Methods in Molecular Biology 1746

Springer Protocols

Vitantonio Pantaleo
Michela Chiumenti *Editors*

Viral Metagenomics

Methods and Protocols

Humana Press

Methods in Molecular Biology

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Viral Metagenomics

Methods and Protocols

Edited by

Vitantonio Pantaleo and Michela Chiumenti

Institute Sustainable Plant Protection of the CNR, Research Unit of Bari,
Bari, Italy

Humana Press

Editors
Vitantonio Pantaleo
Institute Sustainable Plant Protection of the CNR
Research Unit of Bari
Bari, Italy

Michela Chiumenti
Institute Sustainable Plant Protection of the CNR
Research Unit of Bari
Bari, Italy

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7682-9 ISBN 978-1-4939-7683-6 (eBook)
https://doi.org/10.1007/978-1-4939-7683-6

Library of Congress Control Number: 2018933182

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Viruses are obligate entities that infect the cells of living organisms of every kingdom of life, including unicellular prokaryotes and eukaryotes, fungi, plants, and animals. They are the most abundant biological entities on earth and the major agents of disease and mortality as well as the drivers of global processes. Furthermore, viruses likely represent the most extensive genetic and biological diversity on the planet, as revealed by genomic data. The diversity is defined by a wide array of virus types, including those with double-stranded (ds)DNA, single-stranded (ss)DNA, dsRNA, and ssRNA (either of positive or negative orientation) genomes.

Characterizing a viral community in the environment is a complicated task for two reasons: indeed, only a small percentage (<1%) of microbial hosts have been cultivated, and there is no one single gene that is common to all viral genomes. Thus, viral diversity cannot be evaluated using methods that are analogous to ribosomal RNA profiling. Viral metagenomic analyses of uncultured viral communities circumvent these limitations and can provide insights into the composition and structure of environmental viral communities. During the last two decades, deep sequencing experiments have opened new doors of opportunity to the reconstruction of viral populations in a high-throughput and cost-effective manner. Currently, a substantial number of studies have been performed employing next-generation sequencing (NGS) techniques either to analyze known viruses by means of a reference-guided approach or to discover novel viruses using a de novo-based strategy.

Viral metagenomics can be considered to consist of three main processes: (1) sampling for collecting viruses associated with cells, tissues, or environmental samples; (2) sequencing of nucleic acids, mainly using NGS techniques; and (3) bioinformatics analysis for interpreting the data. Once identified, environmental samples of interest are collected through a series of filtering processes to prevent contamination or artifacts. Afterwards, the samples are used for nucleic acid extraction, and specific DNA or cDNA libraries are generated and sequenced, mainly by means of NGS. The most commonly used state-of-the-art NGS techniques are Illumina and Roche 454 pyrosequencing, although third-generation sequencing technologies such as single-molecule real-time sequencing (Pacific Biosciences) and Oxford Nanopore are gaining importance among virus experts. Downstream NGS, viral metagenomics requires specific bioinformatics pipelines that are able to correctly manipulate and interpret data in order to provide answers concerning the viral community associated with the sample and diversity within each viral species. Preparation of the sequencing libraries is usually strictly linked to the sequencing platform. Therefore, the key steps of viral metagenomics are located both upstream and downstream NGS.

Viral metagenomics addresses the relevant queries raised by basic or applied research. Indeed, by means of viral metagenomics it is possible to diagnose known viruses for (1) plant certification or food production, (2) human and animal health, and (3) identifying viral vectors such as insects. Alternatively, viral metagenomics can help researchers to increase the knowledge base of viral communities in the environment, thus enhancing the interpretation of global processes (e.g. viral communities affecting microbiomes).

The chapters presented in this series lend robustness to the protocols developed within a research framework supported by either cutting-edge or driven funding. Several funding sources are the European Research Council (chapters by Cornelissen et al. and Varghese and Van Rji), National Institutes of Health, USA (chapter by Grasis), national research centers (i.e., INRA—France, CNR—Italy), the European intergovernmental organization ELIXIR (research infrastructures, chapter by Balech et al.), Hungarian Scientific Research Fund (OTKA, chapter by Czotter et al.), and the Ministry of the Environment, Government of Japan (Shimura et al.). This book also highlights goal-specific protocols, such as surveys on emergent viruses and vectors in the Mediterranean region (ARIMnet2—EMERAMB, chapter by Gutiérrez-Aguirre et al.), the sanitary status of crops (SaveGraInPuglia for the chapter by Ghasemzadeh et al. and other regional fundings by Regione Puglia as in the case of Navarro and Di Serio chapter), or supported by other fundings (fish farming in the chapter by Økland et al.). A specific chapter provides an overview of viral metagenomics in insects, a key topic that has been underestimated since insects are the principal vectors of plant and animal viruses.

This edition also explores viral metagenomics applied to several diverse specimens such as fish, bacteria, woody or herbaceous plants, pollen, arthropods, water, human blood, and fungi, including arbuscular mycorrhizae and those associated with marine algae. Space is also devoted to the techniques that are required for undertaking the three viral metagenomic steps of sampling, library construction, and interpretation of data from NGS.

A relevant number of contributions were initiated in the frame of the European Research Network (COST-Divas). We wish to thank also professor emeritus John M. Walker for his guidance in finalizing this book. We are also grateful to all the authors for their interesting contributions.

Bari, Italy

Vitantonio Pantaleo
Michela Chiumenti

Contents

Contributors

BACHIR BALECH · *Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, CNR, Bari, Italy*

DÁNIEL BARÁTH · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

BERNARD BERGEY · *UMR 1332, Biologie du Fruit et Pathologie, INRA, Univ. Bordeaux, Villenave d'Ornon, France*

BEN BERKHOUT · *Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands*

FRANÇOIS BLANQUART · *Department of Infectious Disease Epidemiology, Imperial College London, London, UK*

THIERRY CANDRESSE · *UMR 1332, Biologie du Fruit et Pathologie, INRA, Univ. Bordeaux, Villenave d'Ornon, France*

MICHELA CHIUMENTI · *Istituto per la Protezione Sostenibile delle Piante del CNR, Bari, Italy*

MARION CORNELISSEN · *Laboratory of Experimental Virology, Department of Medical Microbiology, Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands*

ADALBERTO COSTESSI · *Bioinformatics department, BaseClear B.V., Leiden, The Netherlands*

NIKOLETTA CZOTTER · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

KRIS DE JONGHE · *Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium*

EMESE DEMIÁN · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

FRANCESCO DI SERIO · *Istituto per la Protezione Sostenibile delle Piante, Consiglio Nazionale delle Ricerche, Bari, Italy*

GIACINTO DONVITO · *National Institute of Nuclear Physics, Bari, Italy*

CHANTAL FAURE · *UMR 1332, Biologie du Fruit et Pathologie, INRA, Univ. Bordeaux, Villenave d'Ornon, France*

EMMANUEL FERNANDEZ · *CIRAD, UMR BGPI, Montpellier, France; BGPI, CIRAD, NRA-UM, Montpellier SupAgro, Univ Montpellier, Montpellier, France*

DENIS FILLOUX · *CIRAD, UMR BGPI, Montpellier, France; BGPI, CIRAD, NRA-UM, Montpellier SupAgro, Univ Montpellier, Montpellier, France*

YOIKA FOUCART · *Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium*

SARAH FRANÇOIS · *NRA-UM, UMR DGIMI, Montpellier, France*

CHRISTOPHE FRASER · *Nuffield Department of Medicine, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK; Department of Infectious Disease Epidemiology, Imperial College London, London, UK*

Astrid Gall · *Department of Veterinary Medicine, University of Cambridge, Cambridge, UK*

Aysan Ghasemzadeh · *Institute for Sustainable Plant Protection, Research Unit of Bari, c/o University Campus of Bari, Bari, Italy; Faculty of Agriculture, Department of Plant Pathology, Tarbiat Modares University (T.M.U.), Tehran, Iran*

Annalisa Giampetruzzi · *Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

Juris A. Grasis · *School of Natural Sciences, University of California, Merced, Merced, CA, USA*

Ion Gutiérrez-Aguirre · *Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*

Marta Małgorzata ter Haar · *Bioinformatics Department, BaseClear B.V., Leiden, The Netherlands*

Annelies Haegeman · *Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium*

Yasunori Koda · *Research Faculty of Agriculture, Hokkaido University, Sapporo, Hokkaido, Japan*

Denis Kutnjak · *Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*

Antoinette van der Kuyl · *Laboratory of Experimental Virology, Department of Medical Microbiology, Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands*

Martine Maes · *Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium*

Giorgio Maggi · *National Institute of Nuclear Physics, Bari, Italy; Politecnico di Bari, Bari, Italy*

Armelle Marais · *UMR 1332, Biologie du Fruit et Pathologie, INRA, Univ. Bordeaux, Villenave d'Ornon, France*

Chikara Masuta · *Research Faculty of Agriculture, Hokkaido University, Sapporo, Hokkaido, Japan*

Ali May · *Bioinformatics department, BaseClear B.V., Leiden, The Netherlands*

Angelantonio Minafra · *Istituto per la Protezione Sostenibile delle Piante del CNR, Bari, Italy*

János Molnár · *Department of Biotechnology, Nanophage-therapy Center, Enviroinvest Corporation, Pécs, Hungary*

Alfonso Monaco · *National Institute of Nuclear Physics, Bari, Italy*

Beatriz Navarro · *Istituto per la Protezione Sostenibile delle Piante, Consiglio Nazionale delle Ricerche, Bari, Italy*

Luca Nerva · *Instituto per la Protezione Sostenibile delle Piante, CNR, Torino, Italy; Department of Life Sciences and Systems Biology, Mycotheca Universitatis Taurinensis (MUT), University of Turin, Torino, Italy*

Are Nylund · *Department of Biology, University of Bergen, Bergen, Norway*

Mylène Ogliastro · *NRA-UM, UMR DGIMI, Montpellier, France*

Arnfinn Lodden Økland · *Department of Biology, University of Bergen, Bergen, Norway*

Vitantonio Pantaleo · *Institute for Sustainable Plant Protection, Research Unit of Bari, c/o University Campus of Bari, Bari, Italy*

Michele Perniola · *National Institute of Nuclear Physics, Bari, Italy*

GRAZIANO PESOLE · *Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, CNR, Bari, Italy; Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "A. Moro", Bari, Italy*

RÉKA PESTI · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

WALTER PIROVANO · *Bioinformatics Department, BaseClear B.V., Leiden, The Netherlands*

NEJC RAČKI · *Lek Pharmaceuticals d.d., Menges, Slovenia*

SIMONE RAMPELLI · *Unit of Microbial Ecology of Health, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy*

MAJA RAVNIKAR · *Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*

RONALD P. VAN RIJ · *Department of Medical Microbiology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands*

PHILIPPE ROUMAGNAC · *CIRAD, UMR BGPI, Montpellier, France; BGPI, CIRAD, NRA-UM, Montpellier SupAgro, Univ Montpellier, Montpellier, France*

MATEVŽ RUPAR · *University of Nottingham, Nottingham, UK*

PASQUALE SALDARELLI · *Istituto per la Protezione Sostenibile delle Piante del CNR, Bari, Italy*

MONICA SANTAMARIA · *Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, CNR, Bari, Italy*

HANAKO SHIMURA · *Laboratory of Horticultural Science, Research Faculty of Agriculture, Hokkaido University, Sapporo, Hokkaido, Japan*

MASSIMO TURINA · *Instituto per la Protezione Sostenibile delle Piante, CNR, Torino, Italy*

SILVIA TURRONI · *Unit of Microbial Ecology of Health, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy*

ÉVA VÁRALLYAY · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

GIOVANNA C. VARESE · *Department of Life Sciences and Systems Biology, Mycotheca Universitatis Taurinensis (MUT), University of Turin, Torino, Italy*

TÜNDE VARGA · *Agricultural Biotechnology Institute, National Agricultural Research and Innovation Center, Gödöllő, Hungary*

FINNY S. VARGHESE · *Department of Medical Microbiology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, HB Nijmegen, The Netherlands*

SAVERIO VICARIO · *Department of Physics, Institute of Atmospheric Pollution Research (CNR), University of Bari "A. Moro", Bari, Italy*

CHRIS WYMANT · *Nuffield Department of Medicine, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK*

Host-Associated Bacteriophage Isolation and Preparation for Viral Metagenomics

Juris A. Grasis

Abstract

Prokaryotic viruses, or bacteriophages, are viruses that infect bacteria and archaea. These viruses have been known to associate with host systems for decades, yet only recently have their influence on the regulation of host-associated bacteria been appreciated. These studies have been conducted in many host systems, from the base of animal life in the Cnidarian phylum to mammals. These prokaryotic viruses are useful for regulating the number of bacteria in a host ecosystem and for regulating the strains of bacteria useful for the microbiome. These viruses are likely selected by the host to maintain bacterial populations. Viral metagenomics allows researchers to profile the communities of viruses associating with animal hosts, and importantly helps to determine the functional role these viruses play. Further, viral metagenomics show the sphere of viral involvement in gene flow and gene shuffling in an ever-changing host environment. The influence of prokaryotic viruses could, therefore, have a clear impact on host health.

Key words Metagenomics, Viral metagenomes, Virome, Microbiome, Prokaryotic virus, Bacteriophage, Host-microbe interactions, Holobiont, Symbiosis

1 Introduction

Prokaryotic viruses (commonly known by the anachronistic term bacteriophage, *see* **Note 1**) infect bacteria and archaea. There are two modes of the prokaryotic viral life cycle, lytic and temperate. The lytic phase involves the infection, replication, and lysis of the bacterium, leading to the death of the cell and escape of viral progeny. The temperate phase involves the integration of the prokaryotic virus into the genome of the bacterium in a proviral form, which when activated at a later time can then become a lytic virus.

The involvement of prokaryotic viruses in host-associated regulation of the microbiome has only recently been appreciated. These viruses likely help to regulate the number of bacteria and the strains of bacteria associating with a host. These associations have an impact on host metabolism [1], immunity [2], as well as animal health and disease [3]. To best evaluate the effects these viruses

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_1, © Springer Science+Business Media, LLC 2018

have on host ecosystems, it is useful to use viral metagenomics to fully assess which genes in the viral genomes have these effects on the microbiome. This protocol will allow the researcher to isolate, purify, and prepare viral nucleic acid for sequencing.

It is important to purify viruses because the total amount of viral nucleic acid is small compared to that of bacterial and host nucleic acid content [4]. The amount of nucleic acid per virus is estimated to be approximately 1 attogram (10^{-18} g). Most current sequencing platforms require at least 1 nanogram (10^{-9} g) of DNA. Therefore, if you have less than 10^9 viruses in your sample, you will need to enrich the viruses and amplify the DNA. Physical enrichment of viruses increases the number of viral sequences. Increased number of viral sequences means greater representation in databases, which will increase the likelihood of database matches. Further, enrichment of viruses for sequencing is necessary, as less than 5% of the unenriched metagenome sequences will contain viral sequences. The most effective purification method is a three-step technique utilizing centrifugation, filtration, and nuclease treatment. Using this technique eliminates more than 80% of host/bacterial contamination as compared to without purification, while using additional purification methods (e.g., density gradients) selects for certain populations of viruses. Similarly, use of viral DNA/RNA extraction kits is not recommended, as they create biases in viral populations [5]. Random shotgun sequencing libraries are recommended, as one can amplify a sample using barcoding primers. Even if you have enough DNA, it is recommended to exercise this protocol to barcode your samples for sequencing. It is not recommended to use multiple displacement amplification for amplifying viral DNA, as the Phi29 polymerase preferentially amplifies circular and single-stranded DNA and can, therefore, amplify contaminating host/bacterial DNA as well as small, circular viral DNA, which can alter the viral community profile prior to sequencing.

The use of a CsCl density gradient for purification of bacteriophages is up to the user. CsCl density gradients are effective at removing host and bacterial DNA, leaving more viruses for sequencing. This comes at a cost, as these gradients select for certain viruses while discriminating against others according to viral specific density. Further, reproducibility of CsCl density gradients is an issue [6]. It is therefore recommended to avoid using CsCl density gradients if accurate viral populations and reproducibility are of concern. If host and bacterial contaminations are of concern, this section is provided as an option. Also, an ultracentrifuge is needed for this optional step. Alternatively, one can repeat nuclease treatment prior to viral nucleic acid purification.

Although only two families of RNA bacteriophages have been described to date, this does not mean that the number and the diversity of RNA bacteriophages are less than those of DNA bacteriophages. More likely, the large percentage of DNA bacteriophages found in databases is due to the purification and selection method used (e.g., CsCl gradient selection) and consequent preferential sequencing of these selected viruses. Further, working with RNA samples is difficult due to the labile nature of RNA and the ubiquitous presence of RNases. Therefore, it is recommended to split the samples into DNA and RNA fractions to sample both DNA and RNA bacteriophages. Make sure that the working space, equipment, and reagents are as RNase free as possible. Keep in mind that the low number of characterized RNA bacteriophages will yield low numbers of hits in databases. However, as researchers sequence more RNA bacteriophages, more of these viruses will be characterized, and the RNA bacteriophage database will increase. Therefore, it is recommended to sequence both DNA and RNA bacteriophages.

This protocol takes the user from the isolation of bacteriophages (and other viruses, *see* **Note 2**) in host-associated environments, purification of bacteriophages, enumeration of viral-like particles (VLPs) through epifluorescence microscopy [7], characterization of VLPs through transmission electron microscopy [8, 9], extraction of viral DNA (vDNA) [10] and viral RNA (vRNA) [11, 12], and barcoding and amplification of viral nucleic acids for sequencing using Illumina sequencing technologies.

2 Materials

2.1 Extraction of Bacteriophages from Host Tissue

1. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
2. Saline magnesium (SM) buffer, pH 7.5, per 100 mL: To 80 mL molecular grade dH_2O add 0.58 g NaCl (100 mM), 5 mL 1 M Tris–HCl pH 7.4 (50 mM), 0.25 g $MgSO_4$ $7H_2O$ (10 mM), 0.1 g gelatin (0.1%, optional), adjust to pH 7.4, fill to 100 mL with molecular grade dH_2O, autoclave to sterilize, store at room temperature.
3. Microcentrifuge pestle (cleaned with 70% ethanol and RNA-Zap).
4. Handheld electric tissue homogenizer (cleaned with 70% ethanol and RNA-Zap).
5. 5 mL Sterile syringes.
6. 0.45 μm Syringe filters, PVDF, 28 mm diameter, sterile.
7. 0.02 μm Syringe filters, 25 mm diameter, sterile (Whatman Anotop-25).
8. Chloroform, use in a chemical hood.

9. DNase I (10 U/μL) and 10× DNase buffer (100 mM Tris pH 7.5, 5 mM $CaCl_2$, 25 mM $MgCl_2$).
10. RNase I (10 U/μL).
11. 4% Paraformaldehyde (PFA), per 10 mL: To 10 mL molecular grade dH_2O add 0.4 g PFA, heat at 50 °C for 1 h, fill to 10 mL with molecular grade dH_2O, 0.02 μm filter sterilize, good for 1 month at 4 °C.
12. 5% Formaldehyde (FA), per 10 mL: To 8 mL molecular grade dH_2O add 0.5 g formaldehyde, mix, and fill to 10 mL with molecular grade dH_2O, 0.02 μm filter sterilize, good for 1 month at 4 °C.

2.2 Purification of Bacteriophages (Optional)

1. Cesium chloride (CsCl) density step gradient—made in buffer used for resuspension of sample.
   (a) 1.7 g/mL (w/v): To 7.5 mL of resuspension buffer add 9.5 g CsCl, weigh 1 mL to adjust to final concentration of 1.7 g/mL with buffer or CsCl.
   (b) 1.5 g/mL (w/v): To 8.2 mL of resuspension buffer add 6.7 g CsCl, weigh 1 mL to adjust to final concentration of 1.5 g/mL with buffer or CsCl.
   (c) 1.35 g/mL (w/v): To 9.9 mL of resuspension buffer add 5.0 g CsCl, weigh 1 mL to adjust to final concentration of 1.35 g/mL with buffer or CsCl.
   (d) 1.2 g/mL (w/v): To 9.9 mL of resuspension buffer add 2.4 g CsCl, weigh 1 mL to adjust to final concentration of 1.2 g/mL with buffer or CsCl.
2. Ultracentrifuge tubes (Ultraclear 14 × 89 mm, Beckman Coulter).
3. SW41 Ti swinging bucket ultracentrifuge rotor.
4. Ultracentrifuge (Beckman Coulter).
5. 18-gauge needles.
6. 5 mL Syringes.

2.3 Viral-Like Particle Counts Using Epifluorescent Microscopy

1. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
2. 0.02 μm Anodisc.
3. SYBR Gold (10,000×).
4. Mounting solution, 0.1% ascorbic acid, 50% glycerol. Per 10 mL: To 4.9 mL molecular grade dH_2O add 100 μL of 10% ascorbic acid, then slowly add 5 mL glycerol, mix thoroughly, divide into 1 mL aliquots, and store at −20 °C.
5. Vacuum pump.
6. Microscopy slides.

7. Microscopy coverslips.
8. Forceps.
9. Petri dish.
10. Epifluorescent microscope with standard objective orientation (63× objective).

2.4 Transmission Electron Microscopy of Viral-Like Particles

1. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
2. 2% Uranyl acetate, pH to 4.5. Per 10 mL: To 9 mL molecular grade dH_2O add 0.2 g uranyl acetate, adjust pH to 4.5 with NaOH, and fill to 10 mL with molecular grade dH_2O; this chemical should be handled with care since it is both toxic and slightly radioactive [8, 9], and stored at 4 °C for up to 2 years.
3. Electron microscopy grids, 200–400 square mesh, copper, stabilized with a 2–10 nm thick carbon-layer Formvar film: Grids can become hydrophobic over time and may show poor adsorption of bacteriophages; therefore, it is recommended to treat the grids with UV light or poly-L-lysine to make the grids hydrophilic again.
4. Parafilm.
5. Forceps.
6. Filter paper.
7. Transmission electron microscope.

2.5 Viral DNA Extraction

1. 0.5 M EDTA, pH 8.0: Per 10 mL: To 8 mL molecular grade dH_2O add 1.86 g EDTA, adjust pH to 8.0 with NaOH (~200 μL), fill to 10 mL with molecular grade dH_2O, sterilize through 0.02 μm filter, and store at room temperature.
2. 2 M Tris–HCl, 0.2 M EDTA, pH 8.5: Per 10 mL: To 3 mL molecular grade dH_2O add 2.4 g Tris, 4 mL 0.5 M EDTA pH 8.0, adjust pH to 8.5 with HCl, fill to 10 mL with molecular grade dH_2O, sterilize through 0.02 μm filter, and store at room temperature.
3. Formamide (10 mL): Store at 4 °C.
4. Glycogen (20 mg/mL): Store at −20 °C.
5. 100% Ethanol (100 mL): Store at room temperature.
6. 1 M Tris–HCl pH 8.0: Per 10 mL: To 8 mL molecular grade dH_2O add 1.21 g Tris, adjust to pH 8.0, fill to 10 mL with molecular grade dH_2O, sterilize through 0.02 μm filter, and store at room temperature.
7. 10 mM Tris, 1 mM EDTA (TE), pH 8.0: Per 100 mL: To 80 mL molecular grade dH_2O add 1 mL 1 M Tris–HCl pH 8.0, add 2 mL 0.5 M EDTA, adjust pH to 8.0, fill to 100 mL with

molecular grade dH_2O, sterilize through 0.02 μm filter, and store at room temperature.

8. 10% Sodium dodecyl sulfate (SDS): Per 10 mL: To 10 mL molecular grade dH_2O add 1 g SDS, store at room temperature.
9. Proteinase K (20 mg/mL): Store at −20 °C.
10. 5 M NaCl: Per 10 mL: To 10 mL molecular grade dH_2O add 2.92 g NaCl, sterilize through 0.02 μm filter, and store at room temperature.
11. 10% Cetyltrimethyl ammonium bromide (CTAB), 700 mM NaCl: Per 10 mL: To 6 mL molecular grade dH_2O add 1 g CTAB, add 0.4 g NaCl, heat to 65 °C to dissolve CTAB/NaCl, fill to 10 mL with molecular grade dH_2O, store at room temperature, and preheat to 65 °C before use.
12. Chloroform (100 mL): Keep in glass container, store at room temperature, and use in a chemical hood.
13. Phenol:chloroform:isoamyl alcohol (25:24:1), pH 8.0 (100 mL): Keep in glass container, store at 4 °C, and use in a chemical hood.
14. Isopropanol (100 mL): Store at room temperature.
15. 70% Ethanol (100 mL).
16. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.6 Viral RNA Extraction

1. RNase-ZAP (Thermo Fisher Scientific) or 1% SDS: Per 100 mL: To 100 mL molecular grade dH_2O add 1 g SDS, store at room temperature.
2. Guanidine isothiocyanate RNA lysis buffer (GITC buffer, TRIzol LS, Thermo Fisher Scientific) (100 mL): Store at 4 °C.
3. 10 mM Dithiothreitol (DTT): Store at −20 °C.
4. Chloroform (100 mL): Keep in a glass container, store at room temperature, and use in a chemical hood.
5. Isopropanol (100 mL): Store at room temperature.
6. Glycogen (20 mg/mL): Store at −20 °C.
7. RNase-free 70% ethanol (100 mL): Use RNase-free molecular grade dH_2O to make 70% ethanol.
8. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.7 Quality Control to Determine Bacterial/Host Contamination

1. 10× *Taq* polymerase buffer: 100 mM Tris–HCl pH 8.4, 500 mM KCl, store at −20 °C.
2. BSA (1 mg/mL): Store at −20 °C.
3. 10 mM $MgCl_2$: Store at −20 °C.

4. 10 mM dNTPs: Store at −20 °C.
5. *Taq* polymerase (10 U/μL): Store at −20 °C.
6. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
7. Prokaryotic primers: Store at −20 °C:
   (a) 1 mM Eub27F—20-mer: 5′—AGR GTT TGA TCM TGG CTC AG—3′.
   (b) 1 mM Eub1492R—19-mer: 5′—GGH TAC CTT GTT ACG ACT T—3′.
8. Eukaryotic primers: Store at −20 °C:
   (a) 1 mM EukF—21-mer: 5′—AAC CTG GTT GAT CCT GCC AGT—3′.
   (b) 1 mM EukR—24-mer: 5′—TGA TCC TTC TGC AGG TTC ACC TAC—3′.

2.8 Viral RNA First-Strand Synthesis

1. Reverse Transcriptase (200 U/μL, SuperScript III, Thermo Fisher Scientific).
2. 5× Reverse transcriptase reaction buffer: 250 mM Tris–HCl pH 8.3, 375 mM KCl, 15 mM $MgCl_2$, store at −20 °C.
3. 10 mM dNTPs: Store at −20 °C.
4. 10 μM Random hexamer primer (N6): Store at −20 °C: 5′—NNN NNN—3′.
5. 50 nM Anchored oligo dT primer (dT18a): Store at −20 °C: 5′—TTT TTT TTT TTT TTT TTT VN—3′.
6. 10 mM Dithiothreitol (DTT): Store at −20 °C.
7. RNase inhibitor (40 U/μL): Store at −20 °C.
8. 10 mM $MgCl_2$: Store at −20 °C.
9. 1 M Dimethyl sulfoxide (DMSO): Store at −20 °C.

2.9 Viral Second-Strand Synthesis

1. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
2. T4 DNA polymerase (5 U/μL).
3. 10× T4 DNA polymerase buffer: 250 mM Tris-acetate pH 7.5, 1 M potassium acetate, 100 mM magnesium acetate, and 5 mM DTT, store at −20 °C.
4. 10 mM dUTPs: Store at −20 °C.
5. RNase H (250 U/μL): Store at −20 °C.
6. 10 mM DTT: Store at −20 °C.

2.10 DNA Cleanup

1. SPRI Beads (Agencourt AMPure XP Beads, Beckman Coulter): Store at 4 °C. Bring to room temperature before use.
2. Magnetic tube rack.

3. 70% Ethanol (70% EtOH).
4. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.11 DNA Shearing

1. Covaris shearing microtubes (microTUBE-50).
2. Covaris focused ultrasonicator (M220).

2.12 DNA End Repair

1. T4 DNA polymerase (5 U/μL).
2. 10× T4 DNA ligase reaction buffer: 500 mM Tris–HCl pH 7.5, 100 mM $MgCl_2$, 10 mM ATP, and 100 mM DTT, store at −20 °C.
3. T4 polynucleotide kinase (10 U/μL): Store at −20 °C.
4. 10 mM dNTPs: Store at −20 °C.
5. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.13 Adenylate 3′ Ends

1. DNA polymerase (5 U/μL), Klenow Fragment (3′–5′ exo-).
2. 10× Klenow reaction buffer: 500 mM NaCl, 100 mM Tris–HCl pH 7.9, 100 mM $MgCl_2$, 10 mM DTT, store at −20 °C.
3. 10 mM dATP: Store at −20 °C.
4. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.14 Adapter Ligation

1. Annealed adapters (see Section 2.18).
2. T4 DNA ligase (20 U/μL).
3. T4 DNA ligase reaction buffer: 500 mM Tris–HCl pH 7.5, 100 mM $MgCl_2$, 10 mM ATP, and 100 mM DTT, store at −20 °C.

2.15 Uridine Removal for Viral RNA Samples (Skip for Viral DNA Samples)

1. Uracil-DNA glycosylase (1 U/μL, UDG).
2. 10× UDG reaction buffer: 200 mM Tris–HCl, 10 mM DTT, 10 mM EDTA, pH 8.0, store at −20 °C.

2.16 Size Selection of Adapter-Ligated Fragments

1. SPRI Beads: Store at 4 °C, bring to room temperature before use.
2. Magnetic tube rack.
3. 70% Ethanol (70% EtOH).
4. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).

2.17 Large-Scale PCR of Size-Selected Adapter-Ligated Fragments

1. Molecular grade dH_2O (sterile, nuclease-free, virus-free dH_2O).
2. High-Fidelity DNA Polymerase (2 U/μL, NEB Q5).
3. 5× High-Fidelity DNA polymerase reaction buffer: Store at −20 °C.
4. 10 mM dNTPs: Store at −20 °C.

5. 1 mM PCR Primer 1 (Illumina TruSeq P5): Store at −20 °C: 5′- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GA—3′.
6. 1 mM PCR Primer 2 (Illumina TruSeq P7): Store at −20 °C: 5′- CAA GCA GAA GAC GGC ATA CGA GAT—3′.

2.18 Preparation of Adapters and Indices

1. Purchase Universal Adapter and Indexed Adapters: Universal adapter must have a 3′ phosphorothioate bond and indexed adapters must have 5′ end phosphorylated: universal adapter (Illumina TruSeq): 5′- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T—3′ {* = Phosphorothioate bond} and indexed adapters (Illumina TruSeq): 5′- P—GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC—NNN NNN—ATC TCG TAT GCC GTC TTC TGC TTG—3′ {5′ end needs to be phosphorylated, Ns are barcode}. Note that the latest Illumina TruSeq adapters are pre-annealed, so these steps may not be necessary. Check the Illumina TruSeq manual to make sure. The following steps are listed in the event the adapters are not pre-annealed.
   (a) Resuspend each adapter in molecular grade dH_2O at 570 μM (volume for 100 μM/5.7, store at −80 °C).
   (b) For annealing, make a working dilution of 10 μM by diluting 5 μL of 570 μM stock with 280 μL TE (store at −20 °C).
   (c) Mix equal volumes of the universal adapter (10 μM) with each of the indexed adapters (10 μM).
   (d) Either boil the mixes for 2 min and then cool slowly to room temperature—OR—use thermocycler program set to 1 cycle (Adapt_Anneal Program):
      - 95 °C for 10 min
      - 72 °C for 5 min
      - 60 °C for 5 min
      - 50 °C for 5 min
      - 40 °C for 5 min
      - 30 °C for 5 min
      - 20 °C for 5 min
      - 10 °C for 5 min
      - 4 °C Hold

   (e) Store the annealed adapters at −20 °C.

3 Methods

Carry out all procedures at 4 °C or on ice unless otherwise specified.

3.1 Extraction of Bacteriophages from Host Tissue

1. Resuspend sample in equal volume of molecular grade dH_2O or SM buffer (*see* **Note 3**) as volume of tissue in appropriate sized tube, e.g., microcentrifuge tube, Falcon tube.
2. Homogenize sample with handheld homogenizer at 4500 × *g* for 60 s or use microcentrifuge mortar and pestle until tissue becomes a slurry.
3. Centrifuge at slow speed (~2500 × *g*) for 20 min at 4 °C to pellet debris (eukaryotic and prokaryotic cells).
4. Optional viral precipitation step (if there is a lot of volume (> 5 mL), *see* **Note 4**). Add 10% (w/v) polyethylene glycol (PEG) 8000 and 1 M (final concentration) NaCl to sample and dissolve. Incubate overnight at 4 °C. Centrifuge to pellet at 4500 × *g* for 20 min at 4 °C. Resuspend pellet in 1 mL of same buffer used as before. Centrifuge at 4500 × *g* for 5 min at 4 °C.
5. Transfer supernatant to a new microcentrifuge tube.
6. Pre-wet 0.45 μm filter with 100 μL molecular grade dH_2O or SM buffer.
7. Filter supernatant through 0.45 μm filter to further remove unwanted debris and cells. Expect some volume loss due to filter retention.
8. Optional chloroform treatment step (*see* **Note 5**):
   (a) Add 0.2 volumes chloroform and mix by inversion or vortex to lyse any cells that have made it through filtration.
   (b) Incubate at 4 °C for 30 min, vortexing every 5 min.
   (c) Centrifuge at 4500 × *g* for 10 min at 4 ° C to pellet chloroform.
9. Add DNase buffer to filtered sample to 1× concentration.
10. Add DNase (final concentration 1 U/mL) (*see* **Note 6**).
11. Optional RNase treatment (final concentration 1 U/mL) (*see* **Note** 7).
12. Incubate for 2 h at 37 °C or overnight at room temperature.
13. Heat-inactivate DNase for 20 min at 65 °C (*see* **Note 8**).
14. Sample for epifluorescent microscopy: in a separate microcentrifuge tube transfer 15 μL sample and add 5 μL 4% PFA for a final concentration of 1% PFA.
15. Save 20 μL sample for electron microscopy. Hold at 4 °C until processed for electron microscopy. Add 5 μL 5% formaldehyde

for a final concentration of 1% formaldehyde to fix sample if not processed for a day or longer.

16. Use remaining ~1.0 mL for nucleic acid extraction and continue immediately to convenient stopping points. Split sample in half for viral DNA (vDNA) extraction (Section 3.5) and viral RNA (vRNA) extraction (Section 3.6).

3.2 Purification of Bacteriophages (Optional) [13]

1. For an overlay density step gradient, lay 1 mL of each step in the centrifuge tube starting with the highest density step (1.7 g/mL) and ending with the lowest density step (1.2 mg/mL). Be careful not to mix any of the steps while preparing the gradient.
2. Mark each layer with an indelible pen to note each layer.
3. Overlay the sample on top of the gradient, and fill to the top using the same buffer used for resuspension of the sample. Be careful not to disrupt the gradient.
4. Carefully balance each tube against each other to ensure that the mass is equivalent (<0.001 g difference between tubes); use resuspension buffer to balance the tubes.
5. Ultracentrifuge at ~82,000 × *g* using SW41 Ti rotor for 2 h at 4 °C; use a slow acceleration and no brake for the deceleration.
6. Carefully remove each tube from the rotor, making sure not to disrupt the gradient.
7. Using an 18-gauge needle, pierce the tube at the 1.5 g/mL layer mark and extract 1.5 mL of the 1.5 g/mL to 1.35 g/mL layer and interface (*see* **Note 9**).
8. Split the sample for vDNA and vRNA extractions (Sections 3.5 and 3.6).

3.3 Viral-Like Particle Counts Using Epifluorescent Microscopy (See Note 10) [7]

1. Dilute 10 μL fixed viral sample into 5 mL molecular grade dH_2O (1:500 dilution).
2. Filter sample onto 0.02 μm Anodisc filter under vacuum pressure of 10 mm Hg (10 psi or ~60 kPa). Filter is unidirectional, so be sure to have the shiny ring side up towards the sample.
3. Remove the filter tower and the filter while still under vacuum pressure.
4. Pipette 100 μL freshly made 1–5× SYBR Gold (to 995 μL molecular grade dH_2O add 5 μL10,000× SYBR Gold (50×, which can be stored at −20 °C for 1 month), then do another 1:10 dilution with molecular grade dH_2O to make 5× SYBR Gold) solution onto a Petri dish, and place the filter sample side up on the droplet to stain for 10 min at room temperature in the dark.

5. Using forceps, lift filter and place on another 100 μL droplet of molecular grade dH_2O on the Petri dish to wash the filter for 10 min at room temperature in the dark.
6. Pipette 10 μL of mounting solution onto a clean microscope slide to hold the filter in place.
7. Place stained and washed filter sample side up on microscope slide.
8. Add 10 μL of mounting solution on top of the filter and cover the filter with a coverslip. Be sure not to leave any air bubbles between the filter and the coverslip.
9. Count bacterial cells and viral-like particles (VLPs) under 485 nm light excitation using standard-orientation epifluorescence microscopy.
10. Count and document at least ten fields per slide. Average the number of VLPs and multiply by dilution factor and field/objective factor to obtain VLPs/mL.
11. Slides can be stored at −20 °C in a light-protected container.

