastn_vs_viral_db.bls
│   │   │       ├── sample_name_cap3_21-22nt_megablast_vs_viral_db.bls
│   │   │       ├── summary_sample_name_cap3_21-
22nt_blastn_vs_viral_db.bls_filtered.txt
│   │   │       ├── summary_sample_name_cap3_21-
22nt_blastn_vs_viral_db.bls_viruses_viroids_ICTV.txt
│   │   │       ├── summary_sample_name_cap3_21-
22nt_megablast_vs_viral_db.bls_filtered.txt
│   │   │       └── summary_sample_name_cap3_21-
22nt_megablast_vs_viral_db.bls_viruses_viroids_ICTV.txt
│   │   ├── blastx
│   │   │   └── NT
│   │   │       ├── sample_name_cap3_21-22nt_blastx_vs_NT.bls
│   │   │       ├── sample_name_cap3_21-22nt_blastx_vs_NT_top5Hits.txt
│   │   │       ├── sample_name_cap3_21-
22nt_blastx_vs_NT_top5Hits_virus_viroids_final.txt
│   │   │       └── summary_sample_name_cap3_21-
22nt_blastx_vs_NT_top5Hits_virus_viroids_final.txt
│   │   └── tblastn
│   │       └── viral_db
│   │           ├── sample_name_cap3_21-22nt_getorf.all.fasta
│   │           ├── sample_name_cap3_21-
22nt_getorf.all_tblastn_vs_viral_db_out.bls
│   │           └── sample_name_cap3_21-
22nt_getorf.all_tblastn_vs_viral_db_top5Hits_virus_viroids_final.txt
│   └── Summary
│       ├──
run_top_scoring_targets_with_cov_stats_with_cont_flag_FPKM_0.01_21-22nt.txt
│       └──
run_top_scoring_targets_with_cov_stats_with_cont_flag_FPKM_0.01_21-
22nt_viral_db.txt
└── 02_VirusDetect
    └── sample_name
    │   ├── blastn.reference.fa
    │   ├── blastn_references
    │   ├── blastx.reference.fa
    │   ├── blastx_references
    │   ├── contig_sequences.blastn.fa
    │   ├── contig_sequences.blastx.fa
    │   ├── contig_sequences.fa
    │   ├── contig_sequences.undetermined.fa
    │   ├── sample_name_21-22nt.blastn.html
    │   ├── sample_name_21-22nt.blastn.sam
    │   ├── sample_name_21-22nt.blastn_spp.txt
    │   ├── sample_name_21-22nt.blastn.summary.filtered.txt
    │   ├── sample_name_21-22nt.blastn.summary.txt
    │   ├── sample_name_21-22nt.blastn.txt
    │   ├── sample_name_21-22nt.blastx.html
    │   ├── sample_name_21-22nt.blastx.sam
    │   ├── sample_name_21-22nt.blastx.summary.txt
    │   └── sample_name_21-22nt.blastx.txt
    └── Summary
        ├── run_summary_top_scoring_targets_virusdetect_21-
22nt_filtered.txt
        └── run_summary_virusdetect_21-22nt.txt
```

Under the **00_quality_filtering** folder:

- a folder is created for each sample which contains zipped quality filtered fastq files, associated QC files and logs
- under the QC_report folder, read size distribution pdf file and read RNA source pdf file are created. The folder also includes a run_qc_report text file:





sample	raw_reads	umi_cleaned_reads	quality filtered reads > 18bp	total filtered bases	q20 bases	q30 bases	gc content	usable reads type	usable reads 18-25 nt	usable reads 21-22nt	usable reads 24nt
CT094	37,500,926	25,177,617	20,768,885	563,514,321	558,123,369	545,321,810	0.56	10,242,824	8,751,495	5,462,336	2,156,106
CT095	37,329,225	23,367,418	18,441,181	500,338,068	494,859,833	482,159,178	0.57	8,837,639	7,593,696	4,441,599	2,039,739
CT096	32,348,340	25,770,492	19,975,000	520,664,085	515,419,191	503,566,413	0.56	11,664,067	10,489,974	7,011,777	2,331,301
CT097	38,991,484	33,423,430	20,545,911	548,672,888	542,835,843	529,769,798	0.54	9,955,178	8,708,427	5,319,253	2,237,084
CT098	44,012,560	38,421,784	26,779,550	734,893,606	727,241,319	709,802,598	0.55	11,024,475	9,379,736	5,935,079	2,171,143
CT099	39,204,915	33,664,856	23,575,060	651,872,254	644,965,353	629,003,244	0.55	9,481,574	8,082,107	4,420,491	2,434,334
CT100	33,408,539	28,340,813	19,547,581	511,963,129	506,689,880	494,493,465	0.55	10,738,186	9,592,671	6,115,353	2,346,177
CT101	37,239,317	30,845,969	28,302,544	702,577,121	694,340,103	675,500,146	0.52	20,503,179	18,904,959	11,644,039	5,183,562
CT102	33,273,925	28,527,723	26,565,358	649,116,858	641,842,721	625,209,117	0.51	20,696,019	19,495,621	11,888,535	5,655,606
CT103	37,173,458	31,987,334	20,241,998	527,759,641	522,394,824	509,782,626	0.54	10,681,607	8,811,645	4,331,547	2,582,013
CT104	37,258,069	26,576,001	23,666,298	589,732,468	583,215,700	568,336,067	0.49	18,246,413	16,542,072	9,193,031	5,599,681
CT105	36,584,577	27,055,391	22,126,164	572,504,484	566,232,843	551,836,328	0.49	14,166,590	12,992,306	6,783,138	4,846,556

01_VirReport folder content:

For each sample:

- **assembly**: results associated with de novo assembly
- **blastn**: megablast results (NCBI NT or viral database PVirDB)
- **blastx**: blastx results against NR
- **tblastn**: tblastn results against viral database PVirDB
- **alignments**: alignment against top reference hit and associated statistic derivation
- **Summary**

Sample	Species	sacc	naccs	length	slen	cov	av-pident	stitle	cumulative_contig_len	Mean read depth	Dedup read count	Dup %	FPKM	PCT_5X	PCT_10X	contamination_flag
MT001	Citrus exocortis viroid	EU564170	1	190	372	0.51	100	Citrus exocortis viroid isolate	219	244.4	931817	36	63367	0.99	0.99	FALSE
MT002	Prunus necrotic ringspot virus RNA1	MF198441	9	488	3332	0.15	99.41	Prunus necrotic ringspot viru	507	9.37	1484	13	23	0.55	0.32	FALSE
MT002	Prunus necrotic ringspot virus RNA2	MH727228	6	331	2591	0.12	100	Prunus necrotic ringspot viru	323	10.11	1244	13	25	0.51	0.31	FALSE
MT002	Prunus necrotic ringspot virus RNA3	MH427285	10	691	1929	0.36	99.3	Prunus necrotic ringspot viru	716	117.41	13039	13	357	0.93	0.93	FALSE
MT003	Citrus dwarfing viroid	HQ219183	1	45	291	0.15	100	Citrus dwarfing viroid compl	46	224.5	17923	33	2111	0.97	0.96	FALSE
MT003	Citrus tristeza virus	EU857538	98	12881	19151	0.66	99.19	Citrus tristeza virus strain SP	13187	225.58	1301993	34	2479	1	1	FALSE
MT003	Citrus vein enation virus	MN187040	1	49	3963	0.01	100	Citrus vein enation virus isol	50	1.93	545	35	3	0.12	0.02	NA
MT003	Hop stunt viroid	KY110716	1	277	302	0.92	100	Hop stunt viroid strain R1409	280	201.07	8870	34	1001	0.98	0.98	FALSE
MT005	Raspberry bushy dwarf virus RNA1	MZ202350	50	3853	3448	0.71	97.81	Raspberry bushy dwarf virus	4138	179.51	211961	48	930	0.98	0.96	FALSE
MT005	Raspberry bushy dwarf virus RNA2	MZ202351	6	443	2230	0.2	99.78	Raspberry bushy dwarf virus	455	199.37	111146	48	1195	0.98	0.97	FALSE
MT005	Rubus yellow net virus	MN245240	13	1320	7844	0.17	90.56	Rubus yellow net virus strair	1447	95.95	94319	48	384	0.57	0.54	FALSE
MT006	Raspberry bushy dwarf virus RNA1	MZ202350	45	2828	3448	0.52	98.37	Raspberry bushy dwarf virus	3061	84.11	30585	42	127	0.89	0.8	FALSE
MT006	Raspberry bushy dwarf virus RNA2	MH800010	2	97	2218	0.04	100	Raspberry bushy dwarf virus	99	117.66	24109	42	246	0.92	0.87	FALSE
MT006	Rubus yellow net virus	MN245240	19	1338	7844	0.14	94.04	Rubus yellow net virus strair	1617	37.79	25053	42	76	0.48	0.4	FALSE
MT007	Citrus dwarfing viroid	EU934020	3	211	293	0.72	100	Citrus dwarfing viroid isolate	224	182.25	6274	24	759	0.95	0.93	FALSE
MT007	Citrus virus A RNA1	MK884372	40	2372	6889	0.35	97.77	Citrus virus A isolate TX NO :	2477	89.47	45671	24	247	0.85	0.76	FALSE
MT007	Citrus virus A RNA2	MZ463040	9	582	2686	0.22	98.19	Citrus virus A isolate B175 se	610	71.13	10413	24	137	0.81	0.73	FALSE

Definitions of terms used in summary report:

- **sacc** Accession number of best homology match recovered
- **av-pident** Average per cent identity of all de novo assembled contigs to the same top reference hit
- **Mean read depth** The mean coverage in bases to the genome/sequence of the best homology match
- **Dedup read count** Read counts after PCR duplicates sharing UMIs are collapsed
- **Dup %** Duplication rate detected using UMIs
- **FPKM:** Fragments Per Kilobase of transcript, per Million mapped reads is a normalised unit of transcript expression. It scales by transcript length to compensate for the fact that most RNA-seq protocols will generate more sequencing reads from longer RNA molecules. The formula is: [deduplicated read count x 10^3 x 10^6]/[total quality filtered reads x genome length]
- **% bases 5X** The fraction of bases that attained at least 5X sequence coverage
- **% bases 10X** The fraction of bases that attained at least 10X sequence coverage
- **Contamination flag**.

Appendix 34 - Policy paper for adoption of HTS for bacteria in PEQ

Natasha Brohier, Fiona Constable, Brendan Rodoni and Rachel Mann

Introduction

Bacterial pathogens can significantly reduce yields of economically important crops (Doddaraju et al., 2019; Loiseau et al., 2019; Malandraki et al., 2015). The increase in global trade has accelerated the movement of organisms from country to country including plant pathogens (Bulman et al., 2018). Diagnosis of these pathogens before entering a new area can greatly reduce the risk of spreading the disease (Bulman et al., 2018; Loiseau et al., 2019). Hence, detection protocols within quarantine areas are essential.

Traditionally, detection of bacterial pathogens has been based on culturing methods, with morphology and biochemical reactions the basis of identification (Gupta et al., 2019). Non-culturable forms of detection and identification have included serology including immunoassays; genotyping which includes amplified fragment length polymorphism (AFLP), PCR-16S rRNA- restriction fragment length polymorphism (RFLP); and amplification methods such as loop-mediated isothermal amplification (LAMP), recombinase polymerase amplification (RPA), PCR coupled with capillary electrophoresis (PCR-CE), and real time PCR (RT-PCR or qPCR) (Doddaraju et al., 2019; Loiseau et al., 2019; Malandraki et al., 2015; Murray et al., 2017; Ordóñez et al., 2019). Some of these protocols can detect pathogen titres as low as 0.1 ng of bacterial genomic/total DNA per reaction (Peňázová et al., 2020). However, running individual tests for the ever-increasing list of both exotic and endemic bacterial pathogens is time consuming, laborious and costly, especially when many plant samples require screening (Anderson et al., 2004; Stukenbrock et al., 2008).

The use of high throughput sequencing (HTS) could provide a rapid and robust approach for detection of multiple bacterial pathogens in a single plant sample in post entry quarantine (PEQ) facilities. The multiplexing capabilities permits sequencing of more than a thousand samples in one run especially when using higher sequencing output platforms such as the Illumina NovaSeq (up to 6 TB per run). Instead of individual tests for each pathogen, with the right technique, HTS would only require one sequencing run to detect all of them. Furthermore, the best outcome from HTS sequencing is its data-rich output providing not only detection, but high taxonomic resolution and origin.

A key component of HTS is the choice of technique used in preparation for sequencing. Some approaches may provide greater sensitivity compared to other bacterial detection methods, via bacterial cell enrichment (Frampton et al., 2018) or through target sequencing (Eichmeier et al., 2019). In this proposal, we discuss and compare multiple HTS techniques that have been used or could be used for plant bacteria diagnostics in PEQ.

Techniques

Whole genome sequencing (WGS) from cultured bacteria