3.4 Transmission Electron Microscopy of Viral-Like Particles [8, 9]

1. Place 100 μL 2% uranyl acetate into a droplet on parafilm.
2. Place another 100 μL molecular grade dH_2O into another droplet adjacent to the uranyl acetate droplet.
3. Place the Formvar-copper grid onto a clean area of the parafilm.
4. Pipette 10 μL of sample onto the copper side of the grid.
5. Allow the sample to adhere to the grid for 5 min.
6. Blot grid with filter paper.
7. Place grid sample side down on uranyl acetate droplet.
8. Allow sample grid to stain for 30 s.
9. Remove grid from uranyl acetate droplet and blot dry.
10. Place grid sample side down on dH_2O droplet.
11. Allow sample grid to wash for 30 s.
12. Remove grid from dH_2O droplet and blot dry.
13. Allow grid to air-dry for 5 min.
14. Visualize by transmission electron microscopy (>40,000×).

3.5 Viral DNA Extraction

1. Add 0.1 volume 2 M Tris–HCl pH 8.5/0.2 M EDTA to sample (e.g., add 50 μL to 500 μL sample).
2. Add 0.01 volume of 0.5 M EDTA to sample (e.g., add 5 μL to 500 μL sample).
3. Add 1 volume formamide (e.g., add 555 μL to 555 μL sample).
4. Add 1 μL glycogen (20 mg/mL) to each sample tube.

5. Incubate at room temperature for 30 min.
6. Split sample into two microcentrifuge tubes.
7. Add 2 volumes of room temperature 100% EtOH to each tube.
8. Centrifuge at 13,800 × *g* for 20 min to pellet.
9. Wash pellet 2× with 250 μL 70% EtOH.
10. Resuspend with 567 μL TE (10 mM Tris, 1 mM EDTA, pH 8.0).
11. *Optional stopping point: Store samples at −20 °C for up to 1 month.*
12. Add 30 μL 10% SDS.
13. Add 3 μL proteinase K (20 mg/mL).
14. Mix and incubate for 1 h at 37 °C.
15. Add 100 μL 5 M NaCl.
16. Add 80 μL CTAB/NaCl solution.
17. Mix and incubate at 65 °C for 10 min.
18. Add equal volume (~780 μL) of chloroform. Be sure to use chloroform in a chemical hood.
19. Mix and centrifuge at 13,800 × *g* for 10 min at room temperature.
20. Transfer top aqueous layer to a new microcentrifuge tube.
21. Add equal volume of 25:24:1 phenol:chloroform:isoamyl alcohol to supernatant. Be sure to use phenol:chloroform in a chemical hood.
22. Mix, and centrifuge at 13,800 × *g* for 10 min at room temperature.
23. Transfer top aqueous layer to a new microcentrifuge tube.
24. Add equal volume of chloroform to supernatant. Be sure to use chloroform in a chemical hood.
25. Mix, and centrifuge at 13,800 × *g* for 10 min at room temperature.
26. Transfer top aqueous layer to a new microcentrifuge tube.
27. Add 0.7 volumes isopropanol to supernatant.
28. Add 1 μL glycogen (20 mg/mL) to each sample tube.
29. Gently mix, and incubate at 4 °C overnight to precipitate DNA.
30. Centrifuge at 13,800 × *g* for 20 min at 4 °C.
31. Wash with cold 70% EtOH.
32. Allow air-drying for 15 min.
33. Resuspend with 50 μL molecular grade dH_2O.

34. Incubate for 5 min at 55 °C to promote resuspension.
35. Check DNA concentration using Qubit (Thermo Fisher Scientific) or spectrophotometer.
36. Check DNA quality using Agilent BioAnalyzer.
37. Check for host/bacterial contamination by 18S/16S PCR (Section 3.7).
38. Samples can be stored at 4 °C for 1–2 days, or at −20 °C for extended periods.

3.6 Viral RNA Extraction

1. Clean area with RNase-ZAP (or use 1% SDS solution) to ensure that working area is RNase free.
2. Split sample into another microcentrifuge tube (~250 μL per tube).
3. Add 3 volumes GITC buffer (e.g., add 750 μL GITC buffer to 250 μL sample).
4. Add 1 μL DTT (10 mM) to each sample tube.
5. *Optional stopping point: Store samples at −80 °C for up to 1 month.*
6. Add 0.2 volume chloroform (e.g., add 200 μL chloroform to 1 mL sample with GITC buffer), vortex, and incubate for 20 min at 4 °C.
7. Centrifuge at 13,800 × *g* for 20 min at 4 °C.
8. Transfer top aqueous layer into an RNase-free tube.
9. Add equal volume of isopropanol (e.g., add 500 μL isopropanol to 500 μL sample).
10. Add 1 μL glycogen (20 mg/mL).
11. Mix and incubate overnight at 4 °C to precipitate RNA.
12. Centrifuge at 13,800 × *g* for 20 min at 4 °C.
13. Wash with 250 μL cold RNase-free 70% EtOH.
14. Centrifuge at 13,800 x *g* for 5 min at 4 °C.
15. Repeat cold 70% EtOH wash.
16. Allow air-drying for 15 min.
17. Resuspend with 50 μL molecular grade dH_2O.
18. Incubate for 5 min at 55 °C to promote resuspension of RNA.
19. Check RNA concentration using Qubit or spectrophotometer.
20. Check quality of RNA using Agilent BioAnalyzer RNA Nano.
21. Check for host/bacterial contamination using 18S/16S PCR (Section 3.7).
22. Samples can be stored at −20 °C for 1–2 days, or at −80 °C for extended periods.

3.7 Quality Control to Determine Bacterial/Host Contamination

1. Standard PCR setup, per sample:
   - (a) 2.5 μL PCR buffer (10×).
   - (b) 1.0 μL BSA (1 mg/mL).
   - (c) 1.0 μL $MgCl_2$ (10 mM).
   - (d) 1.0 μL dNTPs (10 mM).
   - (e) 1.0 μL Taq polymerase (10 U/μL).
   - (f) 1.0 μL Primer 1 (1 mM).
   - (g) 1.0 μL Primer 2 (1 mM).
   - (h) 20 ng sample DNA.
   - (i) Use positive vDNA control and negative control (no template) in separate sample reactions (*see* **Note 11**).
   - (j) Fill to 25 μL with molecular grade dH_2O.
2. Touchdown PCR:
   - (a) 94 °C for 5 min.
   - (b) 94 °C for 30 s.
   - (c) 65 °C for 1 m: −1 °C/cycle.
   - (d) 72 °C for 2 min.
   - (e) Repeat steps (b) to (d) 14 cycles.
   - (f) 94 °C for 30 s.
   - (g) 50 °C for 1 min.
   - (h) 72 °C for 2 min.
   - (i) Repeat steps (f–h) 14 cycles.
   - (j) 72 °C for 10 min.
   - (k) Hold at 4 °C until analysis.
   - (l) Analyze for 18S/16S contamination by gel electrophoresis.

3.8 Viral RNA (vRNA) First-Strand Synthesis

1. Add 10 μL purified vRNA template.
2. Use a negative control of no vRNA template: Add 10 μL molecular grade dH_2O rather than template (*see* **Note 11**).
3. Use a positive control of known vRNA, if possible.
4. Add 4 μL 5× reverse transcriptase (RT) buffer.
5. Heat at 65 °C for 5 min to prime RNA.
6. Cool at 4 °C for 5 min.
7. Quick centrifugation spin to collect precipitation.
8. Primer mix setup, per sample:
   - (a) 1.0 μL dNTPs (10 mM).
   - (b) 1.0 μL N6 Primer (10 μM).
   - (c) 1.0 μL dT18a primer (50 nM) (*see* **Note 12**).

9. Add 3 μL primer mix to each 14 μL sample.
10. Heat at 72 °C for 5 min to anneal primers to template.
11. Cool at 4 °C for 5 min.
12. Quick centrifugation spin to collect precipitation.
13. First Strand Master Mix Setup, per sample:
    (a) 0.5 μL DTT (10 mM).
    (b) 0.5 μL RNase inhibitor (40 U/μL).
    (c) 0.5 μL $MgCl_2$ (10 mM).
    (d) 0.5 μL DMSO.
    (e) 1.0 μL Reverse transcriptase (200 U/μL).
14. Add 3 μL first-strand master mix to each 17 μL sample.
15. Run 1st_Strand program on thermocycler for first-strand synthesis:
    (a) 25 °C for 10 min (to initiate reverse transcription).
    (b) 42 °C for 60 min.
    (c) 50 °C for 1 min.
    (d) 42 °C for 1 min.
    (e) Repeat steps (c–d) 9 cycles.
    (f) 65 °C for 20 min (to inactivate).
    (g) 4 °C for 5 min.
    (h) 4 °C Hold.
16. Quick centrifugation spin to collect precipitation.
17. *Optional stopping point: Store samples overnight at 4 °C.*

3.9 Viral Second-Strand Synthesis

1. Use 20 μL first-strand vRNA converted to cDNA.
2. Use a negative control of no cDNA template (*see* **Note 11**).
3. Use a positive control of cDNA template.
4. Second-strand synthesis reaction per reaction:
    (a) 4.0 μL Molecular grade dH_2O.
    (b) 3.0 μL 10× T4 DNA polymerase buffer.
    (c) 1.0 μL dUTPs (10 mM stock).
    (d) 0.5 μL RNase H (250 U/μL).
    (e) 0.5 μL DTT (10 mM stock).
    (f) 1.0 μL T4 DNA polymerase (150 U/μL).
5. Add 10 μL second-strand master mix to 20 μL sample.
6. Run 2nd_Strand program on thermocycler:
    (a) 4 °C for 2 min.
    (b) 16 °C for 90 min.

   (c) 65 °C for 20 min (to inactivate DNA polymerase).
   (d) 4 °C for 5 min.
   (e) 4 °C Hold.
7. Quick centrifugation spin to collect precipitation.

3.10 DNA Cleanup

1. Allow SPRI beads to equilibrate to room temperature. Vortex to mix well.
2. Add equal volumes of sample and vortexed SPRI beads into microcentrifuge tube (e.g., add 30 μL SPRI beads to 30 μL sample).
3. Vortex briefly and incubate at room temperature for 5 min.
4. Place microcentrifuge tube on magnet holder and allow for beads to bind to magnet at the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
5. Keep tube on magnet and remove all supernatant.
6. While leaving the tube on the magnet, wash twice with 250 μL freshly made 70% EtOH, and remove all liquid between washes.
7. Remove tube from magnet and allow beads to air-dry for 5 min, or until EtOH has fully evaporated.
8. Resuspend beads in 52.5 μL molecular grade H_2O, vortex, and place tube back on magnet. Allow beads to bind to magnet for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
9. Transfer 50 μL of eluted sample from beads and place in a new microcentrifuge tube.

3.11 DNA Shearing

1. Transfer 52.5 μL DNA (vDNA and molecular grade dH_2O to 52.5 μL volume) to Covaris shearing tube.
2. Briefly spin to move sample to the bottom of the tube.
3. Turn on Covaris machine, open SonoLab program, and fill water bath to appropriate level with DI water.
4. Run DNA_Shear_45Sec program to shear DNA:
   (a) Covaris shearing settings:
      - 45 s
      - 50 W peak power
      - Duty factor 20%
      - 200 cycles/burst
      - Room temperature
5. Quick centrifugation spin to collect precipitation.

6. Transfer sheared DNA to PCR tube to continue with end repair.
7. Close SonoLab program, turn off Covaris machine, and clean and dry water bath.

3.12 End Repair

1. End-repair reaction setup, per sample reaction:
   - (a) 6.0 μL 10× T4 DNA ligase buffer (contains ATP)
   - (b) 2.0 μL Molecular grade dH_2O
   - (c) 1.0 μL 10 mM dNTPs
   - (d) 0.5 μL 5 U/μL T4 DNA polymerase
   - (e) 0.5 μL 10 U/μL T4 polynucleotide kinase (PNK)
2. Add 10 μL of end-repair master mix to 50 μL sheared DNA.
3. Run End_Repair program on thermocycler:
   - (a) 4 °C for 2 min
   - (b) 20 °C for 30 min
   - (c) 25 °C for 30 min
   - (d) 4 °C for 5 min
   - (e) 4 °C Hold
4. Add 60 μL SPRI beads. Vortex before use.
5. Repeat DNA cleanup steps (Section 3.10).
6. Elute with 50 μL molecular grade dH_2O.

3.13 Adenylate 3′ Ends

1. Adenylation reaction setup, per sample reaction:
   - (a) 2.0 μL Molecular grade dH_2O
   - (b) 6.0 μL 10× Klenow buffer
   - (c) 1.0 μL 10 mM dATP
   - (d) 1.0 μL 5 U/μL Klenow fragment DNA polymerase (3′-5′ exo-)
2. Add 10 μL of adenylation master mix to 50 μL end-repaired DNA.
3. Run DA_Tail program on thermocycler:
   - (a) 4 °C for 2 min
   - (b) 37 °C for 30 min
   - (c) 4 °C Hold
4. Add 60 μL SPRI beads. Vortex before use.
5. Repeat DNA cleanup steps (Section 3.10).
6. Elute with 50 μL molecular grade dH_2O.

3.14 Adapter Ligation

1. Primary adapter ligation setup, per sample reaction:
   - (a) 6.0 μL 10× T4 DNA ligase buffer (contains 10 mM ATP)

(b) 1.0 μL 10 μM Annealed indexed adapter (*see* **Note 13**)

(c) 0.5 μL 20 U/μL T4 DNA ligase

2. Add 6.5 μL of primary adapter ligation master mix to 50 μL 3′ adenylated DNA + 1 μL each annealed indexed adapter.
3. Run End_Repair program on thermocycler:

   (a) 4 °C for 2 min

   (b) 20 °C for 30 min

   (c) 25 °C for 30 min

   (d) 4 °C Hold
4. Secondary adapter ligation setup, per sample reaction:

   (a) 11.0 μL Molecular grade dH_2O

   (b) 1.0 μL T4 DNA ligase buffer (contains 10 mM ATP)

   (c) 0.5 μL 20 U/μL T4 DNA ligase
5. Add 12.5 μL of secondary adapter ligation master mix to 57.5 μL of primary adapter ligation.
6. Run Adap_Lig program on thermocycler:

   (a) 4 °C for 2 min

   (b) 16 °C for 2 h

   (c) 4 °C Hold
7. Add 70 μL SPRI beads. Vortex before use.
8. Repeat DNA cleanup steps (Section 3.10).
9. Elute with 50 μL molecular grade dH_2O.

3.15 Uridine Removal for Viral RNA Samples (Omit for Viral DNA Samples)

1. Add 5 μL UDG reaction buffer.
2. Add 0.5 μL UDG.
3. Incubate at 37 °C for 30 min.
4. Add 55 μL SPRI beads. Vortex before use.
5. Repeat DNA cleanup steps (Section 3.10).
6. Elute with 50 μL molecular grade dH_2O.

3.16 Size Selection of Adapter-Ligated Fragments

3.16.1 For a 550 bp Insert (for 300 bp Paired-End Sequencing)

1. For each sample, pipette 40 μL SPRI beads into a new microcentrifuge tube. Vortex before use.
2. Dilute beads with 40 μL molecular grade dH_2O.
3. Transfer 50 μL adapter-ligated sample into diluted SPRI bead centrifuge tube.
4. Vortex to mix.
5. Incubate at room temperature for 5 min.
6. Place microcentrifuge tube on magnet holder and allow for beads to bind to magnet at the side of the tube for 5 min or

longer, until bead pellet has formed on the magnet side of tube and solution has cleared.

7. Transfer supernatant to a new microcentrifuge tube. Discard tube with beads.
8. Add 30 μL undiluted and vortexed SPRI beads to supernatant tube.
9. Vortex to mix.
10. Incubate at room temperature for 5 min.
11. Place microcentrifuge tube on magnet holder and allow for beads to bind to magnet at the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
12. Keep tube on magnet and remove all supernatant.
13. While leaving the tube on the magnet, wash twice with 250 μL freshly made 70% EtOH, and remove all liquid between washes.
14. Remove tube from magnet and allow beads to air-dry for 5 min, or until EtOH has fully evaporated.
15. Resuspend beads in 52.5 μL molecular grade H_2O, vortex, and place tube back on magnet. Allow beads to bind to the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
16. Transfer 50 μL of eluted sample from beads and place in a new microcentrifuge tube.

3.16.2 For a 250 bp Insert (for 150 bp Paired-End Sequencing)

1. For each sample, pipette 55 μL SPRI beads into a new microcentrifuge tube. Vortex before use.
2. Dilute beads with 25 μL molecular grade dH_2O.
3. Transfer 50 μL adapter-ligated sample into diluted SPRI bead centrifuge tube.
4. Vortex to mix.
5. Incubate at room temperature for 5 min.
6. Place microcentrifuge tube on magnet holder and allow for beads to bind to magnet at the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
7. Transfer supernatant to a new microcentrifuge tube. Discard tube with beads.
8. Add 30 μL undiluted and vortexed SPRI beads to supernatant tube.
9. Vortex to mix.

10. Incubate at room temperature for 5 min.
11. Place microcentrifuge tube on magnet holder and allow for beads to bind to magnet at the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
12. Keep tube on magnet and remove all supernatant.
13. While leaving the tube on the magnet, wash twice with 250 μL freshly made 70% EtOH, and remove all liquid between washes.
14. Remove tube from magnet and allow beads to air-dry for 5 min, or until EtOH has fully evaporated.
15. Resuspend beads in 52.5 μL molecular grade H_2O, vortex, and place tube back on magnet. Allow beads to bind to the side of the tube for 5 min or longer, until bead pellet has formed on the magnet side of tube and solution has cleared.
16. Transfer 50 μL of eluted sample from beads and place in a new microcentrifuge tube.

3.17 Large-Scale PCR

1. PCR setup, per sample reaction:
   (a) 6.5 μL Molecular grade dH_2O
   (b) 5.0 μL 5× High-fidelity polymerase buffer
   (c) 1.0 μL 10 mM dNTPs
   (d) 1.0 μL 1 μM PCR Primer 1 (Illumina TruSeq P5)
   (e) 1.0 μL 1 μM PCR Primer 2 (Illumina TruSeq P7)
   (f) 0.5 μL 2 U/μL High-fidelity DNA polymerase
2. Add 15 μL of master mix to 10 μL adapter-ligated and size-selected DNA.
3. Run LargeScale program on thermocycler:
   (a) 95 °C for 5 min
   (b) 95 °C for 30 s
   (c) 60 °C for 60 s
   (d) 72 °C for 90 s
   (e) Repeat steps (b–d) using optimal # cycles (*see* **Note 14**)
   (f) 72 °C for 10 min
   (g) 4 °C Hold
4. Reconditioning PCR setup, per sample reaction (increases yield and decreases heteroduplex formation):
   (a) 11.5 μL Molecular grade dH_2O
   (b) 10.0 μL 5× High-fidelity polymerase buffer

   (c) 1.0 μL 10 mM dNTPs
   (d) 1.0 μL 1 μM PCR Primer 1 (Illumina TruSeq P5)
   (e) 1.0 μL 1 μM PCR Primer 2 (Illumina TruSeq P7)
   (f) 0.5 μL 2 U/μL Q5 High-fidelity DNA polymerase

5. Add 25 μL of master mix to 25 μL amplified DNA.
6. Run Recondition program on thermocycler:
   (a) 95 °C for 2 min
   (b) 95 °C for 30 s
   (c) 60 °C for 60 s
   (d) 72 °C for 90 s
   (e) Repeat steps (b–d) 3 cycles
   (f) 72 °C for 10 min
   (g) 4 °C Hold
7. Perform DNA cleanup steps (Section 3.10).
8. Elute in 55 μL molecular grade dH_2O.
9. Quantify by Qubit.
10. Quality control by BioAnalyzer.
11. Samples can be stored at 4 °C for 1–2 days, or at −20 °C for extended periods.

4 Notes

1. The term "bacteriophage" was originally used to describe entities that "ate" bacteria [14]. Hence, the Greek root meaning of "phage" is to eat, or bacteriophage, the "eater of bacteria." Prokaryotic viruses do not "eat" bacteria; they either lyse cells or integrate into genomes upon infection. Therefore, it is more accurate to term these viruses as what they infect, prokaryotic viruses.
2. Although the title of this chapter is for the isolation of bacteriophages from host-associated systems, you will also isolate eukaryotic viruses. This protocol is good for viral metagenome studies; simply substitute the word "virus" for "bacteriophage," and one will be able to obtain a full viral metagenome. If the goal is to select for prokaryotic viruses, it is recommended to use the optional CsCl step density gradient purification (Section 3.2).
3. Other buffers to consider using other than SM buffer are 1% potassium citrate (to 100 mL, add 1 g potassium citrate, 0.14 g Na_2HPO_4 $7H_2O$, and 0.024 g KH_2PO_4, pH to 7) and 10 mM sodium pyrophosphate (to 100 mL, add 0.27 g $Na_2P_2O_7$)

[15]. These buffers are supposed to help detach viruses from tissue. We have not had much success with these buffers, and have had the most success with SM buffer. These alternatives are included in case the user would like to try these buffers.

4. It is important to use small volumes in this protocol to help concentrate the sample and to work with less tubes throughout the procedure. Therefore, if one has large volumes of tissue slurry, it would be useful to precipitate the viruses into a smaller volume using pegylation [16]. Alternatively, Amicon filters (50 kD) can be used to concentrate viruses. Note that this will affect viral number counts, so if these counts are important, do not precipitate viruses until after sub-sampling for epifluorescence microscopy.

5. Chloroform is used at this step to lyse lipid membranes, particularly that of eukaryotic and prokaryotic cells. However, chloroform can also affect (and lyse) enveloped viruses. It is not recommended to use chloroform at this step because it can affect enveloped viruses while increasing the amount of host and bacterial DNA in the sample. Rather, an additional 0.45 μm filtration step can be used to remove eukaryotic and prokaryotic cells while preserving the enveloped viruses. This step is included as optional for those wishing to use this step.

6. Free DNA is common in host-associated samples. Further, released DNA through filtration of cells (unlikely) or chloroform lysis of cells (most likely) can obscure viral DNA isolation later in the protocol. It cannot be stressed enough how important it is to DNase your samples, as it affects viral number counts (*see* **Note 10**), and can affect what is sequenced. If host or bacterial contamination is observed, a second DNase step should be added during the isolation of viruses in future experiments.

7. It is optional to treat the sample with RNase. RNase is useful to remove prevalent rRNAs from a host-associated sample (>80% of the sample). This is particularly necessary when isolating RNA viruses. However, it is difficult to remove RNases from a sample once added. They are not heat-inactivated. It is necessary to use DTT or 2-mercaptoethanol to inactivate the RNases prior to RNA isolation.

8. Heat inactivation works for DNase, but not for RNase. It is necessary to use DTT or 2-mercaptoethanol to inactivate the RNases prior to RNA isolation.

9. After removal of sample with needle, the hole left behind will cause the rest of the gradient to come out of the tube, so have another waste container available to collect it.

10. The enumeration of viruses is rife with pitfalls. Often, viral-like particle (VLP) counts are overestimated in host-associated

systems due to not eliminating host-generated factors (such as vesicles and free DNA) [17]. Therefore, it is important that VLP counts are conducted after filtration, and importantly after DNase treatment to eliminate host factors. When enumerated properly, the VLP numbers are usually equivalent to bacterial numbers in host-associated systems. Please refer to the work by Ortmann and Suttle for reference [7].

11. It is of vital importance to use both positive and negative controls throughout the random amplification library preparation and during the PCRs. A positive control is needed to ensure that your protocol is working throughout. More importantly, a negative control is needed to ensure that contaminating DNA is not infiltrating your protocol and potentially sequenced as a false positive. These controls are often erroneously omitted from the protocol, but if not included false outcomes can creep into your sequencing results.
12. The VN-anchored oligo dT primer allows the primer to only anneal to the 5′ end of the poly(A) tail of mRNA, allowing for more efficient cDNA synthesis.
13. Excess adapters can interfere with sequencing [18]. The adapters have to be diluted relative to the starting material. Using 10 μM adapter stock, for samples >100 ng, do not dilute. For samples 10–100 ng, make a 1:10 dilution of 10 μM adapter stock to 1 μM. For samples 1–10 ng, make a 1:20 dilution, or 500 nM. For samples <1 ng, make a 1:30 dilution, or 250 nM. Add 1 μL of these diluted adapters to each reaction.
14. Use the minimum number of cycles to barcode and amplify the amount of DNA needed to sequence [19]. One should ensure that each fragment of DNA has an adapter on it for sequencing. To achieve this, use the minimum number of cycles of 8, so if starting with more than 1 ng of DNA, use 8 cycles. If one has 1 ng of DNA or less, use 12 cycles. If one has 1 pg of DNA or less, use 21 cycles. Keep in mind that amplifying with too many cycles can lead to amplification bias and sequencing artifacts, so use the minimum number of cycles to obtain the amount of DNA needed to sequence.

Acknowledgements

This work was supported by NIH grant F32AI098418. Special thanks to Marisa Rojas for critical review of the protocol. The author declares no competing financial interests.

References

1. Grasis JA, Lachnit T, Anton-Erxleben F, Lim YM, Schmieder R, Fraune S et al (2014) Species-specific viromes in the ancestral holobiont hydra. PLoS One 9:e109952. https://doi.org/10.1371/journal.pone.0109952
2. Duerkop BA, Clements CV, Rollins D, Rodrigues JL, Hooper LV (2012) A composite bacteriophage alters colonization by an intestinal commensal bacterium. Proc Natl Acad Sci U S A 109:17621–17626. https://doi.org/10.1073/pnas.1206136109
3. Cadwell K (2015) The virome in host health and disease. Immunity 42:805–813. https://doi.org/10.1016/j.immuni.2015.05.003
4. Daly GM et al (2011) A viral discovery methodology for clinical biopsy samples utilizing massively parallel next generation sequencing. PLoS One 6:e28879. https://doi.org/10.1371/journal.pone.0028879
5. Hall RJ et al (2014) Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. J Virol Meth 195:194–204. https://doi.org/10.1016/j.jviromet.2013.08.035
6. Kleiner M et al (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics 16:7. https://doi.org/10.1186/s12864-014-1207-4
7. Ortmann AC, Suttle CA (2009) Determination of virus abundance by epifluorescence microscopy. In: Clokie MR, Kropinski AM (eds) Bacteriophages: methods and protocols, volume 1: isolation, characterization, and interactions, vol 501. Humana Press, New York, pp 87–95. https://doi.org/10.1007/978-1-60327-164-6_10
8. Ackermann H-W, Heldal M (2010) Basic electron microscopy of aquatic viruses. In: Wilhelm SW, Weinbauer MG, Suttle CA (eds) Manual of aquatic viral ecology, vol 18. American Society of Limnology and Oceanography, Waco, TX, pp 182–192. https://doi.org/10.4319/mave.2010.978-0-09845591-0-7.182
9. Ackermann H-W (2009) Basic phage electron microscopy. In: Clokie MR, Kropinski AM (eds) Bacteriophages: methods and protocols, volume 1: isolation, characterization, and interactions, vol 501. Humana Press, New York, pp 113–126. https://doi.org/10.1007/978-1-60327-164-6_12
10. Lim YW et al (2014) Purifying the impure: sequencing metagenomes and metagenomes from complex animal-associated samples. J Vis Exp 94:e52117. https://doi.org/10.3791/52117
11. Culley AI, Suttle CA, Steward GF (2010) Characterization of the diversity of marine RNA viruses. Manual of Aquatic Viral Ecol. 19:193–201. https://doi.org/10.4319/mave.2010.978-0-9845591-0-7.193
12. Weynberg KD et al (2014) Generating viral metagenomes from the coral holobiont. Front Microbiol 5:1–11. https://doi.org/10.3389/fmicb.2014.00206
13. Lawrence JE, Steward GF (2010) Purification of viruses by centrifugation. Manual of Aquatic Viral Ecol 17:166–181. https://doi.org/10.4319/mave.2010.978-0-9845591-0-7.166
14. Summers WC (1999) Felix d'Herelle and the origins of molecular biology. Yale University Press, USA, p 248
15. Williamson KE et al (2003) Sampling natural viral communities from soil for culture-independent analyses. Appl Environ Micro 69:6628–6633. https://doi.org/10.1128/AEM.69.11.6628-6633.2003
16. Hjelmso MH et al (2017) Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. PLoS One 12:e0170199. https://doi.org/10.1371/journal.pone.0170199
17. Forterre P, Soler N, Krupovic M, Marguet E, Ackermann H-W (2013) Fake virus particles generated by fluorescence microscopy. Trends Microbiol 21:1–5. https://doi.org/10.1016/j.tim.2012.10.005
18. Solonenko SA et al (2013) Sequencing platform and library preparation choices impact viral metagenomes. BMC Genomics 14:320. https://doi.org/10.1186/1471-2164-14-320
19. Duhaime MB, Deng L, Poulos BT, Sullivan MB (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environ Microbiol 14:2526–2537. https://doi.org/10.1111/j.1462-2920.2012.02791.x

Chapter 2

Small RNA Isolation from Tissues of Grapevine and Woody Plants

Annalisa Giampetruzzi, Michela Chiumenti, Angelantonio Minafra, and Pasquale Saldarelli

Abstract

A protocol is described to purify small (s)RNA molecules from tissues of grapevine and other woody plants. The protocol has been specifically developed to analyze sRNA populations by high-throughput sequencing. It has been widely used on species of the genera *Prunus* and *Vitis* particularly rich in polyphenols and other enzyme-inhibiting compounds. The high quality of the sRNAs extracted from leaf or phloem tissues makes them suitable for all molecular biology reactions, in particular for next-generation sequencing library preparation.

Key words Small RNA enrichment, Woody plants, Grapevine

1 Introduction

sRNA molecules of 21–24 nucleotides in size are produced during the process of RNA silencing, a pathway that posttranscriptionally regulates mRNA levels in plants [1–4]. Viral small interfering RNAs (vsiRNAs) are by-products of RNA silencing enabling plants to defend against viral invasion [5]. The route, elegantly defined a "plant immune system," degrades RNA molecules of DNA and RNA viruses and viroids to 21–24 small RNAs. The presence of vsiRNAs in plant tissues, which are detected by high-throughput sequencing (HTS), is a hallmark of viral infections [6, 7]. Therefore, the technique generically detects vsiRNAs from known and unknown viruses and viroids. We have developed the present protocol to assess the sanitary status of grapevine and other woody plants by HTS of libraries of sRNAs.

Notoriously, RNA extraction from leaves is largely limited by the presence of polyphenols, polysaccharides, and pigments, which co-precipitate with nucleic acids [8, 9], a phenomenon exacerbated when total RNAs have to be extracted from phloem tissue scrapings of cuttings from woody species. A number of protocols aiming at the

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_2, © Springer Science+Business Media, LLC 2018

limitation of tissue oxidization and elimination of these compounds have been implemented throughout the time to extract grapevine total RNAs [8–11]. The use of an extraction buffer containing chaotropic salts that denature proteins and inactivate RNAses and a polymeric matrix (polyvinylpyrrolidone) to adsorb polyphenols and inhibitory compounds was a breakthrough in the extraction of total RNAs suitable for application in molecular biology techniques [10]. Successive enrichment of low-molecular-weight RNAs is obtained by polyethylene glycol (PEG) precipitation and separation of 21–24 small RNAs is achieved by their selective elution from denaturing polyacrylamide gel [12]. Our protocol has been successfully used to purify high-yield and high-quality sRNAs either from leaves or phloem tissues of grapevines or other woody plants. The obtained sRNAs are free from inhibitory compounds and suitable for synthesizing cDNA libraries to be analyzed by HTS techniques or every other molecular biology techniques.

2 Materials

All solutions are prepared with ultrapure water (double-distilled molecular grade, ddH$_2$O) and analytical grade reagents. Where indicated, solutions are sterilized by autoclaving. Commercial RNase-free water can be used to dissolve nucleic acid pellets. Particular care should be devoted to avoiding RNase contaminations using gloves and RNase-decontaminating surface reagents. These precautions prevent the use of the RNase inhibitor diethyl pyrocarbonate (DEPC) to water solutions, which is toxic and hazardous.

2.1 Total RNA Extraction

1. Extraction buffer (EB), pH 5, per 1 L: To 600 mL molecular grade ddH$_2$O add 472 g guanidine thiocyanate (4 M) and dissolve by stirring; successively add 16.4 g sodium acetate (0.2 M), 9.2 g ethylenediaminetetraacetic acid sodium salt (Na$_2$EDTA, 25 mM), 25 g polyvinylpyrrolidone-40 (2.5%) (*see* **Note 1**); mix by adding water to 800 mL; and adjust pH to 5 with 1 N HCl. Bring the volume to 1 L in a cylinder. Store at 4 °C and protect from light.
2. 20% N-lauroylsarcosine (NLS), for 100 mL: Weight 20 g N-Lauroylsarcosine sodium salt and transfer to a 100 mL beaker (*see* **Note 2**), add water to 90 mL and dissolve by stirring, bring the volume to 100 mL with a cylinder, and store at room temperature.
3. Trizol Reagent (Thermo Fisher Scientific).
4. 2-Mercapthoethanol: Use in a chemical hood.
5. 75% Ethanol: Add 75 mL absolute ethanol to 25 mL water in a cylinder, store at −20 °C.

6. Chloroform: Use in a chemical hood.
7. Isopropanol: Store at −20 °C.
8. Conical tubes (15 mL) (Falcon).
9. Refrigerated bench centrifuge Thermo Scientific Heraeus Multicentrifuge 3SR+ (rotor swingle 6441 15 mL tubes).
10. Ultracentrifuge polycarbonate bottles (Tube 26.3 mL).
11. Ultracentrifuge (Beckman Coulter Avanti J-25 with rotor JA 25–50).
12. Benchtop microcentrifuge.
13. Single-channel pipettes (2–20 μL/20–200 μL/1000 μL).
14. 20 μL Pipette filter tips
15. 200 μL Pipette filter tips.
16. 1000 μL Pipette filter tips.
17. Vortex.
18. Sterilized mortars.
19. Liquid nitrogen.
20. Heat block or water bath at 70 °C.
21. RNase-free water.

2.2 Separation of Low- and High-Molecular-Weight RNAs

1. 20% PEG solution: Dissolve 20 g polyethyleneglycol (PEG) (Molecular Weight 8000) in 60 mL RNase-free water, add 11.6 g sodium chloride, and bring the volume to 100 mL in a cylinder.
2. 10× TBE, gel running buffer: Weight 121 g Tris base, 51.3 g boric acid, and 3.7 g Na_2EDTA in 750 mL RNase-free water, correct pH to 8 with sodium hydroxide pellets, bring the volume to 1 l in a cylinder, autoclave for 20 min at 121 °C, and store at room temperature.
3. 1× TBE, for 100 mL: Add 10 mL of 10× TBE to 90 mL of sterile distilled water.
4. 1.2% Agarose gel: Weigh 1.2 g agarose electrophoresis grade and add to 100 mL 1× TBE in a 250 mL flask. Dissolve completely the agarose by heating in a microwave. Pour the gel into a tray with combs and wait until it solidifies.
5. 6× DNA loading dye containing xylene cyanol and bromophenol blue.
6. Spin-X Centrifuge Tube Filters, 0.45 μm (Costar, Corning Inc.).
7. Tray with comb.
8. Electrophoresis apparatus.
9. Absolute ethanol.

10. 75% Ethanol.
11. Microtubes (2 mL) (Eppendorf).
12. Ice.
13. Refrigerate benchtop microcentrifuge.
14. Single-channel pipettes (2–20 μL/20–200 μL/1000 μL).
15. 20 μL Pipette filter tips
16. 200 μL Pipette filter tips.
17. 1000 μL Pipette filter tips.
18. Vortex.
19. Heat block or water bath at 70 °C.
20. Formamide.
21. UV transilluminator.

2.3 Purification of 18–30 nt sRNAs from the LMW RNA Fractions

1. 10% Ammonium persulfate (APS): 10% Solution (w/v) in water (*see* **Note 3**).
2. N,N,N′,N′-tetramethylethylenediamine (TEMED): Store at 4 °C.
3. 15% Acrylamide, 8 M UREA gel in 0.5× TBE, for 12.5 mL: Dissolve 5.25 g urea in 3 mL water, add 0.625 mL 10× TBE and 4.7 mL 40% acrylamide (19:1 acrylamide: bis-acrylamide) (*see* **Note 4**), add water to a volume, finally add 87.5 μL 10% APS and 4.4 μL TEMED, and pour immediately the gel.
4. DNA ladder: 21 nt long single-stranded DNA at 20 ng/μL.
5. 0,3 M NaCl elution buffer, for 100 mL: Dissolve 1.74 g NaCl in 100 mL water, sterilize by autoclaving, and store at room temperature.
6. Resuspension buffer, 10 mM Tris–HCl (pH 8.5): Weigh 0.12 g Tris–HCl and transfer to the cylinder. Add water to a volume of 90 mL. Mix and adjust pH with 1 N HCl (*see* **Note 5**), make up to 100 mL with water, sterilize by autoclaving, and store at 4 °C.
7. Glasses of a Mini-Protean Tetra gel 1 mm and 10 × 8 cm size (Biorad).
8. Mini-Protean electrophoresis system (Biorad).
9. Razor blades.
10. 10 mg/mL Ethidium bromide.
11. Plastic vessel slightly bigger than the gel.
12. Microtubes (0.5, 1.5, and 2 mL) (Eppendorf).
13. 6× DNA loading dye containing xylene cyanol and bromophenol blue.

14. Benchtop microcentrifuge.
15. Single-channel pipettes (2–20 μL/20–200 μL/1000 μL).
16. 20 μL Pipette filter tips
17. 200 μL Pipette filter tips.
18. 1000 μL Pipette filter tips.
19. Vortex.
20. Heat block or water bath at 65 °C.
21. UV transilluminator.
22. Isopropanol.
23. 70% Ethanol.
24. GlycoBlue Coprecipitant (15 mg/mL) (Thermo Fisher Scientific).
25. 21-Gauge needle.
26. Rotating shaker.

3 Methods

3.1 Total RNA Extraction

All procedures are performed at room temperature. Mature grapevine cuttings are used to isolate phloem tissues. Bark is removed with a knife or a scalpel until the soft and green phloem is exposed. Phloem tissues are scraped by a scalpel to obtain small chips. Proceed with the scraping until the internal hard wood is reached. Immediately process the phloem scrapings to avoid tissue oxidization. Alternatively, store the tissues at −80 °C.

1. Weight 1 g of grapevine plant tissues (leaves with petioles or phloem scrapings). Grind tissues with liquid nitrogen in a mortar to obtain a fine powder.
2. Transfer the powder to a 15 mL conical tube (*see* **Note 6**) and add 10 mL/1 g of EB extraction buffer containing 1% 2-mercaptoethanol (*see* **Note 7**). Mix by vortexing for 30 s.
3. Add 1 mL of 20% N-lauroylsarcosine. Transfer tubes in a water bath and incubate at 70 °C for 10 min with intermittent shaking.
4. Transfer on ice and incubate for 5 min.
5. Centrifuge at 2500 g for 15 min at 4 °C, in a benchtop centrifuge.
6. Transfer the supernatant to a new 15 mL plastic tube and add 1/2 volume of Trizol.
7. Vortex for 5 min and centrifuge at 2500 g for 15 min at 4 °C, in a benchtop centrifuge.

8. Transfer supernatant to a new 15 mL plastic tube and add 1/3 volume of chloroform.
9. Vortex for 5 min and centrifuge at 2500 g for 5 min at 4 °C in a benchtop centrifuge. Repeat **steps 8** and **9** if the supernatant is not clear.
10. Transfer the supernatant to a new 15 mL plastic tube and add 1 volume of cold (−20 °C) isopropanol. Store the tubes overnight at −20 °C or 1 h at −80 °C.
11. Transfer the solution to 2 mL tubes and *centrifuge* at 18,000 g for 20 min at 4 °C, in a benchtop microcentrifuge.
12. Gently discard the supernatant and add 0.5 mL 75% ethanol to each pellet. Centrifuge at 18,000 g for 5 min at 4 °C, in a benchtop microcentrifuge.
13. Gently discard the supernatant without disturbing the pellet. Air-dry pellet at room temperature (*see* **Note 8**).
14. Dissolve pellets in 0.75 mL RNase-free water.

3.2 Separation of Low- and High-Molecular-Weight RNAs

1. In a 1.5 mL Eppendorf tube add 0.75 mL of PEG precipitation solution to 0.75 mL of purified total RNAs, mix by vortexing, and incubate on ice for 30 min.
2. Centrifuge at 16,000 g for 20 min at 4 °C, in a benchtop microcentrifuge.
3. Transfer the supernatant containing low-molecular-weight RNAs (LMW) to a new 1.5 mL tube. The pellet, which contains the high-molecular-weight RNAs (HMW), is dissolved in 0.1 mL 90% formamide (*see* **Note 9**).
4. Add 3 volumes of cold (−20 °C) absolute ethanol to the supernatant and store overnight at −20 °C or 1 h at −80 °C.
5. Centrifuge at 16,000 g for 25 min at 4 °C, in a benchtop microcentrifuge. Gently discard the supernatant and add 0.5 mL 75% ethanol to each pellet. Centrifuge at 18,000 g for 5 min at 4 °C, in a benchtop microcentrifuge.
6. Air-dry the pellet and resuspend in 0.1 mL 90% formamide. LMW RNAs in this fraction have a size lower than 300 nucleotides.
7. Check the quality of LMW and HMW RNAs by semi-denaturing gel electrophoresis in 1.2% agarose/TBE (*see* **Note 10**).
8. Add 2 μL 6× gel loading dye to 10 μl LMW or HMW RNAs: denature at 65° for 5 min, quickly chill in ice, and load in the agarose gel.
9. Run the gel at 100 volts constant voltage until the bromophenol blue is at 2 cm from the end of the gel.

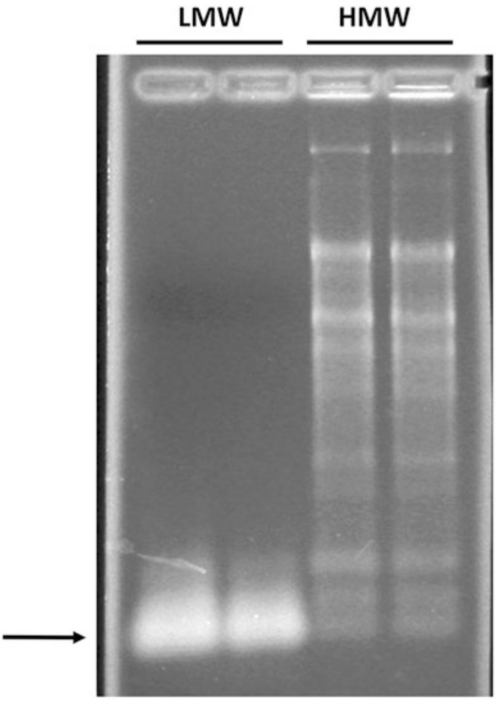


Fig. 1 1%/TBE agarose gel showing LMW and HMW RNAs after the PEG precipitation step. Arrow points to LMW RNAs

10. Stain the gel in a 0.5 μg/mL ethidium bromide solution in water. Submerge the gel and agitate for 15 min. Recover the solution and wash the gel for 5 min in water (*see* **Note 11**).
11. Observe the gel using UV transilluminator. LMW RNAs migrate as a bulk to the bottom of the gel whereas in HMW RNAs a distinct pattern consisting of major ribosomal (5S, 18S, and 25S rRNAs) RNAs is visible (Fig. 1) (*see* **Note 12**).

3.3 Purification of 18–30 nt Small RNAs from the LMW RNA Fraction

1. Prepare the 15% 8 M UREA gel using Mini-Protean Tetra gel 1 mm and 10 × 8 cm size (*see* **Note 13**).
2. Pre-run the gel for 30 min at 140 V.
3. Before loading the samples, wash the wells using 0.5× TBE.
4. Add 2 μL 6× loading dye to 10 μL of purified LMW RNAs. Heat the LMW at 65 °C and put on ice (*see* **Note 14**).
5. In a separate 1.5 mL tube prepare 10 μL of the DNA ladder (*see* **Note 15**) at 20 ng/μL concentration and add 2 μL 6× loading dye.
6. Load immediately the denatured samples and run the gel until the xylene cyanol reaches 2/3 of the run of the gel.

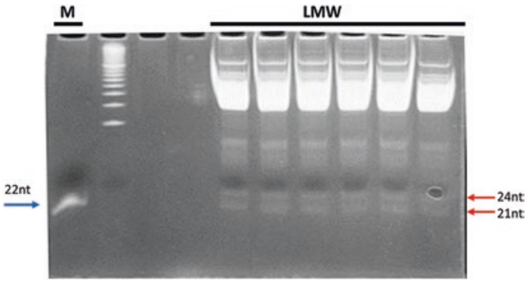


Fig. 2 15% 8 M Urea gel showing grapevine LMW RNAs extracted from about 1.5 gr leaves and petiole tissue. Blue arrow indicates the band of the used oligonucleotide (22 nt). Red arrows point the grapevine siRNA duplex (21–24 nt)

7. Stain the gel in a 0.5 μg/mL ethidium bromide solution in water. Submerge the gel and agitate for 15 min. Wash the gel for 5 min in water (*see* **Note 11**).
8. Identify 18–30 nt small RNA molecules by comparison with the ladder (Fig. 2). Cut the gel containing sRNAs with a razor blade and transfer gel slices into a 0.5 mL tube, previously punctured at the bottom in 3–4 points by a 21-gauge needle.
9. Put this tube into a 2 mL tube and centrifuge at 18,000 g for 2 min (*see* **Note 16**).
10. Add 2 volumes of 0.3 M NaCl elution buffer to the fragmented gel pieces and elute the small RNAs by shaking the tube gently and rotating overnight at 4 °C.
11. Transfer the fragmented gel pieces and the eluate into a Spin-X filter column and centrifuge for 2 min at 16,000 g in a bench microcentrifuge.
12. Continue by adding 100 μL of 0.3 M NaCl to the fragmented gel pieces and spin for another 2 min at max speed in a bench microcentrifuge.
13. Transfer the eluate into a 1.5 mL tube and add 1 volume of cold isopropanol and 2 μL of GlycoBlue at 2.5 μg/μL. Keep at −80 °C for at least 2 h.
14. Spin at 18,000 g for 25 min to pellet the sRNAs.
15. Add to the pellet 750 μL of 70% EtOH, air-dry, and resuspend in 10 μL of sterile resuspension buffer.

4 Notes

1. Wait until each reagent is dissolved before adding the successive.
2. N-lauroylsarcosine is toxic by inhalation. Use a protecting mask.

3. The 10% APS solution should be freshly prepared. It can be stored for several months at −20 °C.
4. Urea solubilizes with an endothermic reaction; thus slightly increase the temperature for dissolving it (not more than 30 °C).
5. Dilute concentrated HCl (12 N) 1:12 with water and regulate pH by adding small volumes to avoid a sudden decrease below the required pH.
6. Conical plastic tubes resistant to organic solvents should be used.
7. Add 2-mercaptoethanol to the EB buffer just before the use.
8. Air-dried pellets containing nucleic acids are easier resuspended in water or buffer than those dried under vacuum.
9. This volume is indicative and originates from our experience. It can be modulated according to the consistency of the pellet to keep the RNA concentration as high as possible.
10. For semi-denaturing gel is intended as a simple agarose 1.2% gel, where the samples are resuspended with formamide and denatured, before being loaded on the gel.
11. Ethidium bromide is a potent mutagen. Wear nitrile gloves during manipulation.
12. Quality of LMW and HMW RNAs is crucial for every successful molecular biology applications. Clear and slightly smeared bands corresponding to the rRNAs should be visible in the HMW RNA fraction: we consider it as a guarantee of the good quality of the LMW RNAs, which are not resolved in the agarose gel.
13. Acrylamide is a potent *neurotoxic*; we currently use *commercial* solution of acrylamide:bis-acrylamide ready to use to avoid risks deriving from the manipulation of the powder. Solutions should be stored at 4 °C.
14. Several wells should be loaded with LMW-RNAs to obtain sufficient amount of purified sRNAs for library preparation (Fig. 2).
15. Any 18–30 nt long oligonucleotide can be used as ladder.
16. The gel slices are minced when forced to pass through the holes, which makes the contained small RNAs to be successively extracted.

References

1. Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297(5589):2053–2056
2. Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis Thaliana. Curr Biol 12(17):1484–1495

3. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16(13):1616–1626
4. Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. Cell 136(4):669–687
5. Ding SW (2010) RNA-based antiviral immunity. Nat Rev Immunol 10:632–644
6. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S et al (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7
7. Pirovano W, Miozzi L, Boetzer M, Pantaleo V (2014) Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to a Cicer Arietinum population. Front Microbiol 5:790. https://doi.org/10.3389/fmicb.2014.00790
8. Iandolino AB, Goes da Silva F, Lim H, Choi H, Williams LE, Cook DR (2004) High-quality RNA, cDNA, and derived EST libraries from grapevine (Vitis Vinifera L.) Plant Mol Biol Rep 22(3):269–278
9. Meisel L, Fonseca B, González S, Baeza-Yates R, Cambiazo V, Campos R, Gonzalez M, Orellana A, Retamales J, Silva H (2005) A rapid and efficient method for purifying high quality total RNA from peaches (*Prunus persica*) for functional genomics analyses. Biol Res 38:83–88
10. MacKenzie DJ, Mclean MA, Mukerji S, Grenn M (1997) Improved RNA extraction from woody plants for the detection of viral pathogens by reverse transcription-polymerase chain reaction. Plant Dis 81:222–226
11. Gambino G, Perrone I, Gribaudo I (2008) A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem Anal 19(6):520–525
12. Lu C, Meyers BC, Green PJ (2007) Construction of small RNA cDNA libraries for deep sequencing. Methods 43(2):110–117

Chapter 3

Double-Stranded RNA-Enriched Preparations to Identify Viroids by Next-Generation Sequencing

Beatriz Navarro and Francesco Di Serio

Abstract

Approaches based on next-generation sequencing (NGS) coupled with bioinformatics tools have been developed for detecting viruses and viroids infecting herbaceous and woody plants. Here we describe a protocol to extract nucleic acids from citrus bark and enrich them in double-stranded RNAs. These preparations can be efficiently used for generating cDNA libraries that, after pair-end sequencing and bioinformatics analyses, allow efficient identification of the viroids infecting the source plant.

Key words dsRNA, Viroids, Citrus, NGS

1 Introduction

Viroids are small, circular, non-protein coding RNAs infecting herbaceous and woody plants [1]. They may induce cytopathic and physiological alterations in the infected hosts [2], which in turn may develop severe diseases [3]. Replication of viroids is mediated by a symmetric or an asymmetric rolling-circle mechanism and is dependent on the activity of host-encoded DNA-dependent RNA polymerases that are forced to use viroid RNAs as template to generate the mature circular RNA forms through double-stranded (ds) RNA intermediates [4]. Viroid-derived dsRNAs are likely also generated by host RNA-dependent RNA polymerases (RDRs) involved in plant RNA silencing pathways [5].

In the infected plant tissues, 21–24 nt viroid-derived small RNAs (vd-sRNAs) are associated with viroids and are generated by DICER-LIKE proteins likely targeting viroid dsRNAs for degradation [6]. Vd-sRNAs are phosphorylated at their 5′-end and methylated at the 3′-terminus [7], thus being structurally similar to the host-derived micro-RNAs (miRNAs). As for miRNAs, vd-sRNAs are selectively loaded into ARGONAUTE (AGO) proteins according to their size and the nucleotide at the 5′-end [8]. Directing specific degradation of complementary viroid or host RNAs [9],

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_3, © Springer Science+Business Media, LLC 2018

vd-sRNAs may be involved in plant antiviroid defense and pathogenesis [6]. The characterization of vd-sRNAs by next-generation sequencing (NGS) was first carried out in 2009 in peach trees and grapevine infected by one and two viroids, respectively, thus contributing to further dissection of vd-sRNA biogenesis [10, 11]. Since then, NGS of small RNA libraries has been deeply used for further dissecting the molecular interplay between viroids and their hosts in both herbaceous and woody hosts [12]. Coupled with bioinformatics tools to assemble contigs and search for homologous viroid sequences in databases, NGS of small RNA libraries has been largely and effectively used to identify viroids already known and discover new ones [12]. Recently, specific computational algorithms were developed, which allowed the homology-independent identification of three novel viroids or viroid-like RNAs, one from apple and two from grapevine [13, 14], thus confirming the efficiency of NGS of small RNA libraries for discovering new viroids.

As an alternative, dsRNA-enriched preparations, which actually contain viroid replication intermediate or dsRNAs generated by host RDRs, can also be efficiently used for viroid detection through NGS [15]. In 2013, a novel viroid infecting persimmon has been identified by NGS of a library generated from a dsRNA-enriched preparation [16], thus confirming the potentiality of this approach in viroid identification. In the frame of a study on viroids infecting citrus, we observed that libraries from dsRNA-enriched preparations provided more exhaustive coverage of the viroid genomic RNAs than the libraries of small RNAs extracted from the same samples. We also observed that preparations from bark tissues (including the phloem) are more appropriate to study citrus viroids than those from leaves, likely because citrus viroids are phloem-restricted in most natural hosts [17].

Seven viroids have been reported in citrus so far (citrus exocortis viroid, CEVd; hop stunt viroid, HSVd; citrus bark cracking viroid, CBCVd; citrus bent leaf viroid, CBLVd; citrus viroid V, CVd-V; citrus viroid VI, CVd-VI; citrus dwarfing viroid, CDVd), some of which determine severe disease [18]. Here, we report the protocol used in our laboratory to identify and characterize viroids infecting citrus through NGS. In this case, dsRNA-enriched preparations are the source material for generating cDNA libraries that can be efficiently sequenced by pair-end (2–125) Illumina NGS. By this method, we generated a cDNA library of dsRNAs from citrus bark in which three viroids (CEVd, HSVd, and CDVd), simultaneously infecting the same sample, were efficiently detected.

2 Materials

All solutions must be prepared using ultrapure water and autoclaved (121 °C, 20 min, 1 bar). Solutions that need to be filtered or do not require sterilization are indicated. Mortars and pestles, phenol-resistant tubes (50 mL and 250 mL) for centrifuge, and beakers must be used after sterilization in an autoclave (121 °C, 20 min, 1 bar).

The protocol reported below refers to 20 g of bark tissue from young (green) stems of citrus trees (*Citrus sinensis* (L.) Osbeck).

2.1 Total Nucleic Acid (TNA) Extraction

1. Water-saturated phenol neutralized with NaOH at pH 7 (*see* **Note 1**).
2. 0.2 M Tris–HCl pH 8.9, for 1 L: Dissolve 24.23 g of Tris base (Mw: 121.14) in 900 mL of water, adjust pH at 8.9 with HCl, and make up to 1 L with water.
3. 0.1 M EDTA pH 8, for 100 mL: Dissolve 3.72 g of EDTA in 90 mL of water, adjust the pH with pellets of NaOH, and make up to 100 mL with water.
4. 2% SDS (w/v) in water, for 100 mL: Dissolve 2 g of sodium dodecyl sulfate (SDS) in 90 mL of water and then adjust to 100 mL with water, sterilize by filtering through a 0.45 μm filter.
5. Extraction solution for 20 g of tissue: Mix 120 mL of phenol water-saturated and neutralized, 30 mL of 0.2 M Tris–HCl pH 8.9, 18.75 mL of 2% SDS, and 7.5 mL of 0.1 M EDTA pH 8, before using add 1.5 mL of β-mercaptoethanol.

2.2 dsRNA Enrichment by Cellulose

1. 10× STE (500 mM Tris–HCl, 1 M NaCl, 10 mM EDTA), pH 7.2, for 1 L: Dissolve 58.5 g NaCl, 60.6 g Tris, and 3.72 g EDTA in 900 mL H20; adjust pH to 7 with HCl; and make up to 1 L with water.
2. 1× STE/16% ethanol, for 1 L: Mix 100 mL of 10× STE, 160 mL of ethanol, and 740 mL of sterile water.
3. CF-11 non-ionic cellulose powder (Whatman) (*see* **Note 2**).
4. 3 M Sodium acetate, pH 5.5, for 25 mL: Dissolve 10.205 g of sodium acetate with 15 mL of water, and add approximately 5.8 mL of glacial acetic acid to adjust pH 5.5.
5. Glass centrifuge tubes.
6. One refrigerated centrifuge (Beckman JA–14 and JA–20) and one centrifuge with swing-out rotor.

2.3 Quantification of Nucleic Acids

1. Spectrophotometer NanoDrop 2000 (Thermo Scientific).

2.4 Elimination of DNA Contaminants

1. Turbo DNA-free Kit (Ambion by Life Technologies, USA).

3 Methods

The protocol for dsRNA enrichment is based on the original method reported by Morris and Dodds [19] with some modifications. The protocol is set for 20 g of bark tissue from young citrus stems. For different amounts of starting material, the volumes of buffers must be scaled accordingly. For plant tissues containing high level of polysaccharides (which is the case of many woody plants) the volumes of the extraction solutions can be slightly increased.

3.1 Preparation of Bark Tissue

1. Eliminate the leaves from the young (still green) stems and remove the bark tissues (including the phloem) using a surgical blade.
2. Collect the bark in an aluminum foil located on ice.
3. Use the bark immediately for nucleic acid extraction or store it at −80 °C.

3.2 Total Nucleic Acid (TNA) Extraction

1. Crash and powder the plant tissue in a mortar and a pestle using liquid nitrogen (*see* **Note 3**).
2. Add the pulverized material to the extraction solution and shake for 30 min at room temperature for facilitating the nucleic acid extraction. If a homogenizer is used this step can be avoided.
3. Transfer the extract to centrifuge tube(s) and centrifuge for 15 min at 7500 × *g* at 4 °C.
4. Transfer the supernatant to a new tube avoiding to touch or disturb the interphase and add half volume of water-saturated and neutralized phenol (*see* **Note 4**).
5. Centrifuge for 15 min at 7500 × *g* at 4 °C.
6. Transfer the supernatant to a 50 mL sterilized tube as reported at point 4.

3.3 dsRNA Enrichment by Cellulose

1. Add water up to a volume of 40 mL.
2. Add 5 mL of 10× STE solution, 8 mL ethanol, and 2.5 g of CF11 cellulose powder.
3. Shake for at least 4 h at room temperature or overnight if more convenient.
4. Centrifuge for 5 min at 1500 × *g* in a centrifuge with a swing-out rotor at room temperature.
5. Eliminate the supernatant by decantation.
6. Wash the pellet adding 30 mL of 1× STE/16% ethanol and vortexing for 2 min.

7. Centrifuge for 5 min at 1500 × *g* in a centrifuge with a swing-out rotor and eliminate the supernatant by decantation.
8. Repeat **steps** from **5** to 7 twice.
9. Eliminate as much as possible the washing buffer with a pipette without disturbing the pellet.
10. Eluted the dsRNA from the cellulose by adding 3.3 mL of sterilized water and vortexing vigorously for 2 min.
11. Centrifuge for 5 min at 1500 × *g* in a centrifuge with a swing-out rotor.
12. Transfer the supernatant to a new tube precooled and keep it on ice.
13. Repeat **steps** from **10** to **12** twice to recover a final volume of 10 mL in the same tube on ice.
14. For elimination of possible traces of cellulose, centrifuge again the eluted solution for 1 min at 1500 × *g*.
15. Transfer the supernatant to a glass centrifuge tube on ice and discard the pellet.
16. Precipitate the dsRNA by adding 3 volumes of ethanol and 0.1 volume of 3 M sodium acetate, mixing and maintaining the tube at −20 °C for more than 4 h (overnight if convenient).
17. Centrifuge for 30 min at 7700 × *g* at 4 °C, eliminate the supernatant, and dry the pellet.
18. Resuspend the pellet in 250 μL of sterilized water and maintain the tube on ice.

3.4 Quantification of the Nucleic Acid Preparation

Use 1 μL of the dsRNA-enriched preparation for the RNA quantification by NanoDrop (*see* **Note 5**).

3.5 Elimination of DNA

Traces of DNA present in the dsRNA preparation are eliminated using Turbo DNA-free Kit following the manufacturer's protocol.

1. Transfer approximately 40 μL of dsRNA preparation (3.5 μg) in a 0.5 mL Eppendorf tube.
2. Add 4 μL (0.1 vol) of 10× TURBO DNase Buffer and 1 μL TURBO DNase and mix gently.
3. Incubate at 37 °C for 30 min.
4. Add 4 μL (0.1 vol) of DNase inactivation reagent that must be well resuspended before using.
5. Incubate for 5 min at room temperature mixing occasionally, since the DNase inactivation reagent tends to sediment in the bottom of the tube. Therefore, flick the tube several times during the incubation time.

6. Centrifuge at 10,000 × *g* for 1.5 min at 4 °C.
7. Transfer, quick and carefully, the supernatant (that contains the dsRNA) to a fresh tube maintained on ice. During this procedure, be careful not to touch the pellet that is easily detached from the tube.
8. Quantify of RNA (DNA free) by NanoDrop (*see* **Note 6**).

3.6 Preparation of cDNA Library and NGS Sequencing

Purified and enriched dsRNAs are used by specialized services for RNA-seq cDNA library preparation upon removing ribosomal RNAs. Libraries are sequenced (pair-end 2 × 125) according to standard Illumina procedures.

4 Notes

1. For preparation of water-saturated phenol, heat the phenol at 65 °C, add an equal volume of sterilized water, and mix. Centrifuge at 5500 × *g* for 5 min and remove water phase. Then add an equal volume of sterilized water, mix, centrifuge at 5500 × *g* for 5 min, and remove the water phase. Store at 4 °C protected from the light. Phenol and phenol-containing solutions are harmful and must be handled with appropriate protections as gloves and lab coat and under an extractor hood. For phenol neutralization use a concentrate solution of NaOH and litmus paper for pH verification. No sterilization.
2. As an alternative to CF-11 (Whatman), Cellulose C6288 (Sigma) could be used.
3. Alternatively, tissues can be directly immersed in the extraction mix and homogenized in a homogenizer.
4. The aqueous phase after the phenol extraction could be very dense; aspiration without disturbing the organic phase can be facilitated using 1 mL tips cut at 1 cm from the end of the tip.
5. A nucleic acid concentration of about 85 ng/μL (A280/A260 = 2) is generally obtained.
6. In our experience, RNA concentration of about 60–70 ng/μL (A280/A260 = 1.86) is generally obtained.

Acknowledgments

This work was partially supported by the project INNOCI—DGR n. 903 del 15/05/2012—Linee guida per la ricerca e la sperimentazione in agricoltura—Regione Puglia.

References

1. Navarro B, Gisel A, Rodio ME, Delgado S, Flores R, Di Serio F (2012) Viroids: how to infect a host and cause disease without encoding proteins. Biochimie 94:1474–1480
2. Di Serio F, De Stradis A, Delgado S, Flores R, Navarro B (2013) Cytopathic effects incited by viroid RNAs and putative underlying mechanisms. Front Plant Sci 3:288
3. Flores R, Serra P, Minoia S, Di Serio F, Navarro B (2012) Viroids: from genotype to phenotype just relying on RNA sequence and structural motifs. Front Microbiol 3:217
4. Flores R, Grubb D, Elleuch A, Nohales MÁ, Delgado S, Gago S (2011) Rolling-circle replication of viroids, viroid-like satellite RNAs and hepatitis delta virus: variations on a theme. RNA Biol 8:200–206
5. Di Serio F, Martínez de Alba AE, Navarro B, Gisel A, Flores R (2010) RNA-dependent RNA polymerase 6 delays accumulation and precludes meristem invasion of a nuclear-replicating viroid. J Virol 84:2477–2489
6. Flores R, Minoia S, Carbonell A, Gisel A, Delgado S, Lopez-Carrasco A, Navarro B, Di Serio F (2015) Viroids, the simplest RNA replicons: how they manipulate their hosts for being propagated and how their hosts react for containing the infection. Virus Res 209:136–145
7. Martín R, Arenas C, Daròs JA, Covarrubias A, Reyes JL, Chua NH (2007) Characterization of small RNAs derived from citrus exocortis viroid (CEVd) in infected tomato plants. Virology 367:135–146
8. Minoia S, Carbonell A, Di Serio F, Gisel A, Carrington JC, Navarro B, Flores R (2014) Specific argonautes selectively bind small RNAs derived from potato spindle tuber viroid and attenuate viroid accumulation in vivo. J Virol 88:11933–11945
9. Navarro B, Gisel A, Rodio ME, Delgado S, Flores R, Di Serio F (2012) Small RNAs containing the pathogenic determinant of a chloroplast-replicating viroid guide degradation of a host mRNA as predicted by RNA silencing. Plant J 70:991–1003
10. Di Serio F, Gisel A, Navarro B, Delgado S, Martínez de Alba AE, Donvito G, Flores R (2009) Deep sequencing of the small RNAs derived from two symptomatic variants of a chloroplastic viroid: implications for their genesis and for pathogenesis. PLoS One 4:e7539
11. Navarro B, Pantaleo V, Gisel A, Moxon S, Dalmay T, Bisztray G, Di Serio F, Burgyán J (2009) Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. PLoS One 4:e7686
12. Hadidi A, Flores R, Candresse T, Barba M (2016) Next-generation sequencing and genome editing in plant virology. Front Microbiol 7:1325
13. Wu Q, Wang Y, Cao M, Pantaleo V, Burgyan J, Li WX, Ding SW (2012) Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. Proc Natl Acad Sci U S A 109:3938–3943
14. Zhang Z, Qi S, Tang N, Zhang X, Chen S, Zhu P, Ma L, Cheng J, Xu Y, Lu M, Wang H, Ding SW, Li S, Wu Q (2014) Discovery of replicating circular RNAs by RNA-seq and computational algorithms. PLoS Pathog 10:e1004553
15. Al Rwahnih M, Dolja VV, Daubert S, Koonin EV, Rowhani A (2012) Genomic and biological analysis of grapevine leafroll-associated virus 7 reveals a possible new genus within the family Closteroviridae. Virus Res 163:302–309
16. Ito T, Suzaki K, Nakano M, Sato A (2013) Characterization of a new apscaviroid from American persimmon. Arch Virol 158:2629–2931
17. Bani-Hashemian SM, Pensabene-Bellavia G, Duran-Vila N, Serra P (2015) Phloem restriction of viroids in three citrus hosts is overcome by grafting with Etrog citron: potential involvement of a translocatable factor. J Gen Virol 96:2405–2410
18. Duran-Vila N, Semancik JS (2003) Citrus viroid. In: Hadidi A, Flores R, Randles JW, Semancik JS (eds) Viroids. CSIRO Publishing, Collingwood, Australia
19. Morris TJ, Doods JA (1979) Isolation and analysis of double-stranded RNA from virus-infected plant and fungal tissue. Phytopathology 69:854–858

Chapter 4

Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing

Armelle Marais, Chantal Faure, Bernard Bergey, and Thierry Candresse

Abstract

High-throughput sequencing (or next-generation sequencing—NGS) is an emerging technology that allows the detection of plant viruses without any prior knowledge. Various sequencing techniques and various templates can be used as substrate for NGS. This chapter describes an optimized protocol for the extraction of double-stranded RNAs (dsRNAs) from a wide range of plants and for their random amplification prior to NGS sequencing.

Key words Double-stranded (ds) RNA, NGS, High-throughput sequencing, Metagenomics, Virome, Diagnostic

1 Introduction

In the recent last years, advances in high-throughput sequencing technologies (or next-generation sequencing—NGS) and in bioinformatics allowed the development of new diagnostic tools, in particular in the plant virus field [1–3]. Indeed, the characterization and detection of any virus from a plant sample are now possible, without any a priori knowledge of the viral entities. In contrast with bacteria and fungi metagenomics, where universal gene sequences are available [4, 5], there are no similar molecular markers that can serve such purposes for viruses. A variety of methods have been used to access various types of viral nucleic acid, and using it as starting material for NGS: (1) total DNA or RNA, with or without a ribosomal RNA depletion step [6], (2) virion-associated nucleic acids from semi-purified particles [7, 8], (3) viral derived small interfering RNA [9], and (4) double-stranded (ds) RNAs [10]. DsRNAs are constitutive molecules of genomes of some plant viruses, such as members of the families *Endornaviridae*, *Amalgamaviridae*, *Birnaviridae*, *Chrysoviridae*, *Partitiviridae*,

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_4, © Springer Science+Business Media, LLC 2018

and *Totiviridae*. Moreover, dsRNAs can also be replicative forms of viruses with single-stranded RNA genomes. Therefore, dsRNA sequencing can be used for an efficient discovery and characterization of viruses having RNA genomes, although it cannot be used reliably for viruses with a DNA genome [11, 12].

Here we describe protocols for (1) purification of dsRNAs by CF11 cellulose batch chromatography, modified from [13], and (2) cDNA synthesis and random amplification prior to NGS. The enrichment of viral dsRNAs allows the use of multiplexing strategy and therefore it reduces sequencing costs and improves diagnostic efficiency [14]. Interestingly, this protocol is suitable for the purification and sequencing of dsRNAs from a wide range of plants, fungi, and other organisms [15].

2 Materials

Prepare all solutions using diethyl pyrocarbonate (DEPC)-treated water (*see* **Note 1**).

Use only sterile tips with filter to avoid contamination among samples.

2.1 Double-Stranded RNA Extraction

1. Liquid nitrogen.
2. Mortar, pestle, and funnel kept in liquid nitrogen until use.
3. Extraction buffer: 1 mL 2× STE, 70 μL 20% SDS, 20 μL sodium bentonite, 1.425 mL phenol-TE saturated (*see* **Note 2**).
4. 10× STE: 1 M NaCl, 0.5 M Tris–HCl pH 8, 0.01 M EDTA (*see* **Note 3**).
5. Washing solution: 1× STE + 16% ethanol (v/v) (*see* **Note 4**).
6. 20% SDS (*see* **Note 5**).
7. 40 mg/mL Sodium bentonite solution (*see* **Note 6**).
8. Phenol-TE saturated.
9. CF11 cellulose (*see* **Note** 7).
10. Absolute ethanol.
11. 3 M Sodium acetate, pH 5.2 (*see* **Note 8**).
12. 70% Ethanol.
13. 1 M Magnesium acetate.
14. 1 U/μL DNase RQ1.
15. 10 μg/μL RNase A.
16. 10× SSC.
17. 5 mg/mL Proteinase K.
18. Phenol:chloroform:isoamyl alcohol, TE saturated (25:24:1).
19. Chloroform:isoamyl alcohol (24:1).

20. Agarose.
21. TBE buffer: 89 mM Tris–HCl, 89 mM boric acid, 2 mM EDTA.
22. Horizontal electrophoresis equipment.
23. Sybergreen.
24. 6× Loading buffer.
25. UV transilluminator.

2.2 Random RT-PCR Amplification of dsRNAs

1. dNTP.
2. Superscript II Reverse Transcriptase (Invitrogen).
3. RNase H.
4. RNase inhibitor.
5. DyNAzyme II DNA polymerase (Finnzymes).
6. Primers:

   (a) PcDNA12: 5′ -TGTGTTGGGTGTGTTTGGN$_{(12)}$ -3′.

   (b) MIDGENCO: 5′-**ACGTACACACT**TGTGTTGGGTGTGTTTGG-3′ (*see* **Note 9**).
7. Agarose.
8. TBE buffer: 89 mM Tris–HCl, 89 mM boric acid, 2 mM EDTA.
9. Horizontal electrophoresis equipment.
10. Ethidium bromide.
11. 6× Loading buffer.
12. UV transilluminator.
13. MinElute PCR purification kit (Qiagen).
14. Spectrophotometer.

3 Methods

3.1 Double-Stranded RNA Extraction

(Carry out all procedures and centrifugation steps at room temperature unless otherwise specified).

1. Grind 0.75 g of frozen sample (or 0.075 g of dried sample) in the presence of liquid nitrogen with the precooled mortar and pestle.
2. Transfer with spatula and funnel the frozen powder to a 15 mL Falcon tube containing the extraction buffer. Dispense and thaw the content by vortexing for 1 min.
3. Agitate gently for 30 min on a horizontal shaker (*see* **Note 10**).
4. Centrifuge for 15 min at 3000 × *g*.

5. Transfer the aqueous phase to a 1.5 mL Eppendorf tube and centrifuge at 10,000 × *g* for 20 min.
6. Transfer the aqueous phase to a new 1.5 mL Eppendorf tube and add absolute ethanol to obtain a 16% (v/v) final concentration (*see* **Note 11**). Mix well by vortexing.
7. Binding step: Transfer solution to a new 1.5 mL Eppendorf tube containing 40 mg of CF11 cellulose powder. Mix well by vortexing (*see* **Note 12**).
8. Agitate gently for 30 min on a horizontal shaker (*see* **Note 13**).
9. Centrifuge for 1 min at 5000 × *g*. Remove the supernatant.
10. Washing step: Add 1 mL of washing solution on the cellulose pellet. Mix well and disperse the pellet with a pipette tip. Vortex, and agitate gently for 5 min on a horizontal shaker.
11. Repeat **steps 9 and 10** twice (*see* **Note 14**).
12. Remove completely the supernatant and dry very carefully the pellet by pipetting.
13. Elution step: Add 200 μL of 1× STE to the dried pellet. Mix well and disperse the pellet with a pipette tip. Vortex gently, and agitate slowly for 5 min on a horizontal shaker.
14. Centrifuge at 5000 × *g* for 1 min and collect the supernatant in a new 1.5 mL Eppendorf tube kept on ice.
15. Repeat once **steps 13** and **14**. Collect the supernatant in the same collection tube. Be careful to collect the maximum of liquid (containing dsRNAs) by planting the tip in the pellet and pipetting to drain the pellet.
16. Eliminate the cellulose that may still be present in the pooled eluate by centrifugation for 1 min at 5000 × *g*. Collect the supernatant (about 400 μL), and transfer to a new Eppendorf tube.
17. Ethanol precipitation of dsRNAs: Add 1/10 volume of 3 M sodium acetate pH 5.2 and 0.8 volume of isopropanol. Store overnight at −20 °C or 1 h at −80 °C.
18. Centrifuge at >14,000 × *g* for 30 min at 4 °C. Discard the supernatant.
19. Rinse the pellet with 1 mL of 70% ethanol. Centrifuge again for 15 min. Discard the supernatant.
20. Dry the pellet under vacuum for 10 min and dissolve in 180 μL of DEPC-treated water.
21. Keep an aliquot of 10 μL of untreated dsRNAs.