The cultivation of bacteria on selective media is commonly used for identification using morphology and biochemical reactions (Gupta et al., 2019). Culturing bacteria provides an opportunity to isolate pathogen DNA from its host to enable whole genome sequencing (WGS) with the production of high quality, well characterised genomes that contain taxonomically and functionally informative sequence data. Though culturing may be time consuming, obtaining a culture, where possible, is considered 'gold standard' and will typically improve diagnosis accuracy. Culturing may also support be used to support culture collections, enable the accurate description of new species and can support research into the importance of novel isolated cultures in relation to biosecurity risk and disease management. Therefore, culturing bacteria and the biological data doing so imparts, will continue to be an important part of the diagnosticians toolkit.

Pros:

- Requires low sequencing depth to attain full genome.
- Full genome sequencing can provide high taxonomic resolution, evolution, and tracing information.
- Relatively cheap (low sequencing depth and minimal molecular preparations before sequencing).
- Reduced bioinformatic resources.
- Can identify individual isolates in case there's multiple similar species/genera within a sample.
- Does not require previous genomic data for protocol development.

Cons:

- Will miss unculturable and, in many cases, highly fastidious bacteria
- Bias towards bacteria that readily proliferate under laboratory conditions.
- Time consuming, especially for slow growing and fastidious bacteria.
- Need expertise to know potential bacteria in specific plants.

Metagenomics

Metagenomics is the sequencing of DNA from environmental samples to analyse microbial communities present in a sample (Ghosh et al., 2019). As sequencing technology improves, shotgun metagenomics has become an applicable option for population studies and diagnostics resulting in more microbial genomes being published (Escobar-Zepeda et al., 2015).

Metagenomic sequencing has the capability to identify known and unknown plant pathogens in a sample as total DNA samples are sequenced. This permits the identification of essential genes that may further clarify genus and species and even subspecies or pathovars of bacteria. Furthermore, due to the lack of amplification, metagenomics has the potential to quantify pathogen populations within a sample without bias.

Metagenomic sequencing often uses short read sequencing on platforms such as Illumina. Short reads range from 150 (Illumina HiSeq/MiSeq 2 x 75 bp output) to 600 (Illumina MiSeq 2 x 300 bp paired-end mode) nucleotides and are *denovo* assembled by bioinformatics pipelines into longer contiguous sequences (contigs). *Denovo* assembly can be a bottleneck in the bioinformatic pipeline as it needs high computational resources and time to complete. Furthermore, fragmented draft genome assembly is highly likely, even when the pathogen is at high titre, as there are repetitive and/or monotonous regions within the genome. This can mean large structural variations and genomic rearrangements are unable to be identified and annotation accuracy is impacted.

Long read sequencing, such as Oxford Nanopore Technology (ONT), has also contributed greatly to metagenomics as it enables the assembly of full genomes, or at least the creation of large contigs, thereby linking informative genes and reducing the complexity of downstream data analysis in comparison to short read sequencing (Bronzato Badial et al., 2018; Chalupowicz et al., 2019; Hu et al., 2019).

One of the main limitations of metagenomics is the titre of a bacterium may be too low to detect, because the sequence reads generated are biased towards host genome and other organisms due to their abundance. Consequently, metagenomics is not as sensitive as targeted sequencing.

Another limitation with the use of long read sequencing in metagenomics is the possibility of chimeric reads (1.7%) (Xu et al., 2018). The generation of chimeric reads in sequencing of complex microbial populations with limited depth may incorrectly link important genes used in species differentiation to the incorrect organism and imply the presence of non-existent species. Chimeric reads are the predominant source of cross barcode assignment errors in multiplex long read sequencing. However, promising new chemistries have been released

for ONT which has shown to cause less errors than its previous counterparts.

Pros

- Unbiased identification of bacterial population within a sample.
- Does not require previous genomic data for protocol development.
- Relatively low cost for ONT and Illumina NovaSeq sequencing. Other platforms such as MiSeq can increase the sequencing costs per GB considerably.

Cons:

- Low sensitivity.
- Need to sequence deep which may or may not capture low titre pathogens.
- High bioinformatic resources.
- Unable to differentiate sequences belonging to other bacteria resulting in misassemblies.

Amplicon sequencing

Amplicon sequencing is the isolation and amplification of 'barcodes' using PCR primers on a mixed DNA sample from any origin which will subsequently be sequenced by HTS to identify species composition of the sample. The amplification of these barcodes increases sensitivity and will aid in identifying individual species from complex samples.

Amplicon sequencing can be a swift and cheap method of detection and identification of bacterial species. Depending on targeted regions, such as multilocus sequence analysis (MLSA) and multilocus sequence typing (MLST) genes, amplicon sequences could give high taxonomic resolution on a bacterial pathogen (Ruiz-Padilla et al., 2020; Wicker et al., 2012; Yuan et al., 2010). Unlike MLST genes that have a pre-set of targeted genes, MLSA has no allocated amount of genes and normally uses taxonomically informative genes that occur across and can differentiate diverse species such as gyrase B (*gyrB*), 60 kDa chaperonin (*cpn60*), β-subunit of DNA polymerase (*rpoB*), *recN* and *dnaX* (Barret et al., 2015; Links et al., 2014; Tian et al., 2016; Vos et al., 2012; Zeigler, 2003).

Using MLSA/MLST genes requires each gene to be sequenced individually leading to multiple libraries per genera/species per sample. Accumulatively, more resources and labour costs would be required when testing for more than one bacterial pathogen. Multiplexing PCR primers could alleviate this problem. However, optimization is required as multiple primers are pooled together and can increase the risk of primer dimers and disproportionate amplification of genes due to primer bias (Koh et al., 2014; Nishiki et al., 2018; Onda et al., 2018).

Alternatively, generic 'barcodes', such as the 16S rRNA gene can be used. Within the past three decades, this region has been one of the main barcoding regions used for bacterial species identification. The use of the 16S region has the potential to identify the majority of bacteria present and detect known and possibly unknown pathogens in a sample. However, identifying conserved primer binding regions in these genes across diverse bacterial taxa can be difficult. The primers currently used to amplify many of these barcode regions create bias, only amplifying a reduced taxonomic subset of bacteria at the high taxon level such as genera and families, rather than a diverse microbial population. SNPs may also be present in primer binding sites that result in subsets of bacterial species or strains being inadvertently excluded. To solve this issue, methods incorporating additional barcoding regions within 16S have been developed and may even be combined with other techniques (e.g. metagenomics) (Barb et al., 2016).

Sequencing platforms influence how amplicon sequencing is applied to a sample. When using short read sequencers, amplicon sizes need to be limited to shorter fragments often only as large as ~300nt. This can lead to identification only to genus level due to high similarity between the amplified gene region from closely related species. Long read sequencers can solve this problem and, for example, have been used to sequence near full-

length 16S rDNA for a population of bacteria (Mitsuhashi et al., 2017). However, the high error rate of long sequence reads from ONT could result in misclassification of bacterial species present in a sample (Goodwin et al., 2016). Furthermore, the full length 16S rRNA gene of many plant pathogenic bacteria does not possess enough genetic diversity to differentiate them from closely related bacteria (e.g. *Erwinia amylovora* and *Erwinia pyrifoliae* cannot be distinguished using their 16S rRNA region).

Pros:

- Relatively low cost.
- Relatively fast when detecting single pathogens.
- Highly sensitive. Can detect low titre pathogens.
- Able to give high taxonomic resolution.

Cons:

- Requires the availability of genomic data for target identification
- Producing new primers can be labour intensive, especially with newly discovered bacterial pathogens.
- Can be laborious when detecting multiple bacterial genera.
- Primers can disfavour or exclude organisms as primer binding regions are not 100% conserved, with the consequence that some taxa are systematically ignored.
- General amplicons targets may only identify to genus levels due to high sequence similarities to closely related species.

Hybridization capture HTS

Hybridization capture HTS uses biotinylated oligonucleotide probes to capture targeted DNA regions within a genome. The use of hybridization capture prior to HTS enables capture of target nucleic acids from a broader pool, essentially enriching the sample for a specific target and reducing the amount of sequencing required. Hybridization capture HTS is normally performed after library preparation enabling indexed libraries to be multiplexed before hybridization capture. For bacterial diagnostic purposes, these probes could either target specific bacterial pathogens for direct identification (Gasc et al., 2017) or target a barcoding region such as the 16s rRNA (Gasc et al., 2018).

Hybridization capture HTS has been used for multiple bacteria detection/identification and as a diagnostic tool. Hybridization capture has been used to sequence full bacterial genomes of species such as *Candidatus* Liberibacter asiaticus (Cai et al., 2019), *Mycobacterium tuberculosis* (Nimmo et al., 2017), *Chlamydia psittaci* (Branley et al., 2016) and *Chlamydia pecorum* (Bachmann et al., 2015), without the need for culturing, to analyse diversity of bacterial strains. Hybridization capture HTS has also been used for 16S sequencing to avoid common problems that have risen from primers as described above. These 16S studies found that hybridization capture HTS reduces the bias in 16S rRNA amplicon sequencing of microbial communities (Cariou et al., 2018; Rassoulian Barrett et al., 2020). Capture of the 16S region has also been designed to specifically target the *Ca*. phytoplasma spp. (Trivellone et al., 2022), giving better resolution than amplicon Sanger sequencing.

Capturing full genomes using hybridization capture HTS has been used for viruses, with some research targeting the whole genomes of up to 356 viral species (Metsky et al., 2019). However, application to full bacterial genomes can be costly due to their large genome size compared to many viruses. Hence, to target and identify multiple genera of bacterial pathogens in a sample, only selected gene regions within the genome should be captured and could include the 16S region (Trivellone et al., 2022) as well as other taxonomically informative genes. For instance, using five vertebrate model species, >500 loci in highly conserved regions were developed (Lemmon et al., 2012) which could aid in population genomics and phylogenetic studies. Also referred to as Anchored Hybrid Enrichment, hybridization capture HTS has also been used to successfully investigate loci