3.2 Enzymatic Treatment of dsRNAs

1. Add 20 μL of 1 M magnesium acetate and 10 μL of DNase RQ1 (1 U/μL). Incubate the mixture for 1 h at 37 °C.
2. Add 60 μL of 10× SSC, 1 μL of RNase A (10 μg/mL), and 39 μL of DEPC-treated water. Incubate for 30 min at 37 °C.
3. Add 2.5 μL of 2% SDS and 8 μL of proteinase K (5 mg/mL). Incubate at least for 1 h at 37 °C.

3.3 Phenol/Chloroform Extraction and Ethanol Precipitation

1. Add one volume (300 μL) of phenol:chloroform:isoamyl alcohol (25:24:1), and vortex for 1 min.
2. Centrifuge for 5 min at >14,000 × *g*. Remove the upper aqueous phase, and transfer to a fresh Eppendorf tube.
3. Add one volume of chloroform:isoamyl alcohol (24:1), and vortex for 1 min.
4. Centrifuge for 5 min at >14,000 × *g*. Remove the upper aqueous phase, and transfer to a fresh Eppendorf tube.
5. Add 1/10 volume of 3 M sodium acetate, pH 5.2, and 2 volumes of absolute alcohol.
6. Place the tube at −20 °C overnight or at −80 °C for at least 1 h.
7. Centrifuge at >14,000 × *g* for 30 min at 4 °C. Discard the supernatant.
8. Rinse the pellet with 1 mL of 70% ethanol. Centrifuge again for 15 min. Discard the supernatant.
9. Dry the pellet under vacuum for 10 min and dissolve it in 250 μL of DEPC-treated water.

3.4 Second Round of CF11 Chromatography (See Note 15)

1. Add 40 μL of absolute alcohol to the 250 μL dsRNA and transfer to a new Eppendorf tube containing 40 mg of CF11. Mix well by vortexing.
2. Repeat **steps 8–19** of Subheading 3.1.
3. Dry the pellet under vacuum for 10 min and dissolve it in 20 μL of DEPC-treated water.
4. Load 3 μL of treated dsRNAs previously mixed with 1/6 volume of loading buffer, on 0.8% agarose gel containing 1% of Sybergreen in 1× TBE buffer, together with untreated dsRNAs (**step 21**, Subheading 3.1).
5. Migrate for 1 h at 80 V.
6. Visualize the nucleic acids on a UV transilluminator (Fig. 1) (*see* **Note 16**).

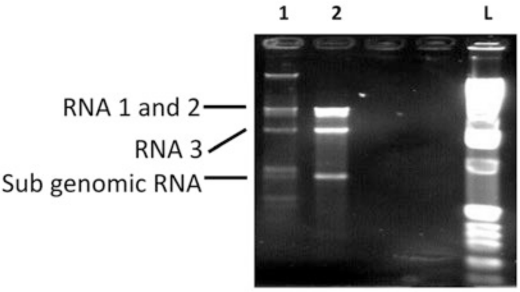


Fig. 1 Agarose gel analysis of double-stranded RNA (dsRNA) extracted from *Cucumber mosaic virus*-infected plant. *Lane 1*: untreated dsRNA. *Lane 2:* DNase- and RNase-treated dsRNA; L: 1 kbp Ladder

3.5 cDNA Synthesis and Random Amplification

1. Denaturate 3 μL of purified dsRNAs by heating at 99 °C for 5 min.
2. Cool on ice for 2 min.
3. Add 2 mM of primer PcDNA12, 0.5 mM of dNTP, 1× reaction buffer, 200 U Superscript II Reverse Transcriptase, and 1 U RNase inhibitor in a 20 μL volume reaction.
4. Incubate at 25 °C for 10 min, and then at 42 °C for 60 min.
5. Inactivate the RT by incubating at 70 °C for 10 min.
6. Add 1.5 U of RNase H and incubate at 37 °C for 20 min (*see* **Note 17**).
7. To 5 μL of RT products, add 1 μM of primer MIDGENCO, 1× reaction buffer, 0.25 mM dNTPs, and 1 U of DyNAzym II DNA polymerase (Finnzymes) in a 50 μL volume reaction (*see* **Note 18**).
8. Amplification conditions: 94 °C for 1 min; 65 °C for 0 s; 72 °C for 45 s, with a slope of 5 °C/s, followed by X cycles of 94 °C for 0 s; 45 °C for 0 s; 72 °C for 5 min, with the same slope; and final steps of 5 min at 72 °C and 5 min at 37 °C (*see* **Note 19**).
9. Visualize the PCR products after loading 10 μL of PCR reactions previously mixed with 1/6 volume of loading buffer, on a 1.5% agarose gel containing 10 μg/mL of ethidium bromide in 1× TBE buffer (*see* Fig. 2).
10. Migrate for 30 min at 100 V.
11. Visualize the nucleic acids on a UV transilluminator (Fig. 1) (*see* **Note 16**). Smears corresponding to PCR products from 100 to 1000 bp are usually visualized.
12. Purify the PCR products as recommended by the manufacturer (MinElute PCR purification kit, Qiagen) and estimate the concentration at 260 nm.
13. Perform deep sequencing of the PCR products by a sequencing platform.

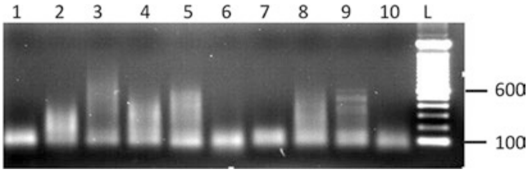


Fig. 2 Agarose gel analysis of the PCR products generated after random amplification. *Lanes 1–10*: various samples; L: 100 bp Ladder

4 Notes

1. DEPC is a strong inhibitor of RNases [16]. Add 1 mL of DEPC to 9 mL of ethanol 95% and adjust to 1 l of water. Shake vigorously and allow to stand overnight at room temperature. Autoclave (15 min, 121 °C) to inactivate DEPC. DEPC will dissolve some plastic pipettes; therefore, glass should be used. DEPC must be handled in a fume hood.

2. Prepare the extraction buffer just before use in a 15 mL Falcon tube. This extraction buffer can be used for a wide variety of plant species. However, for tissues from some woody plants such as grapevine, a modified extraction buffer can be used with better results (3.75 mL of 2× STE; 564 μL of 20% SDS; 60 μL of bentonite solution; 75 μL of 2-mercaptoethanol; 9 μL of NH_4OH concentrate; 1.875 mL of chloroform; 1.875 mL of phenol TE saturated).

3. Diluted solutions of STE:

   (a) 2× STE: Dilute 8 mL of 10× STE in 40 mL of DEPC water.

   (b) 1× STE: Dilute 1 mL of 10× STE in 9 mL of DEPC water.

4. Due to evaporation, it is best to prepare this fresh each time. Washing solution: 4 mL 1× STE, 6 mL absolute ethanol, 30 mL DEPC water.

5. SDS precipitates when the temperatures are too cold. The solution needs to be warmed to room temperature prior to use.

6. Homogenize 10 g of sodium bentonite with 200 mL of 0.01 M Tris–HCl, pH 7.5, in a grinder. Centrifuge for 3 min at 600 × *g* at 4 °C. Collect the supernatant and centrifuge it again for 30 min at 15,300 × *g* at 4 °C. Discard the supernatant and resuspend the pellet in the grinder with 100 mL of 0.01 M Tris–HCl, pH 7.5. Let stand overnight at 4 °C. Repeat the procedure and resuspend the final pellet in 50 mL of 0.01 M Tris–HCl, pH 7.5. To determine the concentration of the solution, 1 mL is dried at 100 °C in an oven and the weight is determined. The bentonite solution is stored at 4 °C.

7. Alternatively, CC41 cellulose (Whatman) or C6288 (Sigma) can be used.

8. Add 40.8 g sodium acetate trihydrate ($CH_3COONa\ 3H_2O$) to 80 mL of water. After dissolution, transfer the bottle in a fume hood and adjust the pH to 5.2 with pure (glacial) acetic acid. Make up to 100 mL with water and autoclave.

9. The bold sequence represents the multiplex identifier (MID) adaptor. In a multiplexing scheme, various primers differing only by the MID sequence can be used.

10. For grapevine extraction, the incubation is performed at 4 °C for 45 min.
11. Be sure to use absolute ethanol at this stage, and not 95% ethanol.
12. For grapevine extraction, two tubes of CF11 are used, due to the larger volume of the aqueous phase obtained.
13. For grapevine extraction, the incubation time is increased to 1 h.
14. The washing steps can be repeated more than twice it needed, until the supernatant becomes colorless.
15. This second run of CF11 chromatography provides a further enrichment of viral dsRNAs and hence a higher proportion of viral reads after NGS.
16. The visualization of bands is not systematic as the concentration of dsRNAs may be limiting in a number of cases.
17. The cDNA can either be immediately used in the amplification step or kept frozen at −20 °C until use.
18. This random amplification procedure allows at the same time the conversion of cDNA to double-stranded cDNA and incorporation of tagging MID adaptors in order to allow the multiplexed sequencing of multiple samples.
19. The number of cycles X should be as low as possible.

References

1. Boonham N, Kreuze J, Winter S, van der Vlugt R, Bergervoet J, TomLinson J, Mumford R (2014) Methods in virus diagnostics: from ELISA to next generation sequencing. Virus Res 186:20–31
2. Massard S, Olmos A, Jijakli H, Candresse T (2014) Current impact and future direction of high throughput sequencing on plant virus diagnostics. Virus Res 188:90–96
3. Massard S, Candresse T, Gil J, Lacomme C, Predajna L, Ravnikar M, Reynard JS, Rumbou A, Saldarelli P, Skoric D, Vainio EJ, Valkonen JPT, Vanderschuren H, Varveri C, Wetzel T (2017) A framework for the evaluation of biosecurity, commercial, regulatory and scientific impacts of plant viruses and viroids identified by NGS technologies. Front Microbiol 8:45
4. Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community diversity. ISME J 6:183–194
5. Xu J (2016) Fungal DNA barcoding. Genome 59:913–932
6. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009) Next generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10:537–545
7. Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois JH, Fernandez E, Martin DP, Varsani A, Roumagnac P (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9:e102945
8. Filloux D, Dallot S, Delaunay A, Galzi S, Jacquot E, Roumagnac P (2015) Metagenomics approaches based on Virion-associated nucleic acids (VANA): an innovative tool for assessing without a priori viral diversity of plants. Methods Mol Biol 1302:249–257
9. Kreuze JF, Pérez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7

10. Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen G, Roe BA (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19:81–88
11. Roossinck MJ, Darren MP, Roumagnac P (2015) Plant virus metagenomics: advances in virus discovery. Phytopathology 105:716–727
12. Wu Q, Ding SW, Zhang Y, Zhu S (2015) Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. Annu Rev Phytopathol 53:425–444
13. Valverde RA, Dodds JA, Heick JA (1986) Double-stranded ribonucleic acid from plants infected with viruses having elongated particles and undivided genomes. Phytopathology 76:459–465
14. Candresse T, Marais A, Faure C, Gentit P (2013) Association of *Little cherry virus 1* (LChV1) with the Shirofugen stunt disease and characterization of the genome of a divergent LChV1 isolate. Phytopathology 103:293–298
15. Marais A, Nivault A, Faure C, Theil S, Comont G, Candresse T, Corio-Costet MF (2017) Determination of the complete genomic sequence of *Neofusicoccum luteum* mitovirus 1 (NlMV1), a novel mitovirus associated with a phytopathogenic *Botryosphaeriaceae*. Arch Virol 162(8):2477–2480
16. Fedorcsak I, Ehrenberg L (1966) Effects of diethyl carbonate and methyl methanesulfonate on nucleic acids and nucleases. Acta Chem Scan 20:107–112

Chapter 5

Workup of Human Blood Samples for Deep Sequencing of HIV-1 Genomes

Marion Cornelissen, Astrid Gall, Antoinette van der Kuyl, Chris Wymant, François Blanquart, Christophe Fraser, and Ben Berkhout

Abstract

We describe a detailed protocol for the manual workup of blood (plasma/serum) samples from individuals infected with the human immunodeficiency virus type 1 (HIV-1) for deep sequence analysis of the viral genome. The study optimizing the assay was performed in the context of the BEEHIVE (Bridging the Evolution and Epidemiology of HIV in Europe) project, which analyzes complete viral genomes from more than 3000 HIV-1-infected Europeans with high-throughput deep sequencing techniques. The goal of the BEEHIVE project is to determine the contribution of viral genetics to virulence. Recently we performed a pilot experiment with 125 patient plasma samples to identify the method that is most suitable for isolation of HIV-1 viral RNA for subsequent long-amplicon deep sequencing. We reported that manual isolation with the QIAamp Viral RNA Mini Kit (Qiagen) provides superior results over robotically extracted RNA. The latter approach used the MagNA Pure 96 System in combination with the MagNA Pure 96 Instrument (Roche Diagnostics), the QIAcube robotic system (Qiagen), or the *m*Sample Preparation Systems RNA kit with automated extraction by the m2000*sp* system (Abbott Molecular). Here we present a detailed protocol for the labor-intensive manual extraction method that yielded the best results.

Key words HIV-1, Nearly complete genome, RNA isolation, QIAamp viral isolation kit, High-throughput deep sequencing

1 Introduction

The molecular analysis of complete viral genomes, including that of HIV, is the new research standard that is made possible by modern high-throughput deep sequencing technologies. The assembly of complete viral genomes from short sequence reads remains a challenge, but many bioinformatics tools have been developed to facilitate this process [1, 2]. An important but largely ignored aspect in the pipeline of sample preparation, sequencing, and data analysis is the optimum isolation of nucleic acids from patient samples. Several variables can have an influence on the success of generation of viral genomes, including the sample storage conditions (temperature, duration of storage, number of freeze-thaw cycles), the RNA/DNA

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_5, © Springer Science+Business Media, LLC 2018

extraction protocol, the initial reverse transcription reaction for RNA viruses, and the method used for PCR amplification, e.g., multiplex PCR [3] or HIV-1 SMART PCR [4]. Recently we have presented a detailed literature overview of nucleic acid isolation methods used for HIV-1 complete genome high-throughput sequencing [5]. Eight similar HIV-1 RNA/DNA extraction methods were compared, all of which used blood plasma samples and not serum. Two studies used robots for RNA isolation [6, 7], five studies exclusively used manual nucleic acid extraction [8–12], and a single study used both robotic and manual extraction [4]. However, no direct comparison between manual and robotic extraction methods was performed.

In previous work, we have performed a direct comparison of manual and robotic extraction methods and scored the amplification efficiency for four overlapping HIV-1 amplicons of ~1.9, 3.6, 3.0, and 3.5 kb [8]. Generation of all four amplicons is essential for successful sequencing of the nearly complete viral genome. We showed that the number of successfully amplified HIV-1 amplicons was significantly greater for RNA preparations purified manually than for those isolated robotically. Among the robotic RNA isolation methods, the QIAamp Viral RNA Kit in combination with the automated QIAcube system performed best. For RNA isolated with the m2000*sp* and MagNa Pure systems, only rarely could more than two RT-PCR amplicons be amplified, precluding complete HIV-1 genome sequencing. Based on these results we decided to select more labor-intensive and time-consuming manual extraction, in combination with the QIAamp Viral RNA Kit, for the BEEHIVE study. Here we provide a complete and detailed protocol for the workup of human blood samples for deep sequencing of HIV-1 genomes.

2 Materials

1. Ethanol (96–100%).
2. 1.5 mL Microcentrifuge tubes.
3. Sterile, RNase-free pipet tips.
4. Microcentrifuge (with rotor for 1.5 and 2 mL tubes).
5. Vortex.
6. The QIAamp Viral RNA Mini kit (Qiagen) (*see* **Note 1**).
7. Material of interest, i.e., plasma/serum sample (*see* **Note 2**).

3 Methods

3.1 RNA Extraction

We performed RNA isolation in a clean room, the pre-PCR room, to prevent contact of the samples with PCR amplicons. Such amplicons could serve as template in the subsequent PCR reactions and thus should be prevented from contaminating the sample, reagents, and tubes used for sample preparation, and all related laboratory equipment. Cross-contamination between individual patient samples was checked by including "no-template controls" during RNA isolation.

1. Thaw the stored plasma in a biohazard hood.
2. Thaw the dissolved carrier RNA from the kit (stored at −20 °C).
3. Prepare two 2 mL Eppendorf tubes for 250 μL plasma.
4. Pipet 560 μL AVL buffer into tube A and 440 μL AVL buffer into tube B (*see* **Note 1**).
5. Add 5.6 and 4.4 μL carrier RNA, respectively, to these tubes, vortex, and centrifuge briefly.
6. Mix the thawed plasma and centrifuge briefly (*see* **Note 2**).
7. Add 140 μL plasma to tube A and the remainder of 110 μL plasma to tube B.
8. Vortex for 15 s.
9. Keep the tubes for 10 min at room temperature.
10. Centrifuge briefly.
11. Add 560 μL and 440 μL 100% ethanol to tubes A and B, respectively (*see* **Note 3**).
12. Vortex briefly and centrifuge the tubes in an Eppendorf centrifuge.
13. Place a QIAamp spin column in a 2 mL tube.
14. Add 630 μL of the lysed plasma onto the column carefully, without wetting the tube itself (*see* **Note 4**).
15. Close the tube and centrifuge for 1 min at 6000 × *g*.
16. Place the column in a new 2 mL tube and discard the filtrate.
17. Repeat **steps 14–17** until all lysed plasma has been applied to the column (*see* **Note 5**).
18. Open the column carefully after the last spin and add 500 μL buffer AW1 (*see* **Note 5**).
19. Centrifuge for 1 min at 6000 × *g*.
20. Place the column in a new 2 mL tube.
21. Open the tube and add 500 μL buffer AW2 (*see* **Note 6**).
22. Centrifuge at maximum speed (20,000 × *g*) for 3 min (*see* **Note 7**).

23. Place the column in a new 2 mL tube and centrifuge for 1 min at maximum speed (20,000 × *g*).
24. Place the column in a new tube and add 40 μL buffer AVE to the column carefully (*see* **Note 6)**.
25. Incubate for 1 min at room temperature to allow elution of the nucleic acids from the column.
26. Centrifuge the column for 1 min at 6000 × *g*.
27. Repeat **steps 25–27** to increase the efficiency of nucleic acid elution (*see* **Note 8)**.
28. Store the nucleic acid (80 μL) at −80 °C for downstream procedures.

3.2 Results

For the initial BEEHIVE pilot study, including 125 plasma samples from HIV-1-positive individuals, we reported that the generation of the complete set of four HIV-1 PCR amplicons is most efficient when the QIAamp Viral RNA Kit is combined with manual RNA extraction [5]. The same kit also performed relatively well when combined with robotic RNA extraction, suggesting that this RNA extraction method outperforms those of other suppliers, at least for HIV-1 RNA extraction from plasma samples. We suggested that manual HIV-1 RNA extraction prevents shearing of the 9 kb viral RNA genome, which may occur during robotic RNA extraction [5].

1. Using this optimized protocol with 616 BEEHIVE plasma samples from HIV-1-positive patients, we reported that the amplification success rate is correlated with the size of the amplicon: 73% for the smallest (1.9 kb) amplicon to 62–65% for the larger (>3 kb) amplicons [5]. Both duration of storage and viral load of the plasma sample also influence the success rate. However, there must be additional variables as we were able to generate complete HIV-1 genomes from some samples with low viral loads, and failed to do so from some samples with high viral loads. Other possibilities include failure of PCR amplification due to sequence variation, or sample degradation caused by other factors than long storage times. Batch effects were detectable in the output.
2. HIV-1 read data generated by our RNA extraction, amplification, and sequencing protocol were recently used to reconstruct complete genomes, although we have to make one reservation. While we and others are usually referring to “complete” HIV-1 genome sequences, parts of the long terminal repeats (LTR) which contain important regulatory elements are often missing [7]. The complete genomes were reconstructed using the new tool shiver [2], which first checks the reads for quality and contamination before mapping them onto a reference specific to the sample, constructed using IVA [13].

Our protocol was also used in a subsequent study which confirms viral genetic variation as a major source of variability of the HIV-1 set-point viral load (Blanquart et al., submitted).

3. In conclusion, manual extraction of HIV-1 RNA from plasma with the QIAamp Viral RNA Mini Kit is the preferred method to generate complete HIV-1 genome sequences. Manual extraction obviates the need for a robot, but is both labor intensive and time consuming, which can be problematic when large sample numbers need to be processed. It remains to be verified whether this method is also optimal for other applications, such as the extraction of the RNA or DNA genomes of other viruses.

4 Notes

1. The QIAamp Viral RNA Mini kit was used for viral RNA extraction from stored plasma samples. RNA was extracted according to the manufacturer's instructions with a few modifications. The first step is the lysis of the sample by guanidinium thiocyanate. Guanidinium thiocyanate is a chaotropic agent that denatures proteins, which also results in deactivation of RNases to ensure isolation of intact viral RNA. The next step is the selective binding of RNA to a silica-based membrane in a spin column. The column is washed extensively to remove contaminants and subsequently the RNA is eluted in an RNase-free buffer. The QIAamp Viral RNA Mini kit includes mini spin columns, carrier RNA, buffers, and 2 mL collection tubes.

   The manufacturer's protocol is optimized for RNA purification from a 140 μL plasma/serum sample. As we increased the sample volume from 140 to 250 μL plasma/serum for manual isolation, the amount of the AVL buffer for the carrier RNA was adjusted proportionally. Carrier RNA and AVE as well as AVL buffer are included in the kit.

   The lyophilized carrier RNA (310 μg) was dissolved in 310 μL AVE buffer to obtain a solution of 1 μg/μL, which was aliquoted and stored at −20 °C. Two solutions were prepared by adding 5.6 μL dissolved carrier RNA to 560 μL AVL buffer (tube A) and 4.4 μL dissolved carrier RNA to 440 μL AVL buffer (tube B).

2. Frozen plasma or serum samples of 250 μL that were stored at −80 °C were thawed and briefly centrifuged. One part of the sample (140 μL) was added to tube A and the remainder (approximately 110 μL) to tube B. Samples and buffer were mixed for 15 s followed by an incubation at room temperature for 10 min to ensure complete lysis of virus particles and deactivation of RNases.

3. The tubes were centrifuged for a second to remove any liquid from the inside of the lid. An equal amount of 100% ethanol was added (560 μL to tube A, 440 μL to tube B) and pulse **vortex**-mixing for 15 s to create a homogenous solution, which is critical to ensure efficient binding of the viral RNA to the QIAamp Mini column.
4. The column was placed inside a 2 mL collection tube, and part of the mixture (630 μL) was added to the column. When pipetting the mixture onto the column it is important to avoid wetting the tube walls, as remnants of the AVL buffer (with guanidinium thiocyanate) will inhibit downstream enzymatic reactions.
5. The caps were closed and the spin column with the collection tube was centrifuged at 6000 × *g* for 1 min. The column was placed into a clean 2 mL collection tube and the filtrate was discarded. Another 630 μL of the mixture was added to the column and centrifuged. This step was repeated four times until the complete lysate was loaded onto the column.
6. The AW1 and AW2 buffers are supplied in the kit as concentrated stock solutions. Before first use, 96–100% ethanol was added as described in the manufacturer's protocol. For 250 isolations, 130 mL ethanol was added to 98 mL AW1 concentrate and 160 mL ethanol to 66 mL AW2 concentrate, respectively. Both buffers are stable for 1 year when stored closed at room temperature. The 310 μg lyophilized carrier RNA was dissolved in 310 μL AVE buffer to obtain a solution of 1 μg/μL, divided into conveniently sized aliquots, and stored at −20 °C. Before starting, samples and AVE buffer were equilibrated to room temperature.
7. After adding 500 μL of the second wash buffer AW2, the column was centrifuged at maximum speed (20,000 × *g*) for 3 min. Centrifugation at maximum speed is important because residual AW2 buffer in the eluate can cause problems in downstream processes. The filtrate of this centrifugation step was discarded and a centrifugation at maximum speed for 1 min at 20,000 × *g* was performed.
8. Lastly, elution of RNA was performed using the AVE buffer. Two elutions with 40 μL of buffer AVE each were performed by centrifugation at 6000 × *g* for 1 min. The viral RNA was stored at −80 °C. The complete protocol takes approximately 2 h.

Acknowledgments

The authors want to thank Suzanne Jurriaans, Matthew Hall, Margreet Bakker, Ard van Sighem, Daniela Bezemer, Swee Hoe Ong, Luuk Gras, Peter Reiss, Paul Kellam, and the BEEHIVE collaboration.

Funding

This work was supported by a European Research Council (ERC) grant 339251 to CF.

References

1. Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M (2016) Challenges in the analysis of viral metagenomes. Virus Evol 2(2):vew022–vew022. https://doi.org/10.1093/ve/vew022
2. Wymant C, Blanquart F, Gall A, Bakker M, Bezemer D, Croucher NJ, Golubchik T, Hall M, Hillebregt M, Ong SH, Albert J, Bannert N, Fellay J, Fransen K, Gourlay A, Grabowski MK, Gunsenheimer-Bartmeyer B, Günthard HF, Kivelä P, Kouyos R, Laeyendecker O, Liitsola K, Meyer L, Porter K, Ristola M, van Sighem A, Vanham G, Berkhout B, Cornelissen M, Kellam P, Reiss P, Fraser C (2016) Easy and accurate reconstruction of whole HIV genomes from short-read sequence data. bioRxiv. https://doi.org/10.1101/092916
3. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, Koblin BA, Heneine W, Lemey P, Jaffe HW (2016) 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature 539(7627):98–101. https://doi.org/10.1038/nature19827
4. Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, Brennan CA (2016) A pan-HIV strategy for complete genome sequencing. J Clin Microbiol 54(4):868–882. https://doi.org/10.1128/jcm.02479-15
5. Cornelissen M, Gall A, Vink M, Zorgdrager F, Binter S, Edwards S, Jurriaans S, Bakker M, Ong SH, Gras L, van Sighem A, Bezemer D, de Wolf F, Reiss P, Kellam P, Berkhout B, Fraser C, van der Kuyl AC (2017) From clinical sample to complete genome: comparing methods for the extraction of HIV-1 RNA for high-throughput deep sequencing. Virus Res 239:10–16. https://doi.org/10.1016/j.virusres.2016.08.004
6. Luk KC, Berg MG, Naccache SN, Kabre B, Federman S, Mbanya D, Kaptue L, Chiu CY, Brennan CA, Hackett J Jr (2015) Utility of metagenomic next-generation sequencing for characterization of HIV and human pegivirus diversity. PLoS One 10(11):e0141723. https://doi.org/10.1371/journal.pone.0141723
7. Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y, Yokomaku Y, Iwatani Y, Sugiura W (2015) Quasispecies analyses of the HIV-1 near-full-length genome with Illumina MiSeq. Front Microbiol 6:1258. https://doi.org/10.3389/fmicb.2015.01258
8. Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, Berry N, Pillay D, Kellam P (2012) Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol 50(12):3838–3844. https://doi.org/10.1128/jcm.01516-12
9. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y, Casali M, Streeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Gunthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereyra F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8(3):e1002529. https://doi.org/10.1371/journal.ppat.1002529
10. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA (2015) Population genomics of intrapatient HIV-1 evolution. elife 4:e11282. https://doi.org/10.7554/eLife.11282
11. Giallonardo FD, Topfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR, Patrignani A, Daumer M, Beisel C, Rusert P, Trkola A, Gunthard HF, Roth V, Beerenwinkel N, Metzner KJ (2014) Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res 42(14):e115. https://doi.org/10.1093/nar/gku537
12. Brener J, Gall A, Batorsky R, Riddell L, Buus S, Leitman E, Kellam P, Allen T, Goulder P, Matthews PC (2015) Disease progression despite protective HLA expression in an HIV-infected transmission pair. Retrovirology 12:55. https://doi.org/10.1186/s12977-015-0179-z
13. Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD (2015) IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics (Oxford, England) 31(14):2374–2376. https://doi.org/10.1093/bioinformatics/btv120

Chapter 6

Monolith Chromatography as Sample Preparation Step in Virome Studies of Water Samples

Ion Gutiérrez-Aguirre, Denis Kutnjak, Nejc Rački, Matevž Rupar, and Maja Ravnikar

Abstract

Viruses exist in aquatic media and many of them use this media as transmission route. Next-generation sequencing (NGS) technologies have opened new doors in virus research, allowing also to reveal a hidden diversity of viral species in aquatic environments. Not surprisingly, many of the newly discovered viruses are found in environmental fresh and marine waters. One of the problems in virome research can be the low amount of viral nucleic acids present in the sample in contrast to the background ones (host, eukaryotic, prokaryotic, environmental). Therefore, virus enrichment prior to NGS is necessary in many cases. In water samples, an added problem resides in the low concentration of viruses typically present in aquatic media. Different concentration strategies have been used to overcome such limitations. CIM monoliths are a new generation of chromatographic supports that due to their particular structural characteristics are very efficient in concentration and purification of viruses. In this chapter, we describe the use of CIM monolithic chromatography for sample preparation step in NGS studies targeting viruses in fresh or marine water. The step-by-step protocol will include a case study where CIM concentration was used to study the virome of a wastewater sample using NGS.

Key words CIM, Monolithic chromatography, Concentration, Water virome, NGS

1 Introduction

Environmental waters (including river, marine, tap water, wastewater, recreational water, irrigation water, and more) represent a confirmed niche for many plant, animal, and bacterial viruses [1–4]. Some of them use this aquatic milieu as a route for transmission to their hosts [4, 5]. These viruses are usually structured as highly stable particles that can persist for a long time in the environment. Enteric viruses such as rotaviruses (RoV) and noroviruses (NoV) are good examples, being one of the major causes of waterborne infections and outbreaks in developing and developed worlds, respectively [2, 6]. There is also an increasing number of plant viruses that have been confirmed to use water as transmission

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_6, © Springer Science+Business Media, LLC 2018

route, examples of which have been extensively described in Mehle et al. (2012) [3]. Due to their stability and highly efficient mechanic transmission, such viruses pose a real risk in hydroponic based vegetable cultures [3, 5]. Aquatic environments also harbor a great diversity of unknown, yet-to-be-discovered viral species [7].

The presence of viruses in water, together with the associated risk, calls for efficient methods to research their occurrence and fate in aquatic media. An ideal method should allow (1) to evaluate the dynamics of pathogenic virus populations in environmental waters, (2) to monitor for selected indicator or pathogenic viruses, and (3) to search for potential new risks. Quantitative polymerase chain reaction (qPCR) or digital PCR (dPCR) are the methods of choice to study specific viruses in water samples, because they are highly sensitive, quantitative, and specific [8, 9]. However, PCR, being a targeted method, does not give generic information on broader virus populations that may be present in a given environmental water sample nor is able to find new viruses. The advent of next-generation sequencing (NGS) technologies is contributing to solve this issue [10]. NGS can be applied to target all the DNA and RNA sequences present in a given water sample, therefore enabling insight into the virus species composition, facilitating the finding of new viruses, and even allowing to study viral populations [7, 11–13]. However, the very generic character of NGS itself can become a disadvantage when sensitivity is required. In such cases the detection of nucleic acids of less abundant species can be hindered by the vast background of more abundant ones, possibly overlooking some of the above-mentioned plant and enteric viruses, which are usually present in water at low, yet infective concentrations. To account for this, different strategies to enrich samples in virus nucleic acids before NGS have been implemented. Different filtrations, PEG precipitation, CsCl ultracentrifugations, DNase/RNase treatments, and inclusion of a preamplification step within the library preparation are some of the options that can be used to increase the sensitivity of NGS when investigating water viromes [11, 12, 14]. None of them is exempt of drawbacks, such as low virus recoveries, nonhomogeneous enrichment of viruses, or bias introduction [11, 12, 14].

An ideal enrichment method should be able to concentrate viruses in a generic way while preferably excluding other abundant organisms such as bacteria. Convective interactive media (CIM) monoliths are chromatographic supports that have demonstrated high efficiency for the concentration of different biomolecules including viruses from a variety of water samples including tap water, river water, bottled water, treated wastewater, and seawater [15–20]. Different (CIM) chemistries allow concentration of viruses based on different interactions [20]. For example in fresh, nonsaline waters, the use of CIM quaternary amine (QA) positively charged columns has enabled concentration of many different human and plant viruses, such as RoV, NoV, astrovirus, sapovirus, hepatitis A virus, and tomato mosaic virus [15, 17, 18]. At a close to neutral pH, in the environment, most viruses are negatively charged and therefore majority of them can bind to the positively charged CIM QA and be subsequently concentrated. In addition, the 1.3–2 μm pore size of the CIM monolith or the inclusion of a prefilter with a smaller pore size (0.8 μm) before the CIM column [21] can aid in removing larger organisms, such as bacteria, from the sample. Recently, a hydrophobic interaction high-density CIM butyl (C4) column has been efficiently used to concentrate RoV and NoV from saline coastal waters [19]. In this case, due to the presence of salt, hydrophobic interactions prevailed to electrostatic ones. In the above-mentioned publications usually 1–5 L water samples were processed using 8 mL CIM columns at 20–100 mL/min flow rates. One of the advantages of the CIM method is that it is scalable and that 80 mL, 800 mL, and even 8000 mL columns are available, which enable processing much higher water volumes at proportionally higher flow rates.

In this chapter, we describe a CIM-based sample preparation method intended for NGS experiments that target viruses in water samples, both fresh and saline. A case study NGS experiment performed with a water sample prepared in such a way is also presented.

2 Materials

2.1 Materials

1. CIMmultus™ QA-8 mL Advanced Composite Column (BIAseparations, Slovenia) for freshwater (Fig. 1) (*see* **Note 1**).
2. CIMmultus™ C4 HLD-8 mL Advanced Composite Column (BIAseparations, Slovenia) for saline water (Fig. 1) (*see* **Note 1**).
3. 50 mM Hepes pH 7 (equilibration buffer for CIM QA 8 mL and elution buffer for CIM C4 8 mL).
4. 50 mM Hepes pH 7, 0.6 M NaCl pH 7 (equilibration buffer for CIM C4 8 mL).
5. 50 mM Hepes, 1 M NaCl pH 7 (elution buffer for CIM QA 8 mL).
6. 20% Ethanol (storage solution for CIM QA 8 mL).
7. 1 M NaOH (sanitization solution).
8. 10 mM NaOH (storage solution for CIM C4 8 mL).
9. 0.22 μm MWCO cellulose acetate filters (Millipore).
10. 0.8 μm MWCO cellulose acetate filters (Sartorius) (Fig. 1c, d) (*see* **Note 2**).

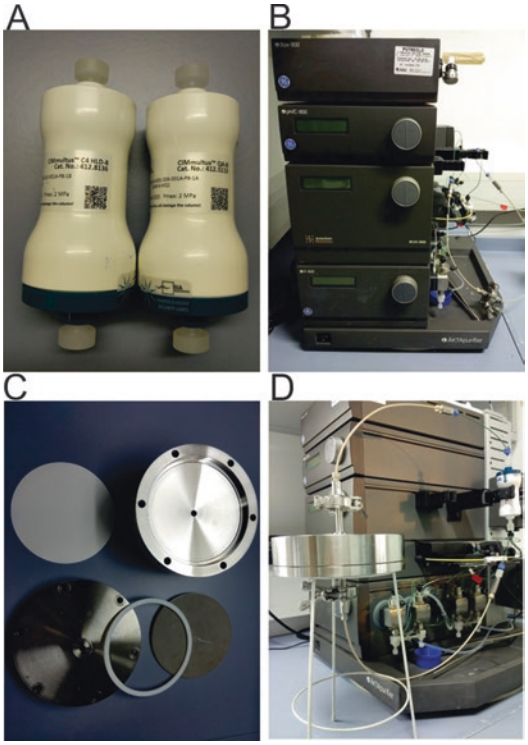


Fig. 1 (**a**) CIM QA and CIM C4 HLD 8 mL columns used for concentration of viruses from fresh and marine water, respectively. (**b**) AKTA purifier 100 is one of the liquid-handling options for concentrating viruses using CIM columns. (**c**) Metallic in-line housing for the 0.8 μm MWCO filter (142 mm in diameter). The two housing components plus the frit, O-ring, and filter are shown. (**d**) AKTA purifier 100 with mounted column (*right*) and in-line filter (*left*)

11. Glycogen for molecular biology (Roche).
12. TRIzol LS (Life Technologies) (*see* **Note 3**).
13. Maxtract high-density 2 mL tubes (Qiagen).

2.2 Equipment

1. A liquid-handling device (FPLC, HPLC, or similar) is required for pumping the water through the CIM column and detecting the elution of the bound viruses. We used an AKTA purifier 100 (GE Healthcare) (Fig. 1d) equipped with pumps, UV detector, and conductivity monitor. Other options are available as well, such as modular pumps and detectors, which may in addition enable onsite applicability of the method [21] (*see* **Note 4**).
2. Nanodrop 1000 (Thermo Scientific) or similar device, to measure spectrophotometrically the amount of nucleic acid after extraction.
3. pH meters and standard buffer filtration systems.

3 Methods

The workflow can be divided into the following steps: (1) preparation of the water sample and CIM column, (2) binding and elution of viruses and column regeneration, and (3) nucleic acid isolation and quantification.

3.1 Preparation of the Water Sample and CIM Column

1. 5–10 L of water sample is collected in appropriate containers. The volume will depend on (a) the type of water sample, (b) expected virus concentrations in the sample, (c) CIM column volume, and (d) potential time limitations (*see* **Note 5**). Usually, water is first filtered through standard cellulose filter paper to remove any larger particles or organic matter present in the sample. Tap water or bottled water does not need any filtration step, but it is essential for, e.g., wastewater, seawater, and ponds. Such type of waters is further filtered through a 0.8 μm MWCO cellulose acetate filter to remove particles or microorganisms larger than 800 μm, which could contribute to a fast column clogging. This can be done previously to the sample loading, or simultaneously using an inline filter holder as the one shown in Fig. 1c, d or similar (*see* **Note 2**).
2. Each concentration run needs an equilibration buffer to prepare the column for the binding of viruses and an elution buffer to elute the bound viruses. Equilibration and elution buffers that are used for fresh and marine water concentration using CIM QA and CIM C4 columns are shown in Sect. 2.1. All buffers are filtered through 0.22 μm MWCO filters before use.
3. Column conditioning is done by flushing in this order: 10 column volumes (CV) of equilibration buffer followed by 10 CV of elution buffer and ending with 10 CV of again equilibration buffer. The column is then prepared for loading the water sample.

3.2 Binding and Elution of Viruses and Column Regeneration

1. The sample is then loaded into the column, directly using the pump instead of superloops or any other injecting device. During the load, parameters as UV absorbance at 280 nm, backpressure, and conductivity are monitored (Fig. 2a). The main limiting factor for choosing the flow rate is the backpressure, which for CIM multus 8 mL columns cannot exceed 2 MPa. When using, e.g., an AKTA purifier 100 and a new column, 5 L of prefiltered wastewater sample can be loaded at 100 mL/min in 50 min. When loading such type of sample at such a flow rate, the backpressure increases from values close to 1.3 MPa to values close to 1.9 MPa. With increasing uses, the backpressure in the CIM column increases faster and it may be necessary to decrease the flow rate at a point, to prevent backpressure from exceeding 2 MPa (*see* **Note 6**) (Fig. 2a).

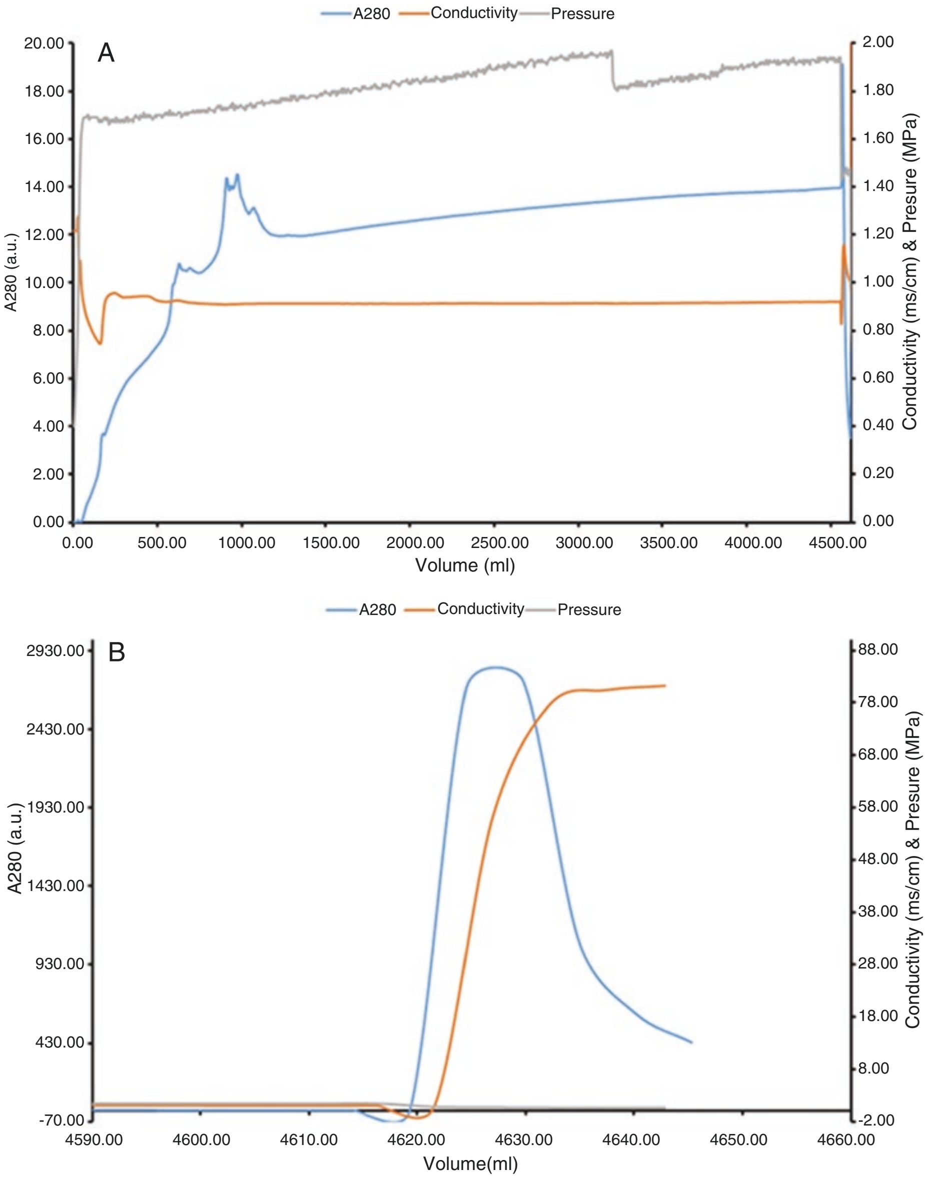


Fig. 2 (**a**) Chromatogram corresponding to the loading of 4.5 L of wastewater effluent into a CIM QA 8 mL column. Note that A280 increases, conductivity remains low, and pressure steadily increases. At cca. 3 L, the flow rate was decreased from 100 mL/min to 80 mL/min to avoid the pressure to reach 2 MPa. (**b**) Elution part of the same chromatogram. Scales are different than in A. Note how the A280 and conductivity increase indicating the elution of bound viruses

2. After loading the sample, both pump and column must be washed using equilibration buffer. The UV signal, which increased during the loading of the sample, needs to decrease, during the washing step, to levels similar as those before the loading (Fig. 2a).

3. For the elution, first, the pump needs to be washed with elution buffer. The flow rate is decreased to 10 mL/min and then also the column is flushed with elution buffer. The conductivity will start to increase in the case of QA column (freshwater) (Fig. 2b) and decrease in the case of C4 column (marine water). At the same time the UV will start to increase, in both QA and C4 columns, indicating the beginning of the elution (Fig. 2b). Typically, for a 5 L water sample, 10–15 mL of elution is collected. A typical chromatographic run is shown in Fig. 2.
4. For regeneration, the column needs first to be sanitized. Pumps are washed with 1 M NaOH solution, which is then flushed through the column for 2 h at low flow rates (i.e., 0.5 mL/min) to achieve sanitization. The column is regenerated by flushing elution buffer until the pH reaches neutral values. Then the column is washed with 10 CV of Milli Q water followed by 10 CV of storage buffer, 20% ethanol for CIM QA, and 10 mM NaOH for C4 column.

3.3 Nucleic Acid Isolation and Quantification

1. There are many different methods available for nucleic acid extraction (*see* ref. [12] and **Note 3**). We used TRIzol LS, following the manufacturer's protocol with the following modifications: (a) 100 μg of glycogen was added to 400 μL of concentrated sample in the beginning of the extraction as a carrier to prevent losses of low-abundance nucleic acids, and (b) Maxtract tubes were used to facilitate the separation between the aqueous and organic phases during the protocol.
2. Purified total nucleic acids need to be quantified before being submitted to the sequencing service. We used Nanodrop spectrophotometer. The highest value that we have measured using Nanodrop was 165 ng/μL of total RNA, after CIM concentration of a raw sewage sample. There are also other options, such as the fluorescence-based Qubit, which is more sensitive and accurate than Nanodrop at lower concentrations. The efficiency of the virus concentration can additionally be tested using qPCR-based quantification of ubiquitous viruses, such as RoV or Pepper mild mottle virus (PMMoV) (*see* **Note 7**).
3. Quantified nucleic acids can now be used for preparation of NGS libraries.

3.4 Case Study: Virome of a Wastewater Effluent Concentrated with CIM QA

1. Following we will present, as an example, the case of a wastewater effluent concentrated in October 2012. 5 L of water was first filtrated through filter paper and then concentrated using a CIM QA 8 mL column with an inline 0.8 μm MWCO cellulose acetate filter. The chromatogram and corresponding elution peak can be seen in Fig. 2a, b.

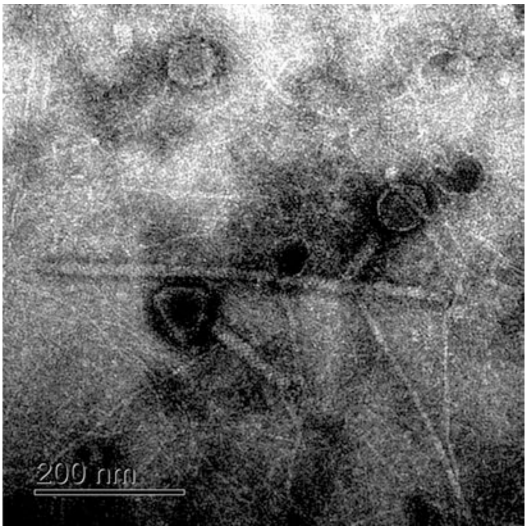


Fig. 3 Electron microscopy micrograph of the concentrated elution from Fig. 2. The complexity of the sample with presence of bacteriophages and other particles such as filaments and rods is evident

2. Electron micrographs from the concentrated fraction showed presence of different viruses, mostly phages and some filaments and rods that could correspond to plant viruses (Fig. 3).
3. Total RNA was isolated with TRIzol LS as explained above. The concentration of isolated total RNA, determined by Nanodrop, was 41 ng/μL. To assess the outcome of the concentration step, we applied RNA purified from the water sample, both before and after concentration, to quantitative PCR (qPCR) assays specific for five different viruses that are typically found in wastewater (rotavirus, norovirus genogroups I and II, astrovirus, and sapovirus) [18]. The reduction of the quantification cycles (Cq) ranged from 6 to 9 depending on the virus, indicating a concentration factor of around 1.5–2.5 orders of magnitude [18].
4. Sequencing libraries for Illumina platform were prepared from isolated total RNA following Illumina's directional RNAseq protocol (15018460 Rev. A), substituting Illumina reagents with other suppliers' reagents as described in Chen et al. [22], and sequenced on HiSeq2000 platform in 2 × 100 bp mode (Fasteris, Switzerland). The sequencing results were first checked for quality using CLC Genomics Workbench and then subjected to shotgun metagenomics analysis using MG-RAST pipeline [23]. A large fraction of the reads in this sample was identified as bacteriophage sequences, a notable amount of plant virus sequences (mostly from *Tobamovirus* genus, family *Virgaviridae*, but also others) were present as well, and also some human enteric viruses were detected. Besides known viruses, a large portion of reads (more than 7%)

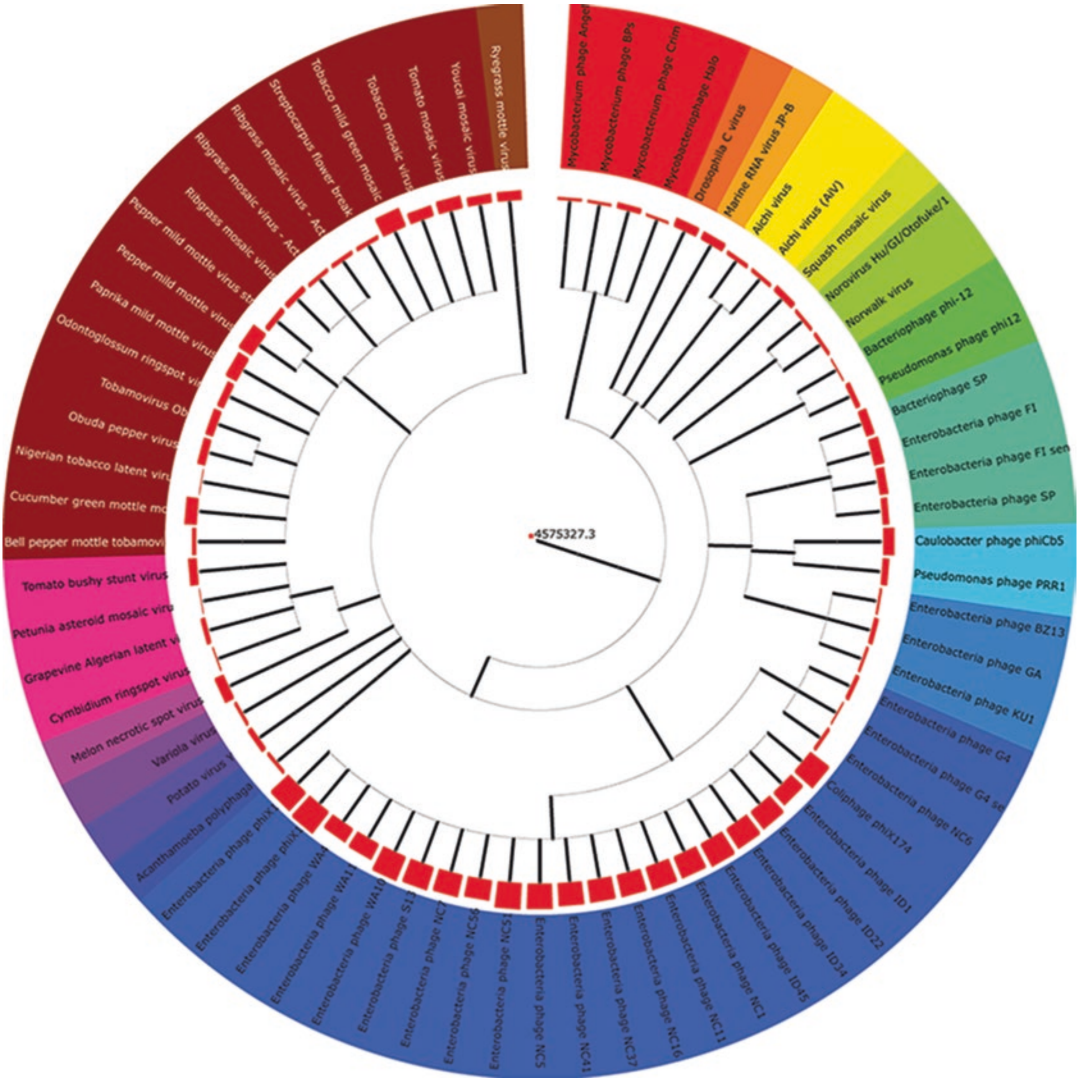


Fig. 4 Example of the visualization of the wastewater sample virome, obtained after the concentration with CIM monoliths and sequencing using Illumina platform. Circular plot was produced using MG-RAST pipeline and it shows the viral taxa and their normalized abundance (red bars next to the taxon labels) in the investigated wastewater sample. Different viral families are color coded

corresponded to yet-unknown virus, with highest similarity to viruses from family *Nodaviridae* (~40% identity of amino acid sequence to the most similar species). An example of the visualization of virome composition is shown in Fig. 4. Here, we showed an example of virome analysis from a concentrated wastewater sample; although Illumina is a current market leader in NGS technologies, other sequencing platforms can also be used, as well as other analysis pipelines (*see* **Note 8**).

4 Notes

1. 8 mL CIM columns, which allow flow rates of 100 mL/min, are suitable for concentration of 1–10 L of water. In addition, the manufacturer BIA Separations (Slovenia) offers scale-up possibilities, with columns of 80, 800, and 8000 mL volumes. The latter allows working at 10 L/min flow rates. In this case, the limiting factor may be having access to a pump with such specifications. Lower volume monoliths of 1 and 0.34 mL are also available for more analytical applications.
2. We use a Sartorius 0.8 μm MWCO filter placed in a metallic in-line housing designed for this particular purpose (see Fig. 2c); however, there are other possibilities for prefiltration. The water sample, instead of in-line, can be previously filtered in a separate step by other means, i.e., with a vacuum-driven system. Alternatively, in-line disposable filters can also be ordered from commercial liquid filtration companies, such as Sartorius or Millipore. Attention should be paid to the material (cellulose acetate results in low nonspecific binding of viruses) and the maximum operating pressure (up to 2 MPa). For some samples, such as tap water or bottled water, this step can be omitted.
3. There are many options available for nucleic acid isolation and each method will surely introduce a certain degree of bias. The selection will depend mostly on the type of virus species that is being targeted, RNA viruses, DNA viruses, or both. Some methods for RNA isolation, for example, include a DNase step. Attention should be paid to the addition of nucleic acid carriers recommended by some kits, because it might increase the background reads after sequencing. In a recent study four different methods were compared in viral metagenomics studies of wastewater [12], including magnetic bead-based and silica column-based methods. They found out that the sequencing outcome depends on the method used. For example, Qiagen Viral RNA mini kit was the method resulting in higher richness of viral species, while Nucleospin RNA XS resulted in highest proportion of viral reads from the total. TRIzol LS was not compared in the study. In conclusion, the nucleic acid isolation method should be chosen in regard to the viruses that are being targeted and having in mind that there will always be an associated bias; thus if results between different samples need to be compared, it is advisable to choose the same isolation method.
4. Any liquid-handling device could, in principle, be used to perform a concentration using CIM columns. A UV detector is strongly recommended as it allows to accurately monitor for the elution of the bound viruses; however, once the elution

parameters have been defined, and the exact volume at which viruses elute is known, the use of the detector can be skipped. Conductivity detector is also advisable to monitor for the ionic strength of the buffers and water samples, but it is not essential, once the conditions of the chromatography have been set up. Knauer (Germany) offers a variety of modular pumps and UV detectors that due to their modularity can also be used on-site. Gutierrez et al. [20] concentrated rotaviruses in the field using a CIM QA column, a modular Knauer UV 200 detector, and a LMI 71 dosing pump from Milton Roy (Milton Roy Europe). Typical laboratory peristaltic pumps usually do not reach the operating backpressures needed to pump, i.e., wastewater through a CIM column at reasonable flow rates.

5. The volume to be concentrated depends largely on the type of water sample. In the case of tap water or bottled water, larger volumes (≥10 L) need to be concentrated to get representative results. On the other hand, for such a complex sample as raw sewage 1–2 L will suffice. Moreover, it is difficult to filtrate 2 L of raw sewage through the 0.8 μm MWCO filter without clogging it, being neccessary in such case the use of more than one filter.

6. Despite sanitization and regeneration, it was observed that the lifetime of a CIM column was of around 6–9 concentration cycles for wastewater effluent (for cleaner water samples this number will surely be higher) [24]. The more times the column is used the faster the backpressure does increase when loading the sample, until it reaches a point where it does not increase any more. From this point ahead, the viruses start to elute in the flow through and a new column has to be used. Rački et al. demonstrated that including a CIM OH 8 mL pre-column prolonged the lifetime of the CIM QA column when working with wastewater, without affecting the binding capacity for rotavirus of the CIM QA column [24]. The use of such column in tandem with the CIM QA column would be advisable when working with dirtier samples such as wastewater; however, there is no data on the binding of viruses other than rotaviruses to the OH column.

7. Targeted quantification before and after the concentration step of one or more specific virus, known to be ubiquitous in the analyzed sample, can serve as a control of the efficiency of method. Sensitive molecular methods such as quantitative PCR or digital PCR are recommended. For example in the case of wastewater, rotavirus or pepper mild mottle virus are almost ubiquitous and can serve as such control. Alternatively, and in particular for cleaner water samples where such viruses are not expected, an easy-to-culture virus of known concentration (for example MS2 phage) can be spiked into the sample before the concentration step and then quantified before and after to assess the correct functioning of the column.

8. Different NGS library preparation approaches and sequencing platforms (e.g., Illumina, Ion Torrent or Ion Proton by Thermo Fisher Scientific, 454 by Roche) can be used to determine the virome of the samples. Illumina is currently the preferred choice by many researchers, since it allows the highest throughputs at lowest prices, thus probably providing, at the moment, the best resolution for virome studies. Different options for library preparation exist for Illumina platform, e.g., Truseq and Nextera approaches, the latter offering simpler and probably the most suitable workflow for most virome studies. For some water samples, for which the yield of isolated nucleic acids after the concentration is very low (e.g., when the concentration is too low to be quantified by spectrophotometer or lower than 2 ng/μL), additional preamplification step is advised before library preparation, as described in [12]. For analysis of sequencing results, several analytical tools can be used, either tools for shotgun metagenomics, such as MG-RAST [23], Kraken [25], BLAST [26], and MEGAN6 [27], or tools specific for virome analysis, e.g., Virome [28], Metavir2 [29], ViromeScan [30], and others.

Acknowledgments

The authors are grateful to the Slovenian Research Agency for the grant P4-0165 and to the FP7 EU project EMERAMB from the ERA-net action ARIMNet2. We are also grateful to Dr. Magda Tušek Žnidarič for her help with the transmission electron microscopy.

References

1. Griffin DW, Donaldson KA, Paul JH, Rose JB (2003) Pathogenic human viruses in coastal waters. Clin Microbiol Rev 16:129–143
2. Sinclair RG, Jones EL, Gerba CP (2009) Viruses in recreational water-borne disease outbreaks: a review. J Appl Microbiol 107:1769–1780
3. Mehle N, Ravnikar M (2012) Plant viruses in aqueous environment - survival, water mediated transmission and detection. Water Res 46:4902–4917
4. Gibson KE (2014) Viral pathogens in water: occurrence, public health impact, and available control strategies. Curr Opin Virol 4:50–57
5. Mehle N, Gutiérrez-Aguirre I, Prezelj N, Delic D, Vidic U, Ravnikar M (2014) Survival and transmission of potato virus Y, pepino mosaic virus, and potato spindle tuber viroid in water. Appl Environ Microbiol 80:1455–1462
6. Ashbolt NJ (2004) Microbial contamination of drinking water and disease outcomes in developing regions. Toxicology 198:229–238
7. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV (2010) New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18:11–19
8. Botes M, de Kwaadsteniet M, Cloete TE (2013) Application of quantitative PCR for the detection of microorganisms in water. Anal Bioanal Chem 405:91–108
9. Rački N, Morisset D, Gutierrez-Aguirre I, Ravnikar M (2014) One-step RT-droplet digital PCR: a breakthrough in the quantification of waterborne RNA viruses. Anal Bioanal Chem 406:661–667
10. Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N (2012) Application of next-generation sequencing technologies in virology. J Gen Virol 93:1853–1868

11. Aw TG, Howe A, Rose JB (2014) Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. J Virol Methods 210:15–21
12. Hjelmsø MH, Hellmér M, Fernandez-Cassi X, Timoneda N, Lukjancenko O, Seidel M, Elsässer D, Aarestrup FM, Löfström C, Bofill-Mas S, Abril JF, Girones R, Schultz AC (2017) Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. PLoS One 12:e0170199
13. Kutnjak D, Rupar M, Gutierrez-Aguirre I, Curk T, Kreuze JF, Ravnikar M (2015) Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population. J Virol 89:4760–4769
14. Rastrojo A, Alcamí A (2017) Aquatic viral metagenomics: lights and shadows. Virus Res 239:87–96. https://doi.org/10.1016/j.virusres.2016.11.021
15. Kramberger P, Petrovic N, Strancar A, Ravnikar M (2004) Concentration of plant viruses using monolithic chromatographic supports. J Virol Methods 120:51–57
16. Gutiérrez-Aguirre I, Banjac M, Steyer A, Poljsak-Prijatelj M, Peterka M, Strancar A, Ravnikar M (2009) Concentrating rotaviruses from water samples using monolithic chromatographic supports. J Chromatogr A 1216:2700–2704
17. Kovac K, Gutiérrez-Aguirre I, Banjac M, Peterka M, Poljsak-Prijatelj M, Ravnikar M, Mijovski JZ, Schultz AC, Raspor P (2009) A novel method for concentrating hepatitis a virus and caliciviruses from bottled water. J Virol Methods 162:272–275
18. Steyer A, Gutiérrez-Aguirre I, Rački N, Beigot Glaser S, Brajer Humar B, Stražar M, Škrjanc I, Poljšak-Prijatelj M, Ravnikar M, Rupnik M (2015) The detection rate of enteric viruses and Clostridium Difficile in a waste water treatment plant effluent. Food Environ Virol 7:164–172
19. Balasubramanian MN, Rački N, Gonçalves J, Kovač K, Žnidarič MT, Turk V, Ravnikar M, Gutiérrez-Aguirre I (2016) Enhanced detection of pathogenic enteric viruses in coastal marine environment by concentration using methacrylate monolithic chromatographic supports paired with quantitative PCR. Water Res 106:405–414
20. Krajacic M, Ravnikar M, Štrancar A, Gutiérrez-Aguirre I (2017) Application of monolithic chromatographic supports in virus research. Electrophoresis 38(22–23):2827–2836
21. Gutiérrez-Aguirre I, Steyer A, Banjac M, Kramberger P, Poljšak-Prijatelj M, Ravnikar M (2011) On-site reverse transcription-quantitative polymerase chain reaction detection of rotaviruses concentrated from environmental water samples using methacrylate monolithic supports. J Chromatogr A 1218:2368–2373
22. Chen Y-R, Zheng Y, Liu B, Zhong S, Giovannoni J, Fei Z (2012) A cost-effective method for Illumina small RNA-Seq library preparation us- ing T4 RNA ligase 1 adenylated adapters. Plant Methods 8:41–45
23. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes. BMC Bioinformatics 9:386–393
24. Rački N, Kramberger P, Steyer A, Gašperšič J, Štrancar A, Ravnikar M, Gutierrez-Aguirre I (2015) Methacrylate monolith chromatography as a tool for waterborne virus removal. J Chromatogr A 1381:118–124
25. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
27. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R (2016) MEGAN Community edition–interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol 12:1–12
28. Roux S, Tournayre J, Mahul A, Debroas D, Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15:76
29. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci 6:427–439
30. Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, Candela M (2016) ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics 17:165

Chapter 7

Viral Metagenomics Approaches for High-Resolution Screening of Multiplexed Arthropod and Plant Viral Communities

Sarah François, Denis Filloux, Emmanuel Fernandez, Mylène Ogliastro, and Philippe Roumagnac

Abstract

Viral metagenomic approaches have become essential for culture-independent and sequence-independent viral detection and characterization. This chapter describes an accurate and efficient approach to (1) concentrate viral particles from arthropods and plants, (2) remove contaminating non-encapsidated nucleic acids, (3) extract and amplify both viral DNA and RNA, and (4) analyze high-throughput sequencing (HTS) data by bioinformatics. Using this approach, up to 96 arthropod or plant samples can be multiplexed in a single HTS library.

Key words Metagenomics, Virus discovery, Diagnostic, Arthropod, Plant, Random amplification, High-throughput sequencing

1 Introduction

While viruses are the most numerous biological entities on Earth, the number of currently classified virus species is probably dramatically underestimated [1, 2]. Several factors can account for this lack of knowledge, including intrinsic characteristics of viruses such as their small size, their rapid rate of evolution, or lack of universally conserved viral genetic markers [3]. In addition, the genetic material recovered from animal or plant samples is mostly of nonviral origin [4], which renders difficult the study of the host virome (the collection of all viruses that are found in or on the host).

Viral metagenomics, which is the direct genetic analysis of viral genomes contained within a sample, has revolutionized the last decade the field of virus discovery [5–7]. Viral metagenomics approaches have targeted four main classes of nucleic acids, including (1) total RNA or DNA, (2) virion-associated nucleic acids (VANA) purified from viral particles, (3) double-stranded RNAs

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_7, © Springer Science+Business Media, LLC 2018

(dsRNA), and (4) virus-derived small interfering RNAs (siRNAs) [8–14]. While each of these approaches has advantages and drawbacks, the VANA approach has gained popularity because it takes advantage of the hardiness of many viral capsids for concentrating and purifying the viral nucleic acids and allows the detection of both RNA and DNA viruses [15–18].

Here, we provide a detailed protocol for the VANA-based metagenomics determination of arthropods and plant viromes. Starting from arthropod and plant samples, we first concentrate viral particles by filtration and centrifugation prior to partially removing the non-encapsidated material by DNase and RNase digestion [19, 20]. Encapsidated DNA and RNA are then extracted and RNA is converted to cDNA using a 26 nt primer (Dodeca Linker) composed by a 14 nt linker linked at 3′ end to N_{12} (Fig. 1). Noteworthy, we present here a set of 96 Dodeca Linkers. Double-stranded DNA is synthetized from single-stranded DNA using large (Klenow) fragment DNA polymerase and the Dodeca Linker used during the reverse transcriptase (RT) step (Fig. 1). Double-stranded DNAs are further amplified using one 24 nt PCR multiplex identifier primer composed by the 14 nt linker used during the RT step linked at 5′ end to a 10 nt tag (Fig. 1). This PCR yields amplicons that are all tagged at both extremities with the same multiplex identifier primer (Fig. 1) [21]. Pools of up to 96 multiplex identifier amplicons can then be mixed, which reduces the cost of library preparation in case of numerous samples. This protocol finally describes the bioinformatic data analysis (data demultiplexing, cleanup, de novo assembly, taxonomic assignment, and read mapping).

2 Materials

Prepare and store reagents according to their individual specifications (room temperature if not specified). 1× HBSS solution should be prepared using sterilized ultrapure water. Special care should be taken to keep enzymes at −20 °C until use. Follow all waste disposal regulations when disposing waste materials. To reduce laboratory contaminations during nucleic acid extraction and amplification, working in a clean environment is recommended.

2.1 Purification of Viral Particles

1. Tissue homogenizer and sterile ceramic beads or sterile mortar, pestle, carborundum, and liquid nitrogen.
2. Conical tubes (15 mL).
3. 10× Hanks’ balanced salt solution (HBSS).
4. 0.45 μm Syringe filters.

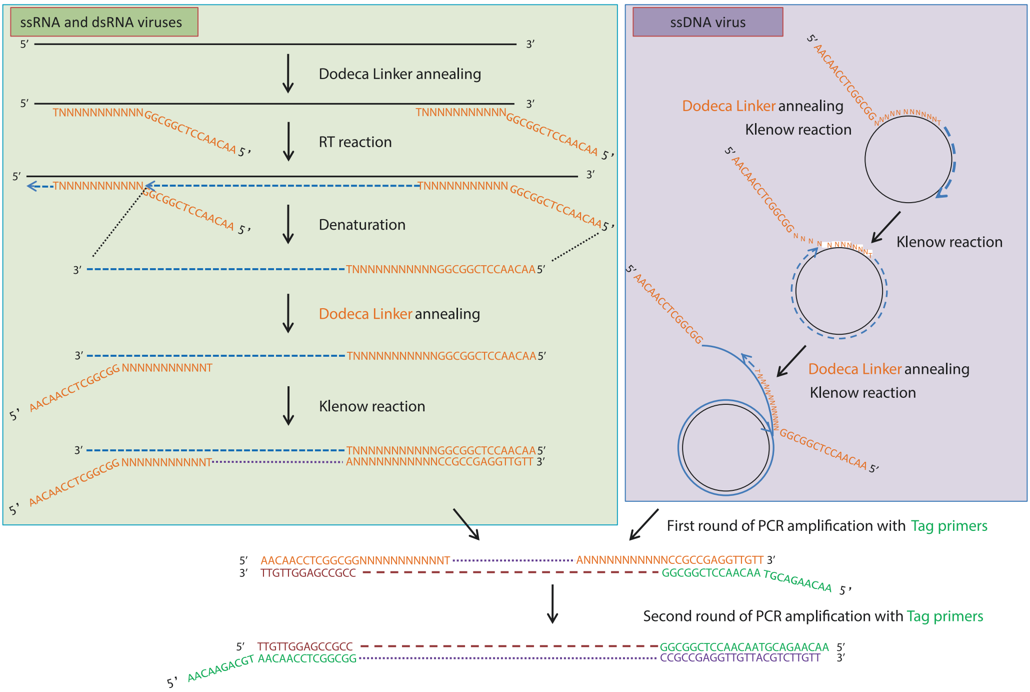

ssRNA and dsRNA viruses
5′
3′
Dodeca Linker annealing
TNNNNNNNNNNNGGCGGCTCCAACAA 5′
RT reaction
Denaturation
3′
TNNNNNNNNNNNGGCGGCTCCAACAA 5′
Dodeca Linker annealing
5′ AACAACCTCGGCGG NNNNNNNNNNNT
Klenow reaction
ANNNNNNNNNNNCCGCCGAGGTTGTT 3′
ssDNA virus
5′ AACAACCTCGGCGG
Dodeca Linker annealing
Klenow reaction
Klenow reaction
Dodeca Linker annealing
Klenow reaction
GGCGGCTCCAACAA 5′
First round of PCR amplification with Tag primers
5′ AACAACCTCGGCGGNNNNNNNNNNNT
ANNNNNNNNNNNCCGCCGAGGTTGTT 3′
3′ TTGTTGGAGCCGCC
GGCGGCTCCAACAA TGCAGAACAA 5′
Second round of PCR amplification with Tag primers
5′ TTGTTGGAGCCGCC
GGCGGCTCCAACAATGCAGAACAA 5′
5′ AACAAGACGT AACAACCTCGGCGG
CCGCCGAGGTTGTTACGTCTTGTT 3′

Fig. 1 Schematic outline of the VANA-based metagenomics method. Top left: Conversion by random priming of viral ssRNA or dsRNA to sequence-ready cDNA, including a reverse transcription step followed by a Klenow reaction step. Top right: Conversion by random priming of viral DNA (e.g., circular ssDNA) to sequence-ready cDNA, including a Klenow reaction step (strand displacement amplification). Bottom: Double-stranded DNA are amplified using one PCR multiplex identifier primer, which yields amplicons that are all tagged at both extremities with the same multiplex identifier primer

5. 5 mL Syringes.
6. Centrifuge Eppendorf 5810R (rotor A-4-62 and F34-6-38 15 mL tubes).
7. Ultracentrifuge polycarbonate bottles (Tube 26.3 mL).
8. Ultracentrifuge.
9. Deionized water.
10. Microtubes (1.5 mL).
11. Single-channel pipettes (0.5–10 μL/10–200 μL/1000–5000 μL).
12. 0.5 mL PCR 8 tube strips.
13. 10 μL Pipette filter tips.
14. 200 μL Pipette filter tips.
15. 1000 μL Pipette filter tips.
16. 5000 mL Pipette filter tips.
17. DNase I.

18. RNase A.
19. Oven.
20. Ice.

2.2 Viral Nucleic Acid Extraction

1. Nucleospin 96 virus Core kit (Macherey Nagel) (*see* **Note 1**).
2. Square-well block.
3. Absolute ethanol.
4. Single-channel pipettes (0.5–10 μL/10–200 μL/100–1000 μL).
5. 10 μL Long pipette filter tips (*see* **Note 2**).
6. 200 μL Pipette filter tips.
7. 1000 μL Pipette filter tips.
8. Centrifuge.
9. Heat block or water bath.

2.3 Reverse Transcription

1. Thermocycler.
2. Single-channel pipettes (0.5–10 μL/10–200 μL/100–1000 μL).
3. 10 μL Long pipette filter tips (*see* **Note 2**).
4. 200 μL Pipette filter tips.
5. 1000 μL Pipette filter tips.
6. PCR plate (96) 0.5 mL.
7. Microtubes (1.5 mL).
8. Nuclease-free water.
9. 10 μM Dodeca Linkers (Table 1).
10. SuperScript III reverse transcriptase (200 U/μL) (Invitrogen).
11. 5× Reverse transcriptase (RT) buffer (supplied with reverse transcriptase).
12. 0.1 M Dithiothreitol (DTT).
13. 10 mM dNTP mix.
14. Ice.

2.4 cDNA Purification

1. RNase A.
2. Thermocycler.
3. QIAquick PCR Purification Kit (Qiagen).
4. Absolute ethanol.
5. Microtubes (1.5 mL).
6. Single-channel pipettes (0.5–10 μL/10–200 μL/100–1000 μL).