diversity in an entire superfamily of insects (Membracoidea) (Dietrich et al., 2017).

Pros

- Unlimited target capture regions.
- Easy to add more probes (to target more genes/genomes) to original probe sets making it a flexible detection/identification system.
- Able to give high taxonomic resolution.
- Can detect multiple bacterial genera and species in one test.
- Can run on multiplexed samples to reduce total hybridization capture HTS reactions.

Cons:

- Can be costly if targeting full genomes.
- Potential to not be as sensitive as amplicon sequencing and qPCR.
- Requires the availability of the genomic data for target identification.

Diagnostic Considerations for Pathogenic Bacteria Diagnostics

HTS diagnostics may show great promise as a one size fits all solution for bacterial detection, but bacterial pathogens have diverse phenotypic and genotypic characteristics that can complicate the application of a single technology (Brbić et al., 2016; Oda et al., 2002; Pradhan et al., 2019). Complexity maybe associated with variability in morphology, niche adaptations, and life cycles of bacterial pathogens within their hosts that can obfuscate their detection. Accurate identification linked to phenotypes may have genetic components that are unknown or unclear. Due to this variability, no single identification protocol has been established as a universal form for diagnostics for phytopathogenic bacteria *in planta*.

Among diagnostic considerations for detection of phytopathogenic bacteria *in planta* is the type of tissue different pathogens inhabit and when sampling should occur throughout the host plant's growing season. Sampling strategies that target the appropriate plant tissues are integral to detecting plant pathogenic bacteria. Optimal tissue type can vary from pathogen to pathogen and across the growing season for the same pathogen. For instance, phytoplasma and *Liberibacter* species are restricted to the sieve tubes in the phloem and often follow a 'source to sink' pattern of distribution within their host plant across the growing season (Tran-Nguyen et al., 2007). In large host plants they can be distributed unevenly and might only be detected at specific times of the year (Constable et al., 2003) (Constable et al., 2003). For this reason, the appropriate tissue for sampling many hosts will vary from planting to maturity. *Clavibacter michiganensis* subsp. *michiganensis* are known to be localized in the leaflets of tomatoes in the early stages of infection (Gleason et al., 1993). *Xyllela fastidiosa* and *Xanthomonas albilineans*, bacterial pathogens of multiple important crop species are located solely within the xylem of a plant (Pieretti et al., 2009). The distribution and concentration of *Xylella fastidiosa* within plants can be variable and is dependent upon plant species type, length of infection, seasonal and environmental factors. To maximize the likelihood of detection, it is recommended that sampling should be performed during the period of active growth of the plants (Hopkins, 1981), usually from late spring to autumn in temperate regions but potentially all year round for tropical hosts such as coffee (EPPO, 2018). Hence, a range of plant tissues would need to be sampled for testing of multiple pathogens in any given host, which may dilute the titre of pathogens that occur in specific tissues with a risk of false negative results. Another problem complicating accurate detection is the titre of bacterial pathogens within the plant. A number of bacterial pathogens are notorious for their low abundance within a plant making them hard to detect (Gleason et al., 1993). Furthermore, some pathogens persist in a latent phase *in planta* (or epiphytically), not expressing pathogenicity factors and inducing symptoms until conditions are desirable (Duarte et al., 2004; Ou et al., 2019). To overcome

this, enrichment, being the increase of target population in comparison to other populations (and host), may be necessary.

Understanding the life cycle of bacterial pathogens to inform timing of sampling is also a key consideration for optimising the probability of detecting plant pathogenic bacteria *in planta*. Pathogens, such as *Erwinia amylovora* and *Pseudomonas syringae* pv. *actinidiae* biovar 3, overwinter in cankers with varying survival rates (Do et al., 2016; Van Der Zwet et al., 1999). In spring, when conditions are favourable, populations of these pathogens increase exponentially, often becoming visual to the naked eye as droplets of ooze. In potatoes, blackleg diseases are caused by *Pectobacterium* species express in cool wet climates due to increased pathogen population sizes (Duarte et al., 2004). In mulberry (*Morus* L.) cultivars, the diversity and community composition of bacterial species were more influenced by season followed by host cultivar (Ou et al., 2019). Hence, a selected season or climate may be able to induce their growth making them easier to detect (Schaad et al., 2004).

Increasing the probability of detecting pathogenic bacteria during sampling is crucial to detect low titre and/or latent or tissue type specific bacteria. Further pathogen cell enrichment for detection can potentially be done before the DNA extraction step (Frampton et al., 2018; Ou et al., 2019; Tran-Nguyen et al., 2007). However, these protocols can be labour intensive and would require more specialized equipment. DNA extraction should be conducted with the addition of controls to confirm successful nucleic acid extraction and identify the presence of inhibitors for downstream library preparation protocols (Bulman et al., 2018). Examples of quality check controls used during DNA extraction include using known infected (positive) and non-infected (negative) plant tissue controls and negative (water) controls (Lear et al., 2018; Pollock et al., 2018). All extracts should be assessed for quality by quantifying extracted nucleic acid and running a quality control PCR to ensure there are no PCR inhibitors (e.g. COX qPCR for plant extracted samples).

Based on the needs of PEQ, hybridization capture HTS is recommended as the best option as it was designed to overcome the limitations of using metabarcoding primers (which can introduce bias and low clarity) and metagenomics (overabundance of host genome).

As mentioned above, hybridization capture HTS has been used on hundreds of loci to target all species within a superfamily and could potentially be used for bacteria (Dietrich et al., 2017). However, for the purpose of PEQ, only exotic pathogenic bacteria need priority for detection and identification. With this in mind, a proof-of-concept hybridization capture HTS method has been developed that targets both unique and MLSA/MLST regions comprised of taxonomically informative genes for bacterial pathogens on the Australian National Priority Plant Pests (NPPP) list (2019). These include unique loci for *Erwinia amylovora* and *Xanthomonas citri* subsp *citri*, MLSA regions for *Candidatus* Liberibacter asiaticus, africanus, americanus and solanacearum, and *Candidatus* phytoplasma 16SrI, and finally both MLST and unique regions for *Xylella taiwanensis* and *fastidiosa* subsp. High specificity for all trialled pathogen isolates has been demonstrated. As MLSA/MLST genes were targeted, off-target capture was observed with closely related species/genera on conservative regions within a gene. This was similar for all closely related species/genus identified where capture would produce incomplete genes, no matter the sequencing depth. All off-target capture can be easily removed through bioinformatic analysis, however, probe design should aim to reduce probability of off-targets as it may reduce sensitivity of the hybridization capture reaction to the real targets.

The proof of concept was in line with some published literature showing that detection limits of the hybridization capture HTS method are on par with qPCR results (Trivellone et al., 2022). However, there are multiple hybridization capture technologies using a variety of bait types that could affect sensitivity. Among those used are single stranded RNA probes (SureSelect, Agilent Technologies), single stranded DNA probes (xGEN, Integrated DNA Technologies), double stranded DNA (Twist Target Enrichment, Twist Bioscience) and double stranded RNA (QuarXeq Human All Exon Probes 1.0, Dynegen Bioscience). Multiple studies have been conducted evaluating and comparing these hybridization capture baits for whole exome capture (Iadarola et al., 2020; Zhou

et al., 2021). Among its findings were that longer reads provide better genotypability, RNA baits have stronger binding power than DNA baits, and both DNA and RNA baits have their own positive characteristics in regard to high GC content and AT dropouts (Iadarola et al., 2020; Zhou et al., 2021). In regard to sensitivity, trialling multiple probe baits may be necessary to assess optimal sensitivity and specificity for hybridization capture HTS.

Conclusion

There are several HTS approaches that can be used to detect and identify bacterial pathogens in a PEQ setting. Metagenomic HTS could be applied when discovery of an unknown pathogen is required. For culturable bacteria, isolation and WGS of pure colonies may be beneficial for confirmatory testing and may be more cost effective when a single culturable bacteria is the target. However, hybridization capture HTS is a targeted approach which overall provides PEQ with the greatest benefits:

- With all probes for every genome in one reaction, hybridization capture HTS is ideal for broad pathogen screening.
- It can be developed for the specific targets that require routine testing in PEQ, reducing the need for policy considerations around the risk of application because novel or unknown pathogens are unlikely to be detected.
- It does not require bacterial isolation and culturing and can be used to detect culturable and non-culturable bacteria.
- It is highly sensitive.
- Interpretation is simple because the analysis pipeline for unique and MLSA/MLST regions is straight forward.
- It is a faster solution than Sanger sequencing multiple amplicons.

The limitation of hybridization capture HTS is the need for genomic information or sequence data of taxonomically informative unique and MLSA/MLST regions, which may not be available for some targets. But this data should be simple to produce in collaboration with international colleagues.

Development, from design to application, would also be faster once quality control and assurance of accurate, reproducible, and consistent results are set in place.

Reference

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., & Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution, 19*(10), 535-544. doi:https://doi.org/10.1016/j.tree.2004.07.021

Bachmann, N. L., Sullivan, M. J., Jelocnik, M., Myers, G. S. A., Timms, P., & Polkinghorne, A. (2015). Culture-Independent Genome Sequencing of Clinical Samples Reveals an Unexpected Heterogeneity of Infections by Chlamydia pecorum. *Journal of Clinical Microbiology, 53*(5), 1573-1581. doi:doi:10.1128/JCM.03534-14

Barb, J. J., Oler, A. J., Kim, H.-S., Chalmers, N., Wallen, G. R., Cashion, A., . . . Ames, N. J. (2016). Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PLOS ONE, 11*(2), e0148047. doi:10.1371/journal.pone.0148047

Barret, M., Briand, M., Bonneau, S., Préveaux, A., Valière, S., Bouchez, O., . . . Jacquesa, M. A. (2015). Emergence shapes the structure of the seed microbiota. *Appl Environ Microbiol, 81*(4), 1257-1266. doi:10.1128/aem.03722-14

Branley, J., Bachmann, N. L., Jelocnik, M., Myers, G. S. A., & Polkinghorne, A. (2016). Australian human and parrot Chlamydia psittaci strains cluster within the highly virulent 6BC clade of this important zoonotic pathogen. *Scientific Reports, 6*(1), 30019. doi:10.1038/srep30019

Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., & Supek, F. (2016). The landscape of microbial phenotypic traits and associated genes. *Nucleic acids research, 44*(21), 10074-10090. doi:10.1093/nar/gkw964

Bronzato Badial, A., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., & King, J. (2018). Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. *Plant Disease, 102*(8), 1648-1652. doi:10.1094/pdis-04-17-0488-re

Bulman, S. R., McDougal, R. L., Hill, K., & Lear, G. (2018). Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. *Australasian Plant Pathology, 47*(5), 467-474. doi:10.1007/s13313-018-0579-3

Cai, W., Nunziata, S., Rascoe, J., & Stulberg, M. J. (2019). SureSelect targeted enrichment, a new cost effective method for the whole genome sequencing of Candidatus Liberibacter asiaticus. *Scientific Reports, 9*(1), 18962. doi:10.1038/s41598-019-55144-4

Cariou, M., Ribière, C., Morlière, S., Gauthier, J.-P., Simon, J.-C., Peyret, P., & Charlat, S. (2018). Comparing 16S rDNA amplicon sequencing and hybridization capture for pea aphid microbiota diversity analysis. *BMC research notes, 11*(1), 461-461. doi:10.1186/s13104-018-3559-3

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., . . . Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology, 68*(2), 229-238. doi:https://doi.org/10.1111/ppa.12957

Constable, F. E., Gibb, K. S., & Symons, R. H. (2003). Seasonal distribution of phytoplasmas in Australian grapevines. *Plant Pathology, 52*(3), 267-276. doi:https://doi.org/10.1046/j.1365-3059.2003.00849.x

Dietrich, C. H., Allen, J. M., Lemmon, A. R., Lemmon, E. M., Takiya, D. M., Evangelista, O., . . . Johnson, K. P. (2017). Anchored Hybrid Enrichment-Based Phylogenomics of Leafhoppers and Treehoppers (Hemiptera: Cicadomorpha: Membracoidea). *Insect Systematics and Diversity, 1*(1), 57-72. doi:10.1093/isd/ixx003

Do, K. S., Chung, B. N., & Joa, J. H. (2016). D-PSA-K: A Model for Estimating the Accumulated Potential Damage on Kiwifruit Canes Caused by Bacterial Canker during the Growing and Overwintering Seasons. *The plant pathology journal, 32*(6), 537-544. doi:10.5423/ppj.Oa.03.2016.0068

Doddaraju, P., Kumar, P., Gunnaiah, R., Gowda, A. A., Lokesh, V., Pujer, P., & Manjunatha, G. (2019). Reliable and early diagnosis of bacterial blight in pomegranate caused by Xanthomonas axonopodis pv. punicae using sensitive PCR techniques. *Scientific Reports, 9*(1), 10097. doi:10.1038/s41598-019-46588-9

Duarte, V., de Boer, S. H., Ward, L. J., & de Oliveira, A. M. (2004). Characterization of atypical Erwinia carotovora strains causing blackleg of potato in Brazil. *J Appl Microbiol, 96*(3), 535-545. doi:10.1111/j.1365-2672.2004.02173.x

Eichmeier, A., Kiss, T., Necas, T., Penazova, E., Tekielska, D., Bohunicka, M., . . . Baldrian, P. (2019). High-Throughput Sequencing Analysis of the Bacterial Community in Stone Fruit Phloem Tissues Infected by “Candidatus Phytoplasma prunorum”. *Microbial Ecology, 77*(3), 664-675. doi:10.1007/s00248-018-1250-9

Escobar-Zepeda, A., Vera-Ponce de León, A., & Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics, 6*. doi:10.3389/fgene.2015.00348

Frampton, R. A., Thompson, S. M., Kalamorz, F., David, C., Addison, S. M., & Smith, G. R. (2018). Draft Genome Sequence of a “<i>Candidatus</i> Liberibacter europaeus” Strain Assembled from Broom Psyllids (Arytainilla spartiophila) from New Zealand. *Genome Announcements, 6*(20), e00430-00418. doi:doi:10.1128/genomeA.00430-18

Gasc, C., Constantin, A., Jaziri, F., & Peyret, P. (2017). OCaPPI-Db: an oligonucleotide probe database for pathogen identification through hybridization capture. *Database, 2017*. doi:10.1093/database/baw172

Gasc, C., & Peyret, P. (2018). Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome, 6*(1), 61-61. doi:10.1186/s40168-018-0442-3

Ghosh, A., Mehta, A., & Khan, A. M. (2019). Metagenomic Analysis and its Applications. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 184-193).

Gleason, M., Gitaitis, R., & Ricker, M. (1993). Recent Progress in Understanding and Controlling Bacterial Canker of Tomato in Eastern North America. *Plant Disease, 77*, 1069-1076. doi:10.1094/PD-77-1069

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351. doi:10.1038/nrg.2016.49

Gupta, S., Mortensen, M. S., Schjørring, S., Trivedi, U., Vestergaard, G., Stokholm, J., . . . Sørensen, S. J. (2019). Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Commun Biol, 2*, 291. doi:10.1038/s42003-019-0540-1

Hopkins, D. (1981). Seasonal concentration of the Pierce's disease bacterium in grapevine stems, petioles, and leaf veins.

Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., & Schwessinger, B. (2019). Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. *Phytobiomes Journal, 3*(2), 92-101. doi:10.1094/pbiomes-01-19-0004-r

Iadarola, B., Xumerle, L., Lavezzari, D., Paterno, M., Marcolungo, L., Beltrami, C., . . . Delledonne, M. (2020). Shedding light on dark genes: enhanced targeted resequencing by optimizing the combination of enrichment technology and DNA fragment length. *Sci Rep, 10*(1), 9424. doi:10.1038/s41598-020-66331-z

Koh, H. S., Kim, G. H., Lee, Y. S., Koh, Y. J., & Jung, J. S. (2014). Molecular Characteristics of Pseudomonas syringae pv. actinidiae Strains Isolated in Korea and a Multiplex PCR Assay for Haplotype Differentiation. *The plant pathology journal, 30*(1), 96-101. doi:10.5423/ppj.Nt.09.2013.0095

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., . . . Hermans, S. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology, 42*(1), 10-50A.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biology, 61*(5), 727-744. doi:10.1093/sysbio/sys049

Links, M. G., Demeke, T., Gräfenhan, T., Hill, J. E., Hemmingsen, S. M., & Dumonceaux, T. J. (2014). Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. *New Phytol, 202*(2), 542-553. doi:10.1111/nph.12693

Loiseau, M., Plaire, M., Renaudin, I., Taylor, R., Fujikawa, T., Griffin, R., . . . Renvoisé, J.-P. (2019). Detecting Spiroplasma citri: a comparison of PCR methods to be used for quarantine diagnostics. *European Journal of Plant Pathology, 155*(1), 71-80. doi:10.1007/s10658-019-01750-x

Malandraki, I., Varveri, C., Olmos, A., & Vassilakos, N. (2015). One-step multiplex quantitative RT-PCR for the simultaneous detection of viroids and phytoplasmas of pome fruit trees. *Journal of Virological Methods, 213*, 12-17. doi:https://doi.org/10.1016/j.jviromet.2014.11.010

Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., . . . Viral Hemorrhagic Fever, C. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology, 37*(2), 160-168. doi:10.1038/s41587-018-0006-x

Mitsuhashi, S., Kryukov, K., Nakagawa, S., Takeuchi, J. S., Shiraishi, Y., Asano, K., & Imanishi, T. (2017). A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Scientific Reports, 7*(1), 5657. doi:10.1038/s41598-017-05772-5

Murray, T. D., Schroeder, B. K., Schneider, W. L., Luster, D. G., Sechler, A., Rogers, E. E., & Subbotin, S. A. (2017). Rathayibacter toxicus, Other Rathayibacter Species Inducing Bacterial Head Blight of Grasses, and the Potential for Livestock Poisonings. *Phytopathology®, 107*(7), 804-815. doi:10.1094/PHYTO-02-17-0047-RVW

Nimmo, C., Doyle, R., Burgess, C., Williams, R., Gorton, R., McHugh, T. D., . . . Breuer, J. (2017). Rapid identification of a Mycobacterium tuberculosis full genetic drug resistance profile through whole genome sequencing directly from sputum. *International Journal of Infectious Diseases, 62*, 44-46. doi:https://doi.org/10.1016/j.ijid.2017.07.007

Nishiki, I., Minami, T., Murakami, A., Hoai, T. D., & Fujiwara, A. (2018). Multilocus sequence analysis of Vibrionaceae isolated from farmed amberjack and the development of a multiplex PCR assay for the detection of pathogenic species. *Journal of Fish Diseases, 41*(8), 1295-1301. doi:https://doi.org/10.1111/jfd.12823

Oda, Y., Wanders, W., Huisman Louis, A., Meijer Wim, G., Gottschal Jan, C., & Forney Larry, J. (2002). Genotypic and Phenotypic Diversity within Species of Purple Nonsulfur Bacteria Isolated from Aquatic Sediments. *Applied and Environmental Microbiology, 68*(7), 3467-3477. doi:10.1128/AEM.68.7.3467-3477.2002

Onda, Y., Takahagi, K., Shimizu, M., Inoue, K., & Mochida, K. (2018). Multiplex PCR Targeted Amplicon Sequencing (MTA-Seq): Simple, Flexible, and Versatile SNP Genotyping by Highly Multiplexed PCR Amplicon Sequencing. *Frontiers in Plant Science, 9*. doi:10.3389/fpls.2018.00201

Ordóñez, N., Salacinas, M., Mendes, O., Seidl, M. F., Meijer, H. J. G., Schoen, C. D., & Kema, G. H. J. (2019). A loop-mediated isothermal amplification (LAMP) assay based on unique markers derived from genotyping by sequencing data for rapid in planta diagnosis of Panama disease caused by Tropical Race 4 in banana. *Plant Pathology, 68*(9), 1682-1693. doi:https://doi.org/10.1111/ppa.13093

Ou, T., Xu, W.-f., Wang, F., Strobel, G., Zhou, Z.-y., Xiang, Z.-h., . . . Xie, J. (2019). A Microbiome Study Reveals Seasonal Variation in Endophytic Bacteria Among different Mulberry Cultivars. *Computational and Structural Biotechnology Journal, 17*, 1091-1100. doi:https://doi.org/10.1016/j.csbj.2019.07.018

Peňázová, E., Dvořák, M., Ragasová, L., Kiss, T., Pečenka, J., Čechová, J., & Eichmeier, A. (2020). Multiplex real-time PCR for the detection of Clavibacter michiganensis subsp. michiganensis, Pseudomonas syringae pv. tomato and pathogenic Xanthomonas species on tomato plants. *PLOS ONE, 15*(1), e0227559. doi:10.1371/journal.pone.0227559

Pieretti, I., Royer, M., Barbe, V., Carrere, S., Koebnik, R., Cociancich, S., . . . Rott, P. (2009). The complete genome sequence of Xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. *BMC Genomics, 10*(1), 616. doi:10.1186/1471-2164-10-616

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. *Applied and Environmental Microbiology, 84*(7), e02627-02617.

Pradhan, P., & Tamang, J. P. (2019). Phenotypic and Genotypic Identification of Bacteria Isolated From Traditionally Prepared Dry Starters of the Eastern Himalayas. *Frontiers in Microbiology, 10*. doi:10.3389/fmicb.2019.02526

Rassoulian Barrett, S., Hoffman, N. G., Rosenthal, C., Bryan, A., Marshall, D. A., Lieberman, J., . . . Salipante, S. J. (2020). Sensitive Identification of Bacterial DNA in Clinical Specimens by Broad-Range 16S rRNA Gene Enrichment. *Journal of Clinical Microbiology, 58*(12), e01605-01620. doi:10.1128/JCM.01605-20

Ruiz-Padilla, A., Redondo, C., Asensio, A., Garita-Cambronero, J., Martínez, C., Pérez-Padilla, V., . . . Cubero, J. (2020). Assessment of Multilocus Sequence Analysis (MLSA) for Identification of Candidatus Liberibacter Solanacearum from Different Host Plants in Spain. *Microorganisms, 8*(9). doi:10.3390/microorganisms8091446

Schaad, N. W., Postnikova, E., Lacy, G., Fatmi, M. B., & Chang, C.-J. (2004). Xylella fastidiosa subspecies: X. fastidiosa subsp piercei, subsp. nov., X. fastidiosa subsp. multiplex subsp. nov., and X. fastidiosa subsp. pauca subsp. nov. *Systematic and Applied Microbiology, 27*(3), 290-300. doi:https://doi.org/10.1078/0723-2020-00263