Table 1
96 Dodeca Linkers and corresponding PCR primers used to tag each arthropod or plant sample

Dodeca linker	Sequence	Primer	Sequence
1	**AACAACCTCGGCGG**NNNNNNNNNNNT	1	AACAAGACGT**AACAACCTCGGCGG**
2	AACCGAGTCGCGAGNNNNNNNNNNNT	2	AACACACTCAAACCGAGTCGCGAG
3	ATCCGGCGTGATGGNNNNNNNNNNNT	3	AACATGGAGAATCCGGCGTGATGG
4	AACGCTGTCGCAGGNNNNNNNNNNNT	4	AACCAGCCAGAACGCTGTCGCAGG
5	CGGGTCGTCATAGGNNNNNNNNNNNT	5	AAGGAACGTACGGGTCGTCATAGG
6	AAGTGGTACCCGCGNNNNNNNNNNNT	6	AAGGCATGCGAAGTGGTACCCGCG
7	CGTGGAGACTCTGGNNNNNNNNNNNT	7	AAGGTAGAAGCGTGGAGACTCTGG
8	AACGGCCGCCACTANNNNNNNNNNNT	8	AATACACAGCAACGGCCGCCACTA
9	ACCAATGGTCCGGGNNNNNNNNNNNT	9	AATACAGCCTACCAATGGTCCGGG
10	ACGATCCAGCGTCGNNNNNNNNNNNT	10	AATACCGGTAACGATCCAGCGTCG
11	ACGCCATCACACGGNNNNNNNNNNNT	11	AATACTGTGGACGCCATCACACGG
12	ACGGCGTCGGTAGTNNNNNNNNNNNT	12	AATAGCCACAACGGCGTCGGTAGT
13	ACTACCGCACGCTGNNNNNNNNNNNT	13	AATCCGCTCCACTACCGCACGCTG
14	AGAACACCGCGCAGNNNNNNNNNNNT	14	AATTCCTGCTAGAACACCGCGCAG
15	AGAGGATCTGGCGGNNNNNNNNNNNT	15	ACAATTCGAGAGAGGATCTGGCGG
16	AGATGCACCGAGCGNNNNNNNNNNNT	16	ACAGACGTTAAGATGCACCGAGCG
17	AGCCGGATCGTGAGNNNNNNNNNNNT	17	ACCGAACCGTAGCCGGATCGTGAG
18	AGCGAAGGAAGCGGNNNNNNNNNNNT	18	ACCTGATTCTAGCGAAGGAAGCGG
19	AGCTCTCGATCCGGNNNNNNNNNNNT	19	ACGATGAAGTAGCTCTCGATCCGG

(continued)

Table 1
(continued)

Dodeca linker	Sequence	Primer	Sequence
20	AGGACTGGCCGATGNNNNNNNNNNNT	20	ACGCCTCAACAGGACTGGCCGATG
21	AGGCTGGTGCTCAGNNNNNNNNNNNT	21	ACTGGCGCATAGGCTGGTGCTCAG
22	AGTCAACGCGCTCGNNNNNNNNNNNT	22	ACTTACCAAGAGTCAACGCGCTCG
23	AGCTAGGCGCCCTANNNNNNNNNNNT	23	AGAACCACGGAGCTAGGCGCCCTA
24	ATGTCCGCGCTCCTNNNNNNNNNNNT	24	AGAGTGACTTATGTCCGCGCTCCT
25	ATGTGACCGGCTGCNNNNNNNNNNNT	25	AGATAGTGCTATGTGACCGGCTGC
26	CATCCACGCGGTGANNNNNNNNNNNT	26	AGGACATAAGCATCCACGCGGTGA
27	CAAGCGGTAGCCGANNNNNNNNNNNT	27	AGGTGAATAGCAAGCGGTAGCCGA
28	CAAGGCATAGCGCGNNNNNNNNNNNT	28	AGTCCTAATCCAAGGCATAGCGCG
29	CACCTATGCGCCGANNNNNNNNNNNT	29	AGTTGTTGTCCACCTATGCGCCGA
30	CAGAGCGGCACGAANNNNNNNNNNNT	30	ATATCCGCATCAGAGCGGCACGAA
31	CAGCACGTCCGCAANNNNNNNNNNNT	31	ATCTAAGGAGCAGCACGTCCGCAA
32	CAGGAGCGACCTCANNNNNNNNNNNT	32	ATGACGGTAACAGGAGCGACCTCA
33	CCGGCCTACGTCTANNNNNNNNNNNT	33	ATGCGCATATCCGGCCTACGTCTA
34	CCGGTGGTCTCACANNNNNNNNNNNT	34	CAACCGATTGCCGGTGGTCTCACA
35	CACGGGATCGCAGANNNNNNNNNNNT	35	CAACCTCTGACACGGGATCGCAGA
36	CCAAGTACGCGGCANNNNNNNNNNNT	36	CAACGCACAGCCAAGTACGCGGCA
37	CCCGAACGCTGGAANNNNNNNNNNNT	37	CAACGTTAACCCCGAACGCTGGAA
38	CCGACGCTAGGTCANNNNNNNNNNNT	38	CAACTGCTATCCGACGCTAGGTCA
39	CCTGATGCCTACCGNNNNNNNNNNNT	39	CACTACGAATCCTGATGCCTACCG

40	CCCGGTCGCTATCANNNNNNNNNNNT	40	CACTGAGCACCCCGGTCGCTATCA
41	CGAGCTACGCATCGNNNNNNNNNNNT	41	CATTGGCTAACGAGCTACGCATCG
42	CGCCGTTCGCCTTANNNNNNNNNNNT	42	CCACCACACACGCCGTTCGCCTTA
43	CGTCGCCGTATGGANNNNNNNNNNNT	43	CCAGGTCGAACGTCGCCGTATGGA
44	CGTACGAGAGTGCGNNNNNNNNNNNT	44	CCATAACTTGCGTACGAGAGTGCG
45	CGTCCCGTCAACGANNNNNNNNNNNT	45	CCATACTGACCGTCCCGTCAACGA
46	CTAACGGCGGATCGNNNNNNNNNNNT	46	CCGACTTCTCCTAACGGCGGATCG
47	CTCCATAGCGGCCANNNNNNNNNNNT	47	CCGCTATTCGCTCCATAGCGGCCA
48	CTCCCGCTGTACCANNNNNNNNNNNT	48	CCGGAATGCTCTCCCGCTGTACCA
49	CTCGTGCCGGAGTANNNNNNNNNNNT	49	CCGGTCTCTACTCGTGCCGGAGTA
50	GCGCCTGTGGTAGANNNNNNNNNNNT	50	CCGTACGATGGCGCCTGTGGTAGA
51	GAATCCAGCCGCCTNNNNNNNNNNNT	51	CCTCCGTTCTGAATCCAGCCGCCT
52	GACAGATCCGCGCTNNNNNNNNNNNT	52	CCTCTGCGAAGACAGATCCGCGCT
53	GACCACGCACACGTNNNNNNNNNNNT	53	CCTTCCTAGCGACCACGCACACGT
54	GACGCACTGACCGTNNNNNNNNNNNT	54	CGCTTAAGGCGACGCACTGACCGT
55	GCCGTGGACCTAGTNNNNNNNNNNNT	55	CGGACAGAGAGCCGTGGACCTAGT
56	GCAGCGAAGTCGCTNNNNNNNNNNNT	56	CGGTATTAGCGCAGCGAAGTCGCT
57	GCCACACGTGTGCTNNNNNNNNNNNT	57	CGTAGAAGACGCCACACGTGTGCT
58	GCGAACTCACGGCTNNNNNNNNNNNT	58	CGTCCGACTTGCGAACTCACGGCT
59	GCGTGCGACAAGCTNNNNNNNNNNNT	59	CTACTCACTAGCGTGCGACAAGCT

(continued)

Table 1
(continued)

Dodeca linker	Sequence	Primer	Sequence
60	GCCGCCTCAGTCATNNNNNNNNNNNT	60	CTCCACTGAAGCCGCCTCAGTCAT
61	GGAACGCCATCGGTNNNNNNNNNNNT	61	CTCCTATTGTGGAACGCCATCGGT
62	GGACGTTCGGGCTTNNNNNNNNNNNT	62	CTGGAAGTAAGGACGTTCGGGCTT
63	GGACTCACCTCCGTNNNNNNNNNNNT	63	CTGGATTGACGGACTCACCTCCGT
64	GGAGCCGTAGCACTNNNNNNNNNNNT	64	CTTGGAGAACGGAGCCGTAGCACT
65	GGATACGCGTACGGNNNNNNNNNNNT	65	GAATCGCCATGGATACGCGTACGG
66	GGCAACACACCGAGNNNNNNNNNNNT	66	GAATGCTGAAGGCAACACACCGAG
67	GGTCGGGATAGACGNNNNNNNNNNNT	67	GAATTACGGCGGTCGGGATAGACG
68	GGTCCCGCTCATCTNNNNNNNNNNNT	68	GACGGCCTATGGTCCCGCTCATCT
69	GGTCTGTCTCGTCGNNNNNNNNNNNT	69	GATATAGCTCGGTCTGTCTCGTCG
70	GTCGGCAGAGGTGTNNNNNNNNNNNT	70	GATCCAAGAAGTCGGCAGAGGTGT
71	GTCGCGGGTAGAGTNNNNNNNNNNNT	71	GATGATGTTGGTCGCGGGTAGAGT
72	GTCTACCGCGACGTNNNNNNNNNNNT	72	GATGTGACAGGTCTACCGCGACGT
73	GTGACCGACACCGTNNNNNNNNNNNT	73	GCAAGATGTAGTGACCGACACCGT
74	GTGCACGACCACCTNNNNNNNNNNNT	74	GCACCTCTTGGTGCACGACCACCT
75	GTGCATCAGCGGGTNNNNNNNNNNNT	75	GCCATGAGAAGTGCATCAGCGGGT
76	GTGTAGCGGGCTCTNNNNNNNNNNNT	76	GCCGTTCCTTGTGTAGCGGGCTCT
77	TCGCGCCATGCCTTNNNNNNNNNNNT	77	GGAATAAGCATCGCGCCATGCCTT
78	TACAGCGCGGTGCTNNNNNNNNNNNT	78	GGAATTCCAATACAGCGCGGTGCT
79	TACGACCGCTGCACNNNNNNNNNNNT	79	GGCATATACCTACGACCGCTGCAC

80	TAGCTAGCGGTGCCNNNNNNNNNNNT	80	GGCGAAGTATTAGCTAGCGGTGCC
81	CGTCCGGACACATCNNNNNNNNNNNT	81	GGCTGTCTTACGTCCGGACACATC
82	TATGCTCGACCGCCNNNNNNNNNNNT	82	GGTCTTACATTATGCTCGACCGCC
83	TCACCACACCTCGCNNNNNNNNNNNT	S3	GGTTCCTTAATCACCACACCTCGC
84	TCAGCCGGCATACCNNNNNNNNNNNT	84	GTGATTCTCATCAGCCGGCATACC
85	TCATGGCCGTACGCNNNNNNNNNNNT	85	GTTCATTGCCTCATGGCCGTACGC
86	TCCAGGCGGTAGTCNNNNNNNNNNNT	86	GTTGAGCGTATCCAGGCGGTAGTC
87	TCCTCTGATCGGGCNNNNNNNNNNNT	87	GTTGTATGCTTCCTCTGATCGGGC
88	TCCTGCAACGCCTCNNNNNNNNNNNT	88	TAAGCCTCTTTCCTGCAACGCCTC
89	TCGTCGTACACCGCNNNNNNNNNNNT	89	TAGTCCGCTGTCGTCGTACACCGC
90	TCTCTCCAGGCGACNNNNNNNNNNNT	90	TAGTGCAGTCTCTCTCCAGGCGAC
91	TCACGCGGACGATCNNNNNNNNNNNT	91	TATCGTTACGTCACGCGGACGATC
92	TGAGTCCCGGAGACNNNNNNNNNNNT	92	TCCTCTAGTATGAGTCCCGGAGAC
93	TGCCCGTCTGTCTCNNNNNNNNNNNT	93	TCGAGAGAGCTGCCCGTCTGTCTC
94	TGGCCGGCTACTACNNNNNNNNNNNT	94	TGAGGAGTGGTGGCCGGCTACTAC
95	TGGTGAGGCGCTACNNNNNNNNNNNT	95	TGGAATGGAGTGGTGAGGCGCTAC
96	TGTGGTCCTGGCTCNNNNNNNNNNNT	96	TGTTACCTCATGTGGTCCTGGCTC

7. 10 μL Long pipette filter tips (*see* **Note 2**).
8. 200 μL Pipette filter tips.
9. 1000 μL Pipette filter tips.
10. Square-well block.
11. Centrifuge.
12. Ice.

2.5 Klenow Amplification

1. Single-channel pipettes (0.5–10 μL/10–200 μL/ 100–1000 μL).
2. 10 μL Long pipette filter tips (*see* **Note 2**).
3. 200 μL Pipette filter tips.
4. 1000 μL Pipette filter tips.
5. PCR plate (96) 0.5 mL.
6. Microtubes (1.5 mL).
7. Thermocycler.
8. 100 μM Dodeca Linkers (Table 1).
9. 5 U/μL Exo(−) Klenow DNA polymerase I.
10. 10X Exo(−) Klenow Buffer (supplied with Exo(−) Klenow DNA polymerase).
11. Nuclease-free water.
12. 10 mM dNTP mix.
13. Ice.

2.6 PCR Amplification

1. Single-channel pipettes (0.5–10 μL/10–200 μL/ 100–1000 μL).
2. 10 μL Pipette filter tips.
3. 200 μL Pipette filter tips.
4. 1000 μL Pipette filter tips.
5. Microtubes (1.5 mL).
6. PCR plate (96-well) 0.5 mL.
7. HotStar Taq Plus Master Mix kit (Qiagen).
8. 10 μM PCR primers (*see* Table 1).
9. Nuclease-free water.
10. Thermocycler

2.7 Verification of the Composition and Concentration of the PCR Products

1. Microtubes (1.5 mL).
2. Single-channel pipettes (0.5–10 μL/10–200 μL/ 100–1000 μL).
3. 10 μL Pipette filter tips.
4. 200 μL Pipette filter tips.

5. 1000 μL Pipette filter tips.
6. 0.5× TBE buffer: 45 mM Tris–HCl (pH 8.3), 45 mM boric acid, 1 mM EDTA.
7. 1% Agarose (type LE) gel: Prepare in 0.5× TBE buffer. Add GelRed or ethidium bromide for visualization.
8. DNA ladder.
9. 1 kb Plus DNA Ladder (Invitrogen).
10. UV transilluminator.
11. Microtubes (1.5 mL).
12. NucleoSpin gel and PCR clean-up (Macherey Nagel).
13. Absolute ethanol.
14. Nuclease-free water.
15. Qubit 2.0 Fluorometer with Qubit Assay HS Kit for dsDNA (Thermo Fisher Scientific).

2.8 Data Handling and Bioinformatics

1. Intel-based server: 4 × 2 Intel Xeon, 256 GB memory per processor or similar capacity.
2. UNIX-based operating system.
3. FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/).
4. FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).
5. Cutadapt software (https://pypi.python.org/pypi/cutadapt/).
6. SPAdes software (http://cab.spbu.ru/software/spades/).
7. BLAST+ software package (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/).
8. Bowtie2 software (http://bowtie-bio.sourceforge.net/bowtie2/).
9. Integrative Genomics Viewer (http://www.broadinstitute.org/igv/).
10. Web-based resources: NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/).

2.9 Confirmation and Retrieval of Near-Full Genome Sequences

1. Primer3 software (http://primer3.sourceforge.net/).
2. Single-channel pipettes (0.5–10 μL/10–200 μL/100–1000 μL).
3. 10 μL Long pipette filter tips.
4. 200 μL Pipette filter tips.
5. 1000 μL Pipette filter tips.
6. Column-based DNA extraction kit with proteinase K: QIAamp DNA Mini Kit (Qiagen).

7. Column-based RNA extraction kit: RNeasy Mini Kit (Qiagen).
8. First-strand cDNA synthesis kit: SuperScript III reverse transcriptase kit (Invitrogen).
9. Nuclease-free water.
10. PCR plate (96) 0.5 mL.
11. Microtubes (1.5 mL).
12. Thermocycler.
13. HotStarTaq Plus Master Mix kit (Qiagen).
14. Viral-specific PCR primers (10 μM).
15. 0.5× TBE buffer: 45 mM Tris–HCl (pH 8.3), 45 mM boric acid, 1 mM EDTA.
16. 1% Agarose (type LE) gel: Prepare in 0.5× TBE buffer. Add GelRed or ethidium bromide for visualization.
17. DNA ladder.
18. 1 kb Plus DNA Ladder (Invitrogen).
19. UV transilluminator.
20. Ice.

3 Methods

3.1 Purification of Viral Particles (Modified from [15])

1. Grind 200–800 mg of arthropod or plant material in 15 mL tubes containing four sterile ceramic beads using tissue homogenizer. Alternatively, grind material in liquid nitrogen with about 100 mg of carborundum using a pestle and a mortar. Perform all the following steps, until library preparation, on ice.
2. Add 8 mL of 1× HBSS and homogenize.
3. Centrifuge at 4000 × *g* for 5 min at 4 °C.
4. Transfer the supernatant in 15 mL tubes.
5. Centrifuge at 8000 × *g* for 3 min at 4 °C to pellet debris.
6. Use a 5 mL syringe and a 0.45 μm syringe filter (*see* **Note 3**) to transfer the supernatant in 26.3 mL ultracentrifuge polycarbonate bottles to remove any remaining debris.
7. Fill the ultracentrifuge polycarbonate bottles with 1× HBSS solution.
8. Centrifuge at 148,000 × *g* for 2 h and 30 min at 4 °C.
9. Discard the supernatant by pipetting. Be careful not to remove the pellet.
10. Add 200 μL of 1× HBSS. Tilt the bottles so that the pellet is immersed in the buffer.

11. Keep the tubes overnight at 4 °C to resuspend the pellet.
12. Transfer 150 μL of the viral particle suspension to 0.5 mL PCR 8 tube strips.
13. Dilute the DNase I and the RNase A to the working concentration using nuclease-free water. Gently mix by pipetting or by flicking the tube a few times. Keep on ice until use.
14. Add 1 μL of DNase I and 2 μL of RNase A and 47 μL of 1× HBSS solution. Incubate at 37 °C for 1 h and 30 min. Store at −80 °C or proceed directly to stop DNA and RNA degradation.

3.2 Viral Nucleic Acid Extraction

1. Extract DNA and RNA from the total volume of viral particle-digested suspension obtained at the previous step (approximately 200 μL) with the NucleoSpin 96 Virus Core Kit according to the manufacturer's protocol (*see* **Note 1**).
2. Store at −80 °C or proceed directly.

3.3 Reverse Transcription

1. Dilute Dodeca Linkers to 10 μM with nuclease-free water. Add 1 μL of Dodeca Linkers in 10 μL of extracted viral nucleic acid.
2. Denature at 85 °C for 2 min in a thermal cycler and chill on ice for 2 min.
3. Add 2 μL of DTT, 1.25 μL of dNTP mix, 4 μL of SuperScript buffer, and 1 μL of SuperScript III and 0.75 μL of nuclease-free water. Mix gently.
4. Perform the reverse transcription by incubating in a thermocycler at 25 °C for 10 min, 42 °C for 60 min, 70 °C for 5 min, and 4 °C for 2 min. Store at −80 °C or proceed directly.

3.4 cDNA Purification

1. Add 1 μL of RNase A and mix gently.
2. Incubate at room temperature for 15 min.
3. Heat at 85 °C in a thermocycler for 2 min and keep at room temperature.
4. Purify the cDNA using the QiaQuick PCR cleanup kit according to the manufacturer's protocol. Store at −80 °C or proceed directly.

3.5 Klenow Amplification

1. Put 20 μL of cleaned cDNA in PCR plate. Add 0.5 μL of Dodeca Linker. Mix.
2. Place the plate in a thermocycler at 95 °C for 2 min and immediately at 4 °C for 2 min.
3. Add 0.5 μL of Klenow DNA polymerase, 2.5 μL of 10× Klenow reaction buffer, 1 μL of dNTP mix, and 0.5 μL of nuclease-free water.

4. Incubate in a thermocycler at 37 °C for 60 min followed by 75 °C enzyme heat inactivation for 10 min. Store at −80 °C or proceed directly.

3.6 PCR Amplification

1. Put 5 μL of the Klenow product in a PCR plate. Add 4 μL of PCR primer (diluted to 10 μM in nuclease-free water), 10 μL of HotStar Taq Plus Master Mix, and 1 μL of nuclease-free water.
2. Place the plate in a thermocycler and perform the following PCR cycling conditions: 1 cycle of 95 °C for 5 min, 5 cycles of 95 °C for 1 min, 50 °C for 1 min, 72 °C for 1.5 min and 35 cycles of 95 °C for 30 s, 50 °C for 30 s, 72 °C for 1.5 min +2 s at each cycle. Perform an additional final extension for 10 min at 72 °C.

3.7 Verification of the Composition and Concentration of the PCR Products

1. Verify the yield of the PCR products by the migration of 6 μL of PCR products loaded with 1 μL of DNA ladder to a 1% agarose gel. Migrate at 100 V for 45 min. Visualize PCR products under UV after staining with ethidium bromide or GelRed (Fig. 2).
2. Pool 2–6 μL of each PCR product in a 1.5 mL tube according to the smear intensity.
3. Clean the pooled PCR products using the NucleoSpin Gel and PCR cleanup according to the manufacturer's protocol (*see* **Note 4**).
4. Measure the DNA concentration of cleaned PCR products using the Qubit dsDNA HS Assay kit according to the manufacturer's protocol.

3.8 Library Construction and Sequencing

1. Send the cleaned PCR products to an external HTS provider which carries out the library construction and the sequencing. The PCR products can be sequenced on 454 pyrosequencing platform as well as on a number of different Illumina platforms. Most existing HTS platforms have their own protocols to convert PCR products into a sequencing library suitable for subsequent cluster generation and sequencing. These protocols differ in the quantity and quality of the starting material,

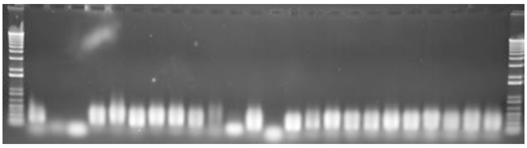


Fig. 2 Agarose gel analysis of PCR amplicons obtained from arthropod and plant samples using the VANA-based metagenomics approach. Lanes 1 and 25: 1 kb plus ladder size marker (Invitrogen); lanes 2–24: VANA-based metagenomics amplicon smears

but they are usually comprised of end repair, modification, and ligation of adapters, which enable DNA amplification by adapter-specific primers and size selection of DNA molecules with a length optimal for the sequencing strategy. The example reported here used the Illumina MiSeq platform 300 pb as paired-end reads. The sequencing generated about 18 million paired-end reads.

3.9 Data Handling and Bioinformatics (Fig. 3)

3.9.1 Data Demultiplexing and Cleanup

1. Verify the number of reads and evaluate their average quality using FastQC software.
2. Identify each PCR primer in each raw read using the "agrep" command [22] in order to assign them to the particular samples from which they originated (demultiplexing) (*see* **Note 5**).
3. Remove the Illumina adaptors and the PCR primers, and perform a quality filtering of the reads (remove sequence regions with quality score <q30 and reads smaller than 15 nt) using Cutadapt version 1.9 [23].

3.9.2 De Novo Assembly

1. Assemble the reads into longer continuous sequences (contigs) using the SPAdes assembler 3.6.2 [24] (or similar software). K-mer length can be modified to improve the assembly. Consult the assembler manual for suggested settings of data.

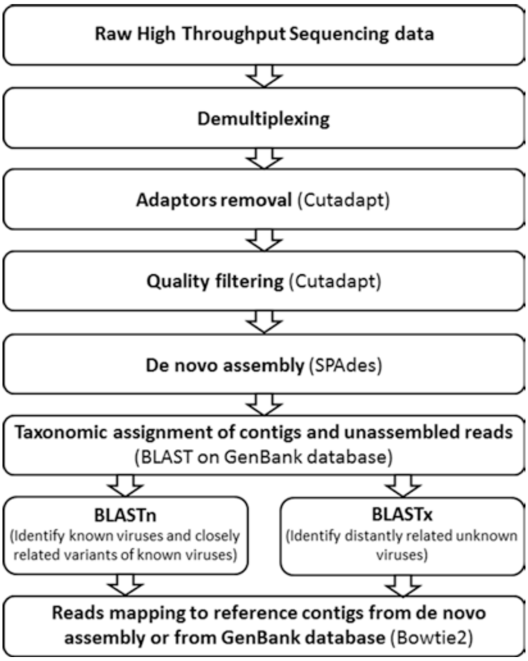

Raw High Throughput Sequencing data
Demultiplexing
Adaptors removal (Cutadapt)
Quality filtering (Cutadapt)
De novo assembly (SPAdes)
Taxonomic assignment of contigs and unassembled reads (BLAST on GenBank database)
BLASTn (Identify known viruses and closely related variants of known viruses)
BLASTx (Identify distantly related unknown viruses)
Reads mapping to reference contigs from de novo assembly or from GenBank database (Bowtie2)

Fig. 3 General workflow of the bioinformatics analysis of high-throughput sequencing data

3.9.3 Taxonomic Assignment

1. Perform BLASTn and BLASTx searches [25] against local homologs of NCBI nucleotide and protein databases using NCBI's BLAST program for taxonomic classification (*see* **Note 6**). This can be performed for both reads and contigs.
2. For potential viruses identified during the evaluation of BLAST results, retrieve candidate reference genomes from GenBank in FASTA format.

3.9.4 Read Mapping

1. Cleaned unassembled reads can be mapped on viral contigs produced by de novo assembly, or on viral reference sequences that can be found in GenBank. Map reads and/or contigs using Bowtie 2.1.0 (options end-to-end very sensitive) [26, 27] (or similar software) against the reference genomes or contigs to allow analysis and visualization of similarities and coverage distribution. The results from alignment can be checked using the Integrative Genomics Viewer (IGV) or similar viewers (Fig. 4) [28].

3.10 Confirmation and Retrieval of Near-Full Genome Sequences

1. Based on the results from the alignments, use the Primer3 program [29] (or similar software) to design specific PCR primers to confirm the presence of virus in the original material and to close gaps.
2. Extract DNA and/or RNA from the original material using QIAamp DNA Mini Kit or RNeasy MiniKit following the

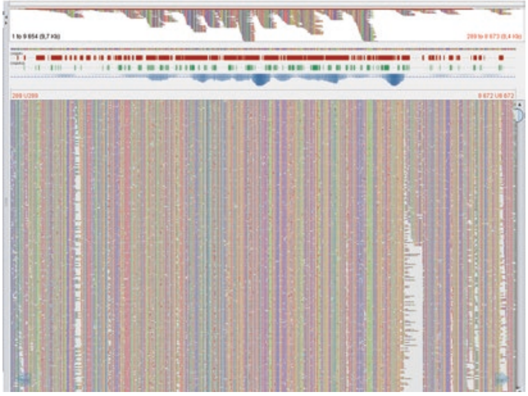


Fig. 4 An example of viral genome reconstitution using the protocol presented here. This viral contig was created using SPAdes 3.6.2 with standard parameters. It represented a nearly complete genome of 9654 nucleotides in length. BLASTn analysis showed that this viral contig shared 98% of nucleotide identity with the Aphid lethal paralysis virus (*Dicistroviridae*, accession number KX884276). 119,474 reads from an *Acyrthosiphon pisum* virome were mapped against this viral contig, using Bowtie 2.1.0 options end-to-end very sensitive, which corresponded to an average coverage of 400×

manufacturer's protocols. In case of RNA extraction, generate cDNA using a SuperScript III reverse transcriptase kit with random hexamers according to the manufacturer's instructions.

3. Amplify the viral nucleic acid using a PCR kit such as the HotStarTaq Plus Master Mix kit according to the manufacturer's protocol.
4. Visualize a fraction of the amplified products on an agarose gel and perform Sanger sequencing from the remaining volume of the PCR products. Otherwise, extract the DNA bands of interest by using a UV transilluminator and a scalpel, purify the PCR products using a column-based gel extraction kit, and perform Sanger sequencing. PCR with overlapping primers can be used to sequence PCR fragments that are too long to be sequenced in a single round of Sanger sequencing.

4 Notes

1. Nucleospin 96 virus Core kit is well suited for the simultaneous extraction of encapsidated DNA and RNA nucleic acids.
2. Only 10 μL long pipette filter tips allow pipetting small quantities of solution in MN square-well blocks.
3. The use of a 0.45 μm filter may prevent the recovery of giant viruses [30].
4. Libraries can be home made using illumina kits.
5. The NucleoSpin Gel and PCR cleanup allows size selection of the PCR products. Consult the kit manual for further details.
6. Using the multiplexing method detailed in this chapter, about 50% of Illumina MiSeq 300 bp paired-end raw reads are correctly assigned to their samples of origin.
7. The BLAST software suite released by NCBI can perform a number of different types of homology searches using nucleotide sequences as query against databases of nucleotide and amino acid sequences (e.g., BLASTn, BLASTx). Updated and preformatted nucleotide and protein databases can be obtained from NCBI as compressed archives (ftp://ftp.ncbi.nih.gov/blast/db/). When performing a BLAST search, it is possible to configure parameters of the BLAST search to specify BLAST algorithm, database(s) to be searched, output format, cutoff levels, etc. For more information, *see* the manual at http://www.ncbi.nlm.nih.gov/books/NBK1763/.

Acknowledgments

S.F. was supported by a scholarship from the National Institute of Agronomical Research (INRA) and from the University of Montpellier (UM). P.R. has received an EU grant FP7-PEOPLE-2013-IOF (N°PIOF-GA-2013-622571).

References

1. Suttle C (2007) Marine viruses (mdash) major players in the global ecosystem. Nat Rev Microbiol 5:801–812
2. Brum JR, Sullivan MB (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat Rev Microbiol 13:147–159
3. Koonin EV, Dolja VV, Krupovic M (2015) Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology 479–480:2–25
4. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol 10:607–617
5. Chiu CY (2013) Viral pathogen discovery. Curr Opin Microbiol 16:468–478
6. Rosario K, Breitbart M (2011) Exploring the viral world through metagenomics. Curr Opin Virol 1:289–297
7. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77
8. Roossinck MJ, Martin DP, Roumagnac P (2015) Plant virus metagenomics: advances in virus discovery. Phytopathology 6:716–727
9. Tokarz R, Williams SH, Sameroff S, Sanchez Leon M, Jain K, Lipkin WI (2014) Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. J Virol 88:11480–11492
10. Alquezar-Planas DE et al (2013) Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing. Sci Rep 3:2468
11. Angly FE et al (2006) The marine viromes of four oceanic regions. PLoS Biol 4:e368
12. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D (2014) High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl Environ Microbiol 80:6888–6897
13. Whon TW, Kim M-S, Roh SW, Shin N-R, Lee H-W, Bae J-W (2012) Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J Virol 86:8221–8231
14. Poojari S, Alabi OJ, Fofanov VY, Naidu R (2013) A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family geminiviridae implicated in grapevine redleaf disease by next-generation sequencing. PLoS One 8:e64194
15. Candresse T et al (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9:e102945
16. Bernardo P et al (2016) Molecular characterization and prevalence of two capulaviruses: alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa. Virology 493:142–153
17. Palanga E et al (2016) Metagenomic-based screening and molecular characterization of cowpea- infecting viruses in Burkina Faso. PLoS One 11:1–21
18. Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434:162–174
19. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci U S A 98:11609
20. Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E (2009) Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol 83:4642–4651
21. Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarría F, Shen G, Roe B (2010) Ecogenomics: using massively parallel

pyrosequencing to understand virus ecology. Mol Ecol 19:81–88

22. Wu S, Manber U (1992) Agrep – a fast approximate pattern-matching tool. In Proceedings of USENIX Technical Conference. USENIX Association, Berkeley, CA, USA. 153–162
23. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB 17:10–12
24. Bankevich A et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477
25. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 5:403–410
26. Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11, Unit 11 7
27. Toland AE, Çatalyürek ÜV, Hatem A, Bozda D (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184
28. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
29. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40:e115
30. Halary S, Temmam S, Raoult D, Desnues C (2016) Viral metagenomics: are we missing the giants? Curr Opin Microbiol 31:34–43

Chapter 8

Different Approaches to Discover Mycovirus Associated to Marine Organisms

Luca Nerva, Giovanna C. Varese, and Massimo Turina

Abstract

Here we describe the protocols to characterize the virome associated to fungi isolated from marine organisms assessed on the seagrass *Posidonia oceanica* and on the marine animal *Holothuria poli*. We provide detailed protocols for fungal isolation, fungal growth, and total RNA extraction. Ribosomal RNA depletion, cDNA library synthesis and normalization, and sequencing runs on different platforms are part of the protocols that are generally outsourced and therefore are not described in this chapter. We describe, instead, how raw reads are assembled into contigs and how to search for putative viral sequences. Furthermore, we detail qualitative checks to infer the existence of the virus as a replicative biological entity.

Key words Mycovirus, RNA-Seq, *Posidonia oceanica*, *Holothuria poli*, Viral genome assembly

1 Introduction

Interest in mycoviruses has been mostly originated from their ability to change the virulence phenotype of their pathogenic fungal host [1]. Furthermore, in some specific model systems, mycoviruses have been shown to have an important role for adaptation to specific ecological niches [2] and the study of their molecular and biological properties has allowed to infer their possible evolutionary trajectories [3]. The characterization of the virome associated to marine microorganisms (in specific fungi) is important for their ecological and socioeconomical relevance of marine ecosystems and for the potential biotechnological application of metabolites and enzymes they produce [4]. Mycoviruses could indeed have an effect on their fungal host by stimulating or inhibiting production of specific enzymes and metabolites involved in adaptation to high-salinity environments.

The number of reported mycoviruses is increasing exponentially because of the current ability to detect new sequences by next-generation sequencing (NGS) approaches; a high number of mycovirus genomes are built in silico using data from fungal

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_8, © Springer Science+Business Media, LLC 2018

transcriptome projects [5, 6]. NGS techniques for virome characterization can be performed on different template RNA molecules: total RNA depleted from ribosomal RNA (rRNA) or small RNA (sRNA) fragments [7]. These two approaches have specific features resulting in slightly different qualitative and quantitative outputs. In total RNA extraction it is possible to identify all kind of viruses independently from the genome type (because all the proteins encoded by viral genomes are translated from RNA) but the amount of data to be analyzed is difficult to be managed. On the other hand, NGS of sRNA fragments gives a smaller amount of data that can be used to build most viral genomes [5]. However, sRNA approach does not always provide enough information to build full-length viral genomes possibly because not all the viruses are equally well targeted along their genomes by the short interfering RNA (siRNA) pathway [8]. Moreover, for viruses located inside organelles (as is the case of fungal mitoviruses inside mitochondria) the viral sRNA accumulation could be insufficient. Figure 1 provides a general workflow for characterizing the virome of fungi associated with marine organisms. Library construction, standardization, and sequencing that could include ribosome depletion and small RNA purification are generally outsourced; therefore, those parts of the protocol are not described in this chapter.

Almost all the programs used to manage NGS data are executed from command line and require a basic familiarity with Unix environment. The main program used to assemble sequences from total RNA datasets is Trinity [9, 10] and it requires a multicore server with high memory (at least 1 GB of RAM per million of pair-end reads). The second step to identify new viral sequences into the assembled transcriptomes is to blast it to a custom database with viral sequence. This is a crucial step because the more the custom database is complete and inclusive the higher will be the possibilities to identify any conserved viral sequence in the assembled data. The presence of in silico-assembled genomes needs to be verified in the original sample with other methodologies (two examples are provided here), and possibly evidence of the existence of the virus as a biological entity should be provided beyond the existence of its mere genomic sequence.

2 Materials

Prepare all solutions using deionized water and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Diligently follow all waste disposal regulations when disposing waste materials.