Stukenbrock, E. H., & McDonald, B. A. (2008). The Origins of Plant Pathogens in Agro-Ecosystems. *Annual Review of Phytopathology, 46*(1), 75-100. doi:10.1146/annurev.phyto.010708.154114

Tian, Q., Zhao, W., Lu, S., Zhu, S., & Li, S. (2016). DNA Barcoding for Efficient Species- and Pathovar-Level Identification of the Quarantine Plant Pathogen Xanthomonas. *PLOS ONE, 11*(11), e0165995. doi:10.1371/journal.pone.0165995

Tran-Nguyen, L. T., & Gibb, K. S. (2007). Optimizing Phytoplasma DNA purification for genome analysis. *J Biomol Tech, 18*(2), 104-112.

Trivellone, V., Cao, Y., & Dietrich, C. H. (2022). Comparison of Traditional and Next-Generation Approaches for Uncovering Phytoplasma Diversity, with Discovery of New Groups, Subgroups and Potential Vectors. *Biology, 11*(7), 977. Retrieved from https://www.mdpi.com/2079-7737/11/7/977

Van Der Zwet, T., Beer, S. V., United States. Agricultural Research, S., & Cornell, U. (1999). *Fire blight--its nature, prevention, and control : a practical guide to integrated disease management* ([Slightly rev. June 1999]. ed. Vol. no.631). Washington, DC: U.S. Dept. of Agriculture.

Vos, M., Quince, C., Pijl, A. S., de Hollander, M., & Kowalchuk, G. A. (2012). A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. *PLOS ONE, 7*(2), e30600. doi:10.1371/journal.pone.0030600

Wicker, E., Lefeuvre, P., de Cambiaire, J. C., Lemaire, C., Poussier, S., & Prior, P. (2012). Contrasting recombination patterns and demographic histories of the plant pathogen Ralstonia solanacearum inferred from MLSA. *Isme j, 6*(5), 961-974. doi:10.1038/ismej.2011.160

Xu, Y., Lewandowski, K., Lumley, S., Pullan, S., Vipond, R., Carroll, M., . . . Crook, D. (2018). Detection of Viral Pathogens With Multiplex Nanopore MinION Sequencing: Be Careful With Cross-Talk. *Frontiers in Microbiology, 9*. doi:10.3389/fmicb.2018.02225

Yuan, X., Morano, L., Bromley, R., Spring-Pearson, S., Stouthamer, R., & Nunney, L. (2010). Multilocus Sequence Typing of Xylella fastidiosa Causing Pierce's Disease and Oleander Leaf Scorch in the United States. *Phytopathology®, 100*(6), 601-611. doi:10.1094/phyto-100-6-0601

Zeigler, D. R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International Journal of Systematic and Evolutionary Microbiology, 53*(6), 1893-1900. doi:https://doi.org/10.1099/ijs.0.02713-0

Zhou, J., Zhang, M., Li, X., Wang, Z., Pan, D., & Shi, Y. (2021). Performance comparison of four types of target enrichment baits for exome DNA sequencing. *Hereditas, 158*(1), 10. doi:10.1186/s41065-021-00171-3

Appendix 35 - Standard Operating Procedure for culturable bacteria using Illumina sequencing (end-to-end)

Natasha Brohier, Fiona Constable, Brendan Rodoni and Rachel Mann

INTRODUCTION

Culturing pathogenic bacteria for whole genome sequencing is a means to isolate pathogen DNA from host and environmental DNA. Sequencing pure cultures reduces the need for higher sequencing depth, bioinformatic bottlenecks and technical resources required for sequencing. Shallow short read sequencing (~300-500 MB) of genomic DNA extracted from cultured bacteria has shown to be sufficient for taxonomic identification (e.g. subspecies, pathovar or biovar). At this depth of sequencing, mean read depth would be higher than 100X meeting minimum requirements for novel species/taxon identification (Chun, 2014).

Problems that could prevent accurate identification of bacterial pathogens include:

- Contamination with other organisms during the isolation step.
- Presence of smaller fragments (index or primer dimers) during the library preparation step.
- Index hopping
- Incomplete/inaccurate reference genomes used for phylogenetic analysis.

To alleviate problems that may occur during the HTS pipeline:

- Sequence a minimum of 500 MB depth with 1 GB preferred in case of contamination.
    - Bioinformatic protocols provided below will aid in identifying contamination and select contigs required for phylogenetic analysis.
- Use known genomes as references; from current literature or from curated databases such as ezbiocoud. NCBI may not have updated species to their current taxonomy. Species from genera such as *Pseudomonas* and *Erwinia* are still being reclassified, hence need to be constantly updated.
- Follow support protocols and controls to ensure high quality sequencing libraries and output sequences for bioinformatic analysis.
- Steps to reduce index hopping should be taken whenever possible. This includes using unique dual indexes (UDIs) and using FAB reagents (Support protocol 2.2) on pooled libraries.

Workflow for high throughput protocols for detection of key culturable plant pathogenic bacteria are described below (Figure 35.1).

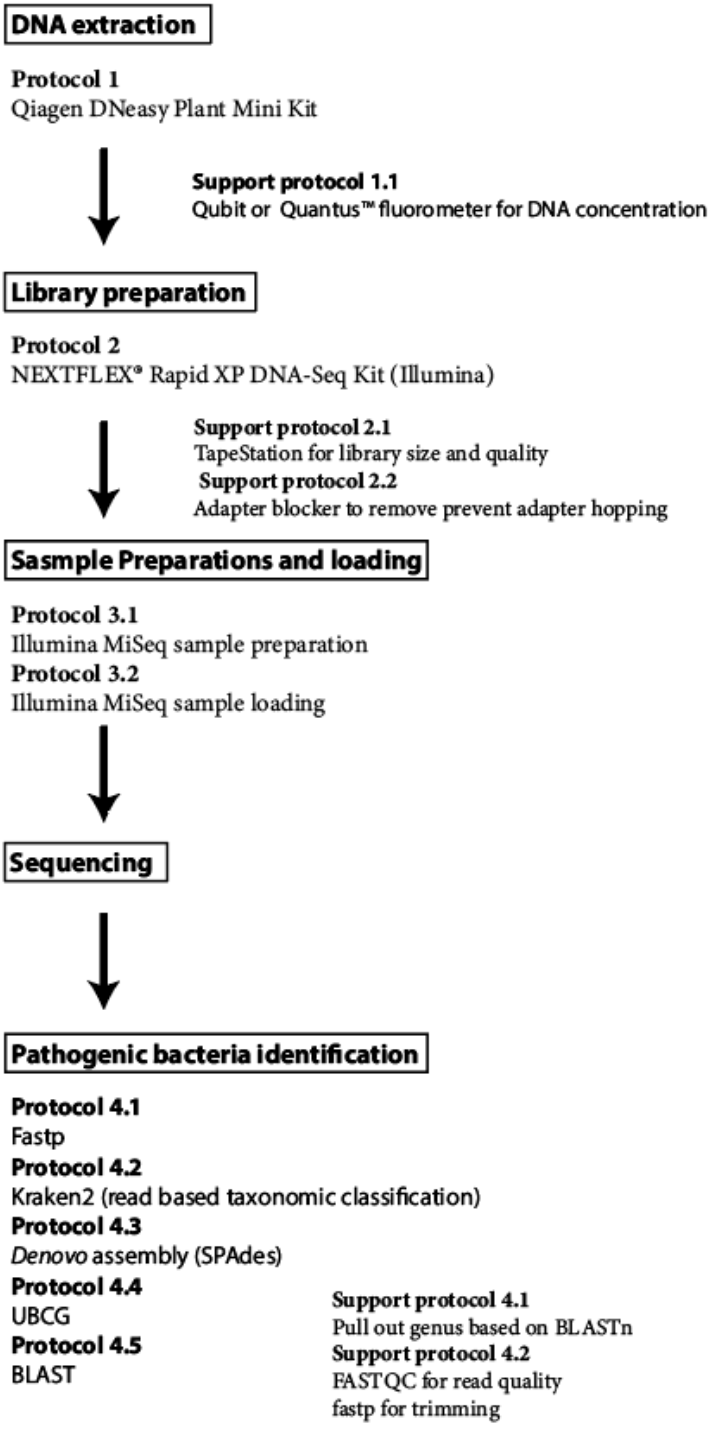


Figure 35.1 Pipeline used to identify culturable plant pathogenic bacteria

1. DNA extractions

This kit can be used for bacteria growing in both liquid and on agar. Bacterial genomic DNA is extracted using the DNeasy Blood & Tissue Kit (Qiagen), following the instructions provided by the manufacturer with minor amendments. Options for lysing bacterial cells for 1 hour and RNase A treatment are included. DNA is eluted from the column with half the suggested elution volume (from 100 to 200 µL) to increase the concentration of gDNA extracted.

Protocol 1: Qiagen DNeasy® Blood & Tissue kit for Illumina sequencing

Things to do before starting

- Buffer ATL and Buffer AL may form precipitates upon storage. If necessary, warm to 56°C until the precipitates have fully dissolved.
- Buffers AW1 and AW2 in the DNeasy® Blood & Tissue kit are supplied as a concentrate – add the appropriate amount of 100% ethanol as indicated on the bottles to obtain a working solution. Label the bottle with the date and initials.
- Heat the UltraPure™ DNase/RNase-Free Distilled Water (Thermo Fisher Scientific) water to 65 °C prior to elution for an increased final yield.
- Preheat a heating block to 56°C.

1) Pick a growing bacteria colony from plate and place in a labelled 1.5 ml microcentrifuge tube filled with 180 µL Buffer ATL. If using liquid cultures, harvest cells from overnight cultures (maximum 2 x 10^9 cells) in a microcentrifuge tube by centrifuging for 10 min at 5000 x g (7500 rpm). Discard supernatant. And resuspend pellet in 180 µl Buffer ATL
2) Add 20 µL proteinase K. Mix thoroughly by vortexing and incubate at 56°C for 1 hr. Vortex occasionally during incubation to disperse the sample.
3) Add 4 µL RNase A (100 mg/ml), mix by vortexing, and incubate for 2 min at room temperature before continuing with step 3.
4) Vortex for 15 s. Add 200 µL Buffer AL to the sample and mix thoroughly by vortexing. Then add 200 µL ethanol (96–100%) and mix again thoroughly by vortexing.
5) Pipette the mixture from step 4 (including any precipitate) into the DNeasy Mini spin column placed in a 2 ml collection tube (provided). Centrifuge at 6000 x g (8000 rpm) for 1 min. Discard flow-through and collection tube.
    - Flow-through contains Buffer AL or Buffer AW1 and is therefore not compatible with bleach. See page 8 for safety information.
6) Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µL Buffer AW1, and centrifuge for 1 min at 6000 x g (8000 rpm). Discard flow-through and collection tubes.
7) Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µL Buffer AW2, and centrifuge for 3 min at 20,000 x g (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube.
    - Note: It is important to dry the membrane of the DNeasy Mini spin column, since any residual ethanol may interfere with subsequent reactions.
8) Place the DNeasy Mini spin column in a clean 1.5 ml or 2 ml microcentrifuge tube, and pipette warmed 100 µL UltraPure water directly onto the DNeasy membrane. Incubate at room temperature for up to 5 min, and then centrifuge for 1 min at 6000 x g (8000 rpm) to elute.
9) DNA is now ready for downstream testing, or it can be stored at –20°C until use.

Controls

Negative extraction control is used to monitor contamination during nucleic acid extraction as well as cross-contamination among samples. The control should be an extraction with no samples.

Support protocol 1.1 DNA quantification

DNA was quantified using the Quantus™ fluorometer (Promega) according to the manufacturer's instructions. Qubit™ fluorometer (Thermo Fisher Scientific) can also be used for DNA quantification. Where possible, the same protocol, reagents, and kits should be used for all samples within a sequencing run.

2. Library preparation

The NEXTFLEX Rapid XP DNA-Seq Kit (Perkin Elmer) is recommended as it is cost effective, and it produces longer fragments that aid with genome assembly. Half reactions can be used for sequencing on the Illlumina MiSeq, HiSeq and NovaSeq platforms. Reactions are prepared according to the manufacturer's instructions, except volumes of the components are halved. This protocol is dependent on the amount of input DNA and should be adjusted accordingly. Minimum input amount of DNA for the half reaction protocol is 2 ng of DNA. Targeted fragment size was 200-300 bp with final library size ranging between 360-460 bp which is sufficient for 2X150 bp sequencing and higher.
As of June 2022, a new chemistry (NEXTFLEX® Rapid XP V2) has been released by Perkin Elmer which promises to accommodate low input DNA (100 pg – 1 µg) and low adapter dimer formation. Protocols are subject to change in the future once the new kit has been optimized.

Protocol 2: NEXTFLEX Rapid XP DNA-Seq Kit (half reaction) for standard use

NEXTFLEX Rapid XP DNA-Seq Kit is split into 4 steps;

A. Fragmentation, end-repair and adenylation
B. Adapter ligation
C. Bead size selection
D. PCR amplification

NEXTFLEX Rapid XP DNA-Seq Kit protocol Option 2 (from page 14) (bead size selection) of the protocol is suggested for DNA above 2 ng as it gives single sharp peaks on the TapeStation indicating lack of primer/adapter dimers and smearing.

It is recommended to use UDIs for sequencing (NEXTFLEX® Unique Dual Index Barcodes – 96 (Cat # 514150, 514151)). Other indexes that use A-tailing ligation, (such as UDIs from NEB) can be adapted to this protocol but protocol optimization would be required.

UDIs must be ordered separately.

The PCR amplification step has been modified to use real-time PCRs to observe optimal PCR amplification cycles.

Notes:

All input DNA would need to be normalized (same amount) before running the NEXTFLEX Rapid XP DNA-Seq Kit. Quantification of DNA can be done using Support Protocol 1.1 and input concentrations adjusted for even DNA input.

EDTA disrupts the fragmentation step in this protocol. It is suggested to either clean the library with AMPure XP beads (Beckman Coulter) and resuspend with UltraPure water or to use their condition solution (page 24). EDTA is found in most extraction kit elution buffers (i.e. DNeasy Bloody and Tissue kit). however, if the DNA extraction protocol in this SOP was used, then DNA would have already been eluted in water and no extra steps would be necessary.

A. FRAGMENTATION, END-REPAIR & ADENYLATION

Materials supplied by Bioo Scientific.

CLEAR CAP - NEXTFLEX® Fragmentation Buffer, NEXTFLEX® Fragmentation Enzyme Mix
WHITE CAP - Nuclease-free Water

User Supplied
DNA in 34 µL (or less) nuclease-free water
Thermal Cycler
96 well PCR Plate