2.1 Sample Collection and Fungal Isolation

1. Laminar flow sterile hood.
2. Sterile Petri dishes.
3. Sterile scalpels and tweezers.

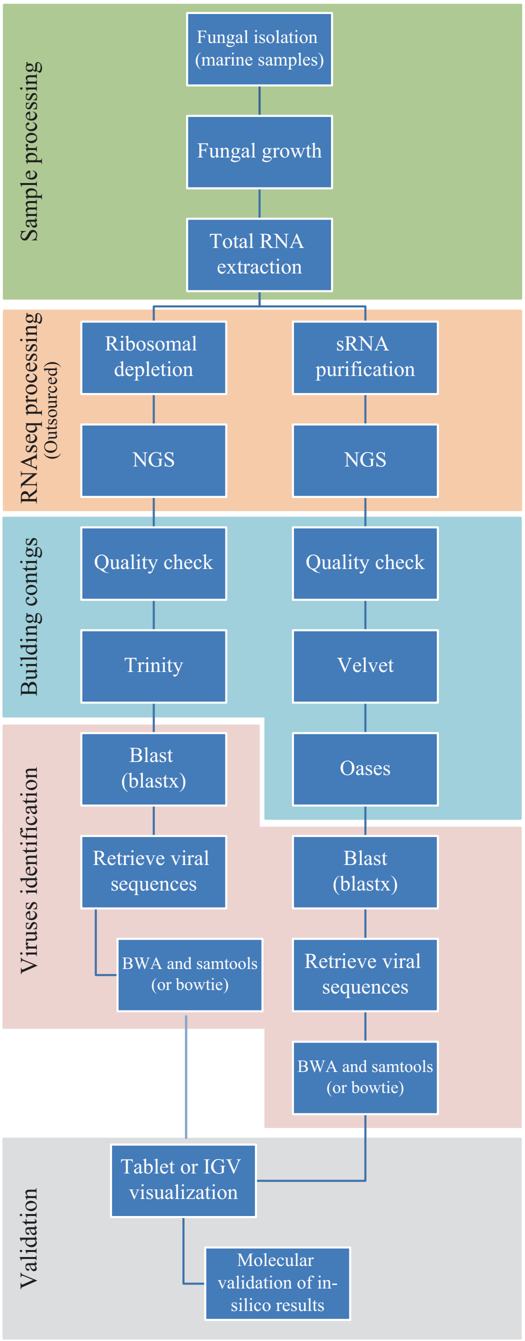


Fig. 1 Suggested workflow chart for the characterization of the virome of fungi associated to marine organisms

4. Sterile water (*see* **Note 1**).
5. Sterile sealed bags and Falcon tubes.
6. Glucose peptone yeast extract seawater agar (GPYSA): Dissolve 1 g glucose, 0.5 g peptone, 0.1 g yeast extract, 18 g agar in 500 mL seawater, sterilize in autoclave, and before solidification (temperature around 50 °C) bring to 1 L with more seawater sterilized by filtration (*see* **Note 2**).
7. Corn meal seawater agar (CMSA): Dissolve 17 g CMA Oxoid in seawater 500 mL; after sterilization in autoclave and before solidification (around 50 °C) bring to 1 L with more seawater sterilized by filtration (*see* **Note 2**).
8. Agar Posidonia (AP): Blend 20 g (fresh weight) of *P. oceanica* tissues in 100 mL of filtered seawater, heat to 60 °C for 30′, and filter with filter paper to eliminate any debris. Bring to 500 mL using seawater and add 18 g of agar; after autoclave sterilization at 121 °C for 20′ and before solidification (around 50 °C), bring to 1 L with more seawater sterilized by filtration (*see* **Note 2**).
9. Antibiotics: 500 mg/L Gentamicin and 50 mg/L chloramphenicol (*see* **Note 3**).
10. Incubators at different temperatures (15 and 24 °C).
11. NucleoSpin kit (Macherey-Nagel GmbH).

2.2 Fungal Growth and Maintenance

1. MEA medium, pH 5.3: 3% Malt extract, 2% D-glucose, 0.1% peptone, 0.0005% $CuSO_4$, 0.001% $ZnSO_4$, 2% agar (all percentages should be considered as w/v).
2. MEA 3% medium, pH 5.3: 3% Malt extract, 2% D-glucose, 0.1% peptone, 0.0005% $CuSO_4$, 0.001% $ZnSO_4$, 30 g sea salts, 2% agar (all percentages should be considered as w/v).
3. CYA medium, pH 6.3: 0.3% $NaNO_3$, 0.5% yeast extract, 3% sucrose, 0.13% K_2HPO_4, 0.05% $MgSO_4$, 0.05% KCl, 0.001% $FeSO_4$, 0.0005% $CuSO_4$, 0.001% $ZnSO_4$, 1.5% agar (all percentages should be considered as w/v).
4. ME broth, pH 5.3: 3% Malt extract, 2% D-glucose, 0.1% peptone, 0.0005% $CuSO_4$, 0.001% $ZnSO_4$ (all percentages should be considered as w/v).
5. CY broth, pH 6.3: 0.3% $NaNO_3$, 0.5% yeast extract, 3% sucrose, 0.13% K_2HPO_4, 0.05% $MgSO_4$, 0.05% KCl, 0.001% $FeSO_4$, 0.0005% $CuSO_4$, 0.001% $ZnSO_4$ (all percentages should be considered as w/v).
6. Virtis blade-type homogenizer (Labequip).
7. Thermostatic incubators for solid fungal cultures and orbital shakers for liquid cultures.
8. Suitable glassware (flasks) for fungal cultures in liquid substrates.

2.3 RNA Extraction

1. Buchner funnel.
2. Sterile Miracloth (Merck Millipore) and sterile paper towels.
3. Freeze dryer apparatus (*see* **Note 4**).
4. Two milliliter sterile tubes with screw caps.
5. Spectrum Plant Total RNA Kit (Sigma-Aldrich).
6. Zirconia beads 0.5 mm.
7. FastPrep24 (MP Biomedicals) (*see* **Note 5**).
8. Direct-zol RNA MiniPrep (Zymoresearch).
9. Tabletop microfuge.
10. Sterile tips (1000, 100, and 10 μL).
11. NanoDrop spectrophotometer (Thermo Fisher Scientific).
12. Trizol reagent.
13. Absolute ethanol.
14. PEG8000.
15. NaCl.
16. Agarose.
17. Acrylamide.
18. Urea.

2.4 Transcriptome Assembly and Analysis

1. High memory and multicore server: The assembly step uses Trinity, which requires ca. 1 GB of RAM for one million of reads during the first stage and then a high parallel computing capacity during the final assembly stage.
2. Custom viral database: An important aspect of a viral database is that it should not contain viral sequences that show high homology with the host genome. For this reason, we suggest to exclude viruses that show high number of endogenized sequences into the host genome or, the opposite, host genes into their genomes. An example of these sequences are those present in the eukaryotic nucleocytoplasmic large DNA viruses (NCLDV) with extensive regions of their genomes harboring host sequences (i.e., Mimiviridae, Ascoviridae, Iridoviridae, Megaviridae, Pandoraviridae, Poxiviridae) [11].
3. Software (*see* **Note 6**):
   (a) Blast + suite 2.6.0;
   (b) BWA 0.7.15-r1140;
   (c) Samtools 1.3;
   (d) Trinity 2.3.2;
   (e) Velvet 1.2.10;
   (f) Oases 0.2.08.

2.5 Validation of Virus In Silico Assembly Results

1. Thermal cyclers.
2. Gel electrophoresis apparatus for agarose gels.
3. Real-time PCR apparatus.
4. Platinum SuperFI PCR Master Mix (Thermo Fisher Scientific).
5. Electron microscope.
6. PELCOTEM grids (Formvar carbon- and silicon-coated copper grids 200 mesh) (Ted Pella Inc).
7. Uranyl acetate 2% aqueous solution.
8. Tris-buffered phenol, pH 8.
9. Two milliliter sterile tubes.
10. Chloroform.
11. Isopropanol.
12. DNase I.
13. RNase A.
14. TAE buffer (40 mM Tris, 20 mM acetic acid, 2.5 mM EDTA, pH 8).
15. Agarose.
16. TE buffer (10 mM Tris pH 8, 1 mM EDTA, pH 8).
17. 0.5 mm Glass beads (Biospec).
18. Vacuum concentrator.
19. FastPrep24 (MP Biomedicals) (*see* **Note 5**).

3 Methods

3.1 Fungal Isolation

1. Professional divers collect the marine samples (seagrasses, algae, animals, etc.) (*see* **Note** 7).
2. Each sample is opened under a laminar flow sterile hood using sterile Petri dishes, eventually subdivided into different parts (*see* **Note 8**) using sterile cutters or scissors and put in sterile 50 mL Falcon tubes containing the proper amount of sterile seawater.
3. To eliminate all the debris and possible propagules of microorganisms casually present on the surface of the sample and allow the isolation of fungi intimately associated with the host, samples are subjected to serial washes: samples are sonicated at low intensity for 30 s, dried onto sterile filter papers, and transferred in new sterile tubes containing sterile seawater. The procedure is repeated five times (*see* **Note 9**).
4. Five gram (fresh weight) of each sample is homogenized in 100 mL sterile seawater (*see* **Note 10**).

5. Homogenates are serially diluted 1:10 using sterilized seawater (*see* **Note 11**). The final dilutions of each sample are plated (1 mL per plate) in 15 cm diameter Petri dishes containing 40 mL of the media (*see* **Note 2**).

6. Plates are incubated at 15 and 24 °C for 30 days to allow the development and isolation of slow-growing colonies (*see* **Note 12**).

7. At regular time intervals, count the colony-forming units (CFUs) and isolate the different fungal morphotypes in pure culture cutting a small part of the colony and transplanting it on agar plates of the medium (*see* **Note 2**). Express fungal load as CFU/g dry weight of sample.

8. Fungi should be identified with a polyphasic approach, which couples morphophysiological features with molecular studies. After determination of genera according to macroscopic and microscopic features, transfer the fungal strains to the media recommended by the authors of selected genus monographs for species identification by morphophysiological tools (*see* **Note 13**).

9. Molecular identification is performed by amplification and sequencing of the appropriate DNA region (i.e. ITS, α-actin, and β-tubulin) according to the specific fungal genera. Extract genomic DNA from about 100 mg of mycelium scraped from MEA Petri dishes using the NucleoSpin kit according to the manufacturer's instructions. Measure the quality and quantity of DNA samples with the NanoDrop-1000 Spectrophotometer or similar instruments. DNA extracts can be stored at −20 °C. For the isolates morphologically identified as *Aspergillus*, *Eurotium*, *Penicillium*, and *Talaromyces* species, perform the amplification of the β-tubulin gene using primers Bt2a/Bt2b (*see* **Note 14**); for fungi belonging to the genus *Cladosporium* perform the amplification of the α-actin gene using the primer pair ACT-512F/ACT-783R (*see* **Note 15**). For the other fungal genera perform the amplification of the ITS1-5,8S-ITS2 region using the primer pair ITS1/ITS4 (*see* **Note 16**). Reaction mixtures consist of 30 ng genomic DNA, 1 μM each primer, 1 U Taq DNA polymerase, 1× buffer, and 200 μM each dNTP. DNA amplifications are performed using a thermal cycler. Send PCR products to specialized laboratories for purification and sequencing.

10. Compare the resulting sequences with reference sequences in online databases, i.e., the NT database provided by the NCBI National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/guide/dna-rna/).

11. The fungal sequences and corresponding species identifications should be deposited in GenBank following the institution instructions. The fungal strains should be deposited in a public culture collection (*see* **Note 17**).

3.2 Fungal Maintenance and Growth

1. Maintenance: Identified fungi are deposited and preserved with different techniques (cryopreserved, lyophilized, and/or as actively growing axenic cultures) in public culture collections and made available to academic and industrial users usually as fungal colonies actively growing on 90 mm Petri dishes containing the suitable medium.
2. Growth: From agar plates, inoculation of liquid cultures is performed cutting a sector of the fungal colony (a quarter of a 90 mm diameter colony) in small squares and then homogenized in a sterile 50 mL Falcon tube using a Virtis homogenizer with 30 mL of liquid media for 20 s at highest setting (*see* **Note 18**). Five milliliter of fungal homogenate is used to inoculate 100 mL cultures in 250 mL conical flasks using media appropriate to each fungal isolate. Fungal growth for RNA extraction is generally carried out at 24 °C for 48–72 h at 120 rpm on a rotary shaker; most fungi are grown in liquid cultures at 24 °C in malt extract broth with 3% of sea salts and produce a good amount of mycelia in 4 days.

3.3 Total RNA Extraction

There are many possibilities for total RNA extraction and here we suggest a kit-based method that allows obtaining good amount of total RNA of good quality from carbohydrate-rich matrices as those of filamentous fungi. With most methods, special care should be taken in reducing the amount of lyophilized mycelia to 20–30 mg for each sample.

Fungi are harvested at the beginning of their stationary growth phase, filtered through Miracloth in the Buchner funnel, supplied with vacuum filtering when available and necessary. The mycelial pad is dried with sterile paper towels, frozen at −80 °C, and freeze dried.

The protocol is based on the Spectrum Plant Total RNA Kit:

1. Put 30–40 mg of lyophilized mycelia in a 2 mL screw-cap tube with 0.5 mL of glass/zirconia beads (*see* **Note 19**).
2. Homogenize the mycelia using FastPrep24 for 30 s at the maximum speed until a fine powder is produced.
3. Add 1 mL of lysis buffer and homogenize again (30 s at the maximum speed).
4. Centrifuge for 5 min at 13,000 × *g*.
5. Transfer 450 μL of supernatant to the filtration column (blue ring).
6. Centrifuge for 30 s at 13,000 × *g*.
7. Add the same volume of "binding solution" to the filtered liquid and apply to the binding column (red ring).
8. Centrifuge for 30 s at 13,000 × *g*. Discard the flow-through (*see* **Note 20**).
9. Add 400 μL of wash solution 1.

10. Centrifuge for 30 s at 13,000 × *g*. Discard the flow-through.
11. Add 400 μL of wash solution 2.
12. Centrifuge for 30 s at 13,000 × *g*. Discard the flow-through. Repeat the last washing step.
13. Centrifuge for 60 s at 13,000 × *g* to remove completely traces of ethanol.
14. Elute RNA in a new tube with 40 μL of elution buffer.
15. Quantify RNA with spectrophotometer.
16. Total RNA quality control before NGS (*see* **Note 21**).

3.4 sRNA Extraction

1. Put 30–40 mg of lyophilized mycelia in a 2 mL screw-cap tube with 0.5 mL of glass/zirconia beads.
2. Homogenize the mycelia using a mill (i.e., FastPrep24).
3. Add 1 mL of Trizol reagent and homogenize again.
4. Centrifuge for 5 min at 13,000 × *g*.
5. Collect 450 μL of supernatant and add 450 μL of ethanol.
6. Follow the manufacturer's instructions of Direct-zol RNA MiniPrep until elution in 40 μL of water.
7. Quantify RNA with spectrophotometer.
8. Dissolve RNA in RNase-free water to a final concentration of 1 μg/μL. For a good yield of sRNA it is necessary to start from at least 200 μg of total RNA.
9. Precipitate high-molecular-weight RNA (rRNA and mRNA) by adding 5% of PEG8000 and NaCl to a final concentration of 0.5 M.
10. Mix well and put on ice for 30–40 min.
11. Centrifuge at 13,000 × *g* for 10 min.
12. Collect supernatant (high-molecular-weight RNA will be in the pellet) and add three volumes of ethanol. Place at −20 °C overnight.
13. Centrifuge at 13,000 × *g* for 30 min to pellet the sRNAs.
14. Discard the supernatant and wash the pellet with 1 mL of 75% ethanol.
15. Dry the pellet and dissolve in 10–20 μL of DPEC water.
16. Abundance and integrity of sRNA can be optionally checked by running an aliquot on 12% denaturing polyacrylamide gel (*see* **Note 22**).

3.5 RNA Sequencing

Once RNA is obtained (total or small fragments), the following passages are outsourced. For total RNA sequencing, we suggest to require ribosomal RNA depletion using the Ribo-Zero Gold rRNA Removal Kit for Human/Mouse/Rat (Illumina) able to remove both cytoplasmic and mitochondrial rRNAs also in fungi.

Commercial specific kits for cDNA libraries (using rRNA-depleted total RNA or sRNA as template) are also used. The most common sequencing platforms we used are Illumina HiSeq 2000 or Illumina HiSeq 4000 giving outputs of 60–150 millions of 100–150 bp pair-end reads for each sequencing lane. MiSeq platforms were used for sRNA sequencing.

3.6 Transcriptome Assembly from Total RNA Sequencing

1. To assemble reads obtained from total RNA-seq outputs we suggest to use Trinity [9], designed for Unix-type operating system, able to de novo assemble a wide range of samples. The protocol requires users to supply short read data in either fastq or fasta formats. The reads can be either pair end (Trinity can identify reads corresponding to opposite end of a single sequenced molecule) or single end (*see* **Note 23**).
2. Although sequence quality control steps are not required, performing the following steps can improve the results:
   (a) All barcodes must be removed before running Trinity using Trimmomatic [12] (included in Trinity).
   (b) Removing reads that probably include sequencing errors (hence reads with low-quality score) may reduce RAM usage and program runtime (Trimmomatic can easily do it).
   (c) If more than 200 million paired-end reads are to be assembled, the user may consider performing an in silico normalization of the sequencing reads (Trinity includes an in silico read normalization utility: no more than 50 identical reads present in the sequencing output will be considered during assembly).
3. A simple command line for single-end data (100M reads) should be:

   Trinity.pl --trimmomatic --seqType fa/fq --single single.fa/fq --max_memory 100G --CPU 30 (*see* **Note 24**).

   If you are working with pair-end sequences, the command should be:

   Trinity.pl --trimmomatic --seqType fa/fq --left left.fa/fq --right right.fa/fq --max_memory 100G --CPU 30

   The only difference is that you need to specify two datasets (a *left* and a *right* file) instead of a single read file.
4. If RNA-seq data contain hundreds of millions to billions of reads, in silico normalization ("on" by default in the latest version of Trinity) is necessary to lower the memory and computing requirements and reduce runtimes.
5. When assembly is finished, a *Trinity.fasta* file will be present inside the *trinity_out_dir* folder. Such file contains all the assembled contigs, hence the host transcripts and the viral transcripts/genomes (if present).

3.7 Assembly of sRNA Reads from sRNA Sequencing Outputs

1. To assemble contigs from sRNA sequencing outputs we suggest to use a combination of Velvet [13] and Oases [14]. As in the previous approach, we also suggest to clean reads by using Trimmomatic.
2. The first command is *velveth* to build the dataset which will be used by the next command (*velvetg*):

   velveth kmer 13,25,2 -short -fasta sRNA.fasta (*see* **Note 25**).
3. The second command, hence *velvetg*, is going to build the contigs:

   for((n=9; n<=23; n=n+2)); do velvetg kmer_"$n" –read_trkg yes; done (*see* **Note 26**).
4. The third and last command is Oases, which can maximize the assembly of contigs from the ones already assembled with Velvet:

   for((n=13; n<=23; n=n+2)); do oases kmer13-23_"$n"; done (*see* **Note 27**).

Output fasta files will be inside each folder created by the two previously used programs.

3.8 Viral Genome Identification

1. A general database can be obtained in NCBI, with all the viral RefSeq sequences, but a custom database will enhance the probability of virus detection.
   - (a) To build the protein database:
     makeblastdb -dbtype prot -in database.fasta
   - (b) Use the database to run *blastp*:
     blastx -query trinity_output.fasta -db database.fasta -evalue 10e-5 > blast_result.fasta (*see* **Note 28**).
   - (c) Identify and retrieve which contigs showed similarity to viral sequences. To do so:
     grep -A 100 -B 100 'significant' blast_result.fasta > blast_significant.fasta (*see* **Note 29**).
   - (d) Now it is possible (due to smaller size) to open the *blast_significant.fasta* file and check one by one the alignments (*see* **Note 30**).
   - (e) A final step is to retrieve contig sequences from the Trinity.fasta file:
     grep -A 100 "contig_name" Trinity.fasta > contig_name.fasta
   - (f) Then it is necessary to blast again each retrieved sequence, using the online suite, against the complete nonredundant protein sequence database to confirm if contigs belong to virus or host genome.
2. When viral sequences are identified, it is necessary to confirm the number of reads mapping against the genome (a quantita-

tive indirect estimation of virus concentration in the sample). Two software are used in succession: BWA and samtools (*see* **Note 31**):

(a) *bwa index -a bwtsw contig.fasta*

(b) *bwa aln -t 4 contig.fasta raw_reads.fastq > raw_on_contig.sai* Option *-t* is not necessary; it parallelizes the process using the set number of cores (the higher is this number the faster will be the process).

(c) *bwa samse* (or *sampe* if pair-end) *bwtsw contig.fasta raw_on_conting.sai raw_reads.fastq> raw_on_contig.sam*

(d) *samtools view -bS raw_on_contig.sam -o raw_on_contig.bam*

(e) *samtools sort -@4 -m 2G raw_on_contig.bam –o raw_on_contig.sort.bam*
Option -@ parallelizes work on indicated number of cores and option *-m* gives the selected number of gigabytes to each core. These two options are not mandatory but will speed the process.

(f) *samtools index raw_on_contig.sort.bam*

(g) At this point it is possible to use the *.sort.bam* file in a graphical viewer for next-generation sequence assemblies and alignments (*see* **Note 32**).

3.9 Validation of Virus In Silico Assembly Results

3.9.1 PCR-Based Analysis (*See* Note 33)

1. Put 60 mg of lyophilized mycelia in a 2 mL tube with 0.5 mL of 0.5 mm glass beads.
2. Grind the mycelia using a Fast-Prep24 homogenizer.
3. Add 500 μL TE buffer and 500 μL phenol.
4. Homogenize again in Fast-Prep24 homogenizer (20 s maximum setting).
5. Centrifuge for 10 min at 13,000 × *g*.
6. Collect supernatant, transfer to a new Eppendorf tube, and add an equal volume of chloroform.
7. Repeat **steps 5** and **6**.
8. Centrifuge at 13,000 × *g* for 10 min.
9. Collect supernatant and add an equal volume of ice-cold isopropanol.
10. Gently invert tubes for 30–60 s and leave on ice for 20 min.
11. Centrifuge at 13,000 × *g* for 10 min, and discard the supernatant.
12. Wash with 1 mL of 70% ethanol, centrifuge at 13,000 × *g* for 10 s, and discard the supernatant.
13. Dry the pellet in a vacuum concentrator.
14. Resuspend supernatant in 60 μL of TE buffer.

15. Divide the total nucleic acid extraction into two tubes.
16. In the first tube perform a DNA digestion with DNaseI (following the manufacturer's protocols).
17. In the second tube perform an RNA digestion with RNase A (following the manufacturer's protocols).
18. Perform a RT reaction on the DNase-treated sample (use any cDNA kit following the manufacturer's instructions).
19. Perform a PCR reaction on both samples (cDNA from DNase-treated RNA and DNA from RNase-treated total nucleic acid) with specific primer designed on the in silico-assembled putative viral genome segments (*see* **Note 30**) using Platinum SuperFI PCR kit as suggested by the manufacturer.
20. Run the PCR products on a 1% TAE agarose gel.

If the PCR products will be present only in the DNase-RT-treated sample it means that the sequence is present only as a RNA molecule. We suggest to clone the obtained sequence in a plasmid with the T7 promoter to further use it as ribo-probe in northern blot hybridization (*see* **Note 34**) and to compare to the in silico-assembled sequence.

On the contrary, if both samples will display a specific PCR product it means that the detected sequence is present also in DNA form (probably as a genome-integrated sequence).

3.9.2 Direct Particle Observation with Electron Microscopy

1. Put 50–100 mg of lyophilized mycelia in a 2 mL tube with 0.5 mL of 0.5 mm glass beads.
2. Add 1 mL of TE buffer and grind the mycelia using the FastPrep24 bead beaters.
3. Centrifuge for 10 min at 10,000 × *g*.
4. Add 10 μL of supernatant on a Formvar- and carbon-coated copper electron microscopy grids, and let air-dry. Add 10 μL of the uranyl acetate stain (2% aqueous solution) and let it air-dry.
5. The sample is ready to be observed by the transmission electron microscope (TEM) (*see* **Note 35**).

4 Notes

1. In case of marine samples, you can use sterile seawater to avoid any osmotic stress to the samplings. To sterilize seawater, use 0.2 μm diameter filters. In the media preparation, the use of filtered seawater may confer to the fungus specific micronutrients, quorum-sensing molecules, etc. essential for the growth of some fungi.

2. Select the most suitable media with respect to the matrices to be analyzed. Here are reported two general media (GPYSA and CMSA) and one medium (AP) specifically designed to mimic as much as possible the natural environment. You can design your one medium based on the different substrates you are analyzing.
3. The addition of different antibiotics inhibits bacterial growth, enhancing the capability to isolate fungi, especially the slow-growing ones. Gentamicin and chloramphenicol are cheap and heat stable, and thus resist autoclaving. Other antibiotics that can be used are kanamycin, ampicillin, and rifampicin.
4. RNA can be extracted also from freshly harvested material, prior to lyophilization, and in this case, 300 mg of mycelia filtered through a Buchner fennel and pad dried with paper towels are used for extraction.
5. A good alternative to a homogenizer is the use of mortar and pestle assisted by liquid nitrogen break of fungal cell walls.
6. Software is constantly updated with new versions, and sometimes command lines and specific options differ among older and newer version. The command lines we specify are referred to the software version we provide.
7. Samples collected during diving are immediately enclosed in sterile and labeled plastic containers (sterile sealable plastic bags, Falcon tubes, etc.) according to sample sizes. On the ship and during transportation, samples are stored in refrigerators at about 4 °C. Samples should be processed for fungal isolation as fast as possible (within 48 h).
8. For *Posidonia oceanica*, each sample is divided into four districts: leaves, rhizomes, roots, and matte. Each sample of *Holothuria poli* has been sectioned to divide the body wall from the intestine.
9. The number of serial washes can be different according to the sample. A final surface sterilization can be performed using different reagents, i.e., 70% ethanol, sodium hypochlorite (laundry bleach diluted to 10–20%), or 30% hydrogen peroxide After the material is sterilized, it must be rinsed thoroughly with sterile water. Typically three to four separate rinses are performed.
10. The amount of fresh weight to be used can be changed depending on the sample availability. Keep in mind that an appropriate sample size should also be used to calculate the sample’s dry weight useful to estimate the fungal load.
11. The proper dilution must be determined empirically for each type of sample.

12. Select the temperature of incubation according to the specific ecological niche under scrutiny. Different incubation temperatures enhance the isolation of psychrophilic (low temperatures), mesophilic (around 24 °C), and eventually thermophilic (37 or 45 °C) fungi.

13. List of references useful to identify the main fungal genera [15–17].

14. List of reference useful for the amplification of the β-tubulin gene [18–20].

15. Reference useful for the amplification of the α-actin gene using the primer pair ACT-512F/ACT-783R [21].

16. Reference useful for the amplification of the ITS1-5,8S-ITS2 region using the primer pair [22].

17. To find the closest public culture collection, please consult the World Federation Culture Collections database (http://www.wfcc.info/).

18. For conidia-producing isolates, we need to avoid the risk of inoculating only with conidia, since some mycoviruses are not transmitted to the conidial progeny. Therefore we suggest to homogenate young colony agar plugs as inoculation material for liquid cultures.

19. A suitable alternative is the use of up to 300 mg of fresh mycelia. Prior to adding the beads it is necessary to freeze the sample in liquid nitrogen; if a homogenizer is not available, immediately homogenize with mortar and pestles.

20. Part of the solution often does not pass through the filter due to polysaccharide residues from fungal mycelia. In this case it is possible to repeat the centrifuge step by increasing the time up to 60 s. If solution is still present on top of the filter discard it and proceed with washes.

21. In general, an automatic RNA integrity number (RIN) analysis is performed before NGS analysis, and quality thresholds are suggested: such analysis is based on rRNA integrity, and often abundant virus infections (such as those caused by some mycoviruses) can considerably alter the rRNA profile. In our experience, suboptimal RIN values do not impair the possibility to characterize the virome associated to a sample.

22. Further selection of sRNA fragment can be done on polyacrylamide/urea gel. Urea polyacrylamide gel electrophoresis employs urea to denature secondary RNA structures and it is used for separation in a polyacrylamide gel matrix based on molecular weight.

23. If multiple sequencing runs are given for a single experiment, these reads may be concatenated into a single-read file for single-end sequencing or into two files (e.g., merging all “left”

and all "right" reads into single "left.fq" and "right.fq" files, respectively) in the case of paired-end sequencing. Trinity may be used with data of any read length commonly produced by next-generation sequencer.

24. In this command line:
    (a) Trimmomatic will clean reads from adaptor and/or barcoding.
    (b) seqType parameter will be *fa* in case of fasta format or *fq* in case of fastq format.
    (c) After *--single* specifies the name of input reads file (in case of different folder specifies the path).
    (d) *--max_memory* is the maximum amount of RAM memory that can be used.
    (e) *--CPU* is the number of CPU that Trinity can use.

25. This command will create different datasets starting from *k-mer* of 13 and increasing the *k-mer* value by 2 until a value of 25. In this way the program is going to build different datasets that as in a previous work we observed optimize identification of viral sequences [5].

26. With this command line *velvetg* creates one folder for each *k-mer* used; the corresponding file with the assembled sequences will be placed in each folder.

27. Also for this command, it is possible to use more than one *k-mer* value to maximize the ability of the program to assemble longer contigs.

28. The option *-evalue* can be changed to be more stringent. It is possible to make this passage faster by reducing the number of alignments using these two options: *-num_alignemnts* 10 *-num_descriptions* 10. In this way blastp will align and report only ten hits for each contig (that is sufficient to identify viruses).

29. With this command you get a file containing only contigs with significant alignments (over the e-value threshold).

30. The file contains a number of graphical alignments and further selection is generally based on personal experience, but some guidelines can be given: exclude small fragments with e-value close to threshold. If you are unsure about not considering some significant alignment, generally the further step of comparing the sequence to the *nr* general database will indicate if the sequence is really viral or more similar to sequences not of viral origin.

31. A suitable alternative is the use of the software Bowtie.

32. In order to view graphically reads aligned to identified genomes, a number of alternatives are available also for Microsoft operating systems, such as Tablet 1.16.09.06 and IGV 2.3: in both cases, the output files .sort.bam and sort.bam.bai must both be present in the same directory.
33. Although so far only one DNA virus has been described to be able to infect fungi, the same assay with different controls can be applied to identify viral DNA genomes.
34. Northern analysis using riboprobes obtained through in vitro transcription is our method of choice to validate genomes assembled in silico: Northern blots are quantitative, they can confirm predicted size of RNA, they can reveal possible subgenomic RNAs (evidence of replication), and they can confirm the replicative nature of the molecule (strand-specific probes can provide evidence of RNA present in both orientations).
35. For low-titer viruses, a more complete differential centrifugation protocol for virus purification should be applied.

References

1. Pearson MN, Beever RE, Boine B, Arthur K (2009) Mycoviruses of filamentous fungi and their relevance to plant pathology. Mol Plant Pathol 10:115–128
2. Marquez LM, Redman RS, Rodriguez RJ, Roossinck MJ (2007) A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. Science 315:513–515
3. Nerva L, Varese GC, Falk BW, Turina M (2017) Mycoviruses of an endophytic fungus can replicate in plant cells: evolutionary implications. Sci Rep 7:1908
4. Panno L, Voyron S, Anastasi A, Varese GC (2010) Marine fungi associated with the sea grass *Posidonia oceanica* L.: a potential source of novel metabolites and enzymes. J Biotechnol 150:S383–S384
5. Nerva L, Ciuffo M, Vallino M, Margaria P, Varese GC, Gnavi G, Turina M (2016) Multiple approaches for the detection and characterization of viral and plasmid symbionts from a collection of marine fungi. Virus Res 219:22–38
6. Marzano SYL, Nelson BD, Ajayi-Oyetunde O, Bradley CA, Hughes TJ, Hartman GL, Domier LL (2016) Identification of diverse mycoviruses through metatranscriptomics characterization of the viromes of five major fungal plant pathogens. J Virol 90:6846–6863
7. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7
8. Nakayashiki H, Nguyen QB (2008) RNA interference: roles in fungal biology. Curr Opin Microbiol 11:494–502
9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Chen Z (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
10. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, MacManes MD (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512
11. Yutin N, Colson P, Raoult D, Koonin EV (2013) Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol J 10:106
12. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
13. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829
14. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092

15. Domsch KH, Gams W, Anderson TH (eds) (1980) Compendium of soil fungi, vol 1. Academic Press, London
16. Kiffer E, Morelet M (eds) (1997) Les deutéromycètes: Classification et clés d'identification générique. INRA, Paris
17. Von Arx JA (ed) (1981) The genera of fungi sporulating in pure culture. J. Cramer, Vaduz
18. Geiser DM, Frisvad JC, Taylor JW (1998) Evolutionary relationships in Aspergillus section Fumigati inferred from partial β-tubulin and hydrophobin DNA sequences. Mycologia 90:831–845
19. Glass NL, Donaldson GC (1995) Development of primer sets designed for use with the PCR to amplify conserved genes from filamentous ascomycetes. Appl Environ Microbiol 61:1323–1330
20. Samson RA, Seifert KA, Kuijpers AFA, Houbraken JAMP, Frisvad JC (2004) Phylogenetic analysis of Penicillium subgenus Penicillium using partial β-tubulin sequences. Stud Mycol 49:175–200
21. Carbone I, Kohn LM (1999) A method for designing primer sets for speciation studies in filamentous ascomycetes. Mycologia 91:553–556
22. White TJ, Bruns T, Lee SJWT, Taylor JW (eds) (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: PCR protocols: a guide to methods and applications, 18. Academic Press, New York, p 315–322

Chapter 9

Use of siRNAs for Diagnosis of Viruses Associated to Woody Plants in Nurseries and Stock Collections

Nikoletta Czotter, János Molnár, Réka Pesti, Emese Demián, Dániel Baráth, Tünde Varga, and Éva Várallyay

Abstract

Woody perennial plants like grapevine and fruit trees can be infected by several viruses even as multiple infections. Since they are propagated vegetatively, the phytosanitary status of the propagation material (both the rootstock and the variety) can have a profound effect on the lifetime and health of the new plantations. The fast evolution of sequencing techniques provides a new opportunity for metagenomics-based viral diagnostics. Viral derived small RNAs produced by the host immune system during viral infection can be sequenced by next-generation techniques and analyzed for the presence of viruses, revealing the presence of all known viral pathogens in the sample. This method is based on Illumina sequencing of short RNAs and bioinformatics analysis of virus-derived small RNAs in the host. Here we describe a protocol for this challenging technique step by step with notes, in order to ensure success for every user.

Key words Virus, Diagnostics, Small RNA, Next-generation sequencing

1 Introduction

Next-generation sequencing (NGS) methods and discovery of RNA interference opened new possibilities in virus diagnostics [1–3]. During virus infection, small interfering RNAs of viral origin (21–25 nt long) representing the exact sequence of the infecting viruses are formed by the plant immune system [4]. Deep sequencing and bioinformatics analysis of the small RNA population extracted directly from field-grown plants offer a unique opportunity to reveal the presence of any virus or viroid present in the sample [5, 6], even if they were not described before [2, 7]. Small RNA NGS can also be used to test in parallel for the presence of all quarantined viruses in the sample. The probable decrease in the price of sequencing and more experience in the use and validation of this new method will revolutionize virus diagnostics by the authorities in the near future [8]. We used this technique to filter out virus infection in grapevine and fruit trees even in the stock collections and rootstock plantations.

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_9, © Springer Science+Business Media, LLC 2018

Total RNA is extracted from the collected sample and the small RNA fraction is purified. In order to make the small RNAs ready for sequencing, adapters must be ligated to both RNA ends, thus allowing reverse-transcription and PCR-based library generation. After sequencing and quality control, sequenced reads are aligned to viral reference genomes by bioinformatics pipelines, which reveal the presence of viral pathogens in the sample.

2 Materials

2.1 RNA Extraction and Library Preparation

1. 40% Acrylamide:bis-acrylamide (19:1).
2. 10% Ammonium persulfate (APS): Store aliquots at −20 °C.
3. Chloroform-isoamyl alcohol (24:1 v/v): Store at 4 °C.
4. Ethanol: 100% Ethanol, 70% ethanol, store at −20 °C.
5. 10 mg/mL Ethidium bromide.
6. Extraction buffer: 2% Cetyltrimethylammonium bromide (CTAB), 2.5% PVP-40 (polyvinylpyrrolidone), 100 mM Tris–HCl (pH 8.0), 25 mM EDTA (pH 8.0), 2 M NaCl, store at room temperature.
7. FDE loading dye: Dissolve 10 mg bromophenol blue and 10 mg xylene cyanol in 10 mL formamide, add 200 μL 0.5 M EDTA (pH 8.0), store in aliquots at −20 °C.
8. 15 mg/mL GlycoBlue coprecipitant: Store at −20 °C.
9. Illumina TruSeq Small RNA Library Prep Kit: Ligation Buffer (HML), Stop Solution (STP), 10 mM ATP, 25 mM dNTP Mix (dilute 12.5 mM dNTP Mix), ultrapure water, RNA RT Primer (RTP), RNA 3′ Adapter (RA3), RNA 5′ Adapter (RA5), RNA PCR Primer (RP1), RNA PCR Primer Index (RPI1-RPI48).
10. 9 M LiCl: Store at 4 °C.
11. Low Molecular Weight DNA Ladder (New England Biolabs Inc.).
12. 99.7% Isopropanol: Store at room temperature.
13. 2% β-Mercaptoethanol: Distribute aliquots in 1.5 mL Eppendorf tubes, store at 4 °C.
14. MilliQ pure water: Store at room temperature in aliquots for single use.
15. 4 M Sodium acetate (pH 5.2): Store at room temperature.
16. 0.3 M Sodium chloride: Store at room temperature.
17. 6× Orange Loading Dye (Thermo Fisher Scientific).
18. O'RangeRuler 20 bp DNA Ladder (Thermo Fisher Scientific).
19. 500 U/μL Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific).