Adhesive PCR Plate Seal
Microcentrifuge
Ice

DO NOT VORTEX NEXTFLEX® Fragmentation Enzyme Mix

1. For each sample, combine the following reagents on ice in a nuclease-free 96 well PCR plate:

__ µL	Nuclease-free Water
__ µL	DNA (2 ng - 1 µg)
2.5 µL	NEXTFLEX® Fragmentation Buffer
19.5 µL	TOTAL

Ensure thorough mixing by pipetting up and down. Proceed with adding the enzyme.

19.5 µL	DNA + NEXTFLEX® Fragmentation Buffer mixture (from above)
5.5 µL	NEXTFLEX® Fragmentation Enzyme Mix (DO NOT VORTEX)
25 µL	TOTAL

Note: Do NOT vortex the final NEXTFLEX® Fragmentation reaction. Mix by pipette only. It is important to mix the

reaction on ice.

2. Apply adhesive PCR plate seal and incubate on a thermal cycler using the following program:

1 min

4 °C

See fragmentation table

35 °C

30 min

65 °C

end

4 °C

Note: The initial 4 °C step is to pre-chill the instrument temperature. Place samples into thermal cycler after the temperature reaches 4 °C and follow the program. If the thermal cycler is preset to 4 °C a full 1 min incubation at 4 °C is not necessary.

The following table lists the recommended incubation times as a guideline for fragmentation. The mode fragment size can be adjusted by changing the duration of incubation at this 35 °C step. These times are recommendations only, and incubation time may need to be optimized for different sample inputs and types to obtain desired mode fragment size

Input DNA	Target Fragment Peak Size			
	200 - 300	300 - 400	400 - 500	500 - 600
	Fragmentation Time (min) at 35 °C			
1 ng	24	15	8	4
10 ng	18	8	5	3
100 ng	15	8	5	3
500 ng	15	8	5	3
1000 ng	14	7	5	3

Note: The final library size will be approximately 120 bp larger than the fragment size.

3. The procedure may be safely stopped at this step with samples stored at -20 °C if needed. To restart the protocol, thaw frozen samples on ice before proceeding to Step B1.

B. Adapter Ligation

Materials supplied by Bioo Scientific.

PURPLE CAP - NEXTFLEX® Ligase Buffer Mix XP, NEXTFLEX® Ligase Enzyme XP
WHITE CAP - Nuclease-free Water, Resuspension Buffer, NEXTFLEX® Cleanup Beads XP

User Supplied

25 µL of Fragmented, End Repaired, and Adenylated DNA (from STEP A)
Thermal Cycler
Adhesive PCR Plate Seal
80% Ethanol, freshly prepared (room temperature)
Magnetic Stand
NEXTFLEX® Unique Dual Index Barcodes – 96 (Cat # 514150, 514151) or NEXTFLEX® Dual Index Barcodes – 96 (Cat # 514160, 514161)

1. Thaw NEXTFLEX® Ligase Buffer Mix XP to room temperature, and vortex for 5-10 seconds. Do not spin down tube, as this may cause components of the mix to separate and this will impact performance.
2. The following table lists recommended barcoded adapter concentration dilutions for the NEXTFLEX® Unique Dual Index Barcodes:
   Each sample will require 1.25 µL of barcoded adapter to be added. Perform barcoded adapter dilutions

Input DNA	Desired Adapter	Adapter Dilution Concentration
1 ng	0.3 µM	1 / 80
10 ng	0.6 µM	1 / 40
100 ng	6.25 µM	1 / 4
250 ng	25 µM	None
500 ng	25 µM	None
1 µg	25 µM	None

   if necessary, with Nuclease-free Water, depending on input amount and starting barcoded adapter concentration.

   The following reaction must be mixed thoroughly. The NEXTFLEX® Ligase Buffer Mix XP is very viscous. Thorough mixing of the reaction below is critical to obtaining optimal results.

   Suggestion: To mix, pipette up and down 15 times; visually inspect tubes to ensure proper homogenization.

Combine the following in the PCR plate and mix thoroughly by pipette:

25 µL

Fragmented, End Repaired & Adenylated DNA (from Step A)

22.25 µL

NEXTFLEX® Ligase Buffer Mix XP*

1.25 µL

NEXTFLEX® Barcoded Adapter

1.5 µL

NEXTFLEX® Ligase Enzyme Mix XP*

50 µL

TOTAL

*These components can be premixed and added in a single step. Adapter should not be premixed in order to prevent excess adapter dimer formation.

3. Apply adhesive PCR plate seal and incubate in a thermal cycler for 15 minutes at 20 °C with heated lid option turned “off” or open, followed by a 4 °C hold.
4. Add 32.5 µL of Nuclease-free Water and 17.5 µL of NEXTFLEX® Cleanup Beads XP to each sample. Mix thoroughly until homogenized. The NEXTFLEX® Cleanup Beads XP and Nuclease-free Water can be premixed and added in a single step.
5. Incubate sample at room temperature for 5 minutes.
6. Place the 96 well PCR plate on the magnetic stand at room temperature for 5 minutes or until the supernatant appears completely clear.
7. Remove and discard clear supernatant taking care not to disturb beads. Some liquid may remain in well.
8. With the plate remaining on the stand, add 200 µL of 80% ethanol to each magnetic bead pellet and incubate plate at room temperature for 30 seconds. Carefully remove ethanol by pipette.
9. Repeat previous step for an additional wash. Ensure all ethanol has been removed.
10. Remove the 96 well plate from the magnetic stand and let dry at room temperature for 3 minutes.
11. Resuspend dried beads with 27.5 µL of Resuspension Buffer. Mix thoroughly until homogenized.
12. Incubate sample at room temperature for 2 minutes.
13. Place the 96 well PCR plate on the magnetic stand at room temperature for 2 minutes or until the supernatant appears completely clear.
14. Do not discard the sample in this step. Transfer 25 µL of clear sample to a new well/plate (this contains your sample). Remove the 96 well PCR plate from the magnetic stand.
15. The procedure may be **safely stopped** at this step with samples stored at -20 °C if needed. To restart the protocol, thaw frozen samples on ice before proceeding to Step C.

C. Bead Size Selection

Materials supplied by Bioo Scientific.

WHITE CAP - Resuspension Buffer, NEXTFLEX® Cleanup Beads XP

User Supplied

25 µL of Adapter Ligated DNA (from STEP B)
80% Ethanol, freshly prepared (room temperature)
96 well PCR Plate
Magnetic Stand

Size Selection may not be optimal for inputs less than 2 ng. The size ranges listed in tables below reflect the total library size, including the insert and NEXTFLEX® Barcode Adapters. NEXTFLEX® Barcode Adapters add ~120bp to the insert length.

The following chart is a general recommendation for certain sizes. Yield and specificity of size selection is affected by size distribution of starting material. It is important to select for an insert size that is compatible with the size range of the starting material. The user should use this chart as a guideline with the expectation that further optimization may be required for their specific application.

Note: During optimization, the user should keep in mind that adding more NEXTFLEX® Cleanup Beads XP at the 1st cleanup step "Bead Volume #1" would decrease the library size for the upper size selection and as a result, the lower size selection as well. Adding more NEXFLEX® Cleanup Beads XP at the 2nd cleanup step "Bead Volume #2" would decrease the library size of the lower size selection only.

The following table lists the appropriate volume of NEXTFLEX® Cleanup Beads XP required to size select for library peak sizes (approximated) below:

Approximate Insert Peak Size (bp)	150 - 250	250 - 350	300 - 500	400 - 600	500 - 700
Approximate Library Peak Size (bp)	270 - 370	370 - 470	420 - 620	520 - 720	620 - 820
Bead Volume #1	17.5	16	15	13.5	12
Bead Volume #2	6	4.5	4	4	4

Ensure all reagents are at room temperature. Vortex the NEXTFLEX® Cleanup Beads XP thoroughly prior to use. Use a fresh dilution of 80% ethanol during wash steps.

1. Add Bead Volume #1 to sample as indicated in the column corresponding to your desired size range in the chart above. Mix thoroughly until homogenized.
2. Incubate sample at room temperature for 5 minutes.
3. Place the 96 well PCR plate on the magnetic stand at room temperature for 5 minutes or until the supernatant appears completely clear.
4. Do not discard the supernatant in this step. Transfer the clear supernatant to a new well. Be careful not to disrupt the magnetic bead pellet or transfer any magnetic beads with the sample.
5. Add Bead Volume #2 to sample as indicated in the column corresponding to your desired size range in the chart above. Mix thoroughly until homogenized.
6. Incubate sample at room temperature for 5 minutes.
7. Place the 96 well PCR plate on the magnetic stand at room temperature for 5 minutes or until the supernatant appears completely clear.
8. Remove and discard clear supernatant, taking care not to disturb beads. Some liquid may remain in wells.

9. With the plate on the stand, add 200 µL of 80% ethanol to each magnetic bead pellet and incubate plate at room temperature for 30 seconds. Carefully remove ethanol by pipette.
10. Repeat previous step for an additional wash. Ensure all ethanol has been removed.
11. Remove the 96 well PCR plate from the magnetic stand and let dry at room temperature for 3 minutes
12. Resuspend dried beads with 14 µL of Resuspension Buffer. Mix thoroughly until homogenized.
13. Incubate sample at room temperature for 2 minutes.
14. Place the 96 well PCR plate on the magnetic stand at room temperature for 2 minutes or until the supernatant appears completely clear.
15. Transfer 11.5 µL of clear sample to a new well.
16. The procedure may be **safely stopped** at this step with samples stored at -20°C, if needed. To restart, thaw the frozen samples on ice before proceeding with Step D2. If input DNA amount was 500 ng or greater, PCR amplification may not be necessary depending on the sequencing application. Users starting with greater than 500 ng of input DNA are advised to quantify their unamplified libraries by qPCR at this point to determine if PCR amplification is necessary. This can be performed using any qPCR quantification kit compatible with Illumina® platforms.

D. PCR Amplification

Materials supplied by Bioo Scientific.

GREEN CAP - NEXTFLEX® PCR Master Mix XP, NEXTFLEX® Primer Mix XP
WHITE CAP - Nuclease-free Water, Resuspension Buffer, NEXTFLEX® Cleanup Beads XP
User Supplied

11.5 µL of Adapter Ligated DNA (from STEP C)
Real-Time PCR Detection System
qPCR Plate Seal
96 Well qPCR Plate
80% Ethanol, freshly prepared (room temperature)
Magnetic Stand
Note: The NEXTflex® Primer Mix that is included in the NEXTflex® NGS Barcodes are NOT compatible with this kit and should NOT be used in place of the Primer Mix XP.
*The following table lists recommended PCR cycles:

	Number of PCR cycles to produce	
Input DNA	100 ng libraries	1 µg libraries
1 ng	10 - 12	13 - 15
10 ng	6 - 8	9 - 11
100 ng	2 - 3	6 - 7
250 ng	1 - 2	4 - 5
500 ng	0	3 - 4
1000 ng	0	2 - 3

Note: Suggested PCR cycles can be used but for the optimal amplification cycles, **qPCR** should be used. Over amplification could reduce detection of low copy regions and/or cause chimeric DNA products. Under amplification would reduce final concentration of libraries.

1. For each sample, combine the following reagents on ice in the qPCR plate. Mix thoroughly.

11.5 µL

Adapter Ligated DNA (from Step C)

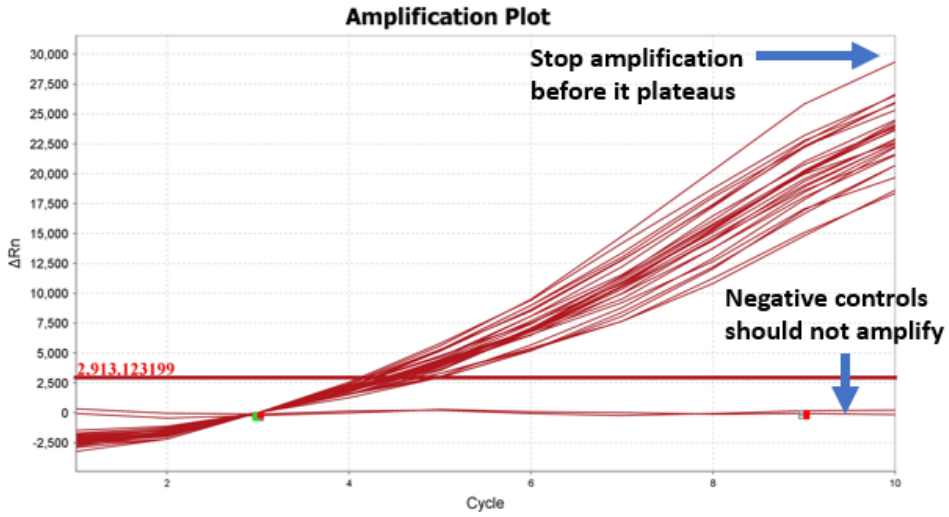
12.5 µL

NEXTFLEX® PCR Master Mix XP*	1 µL
NEXTFLEX® Primer Mix XP*	1 µL
1X SYBR green	26 µL
TOTAL	

* These components can be premixed and added in a single step.

2. Apply adhesive qPCR plate seal and place in thermocycler for the following qPCR cycles:

30 sec 98°C	
15 sec 98°C 30 sec 65°C 30 sec 72°C	Repeat as suggested in above table
2 min 72°C	



Note: Suggested PCR cycles can be used as reference, but cycles should be stopped before the amplification curve plateaus.

3. Add 22.5 µL of NEXTFLEX® Cleanup Beads XP to each sample. Mix thoroughly until homogenized.
4. Incubate at room temperature for 5 minutes.
5. Place the 96 well PCR plate on the magnetic stand at room temperature for 5 minutes or until the supernatant appears completely clear.
6. Remove and discard clear supernatant taking care not to disturb beads. Some liquid may remain in wells.
7. With plate on stand, add 200 µL of 80% ethanol to each magnetic bead pellet and incubate plate at room temperature for 30 seconds. Carefully remove ethanol by pipette.

8. Repeat previous step for an additional wash. Ensure all ethanol has been removed.
9. Remove plate from magnetic stand and let dry at room temperature for 3 minutes.
10. Resuspend dried beads with 33 µL of Resuspension Buffer. Mix thoroughly until homogenized.
11. Incubate resuspended beads at room temperature for 2 minutes.
12. Place the 96 well PCR plate on the magnetic stand at room temperature for 2 minutes or until the supernatant appears completely clear.
13. Do not discard the supernatant in this step. Transfer **30 µL** of clear sample to a new well. Remove the 96 well PCR plate from the magnetic stand.
14. Examine your library by Tapestation (Support protocol 2.1) to ensure proper library sizing and to verify exclusion of contaminating small and large fragments.
15. Quantify DNA library templates using Quantus™ or Qubit™ for optimal cluster density (Support Protocol 1.1).

    The library is now ready for cluster generation per the standard Illumina® protocol. Proceed to cluster generation or seal with adhesive PCR plate seal and store at - 20 °C.

Support protocol 2.1 TapeStation

A detailed protocol to use the 2200 TapeStation system (Agilent technologies) for quality checking of libraries can be found at https://www.agilent.com/cs/library/usermanuals/public/G2964-90000_TapeStation_USR_ENU.pdf. For Illumina sequencing, High Sensitivity D1000 (HSD1000) tape assays are used for DNA concentrations below 30 ng/µL and D1000 tape assays are used DNA concentrations above 30 ng/µL.

Figure 35.2 and Figure 35.3 show examples of a bad and good library, respectively. Libraries should peak at a minimum of 280 bp with no smears or peaks below the 200 bp mark.

Sample Preparation High Sensitivity D1000 and d1000 ScreenTape Assay

1. Allow reagents to equilibrate at room temperature for 30 min.
2. Vortex mix before use.

D1000 protocol

3. If running ladder, prepare by mixing 2 µL High Sensitivity D1000 Sample Buffer ( ) with 2 µL High Sensitivity D1000 Ladder ( )
4. Prepare sample by mixing 2 µL High Sensitivity D1000 Sample Buffer ( ) with 2 µL DNA sample

Or

HSD1000 protocol

3. If running ladder, prepare by mixing 3 µL D1000 Sample Buffer ( ) with 1 µL D1000 Ladder ( )
4. Prepare sample by mixing 3 µL High Sensitivity D1000 Sample Buffer ( ) with 1 µL DNA sample.

5. Spin down, then vortex using IKA vortexer and adaptor at 2000 rpm for 1 min.
6. Spin down to position the sample at the bottom of the tube.
7. Launch the 2200 TapeStation Controller Software.
8. Load HSD1000/D1000 ScreenTape device and loading tips into the 2200 TapeStation instrument.

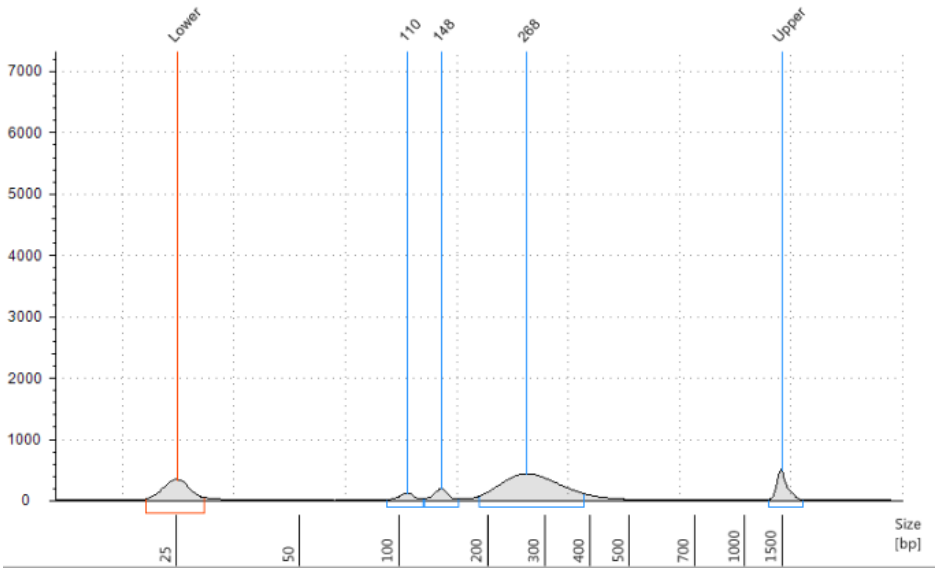


Figure 35.2 Example of a bad library on a HSD1000 tape assay on the TapeStation. The peaks between 100 – 150 bp (circled) indicates the presence of either primer dimers or excess indexes.

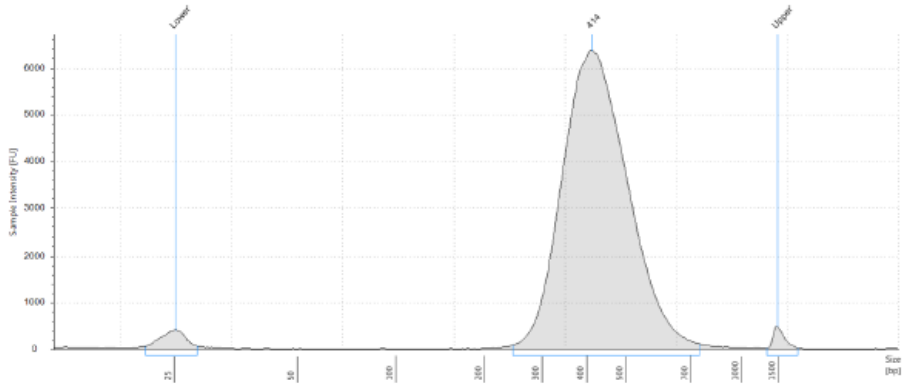


Figure 35.3 Example of a good library on a HSD1000 tape assay on the TapeStation. One strong peak is observed at 414 bp. Absence of primer dimer around the 100-150 bp mark.

Support protocol 2.2 Adapter blocker

Blocking unligated adapters to prevent index hopping is done after pooling libraries following the Illumina Free Adapter Blocking (FAB) Reagent protocol. Ampure XP beads would need to be ordered separately from the kit. Alternate beads can be used (ProNex® Size-Selective Purification System), however, volumes would need to be adjusted to accommodate beads used.

Illumina Free Adapter Blocking Reagent protocol

Things to do before starting

- Thaw FAB at room temperature, and then place on ice.
- RSB and Ampure XP beads; let stand for 30 minutes to bring to room temperature.
- Vortex Ampure XP beads before each use to make sure that beads are evenly distributed

1) Mix FAB reagent by inversion.
2) Centrifuge FAB at 600 × g for 5 seconds.
3) Add 30 µL of library sample pool to PCR tube.
4) Add 30 µL of FAB to each PCR tube and then mix thoroughly by pipetting up and down.
5) Centrifuge briefly to make sure all contents are on the bottom of the tube.
6) Incubate by placing on the thermal cycler and running the FAB program:
   - 38°C for 20 minutes
   - 60°C for 20 minutes
   - Hold at 4°C
   - Preheat lid to 100°C
7) Add 60µL of vortexed SPRIselect beads and mix by pipetting.
8) Incubate at room temperature for 5 minutes.
9) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).
10) Remove and discard all supernatant.
11) Wash 2 times as follows:
    - Add 200µL freshly prepared 80% EtOH.
    - Incubate on the magnetic stand for 30 seconds.
    - Remove and discard all supernatant.
12) Use a 20µL pipette to remove residual EtOH.
13) Air-dry on the magnetic stand for 5 minutes.
14) Add 35µL RSB, remove from the magnetic stand and mix by pipetting
    - Note: if the nM calculated in Step 5 was less than 5nM then lower the elution volume to 25µL to avoid diluting the pooled library.
15) Incubate at room temperature for 2 minutes
16) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).
17) Transfer 33.5µL supernatant to a new tube.
    - Note: this will be 23.5µL if 25µL of RSB was added.

NOTE: Quantify (Support Protocol 1.1) and TapeStation (Support Protocol 2.1) final libraries to finalize concentration and quality of libraries.

3. Sample preparation and loading

Protocol 3.1: Illumina Miseq (MiSeq Reagent Kit v3) sample preparation for loading; denaturing and diluting libraries

To prepare the PhiX calibration control and denature and dilute libraries for loading onto the Miseq sequencer follow the instructions provided by the manufacturer. As MiSeq kit chemistries are constantly being upgraded, be on the lookout for new protocols. The current protocol to date was released in February 2019 (https://sapac.support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-10.pdf). At the Denature and Dilute PhiX Control step it is suggested to make 20 pM PhiX libraries. PhiX libraries can be kept up to a month at -15°C to -25°C. Aim to make between 8 pM to 10 pM libraries and combine with 1% 20 pM PhiX.

Protocol 3.2: Illumina Miseq (MiSeq Reagent Kit v3) Sample loading

Preparing sample sheet

Create the sample sheet in Microsoft Excel and save it as a .csv file. There are three components of the sample sheet. First is the header where details of the run are written. Some of the fields can be changed to suit the current experiment. This includes investigator name, date and assay. In the experiment name, the flowcell ID (Figure 35.5B) is usually incorporated into the name. The two fields that cannot be changed are the workflow field, which needs to be filled with the term “GenerateFASTQ”, and the chemistry field which needs to be filled with the term amplicon.

In the Reads section the amount of basepairs for read 1 and read 2 sequence is listed. Since this protocol is aimed at using the MiSeq Reagent Kit V3, 300 is written as up to 300 sequencing cycles can be completed. In the Settings section, the adapter sequence can be filled, however, it can remain also empty. The Data section is where information of individual samples is written. Example of a sample sheet is shown below.

[Header]									
IEMFileVersion	4								
Investigator Name	Your name								
Experiment Name	J3KH7-your project								
Date	Date								
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	Truseq								
Description	Description								
Chemistry	Amplicon								
[Reads]									
300									
300									
[Settings]									
Adapter	Adapter sequence								
[Data]									
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
S-14RD	S-14RD	1	A1	S747	TAGGAGCT	S524	GCCTTAAC	Pathogen	
S-17RD	S-17RD	1	B1	S716	CGAATTGC	S559	GTAAGGTG	Pathogen	
S-19RD	S-19RD	1	C1	S770	GTCCTAAG	S540	AGCTACCA	Pathogen	
S-36RD	S-36RD	1	D1	S749	CTTAGGAC	S565	CTTCACTG	Pathogen	
S-39RD	S-39RD	1	E1	S784	TCCACGTT	S541	GGTTGAAC	Pathogen	
S-40RD	S-40RD	1	F1	S780	CAACACAG	S530	GATAGCCA	Pathogen	
20-01805-1	20-01805-1	2	A1	UDI0094	CGTCTGCG	UDI0094	TTCACAAT	Pathogen	
20-01805-2	20-01805-2	2	B1	UDI0095	TCGATATC	UDI0095	ACACGAGT	Pathogen	
20-01801	20-01801	2	C1	UDI0278	ATCGTTAG	UDI0278	TGCGAAGT	Pathogen	

Figure 35.4 An example of a sample sheet layout.

Once the sample sheet is completed, the name of the file should be saved as the barcode number on the reagent cartridge. As an example, the name of this sample sheet will be MS9857995-600V3.csv as shown from the Figure 35.5A.

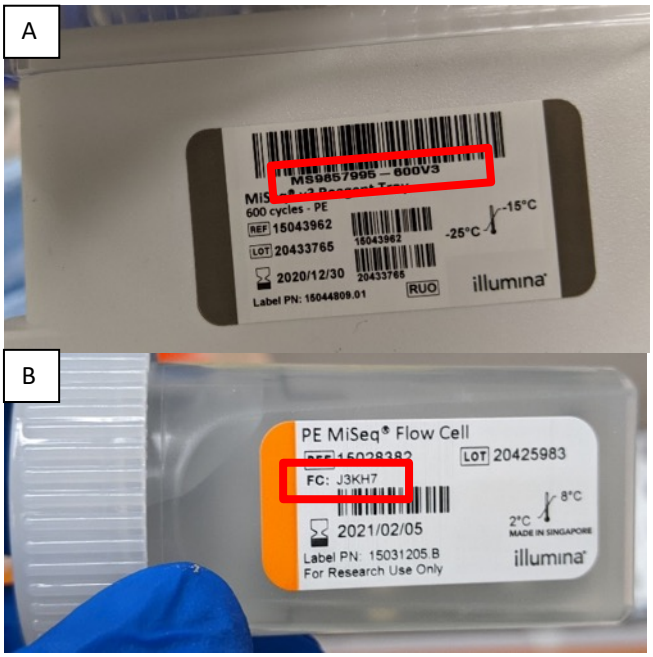


Figure 35.5 Barcodes from the MiSeq kits required to create the sample sheet. A) Reagent cartridge and B) Flow Cell

Sample loading

In depth protocols and troubleshooting can be found in the manufacturer's website (https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-for-miseq-reporter-1000000061014-00.pdf). The sequence of steps are as follows:

9. Inspect the Reagent Cartridge
10. Load Sample Libraries
11. Set Up a Run Using MCS
12. Clean the Flow Cell
13. Load the Flow Cell
14. Load Reagents
    a. Load Sample Libraries can be done here if you prefer to leave you libraries on ice
15. Change Sample Sheet
    a. Can be added through a USB or through email sent to self as you can log on your email on the MiSeq
16. Post-Run Wash

WARNING: of the set of reagents used in this process contains potentially hazardous chemicals. Personal injury can occur through inhalation, ingestion, skin contact, and eye contact. Wear protective equipment, including eye protection, gloves, and laboratory coat appropriate for risk of exposure. Handle used reagents as chemical waste and discard in accordance with applicable regional, national, and local laws and regulations.

4. Bacterial identification

Conda or Miniconda needs to be installed before running most of these scripts. Basic knowledge of bash commands is also required for most of the protocols. The bioinformatic pipeline would require large databases to be set up on the computer, one for Kraken2 and another for BLASTn. Reference genomes would also need to be downloaded for phylogenetic analysis. Use known genomes as references; from current literature or from curated databases such as ezbiocloud. Pay in mind that most databases are not perfect. For instance, even though genomes for ezbiocloud have been curated, they do not classify down to the biovar level for some species. For NCBI, some classifications may be wrong or outdated.

Kraken2 is recommended for a swift analysis to identify genus which is then followed by *de novo* assembly to generate contigs. **U**p-to-date **B**acterial **C**ore **G**ene set (UBCG) is the recommended protocol for phylogenetic analysis used for identification. To accommodate to possible contaminations, an extra BLASTn step was added to aid in isolating contigs from one genus (Figure 35.6).

WORKFLOW:

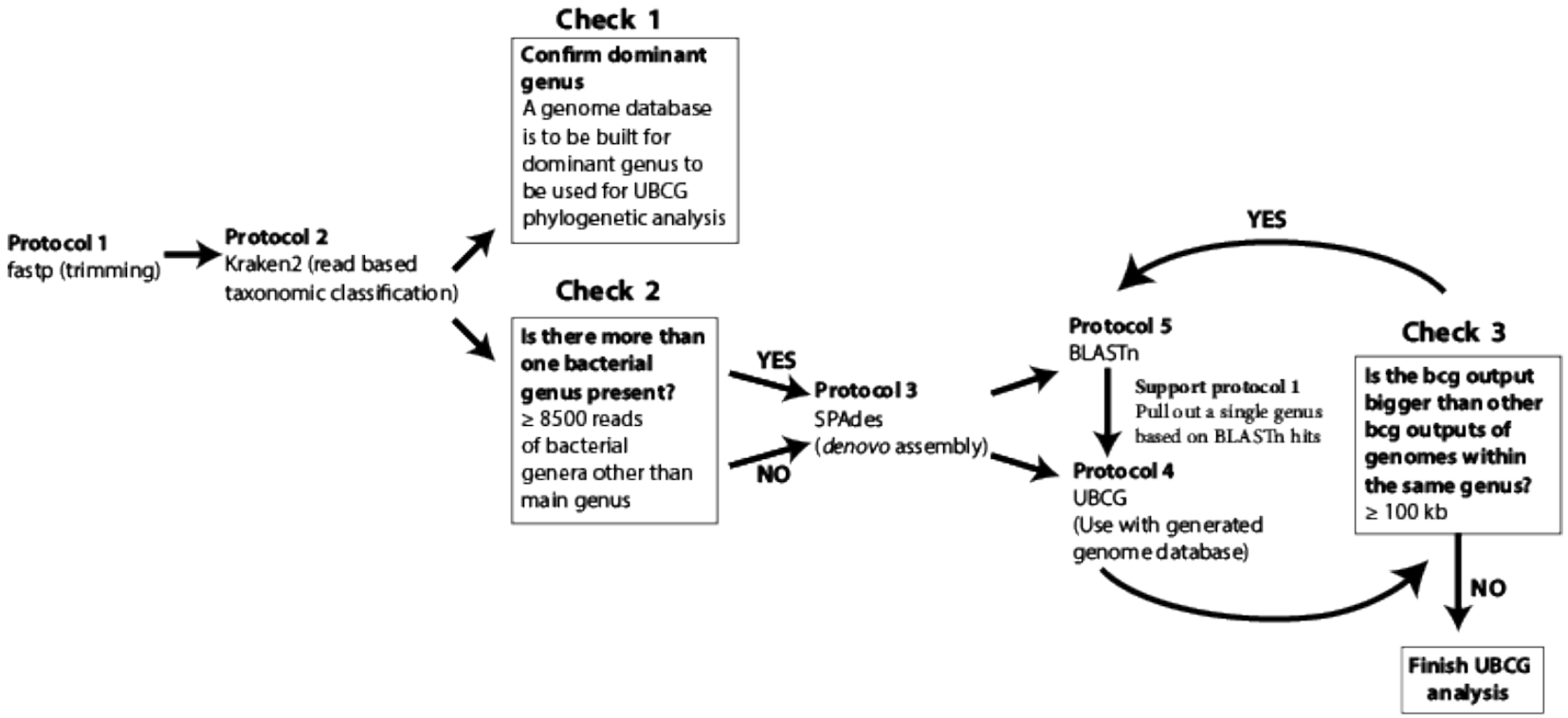


Figure 35.6 Pipeline used to identify culturable plant pathogenic bacteria

Protocol 4.1: fastp

Fastp is available for download in Conda. It can be installed by running the script:

```
conda install --name fastp fastp
```

For paired end reads, you can run the script:

```
fastp -i input_R1_001.fastq.gz -I input_R2_001.fastq.gz -o outpout_R1.fastq.gz -O
outpout__R2.fastq.gz
```

For long reads such as those from ONT, you can run the script:

```
fastp -i input.fastq.gz -o outpout.fastq.gz
```

Protocol 4.2: Kraken2

The initial bioinformatic pipeline uses the k-mer analysis tool - Kraken2. Kraken2 is used to identify genomes to the genus level and this information is used to pull out the genome database associated with that genus for phylogenetic analysis. From a previous report, Kraken2 was able to identify contaminations, which could have occurred during the isolation process. These contaminations could include plant/mammal or fungal contaminants which is defined by low percentage of bacterial reads and high percentage of unclassified reads. From the previous report, this would not affect downstream analysis as long as there are sufficient bacterial reads (≥350,000 reads). Another form of contamination includes other bacterial genera. From the report, contamination of more than 85,000 reads of other bacterial species/genera could affect downstream analysis and more 350,000 reads could be used for individual bacterial identification analysis.

Kraken2 is available for download in Conda. It can be installed by running the script:

```
conda create --name kraken2 kraken2

conda activate kraken2
```

Once installed, you can install the database from here or you can use the script below. If you are using low-memory computing environments, miniKraken2_v1_8GB can be trialled as only the identification of genus is needed.

```
cd $path/to/download/database/to

Bacterial database:
kraken2-build --download-library bacteria --db $DBNAME
Minikraken2 database:
wget ftp://ftp.ccb.jhu.edu/pub/data/kraken2_dbs/minikraken_8GB_202003.tgz
tar zxvf minikraken_8GB_202003.tgz
```

Once database is downloaded, you can run the script:

```
kraken2 --db $/path/to/database --gzip-compressed --paired $*R1.fastq.gz $*R2.fastq.gz --report
$NAME.txt --output report
```

Remove `--gzip-compressed` if not using a compressed file.

For this pipeline, only the sample report output is required to identify the genus with highest percentage of reads.

The first few lines of an example report are shown below

```
0.80   14271   14271   U       0       unclassified
99.20  1776601 49      R       1       root
99.19  1776299 5       R1      131567    cellular organisms
99.18  1776100 2216    D       2           Bacteria
99.01  1773216 267     D1      1783272       Terrabacteria group
98.99  1772772 171     P       1239            Firmicutes
98.97  1772496 1042    C       91061             Bacilli
98.90  1771094 1683    O       1385                Bacillales
98.78  1769063 1474    F       186817                Bacillaceae
98.69  1767327 169155  G       1386                    Bacillus
84.74  1517573 1479594 S       561879                    Bacillus safensis
 2.12  37979   37979   S1      1178541                     Bacillus safensis FO-36b
 2.07  36986   35893   S       1408                      Bacillus pumilus
 0.06  1093    1093    S1      315750                      Bacillus pumilus SAFR-032
 1.19  21300   8       G1      185979                    unclassified Bacillus
 1.07  19198   19198   S       756828                      Bacillus sp. WP8
 0.10  1761    1761    S       2683680                     Bacillus sp. ms-22
 0.01  121     121     S       1071078                     Bacillus sp. NSP9.1
 0.00  34      34      S       2682541                     Bacillus sp. N1-1
 0.00  34      34      S       2567941                     Bacillus sp. DSL-17
```

The sample reports are tab-delimited, with one line per taxon. Each tab, from left to right, are as follows:

1. Percentage of reads covered by the clade rooted at this taxon
2. Number of reads covered by the clade rooted at this taxon
3. Number of reads assigned directly to this taxon
4. A rank code, indicating (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. Taxa that are not at any of these 10 ranks have a rank code that is formed by using the rank code of the closest ancestor rank with a number indicating the distance from that rank. E.g., "G2" is a rank code indicating a taxon is between genus and species and the grandparent taxon is at the genus rank.
5. NCBI Taxonomy ID number
6. Indented scientific name

CHECK 1. Identify dominant genus. This is so that a database can be built for that genus for UBCG analysis. The report example above is an example of a clear genus identification; *Bacillus* genus with read percentage of 98.69%.

```
 2.81  52797   52797   U       0       unclassified
97.19  1828945 70      R       1       root
97.16  1828273 39      R1      131567    cellular organisms
97.12  1827476 483     D       2           Bacteria
96.93  1824032 706     P       1224          Proteobacteria
77.56  1459572 594     C       28211           Alphaproteobacteria
77.47  1457719 1087    O       356               Rhizobiales
77.27  1454035 901     F       82115               Rhizobiaceae
77.13  1451474 12557   F1      227290                Rhizobium/Agrobacterium group
75.98  1429709 34569   G       357                     Agrobacterium
63.33  1191717 1191717 S       359                       Agrobacterium rhizogenes
10.20  191943  4355    G1      1183400                   Agrobacterium tumefaciens complex
 9.13  171729  166078  S       358                         Agrobacterium tumefaciens
 0.20  3677    3677    S1      1300225                       Agrobacterium tumefaciens WRT31
```





```
 0.00  1       1       S1      156889                    Magnetococcus marinus MC-1
19.22  361670  103     C       28216           Betaproteobacteria
19.20  361292  254     O       80840             Burkholderiales
19.13  359998  709     F       506                 Alcaligenaceae
19.00  357566  5734    G       222                   Achromobacter
17.76  334243  324907  S       85698                   Achromobacter xylosoxidans
 0.47  8918    8918    S1      762376                    Achromobacter xylosoxidans A8
 0.02  418     418     S1      1167634                   Achromobacter xylosoxidans NH44784-1996
 0.61  11454   11454   S       217203                  Achromobacter spanius
 0.25  4777    24      G1      2626865                 unclassified Achromobacter
 0.20  3837    3837    S       1881016                   Achromobacter sp. MFA1 R4
```

CHECK 2. There are possibilities of other bacterial contamination and should be looked out for. The report example above shows Agrobacterium as the dominant genus with read percentage of 75.98% with a second dominant genus, *Achromobacter* with 19% reads a read percentage of 19%. Low read percentage of a genus with no competing bacterial genera can be considered as one genus and may be used for UBCG analysis. If contaminated run *denovo* assembly as per normal, then blastn and **SUPPORT PROTOCOL 4.1** before UBCG.

Protocol 4.3: Denovo assembly – SPAdes

SPAdes is available to download from GitHub at https://github.com/ablab/spades#sec1.2. It can be installed by going to the directory in which you wish SPAdes to be installed and using the following command on a Linux terminal:

```
wget http://cab.spbu.ru/files/release3.15.4/SPAdes-3.15.4-Linux.tar.gz
tar -xzf SPAdes-3.15.4-Linux.tar.gz
cd SPAdes-3.15.4-Linux/bin/
```

Once installed, you should test the installed software before starting any analysis by running the script:

```
python /SPAdes-3.15.4-Linux/bin/spades.py --test
```

The log would indicate if your test was successful. If it is, you can start assembly using the script:

```
python /SPAdes-3.15.4-Linux/bin/spades.py -k 127,107,87,67,31 --careful --pe1-1 $*R1.fastq.gz --
pe1-2 $*_R2.fastq.gz -o $/output/folder/
```

SPAdes will create your output folder so making one beforehand is not necessary. If you do make one, an error will occur, and you would either add "-force" to the script or remove the folder.

Protocol 4.4: UBCG

UBCG uses 92 bacterial core genes for phylogenetic analysis. Only two lines of script is required for full analysis; one to extract '.ucg' files and the second for phylogenetic analysis. This makes it a universal method that can supposedly be applied to all bacterial species (Na et al., 2018). This step would require the curated database for each genus to be used for phylogenetic analysis.

UBCG is available to download from http://leb.snu.ac.kr/ubcg2/download. A zip file will be downloaded which would need to be extracted and moved into your Linux system. Before the java scripts can be run, an environment in Conda can be made by running the script:

```
conda create --name UBCG
conda activate UBCG
conda install java-jdk
conda install prodigal
conda install hmmer
conda install mafft
conda install fasttree
conda install raxml
```

This will install all necessary requirements to run the UBCG script. Java would need to be downloaded separately as java-jre package installed through Conda may be broken. Once ready, fasta files of reference genomes and samples are run through the first UBCG script:

```
conda activate UBCG
java -jar UBCG2.jar -i $genome.fasta -ucg_dir $/output/file/ -label $"sample_name" -hmm
hmm/ubcg_v2.hmm
```

This will create ".ucg" files.

CHECK 3. The size of ".ucg" files within a genus has a tendency to be around the same size or slightly lesser if the genome is incomplete. If the size of the file is bigger than the average, there is a probability that there is contamination from other bacterial genera that is affecting the output and BLASTn and the support protocol 4.1 should be used.

Accumulated ".ucg" files will need to be in one folder as the next step is to align and create a phylogenetic tree through the UBCG script:

```
java -jar UBCGtree.jar align -ucg_dir $/output/file/ -run_id $name
```

The output folder will contain both the fasta file and .nwk file of all gene alignment and a combined gene alignment. The .nwk file can then be viewed on FigTree or the MEGA software.

Protocol 4.5: BLAST

If presence of an extra genera was detected (CHECK 2 and CHECK 3) proceed to compare assembled contigs with to the non-redundant database at NCBI using the BLASTn algorithm (Altschul et al., 1990). Installation can be done through Conda, however databases would need to be downloaded from https://ftp.ncbi.nlm.nih.gov/blast/db/ and extracted before use. All nt/nr files would need to be downloaded and extracted before it can be used.

```
conda create --name blast blast

conda activate blast


date
blastn -query $SPAdes/file/name/scaffolds.fasta -db $/blastdb/nt -max_target_seqs
1 -max_hsps 1 -outfmt "6 qseqid stitle slen qlen mismatch gapopen qcovs qcovhsp
pident evalue" -out $blast.name.xml
date
```

Support Protocol 4.1: Pull out Genus based on BLAST hits

```
date
grep chosen_genus $blast.name.xml > $blast.genus.name.txt
awk {'print $1}' $blast.genus.name.txt | uniq > blast.name_list_sorted.txt
perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV'
$blast.name_list_sorted.txt $path/to/blast/output/spades.name/scaffolds.fasta >
$pulled_genus.fasta
date
```

Support protocol 4.2 FASTQC

FASTQC is a common method for checking the quality of reads. It can be downloaded in Conda or as a module on Linux. To run FASTQC on your reads, you can run the script:

```
fastqc -o /output/folder -f fastq *R1.fastq.gz *R2.fastq.gz
```

Data obtained from FASTQC is in a .html file. When opened, a summary is given (Figure 35.7) with details of each error accompanying it.

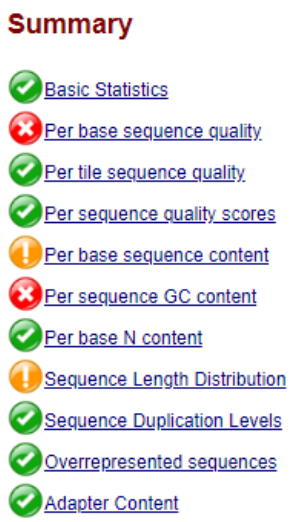


Figure 35.7 Summary as given by FASTQC results.

Interpretation of results

Identification can be confirmed using the phylogenetic tree produced by UBCG. A good reference database will enable accurate identification down to haplotypes, sequence types, subspecies, etc.

Possible reasons sequenced samples do not fall within an appropriate clade for identification:

1. Incomplete reference genomes. Adding more reference genomes would be able to fix this problem.
2. The Kraken2 database could be incomplete and genus identification was wrong. In this case, BLASTn results should be used to confirm genus and reference genomes should be changed accordingly.
3. Mixed samples. This could be due to multiple species of similar genus residing in the same sample. Having more specific species targets instead of genus using Support Protocol 4.1.