20. 10× TBE, pH 8.3: 0.9 M Tris, 0.9 M boric acid, 0.02 M EDTA, store at room temperature, dilute 10× TBE to 1× with sterilized water.
21. *N*,*N*,*N*′,*N*′-tetramethylethylene-1,2-diamine (TEMED): Store at 4 °C.
22. 8% Polyacrylamide gel (PAGE): 8% Acrylamide:bis-acrylamide, 8 M urea, 1× TBE buffer, 0.06% APS, 0.16‰TEMED, make fresh every time.
23. 10,000 U RevertAid H- Reverse Transcriptase (Thermo Fisher Scientific).
24. 40 U/μL RiboLock RNase Inhibitor (Thermo Fisher Scientific).
25. SSTE buffer: 0.5% SDS, 10 mM Tris–HCl (pH 8.0), 1 mM EDTA (pH 8.0), 1 M NaCl, store at room temperature.
26. 5 U/μL T4 RNA Ligase (cloned, Life Technologies).
27. T4 RNA Ligase 2 truncated (New England Biolabs Inc.).
28. 0.5, 1.5, and 2 mL microcentrifuge sterile tubes.
29. 21-gauge needle.
30. SpinX Centrifuge tube filters, 0.45 μm (Costar, Corning Inc.)

2.2 Bioinformatics

1. FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
2. Trimmomatic [9].
3. BBMap/dedupe (https://sourceforge.net/projects/bbmap/).
4. BWA [10].
5. Samtools/idxStats [11].
6. IGV [12].
7. Samtools/bcftools [11].
8. Velvet 1.2.10 [13].
9. Megablast (http://dl.acm.org/citation.cfm?id=1553762).

3 Methods

Small RNA NGS depends on the detection of RNA; therefore it is very important to avoid contamination with RNases. The working environment should be clean and glassware/tubes should be nuclease free. The water and all aqueous solutions should be autoclaved and preferably aliquoted for single usage.

3.1 Sample Collection

High-throughput sequencing gives detailed information about the DNA or RNA content of even a very small amount of sample. To avoid misleading or false conclusions, sample collection must be designed and carried out very precisely. The presence of viruses in perennial woody plants can vary during the vegetation period and can be different in different organs. Phloem-limited viruses further complicate sampling procedure. Our experience showed that combining RNAs from different organs of the same plant and several plants from the same plantation gives a good solution for this problem. Samples are best to be collected in the fast-growing vegetation period. Plant and plantation pools can be generated from the purified RNAs; however, it is also advisable to keep individual RNA extracts for subsequent validation.

1. Collect fresh leaf samples into labeled plastic bags in the field (*see* **Note 1**).
2. Keep the collected samples in a cooler during sample collection and at 4 °C upon arrival to the lab. Samples can be stored at 4 °C for max. 2 days.
3. Pack enough material for RNA extraction individually using aluminum foil, mark it, and freeze it at −70 °C. Samples can be stored at −70 °C for a long period of time (years) without any damage to their RNA content (*see* **Note 2**).

3.2 Total RNA Extraction

High-quality RNA is a prerequisite for RNA sequencing. However, RNA extraction from woody plants is very challenging, because of the high amount of polysaccharides and polyphenols present, which can produce high-molecular-weight complexes via nucleic acid binding and inhibit successful extraction processes.

To protect RNA from ubiquitous RNA-degrading enzymes gloves must be worn, and samples must be kept on ice, homogenized in liquid nitrogen, and centrifuged at 4 °C except when otherwise stated. For further protection a detergent, CTAB preheated to 65 °C, and a reducing agent (β-mercaptoethanol) are applied to destroy the disulfide bridges of the RNases and other proteins. A mixture of chloroform and isoamyl alcohol is used to eliminate phenolic compounds, while LiCl selectively precipitates RNA. As contaminants react differentially with chemically different agents, precipitation by isopropanol and washing with ethanol are included to further purify the extracted RNA [14].

Before starting the procedure make sure that you have thermal block and water bath heated at 65 °C, and the centrifuge cooled to 4 °C. Heat the CTAB extraction buffer to 65 °C in a water bath.

Following is the detailed protocol, whereas a quick protocol is shown in Table 1.

1. Take out samples from the freezer and store in liquid nitrogen until starting their extraction (*see* **Note 3**).
2. Powder the sample (150–200 mg) in liquid nitrogen in a mortar, and then add 850 μL preheated CTAB and 17 μL 2% β-mercaptoethanol (*see* **Notes 4** and **5**).

Table 1
Flowchart of RNA extraction

Process	Time
Leaf homogenization (150–200 mg)	2–3 min/ sample
Incubation at 65 °C	10 min
Extraction with chloroform-isoamyl alcohol + centrifugation[a] (12,000 × *g*)	15 min
Extraction with chloroform-isoamyl alcohol centrifugation[a] (12,000 × *g*)	15 min
Precipitation with LiCl + incubation on ice	30 min
Centrifugation[a] (15,000 × *g*)	20 min
Resuspension of the precipitate in SSTE (65 °C) and extraction with chloroform-isoamyl alcohol centrifugation[a] (12,000 × *g*)	10 min
Precipitation with isopropanol Incubation at room temperature	10 min
Centrifugation[a] (15,000 × *g*)	20 min
Precipitation with 70% ethanol Centrifugation[a] (15,000 × *g*)	5 min
Drying and resuspension in MilliQ pure water	10 min

[a]Perform all centrifugation at 4°C in a cooled centrifuge

3. Pour the homogenized sample into labeled 2 mL Eppendorf tubes, mix thoroughly by vortexing, and place it into the warmed water bath.
4. Incubate the tubes at 65 °C for 10 min, and repeat vortexing every 5 min.
5. Add 850 μL ice-cold chloroform-isoamyl alcohol and mix it thoroughly by inverting the tube several times very gently.
6. Centrifuge at 12,000 × *g* for 10 min at 4 °C.
7. Label new 2 mL Eppendorf tubes, pipette 800 μL chloroform-isoamyl alcohol into them, and store them on ice.
8. Pipette the upper phase from **step 5** into the prepared Eppendorf tubes as a second extraction, and mix gently but thoroughly by inverting the tubes several times (*see* **Note 6**).
9. Centrifuge at 12,000 × *g* for 10 min at 4 °C.
10. Label new 1.5 mL Eppendorf tubes, pipette 250 μL 9 M LiCl into them, and store them on ice.

11. Transfer the upper phase from **step 9** to the LiCl-containing Eppendorf tubes and mix gently but thoroughly by inverting the tubes several times.
12. Incubate the mixture on ice for 30 min.
13. Centrifuge at 15,000 × *g* for 20 min at 4 °C.
14. Discard the supernatant and pipette 450 μL 65 °C preheated SSTE buffer into each tube.
15. Dissolve the precipitate by vortexing vigorously.
16. Add 450 μL chloroform-isoamyl alcohol and mix gently but thoroughly by inverting the tubes several times.
17. Centrifuge at 12,000 × *g* for 10 min at 4 °C.
18. Label new 1.5 mL Eppendorf tubes, and pipette 280 μL isopropanol and 30 μL 4 M sodium acetate into them.
19. Pipette the upper phase from **step 17** into the labeled tubes and mix gently but thoroughly by inverting the tubes several times.
20. Incubate the mixture for 10 min at room temperature.
21. Centrifuge at 15,000 × *g* for 20 min at 4 °C and discard the supernatant.
22. Wash the pellet with 1 mL ice-cold 70% ethanol and centrifuge at 15,000 × *g* for 5 min at 4 °C.
23. Pour off the ethanol, air-dry the tubes for 10 min to remove the residual, and then resuspend the isolated RNA in 25 μL MilliQ pure water.

3.3 Small RNA Library Preparation

Library preparation for Illumina sequencing is based on the use of the Truseq small RNA kit of Illumina. Enzymes and reagents not included in the kit are listed in Subheading 2. According to the kit description, libraries can be prepared from 1 μg total RNA. However, more reads with higher quality can be gained if the library is prepared from gel-purified small RNA fraction [15] prepared from 10–30 μg total RNA. Note that the quality of the reads depends on the quality of RNA and the number of reads (depth of the sequencing) correlates with the sequencing equipment used and the number of combined libraries. On the average, HiSeq2500 is able to produce 160–200 million reads/lane, which allows to combine several libraries in a single lane, depending on the depth required. There are 48 different indexed adapters available; thus a maximum of 48 samples can be combined per each lane. As the number of virus-derived sRNAs in woody plants is usually low, we do not recommend to combine more than 10–12 libraries in a single lane (Table 2).

3.3.1 Purification of sRNA Fraction from Total RNA

1. Prepare 8% TBE denaturing polyacrylamide gel containing urea (*see* **Note 7**).
2. Pre-run the gel at 100 V for 20–30 min. After pre-running, but before loading, wash the wells with 1× TBE.

Table 2
Flowchart of small RNA library preparation

Process	Time
Purification of sRNA fraction from extracted RNA	
PAGE	5–6 h
Elution with NaCl	Overnight
Precipitation	2–3 h
Library preparation	
3′ Adapter ligation	2–3 h
5′ Adapter ligation	2–3 h
Reverse transcription	1–2 h
PCR amplification	1–2 h
Purification of the small RNA library	
PAGE	5–6 h
Elution with NaCl	Overnight
Precipitation	2–3 h
Quality check	

3. Mix 10–30 μg of extracted total RNA with an equal volume of FDE in a microcentrifuge tube.
4. Denature the sample at 65 °C for 20 min, then chill on ice, and spin down briefly.
5. Load the samples on the gel (up to 20 μL/well) and run the gel at a constant voltage of 100 V for 1–1.5 h, until the bromophenol blue dye migrates to the bottom of the gel (*see* **Note 8**).
6. Disassemble the gel apparatus and stain the entire gel for 5 min by soaking in 60 mL 1× TBE containing 3 μL ethidium bromide. Use a separate container for every single gel to avoid cross-contamination.
7. Visualize the gel on a UV transilluminator.
8. Excise the piece of the gel that corresponds to the desired size of small RNA (15–30 nt, usually immediately above the bromophenol blue) with a sterile blade.
9. Puncture the bottom of a sterile, 0.5 mL microcentrifuge tube 3–4 times with a 21-gauge needle. Place the 0.5 mL punctured tube into a 2 mL microcentrifuge tube.
10. Place the excised gel slice into the prepared 0.5 mL punctured tube.

11. Centrifuge the microtubes containing the gel slices at full speed for 2 min at room temperature. Make sure that all of the gel has moved through the holes into the bottom 2 mL tube.
12. Remove and discard the 0.5 mL tube and add 350–400 μL sterile 0.3 M NaCl to the gel debris.
13. Shake the tube gently overnight at 4 °C to elute RNAs.
14. Transfer the eluate and gel debris to a Spin X cellulose acetate filter tube and centrifuge at 16,000 × *g* for 2 min. Repeat this step once more. Remove and discard the Spin X column containing gel debris.
15. Add an equal volume of 100% isopropanol and 1 μL of GlycoBlue to the eluate.
16. Incubate at −70 °C for at least 2–2.5 h (*see* **Note 9**).
17. Centrifuge the precipitated RNA at full speed at 4 °C for 20 min. Carefully discard the supernatant and wash the intact pellet twice with 1 mL of 70% cold ethanol (*see* **Note 10**).
18. Dry the pellet in speed vac machine for 3–5 min at room temperature.
19. Resuspend the pellet in 12 μL MilliQ pure water (*see* **Note 11**).

3.3.2 3′ Adapter Ligation

1. Preheat a thermocycler to 70 °C (*see* **Note 12**).
2. Pipette 2.5 μL purified small RNA into a sterile PCR tube on ice and add 0.5 μL RNA 3′ adapter (RA3).
3. Denature the reaction mixture for 2 min at 70 °C in the thermocycler, and then immediately place the tube on ice.
4. Preheat a thermocycler at 28 °C.
5. Pipette 1 μL ligation buffer (HML), 0.5 μL RNase inhibitor, and 0.5 μL T4 RNA ligase 2 (truncated) in a sterile PCR tube on ice; mix by pipetting up and down several times; and then centrifuge briefly.
6. Add the 2 μL mix to the reaction tube from **step 2**. Gently mix the entire volume by pipetting and incubate at 28 °C for 1 h.
7. Terminate the 3′ adapter ligation reaction by adding 0.5 μL ice-cold stop solution (STP) and mix by pipetting up and down several times.
8. Continue the incubation at 28 °C for 15 min and finally place the tube on ice.

3.3.3 5′ Adapter Ligation

1. Preheat a thermocycler to 70 °C.
2. Pipette 0.5 μL RNA 5′ adapter (RA5) into a sterile PCR tube on ice. Incubate the tube at 70 °C for 2 min and then immediately place on ice.
3. Preheat a thermocycler to 28 °C.

4. Pipette 0.5 μL 10 mM ATP and 0.5 μL T4 RNA ligase into a separate sterile PCR tube on ice.
5. Add the total volume of the denatured 5′ adapter from **step 2** to this mixture. Total volume is 1.5 μL.
6. Add 1.5 μL of the 5′ adapter mixture from **step 5** to the 3′ adapter reaction tube from Subheading 3.3.2, **step 8**.
7. Mix thoroughly by pipetting. The total volume of the reaction is now 6 μL.
8. Place the tube into the preheated thermocycler, incubate at 28 °C for 1 h, and then place the 3′–5′ adapter-ligated reaction on ice.

3.3.4 Reverse Transcription

1. Preheat a thermocycler to 70 °C.
2. Add 1 μL RT Primer (RTP) to the 3′–5′ adapter-ligated reaction (Subheading 3.3.3, **step 8**).
3. Gently pipette the entire volume up and down to mix thoroughly, place it to 70 °C for 2 min, and then place the tube on ice.
4. Preheat a thermocycler to 50 °C.
5. Set up the RT reaction mixture by pipetting 1 μL ultrapure water, 2 μL 5× reaction buffer, 0.5 μL 12.5 mM dNTP mix, 1 μL RNase inhibitor, and 1 μL Revert Aid H-reverse transcriptase into a sterile PCR tube on ice; mix by pipetting up and down several times; and then centrifuge briefly.
6. Add this 5.5 μL RT reaction mixture to the 3′–5′ adapter-ligated/primer reaction mix from Subheading 3.3.4, **step 3**; mix by pipetting up and down several times; and then centrifuge briefly. The total volume of RT reaction is now 12.5 μL.
7. Incubate the RT reaction at 50 °C for 1 h, and then place the cDNA-containing tube on ice (*see* **Note 13**).

3.3.5 PCR Amplification

1. Set up the PCR reaction mixture in a separate, sterile PCR tube on ice by pipetting 4.25 μL MilliQ pure water, 12.5 μL PCR mix (PML), 1 μL RNA PCR primer (RP1), and 1 μL RNA PCR primer index (RPIX) (*see* **Note 14**).
2. Mix the reaction by pipetting up and down several times, centrifuge briefly, and then place the tube on ice.
3. Add 6.25 μL cDNA from Subheading 3.3.4, **step** 7, into the PCR reaction mixture. The total volume is now 25 μL.
4. Pipette the entire volume up and down gently to mix thoroughly, and then place the tube on ice.
5. Denature the reaction in a thermocycler for 30 s at 98 °C, and then in 16 cycles amplify the libraries applying 10 s at 98 °C for denaturation, 30 s at 60 °C for annealing, and 15 s at 72 °C for

elongation. Finalize the reaction by incubating the reaction mixture for 10 s at 72 °C.

6. The amplified small RNA library is now ready for purification (*see* **Note 15**).

3.3.6 Purification of the Small RNA Library

1. Prepare 8% TBE polyacrylamide gel (*see* **Note 16**).
2. Pre-run the gel at 100 V for 20–30 min. After pre-running but before loading, wash the wells with 1× TBE buffer.
3. Mix the 25 μL PCR amplification product with 5 μL 6× Orange DNA loading dye. The total volume is now 30 μL.
4. Load two different size markers, a 20 bp DNA ladder and a 50 bp low-molecular-weight ladder, in the two outermost wells (one on each side of the gel).
5. Load the amplified PCR product from Subheading 3.3.5, **step 6**, in the middle of the gel into two consecutive wells (15 μL/well).
6. Run the gel at a constant voltage of 100 V for 1.5–2 h, until the xylene cyanol dye migrates to the bottom of the gel.
7. Disassemble the gel apparatus and stain the entire gel by soaking, in a separate container, in 60 mL 1× TBE buffer containing 3 μL ethidium bromide for 5 min (*see* **Note 17**).
8. Puncture the bottom of a sterile, 0.5 mL microcentrifuge tube 3–4 times with a 21-gauge needle. Place the 0.5 mL punctured tube into a 2 mL microcentrifuge tube.
9. Visualize the size marker and the amplified sRNA library on a UV transilluminator.
10. Excise the piece of the gel that corresponds to the desired size (145–160 nt) of small RNA library (*see* **Note 18**).
11. Place the gel slice into the prepared 0.5 mL tube from Subheading 3.3.6, **step 8**.
12. Centrifuge at full speed for 4 min at room temperature. Make sure that all the gel has moved through the holes into the bottom 2 mL tube.
13. Remove and discard the 0.5 mL punctured tube and add 350–400 μL sterile 0.3 M NaCl to the gel debris.
14. Shake the tube gently overnight at 4 °C to elute DNA.
15. Transfer the eluate and gel debris to a Spin X cellulose acetate filter tube and centrifuge at 16,000 × *g* for 2 min. Repeat this step once more. Remove and discard the Spin X column containing gel debris.
16. Add 1 μL GlycoBlue to the eluate and precipitate it with 1 mL 100% ethanol.
17. Incubate at −70 °C for at least 2–2.5 h (*see* **Note 19**).

18. Centrifuge the precipitate at full speed at 4 °C for 20 min. Carefully discard the supernatant and wash the intact pellet twice with 1 mL 70% cold ethanol (*see* **Note 20**).
19. Dry the pellet in speed vac machine for 3–5 min at room temperature.
20. Resuspend the pellet in 12 μL sterile 1× TE resuspension buffer.
21. The pure small RNA library is now ready for sequencing, but can be stored at −20 °C or −70 °C for a longer period of time.

3.4 Sequence the Prepared Libraries on Illumina Hiseq Machine

Sequencing can be ordered from different companies (*see* **Note 21**).

3.5 Bioinformatics Analysis of Small RNA Reads

A schematic representation of the bioinformatic pipeline used is shown in Fig. 1.

3.5.1 Collecting Information for Virus Diagnostics

1. Check the quality of the sequenced small RNA libraries using the *FastQC program* (with default settings). This program allows you to gain quality information (i.e., base sequence content, duplicate sequences, base sequence quality). Libraries with high-quality data can be used in further analysis. The format of the input file is fastq.gz (a compressed fastq file that you get from the sequencing machine), whereas the output file format is html.
2. Do trimming on your raw reads using the *Trimmomatic program*. It finds and cuts adapters from the ends of the reads. Both input and output formats are uncompressed fastq.
3. Deduplicate your reads by removing redundant reads using the *BBmap Dedupe Program*. Both input (trimmed) and output (deduplicated = non-redundant) file formats are fastq (*see* **Note 22**).
4. For virus diagnostics use two different strategies: align sRNA reads directly (4(a)) or after contig built (4(b)) to viral reference genomes.

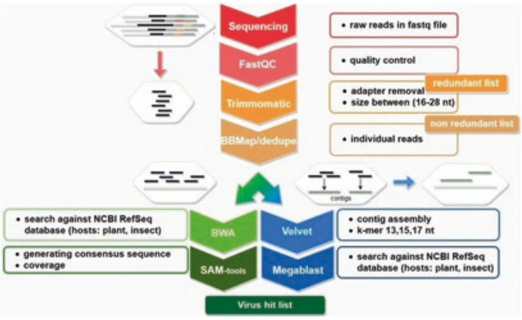


Fig. 1 Schematic diagram of the bioinformatics pipeline

(a) Generate virus hit list from aligning sRNA reads directly to reference genomes.

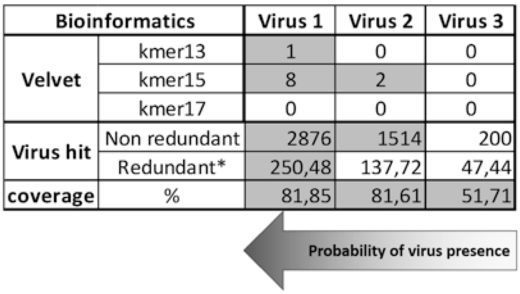
- Map the nonredundant small RNA reads (fastq from **step 3**) to viral Reference Genomes using the *BWA-aln* tool. To do this, download Reference Genomes (plant and insect host) from NCBI into a reference file. Input file format is fastq (nonredundant), and output file format is bam (*see* **Notes 23** and **24**).
- Count mapped reads using the *Samtools idxStats program*. This program uses bam (generated from nonredundant reads) input files and provides the number of viral hit reads as a results.txt output file.
- Prepare consensus sequences using *Samtools/bcftools*. These consensus sequences contain the information from the small RNA reads of your libraries aligned to reference genomes. *Bcftools* calls the nucleotides from the matching reads which align—input file format is bam (generated from nonredundant reads)—to these chosen reference genomes and fill out the gaps between reads with N. As an output you will get a fasta file in the same length as the Reference genome was, with sequence information only about parts from which small RNAs were generated.
- Calculate the coverage of the specified genome by comparing consensus generated at step (c) by the reference genome. Divide the number of positions with sequence information by the number of nucleotide length of the genome and multiple it with 100.

(b) Generate virus hit list from aligning contigs generated from sRNA reads to reference genomes.

- Build longer contigs from small RNA reads by their de novo assembly by the *Velvet program*. Contigs can be assembled from reads by using different *k-mers*. In our practice *k-mer* sizes 13, 15, and 17 work fine. Input file should be fastq (nonredundant), and output is a fasta file which contains the generated contigs.
- Align the contigs assembled by Velvet to Reference Genome sequences using *MegaBLAST*. To set parameters you can use –e1E-10 (e-value = 1^-10), –p95 (in case if the identity of aligned contigs is below this threshold, the read is not present in the output), and –D3 (type of output format). Input file is contigs.fasta; the result is a list of contigs identified as of viral origin with the name and identifier of the reference genome.

3.5.2 Summarizing Bioinformatics Results for Virus Diagnostics

1. Arrange results from bioinformatics analysis in a score table (an example in Fig. 2).



Bioinformatics		Virus 1	Virus 2	Virus 3
Velvet	kmer13	1	0	0
	kmer15	8	2	0
	kmer17	0	0	0
Virus hit	Non redundant	2876	1514	200
	Redundant*	250,48	137,72	47,44
coverage	%	81,85	81,61	51,71

Probability of virus presence

Fig. 2 Arrangement of a virus diagnostics result table with hypothetic values

2. For the highest probability of virus presence, in our personal experiences and for our specific routine applications three parameters have to reach the threshold: presence of at least one assembled contig (at any kmer), high amount of redundant normalized reads (no. of reads/total reads × 1 million) >200, and genome coverage >20%. If any of these parameters is below threshold, the probability of virus presence and the possibility to verify its presence by RT-PCR decrease.
3. For final diagnostics verify the virus presence by RT-PCR using diagnostic primers (*see* **Note 25**).
4. If verification by RT-PCR fails, design new primers based on comparison of the available viral genomes in the database and the consensus sequence generated at Subheading 3.5.1, **step 4 (a) point 3** (the most probable sequence of the strain present in the sample).

4 Notes

1. It is advisable to take photos of each sampled plant, because it can give important information later.
2. If you pack leaves, pack each leaf individually; do not cut them in half, because this damage can activate RNases and other degrading processes.
3. If a relatively large (more than 100 mg) leaf was packed, use only a small part of it for RNA extraction.
4. CTAB powder could cause irritation; therefore make the solution under a fume hood.
5. β-Mercaptoethanol is not only malodorous but can also cause irritations; therefore it is strongly advisable to use it under the fume hood.
6. If working with lots of samples, labeling Eppendorf tubes and pipetting everything needed for the consecutive step can be done in advance to shorten the time of sample handling and further minimize the possibility of the degradation of the sample.

7. It is recommended to prepare a separate gel for each library to avoid cross-contamination.
8. If the concentration of the extracted RNA is low, several wells can be used to reach the required amount of starting material.
9. Precipitated RNA can also be kept at −70 °C overnight.
10. Use of GlycoBlue can help not to discard the pellet accidentally.
11. The isolated small RNA sample can be stored at −70 °C.
12. Using a PCR machine as a thermoblock ensures precise maintenance of the desired temperature.
13. The cDNA prepared here can be stored at −20 °C if it is not directly processed further.
14. Take special care using different indexes for libraries which are planned to be sequenced combined in a single lane.
15. The number of cycles can be increased up to 21 if a very small amount of starting material was used.
16. It is recommended to prepare a separate gel for each library to avoid cross-contamination.
17. Use a separate container for staining each gel to avoid cross-contamination.
18. The exact size of the PCR product depends on the length of the cloned small RNA and the Illumina adapters used.
19. Precipitation can also take place overnight at −70 °C.
20. Use of GlycoBlue can help not to discard the pellet accidentally.
21. Not only the price and the deadline of sequencing, but also the resulting quality and the number of sequenced reads, can be different for different companies. It depends on the type of sequencing machine used and also on experience.
22. In trimmed sRNA libraries, one read can be present several times. Before aligning small RNA reads to reference genomes it is advisable to remove duplicates in order to minimize the number of reads, which reduces processing time.
23. You can map your reads to any sequences that you define, so you will see which reads match the given (genome) sequences.
24. The bam output files can be investigated using the *IGV program*.
25. If one of the three parameters is below the suggested values, the probability of the presence of the virus drops down, but the thresholds depend on both the particular virus and the host.

Acknowledgments

Our work was supported by the Hungarian Ministry of Agriculture and the KTIA_AIK_12 -1-2013-0001 project and by a Hungarian Scientific Research Fund (OTKA; K119783) grant from the National Research, Development and Innovation Office. N.C. and D.B. participate in the Program for Reinforcement of Scientists of the Hungarian Ministry of Agriculture. N.C. is a Ph.D. student of the Festetics Doctoral School at the University of Pannonia. R.P., E.D. and D.B are Ph.D. students of the Doctoral School of Biological Sciences at Szent István University.

The authors acknowledge the contribution of the COST Action FA 1407–DIVAS (Deep Investigation of Virus Associated Sequences).

References

1. Navarro B, Pantaleo V, Gisel A, Moxon S, Dalmay T, Bisztray G, Di Serio F, Burgyan J (2009) Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. PLoS One 4(11):e7686. https://doi.org/10.1371/journal.pone.0007686
2. Giampetruzzi A, Roumi V, Roberto R, Malossini U, Yoshikawa N, La Notte P, Terlizzi F, Credi R, Saldarelli P (2012) A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus Res 163(1):262–268. https://doi.org/10.1016/j.virusres.2011.10.010
3. Pantaleo V, Saldarelli P, Miozzi L, Giampetruzzi A, Gisel A, Moxon S, Dalmay T, Bisztray G, Burgyan J (2010) Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine. Virology 408(1):49–56. https://doi.org/10.1016/j.virol.2010.09.001
4. Baulcombe D (2004) RNA silencing in plants. Nature 431(7006):356–363. https://doi.org/10.1038/nature02874
5. Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJ, Burger JT (2010) Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virology 400(2):157–163. https://doi.org/10.1016/j.virol.2010.01.023
6. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388(1):1–7. https://doi.org/10.1016/j.virol.2009.03.024
7. Wu Q, Wang Y, Cao M, Pantaleo V, Burgyan J, Li WX, Ding SW (2012) Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. Proc Natl Acad Sci U S A 109(10):3938–3943. https://doi.org/10.1073/pnas.1117815109
8. Massart S, Candresse T, Gil J, Lacomme C, Predajna L, Ravnikar M, Reynard JS, Rumbou A, Saldarelli P, Skoric D, Vainio EJ, Valkonen JP, Vanderschuren H, Varveri C, Wetzel T (2017) A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. Front Microbiol 8:45. https://doi.org/10.3389/fmicb.2017.00045
9. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170
10. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324
11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

12. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26. https://doi.org/10.1038/nbt.1754

13. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821–829. https://doi.org/10.1101/gr.074492.107

14. Gambino G, Perrone I, Gribaudo I (2008) A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem Anal 19(6):520–525. https://doi.org/10.1002/pca.1078

15. Nagy T, Kis A, Poliska S, Barta E, Havelda Z, Marincs F (2016) [Letter to the Editor] Comparison of small RNA next-generation sequencing with and without isolation of small RNA fraction. BioTechniques 60(6):273–278. https://doi.org/10.2144/000114423

The Use of High-Throughput Sequencing for the Study and Diagnosis of Plant Viruses and Viroids in Pollen

Kris De Jonghe, Annelies Haegeman, Yoika Foucart, and Martine Maes

Abstract

This protocol details the wet lab preparation, extraction of fruit pollen samples, and analysis of the sequencing data following Illumina NextSeq small and total RNA sequencing. The protocol was developed for virus and viroid detection using NGS sequencing and was based on the results of a comparison between different extraction methods followed by yield, RNA purity, and integrity assessment. Moreover, the advantage of an additional ribosomal (r)RNA depletion step to the total RNA extraction protocol was evaluated. The smallRNA procedure is the preferred method of choice. If the total RNA protocol is chosen, the use of the mirVana kit followed by an rRNA depletion step is the best option. The library preparation and sequencing steps were outsourced. As a final step in the data analysis, the VirusDetect software was used to detect the viruses and viroids in the pollen samples.

Key words NGS sequencing, Pollen, Viruses, Viroids, VirusDetect

1 Introduction

The use of high-throughput sequencing as a diagnostic screening tool for full-range testing of viruses and viroids in plant materials offers many advantages over the currently routinely used biological indexing, ELISA, LAMP, PCR-based, and Sanger sequencing methods. Over time, each of the new or improved diagnostic tools significantly improved our ability to detect and identify virus infections [1, 2]. Development of tests offering extra assets, such as usefulness in resource-poor and non-laboratory environments (e.g., loop-mediated isothermal amplification; LAMP), extremely short output times (e.g., lateral flow devices; LFD), and multiplexing (e.g., array-based tests), continuously contributed to the plant virus diagnostic innovations [3, 4]. With the introduction of the high-throughput sequencing technology (next-generation sequencing, NGS), the scope is now also widened to the discovery of non-target and new viruses in the test samples [4]. Now that NGS is becoming cheaper, and therefore more accessible, the introduction of this

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_10, © Springer Science+Business Media, LLC 2018

technology in routine plant health diagnostics is nearby, thereby replacing the routinely used automated Sanger sequencing. This protocol specifically focuses on the methodology to use NGS for the study and diagnosis of plant viruses and viroids in pollen matrices, more specifically on the sample preparation, RNA extraction in preparation of the actual sequencing, and data analysis.

2 Materials

2.1 Samples

1. This protocol is specifically focusing on dry powdered pollen. The protocol was tested on fruit tree pollen, from *Prunus avium* L. (cherry), *Malus domestica* L. Borck (apple), and *Pyrus communis* L. (pear).

2.2 RNA Extraction

1. Total RNA extraction: RNeasy Plant Mini (Qiagen) and mirVana miRNA with phenol (Ambion) kits.
2. small RNA extraction: mirVana kit.
3. RNaseZap®, RNase decontamination solution (Ambion).

2.2.1 Total RNA Extraction by Means of RNeasy Plant Mini Kit (Qiagen)

Materials not provided in the kit:

1. (Blue) polypropylene pellet pestles (Sigma Aldrich).
2. Liquid nitrogen.
3. RNase-free 2 or 1.5 mL microfuge tubes.
4. Adjustable pipettes and RNase-free tips.
5. ≥99.8% Ethanol.
6. Microcentrifuge capable of at least 10,000× *g*.
7. Vortex Genie2 (Fiers).
8. 2-Mercaptoethanol.

2.2.2 mirVana miRNA Isolation and Total RNA Extraction (Ambion)

Materials not provided in the kit:

1. (Blue) polypropylene pellet pestles (Sigma Aldrich).
2. Liquid nitrogen.
3. RNase-free 2 or 1.5 mL microfuge tubes.
4. Adjustable pipettes and RNase-free tips.
5. ≥99.8% Ethanol.
6. Microcentrifuge capable of at least 10,000× *g*.
7. Vortex Genie2 (Fiers).
8. Eppendorf Thermomixer comfort (VWR International).

2.3 Ribosomal RNA Depletion

1. Ribosomal RNA depletion was done by means of the RiboMinus™ Plant Kit for RNA-Seq kit (Thermo Fisher Scientific)

Materials not provided in the kit:

2. RNase-free 2 or 1.5 mL microfuge tubes.
3. Adjustable pipettes and RNase-free tips.
4. ≥99.8% Ethanol.
5. Disobit (ethanol, denatured with 3% IPA (isopropyl alcohol) + bitrex (denatonium benzoate)).
6. Magnetic particle separator.
7. Eppendorf Thermomixer.
8. Waterbath Julabo TW20 (Omnilabo International).
9. Glycogen, 20 μg/μL (Roche Diagnostics).
10. 3 M Sodium acetate (NaOAc), pH 5.2 (Sigma Aldrich).

2.4 Ethanol Precipitation (RNA/DNA Concentrating and Desalting Procedure)

1. RQ1 RNase-Free DNase-kit (Promega).
2. RNase-free 2 or 1.5 mL microfuge tubes.
3. Adjustable pipettes and RNase-free tips.
4. Cooled Eppendorf Centrifuge 5417R (Eppendorf).
5. 3 M Sodium acetate (NaOAc), pH 5.2 (Sigma Aldrich).
6. ≥99.8% Ethanol.
7. Disobit (ethanol, denatured with 3% IPA + bitrex).

2.5 Cleanup and Concentrate Pre-purified RNA Samples

1. Nucleospin® RNA Cleanup XS kit (Macherey-Nagel).
2. RNase-free 2 or 1.5 mL microfuge tubes.
3. Adjustable pipettes and RNase-free tips.
4. Microcentrifuge capable of at least 10,000× *g*.
5. Vortex Genie2 (Fiers).
6. ≥99.8% Ethanol.

2.6 Bleach Gel Electrophoresis for RNA Integrity Evaluation

1. Nanodrop ND-1000 (Thermo Fisher Scientific).
2. Adjustable pipettes and RNase-free tips.
3. Mupid-One Electrophoresis system (Eurogentec).
4. 50× TAE buffer.
5. Agarose molecular biology grade.
6. 6× Orange DNA Loading Dye (Thermo Fisher Scientific).
7. O'GeneRuler 1 kb ladder ready to use (Thermo Fisher Scientific).
8. Midori Green Advanced DNA stain (Filterservice).
9. Azure™ C150 Gel Imaging Workstation (Azure Biosystems).

2.7 Software

1. Data analysis should be done in a LINUX/UNIX environment. The following software need to be installed:

(a) FastQC [5] (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

(b) Cutadapt [6] (http://cutadapt.readthedocs.io/).

(c) PRINSEQ (optional) [7] low-complexity sequence removal (http://prinseq.sourceforge.net/).

(d) PEAR [8] (optional) Paired-End Read Merger: merging of forward and reverse reads; for total RNA datasets only (https://sco.h-its.org/exelixis/web/software/pear/).

(e) sortMeRNA [9] ribosomal RNA filtering; for total RNA datasets only (http://bioinfo.lifl.fr/RNA/sortmerna/).

(f) VirusDetect [10] (http://bioinfo.bti.cornell.edu/cgi-bin/virusdetect/index.cgi).

2.8 Databases

2.8.1 Reference Viral Database

1. The VirusDetect software includes the datasets "vrl_plant" and "vrl_plant_protein". These contain curated and nonredundant DNA and protein sequences of viruses derived from plants [10].
2. On the FTP site of the VirusDetect software (ftp://bioinfo.bti.cornell.edu/pub/program/VirusDetect/virus_database/) additional databases can be downloaded, for example, a database of viruses from invertebrates ("vrl_Invertebrates_217_U100").

2.8.2 Host Database

1. The host database is used for mapping the reads; mapped reads are derived from the host plant and not from the virus and will be removed from the analysis.
2. The host databases need to be added to the subfolder "databases" within the VirusDetect installation folder.
3. In case the full genome sequence of the host plant is publicly available the sequence is downloaded and saved in fasta format, here called "plant_genome.fasta".
4. If there is no reference plant genome, we recommend selecting the closest related plant species for which there is genome information available.
5. The fasta file still needs to be indexed for the mapping program BWA. This can be done as follows: bwa index /path/to/VirusDetect/databases/plant_genome.fasta

3 Methods

3.1 Sample Preparation

Before working with RNA, clean the lab bench, pipettes, and, if applicable, the microcentrifuge tubes with an RNase decontamination solution (e.g., Ambion RNaseZap® Solution). During this procedure it is important to wear gloves and change them frequently; they will protect the RNA from nucleases present on the skin. Use RNase-free pipette tips to handle every step of the

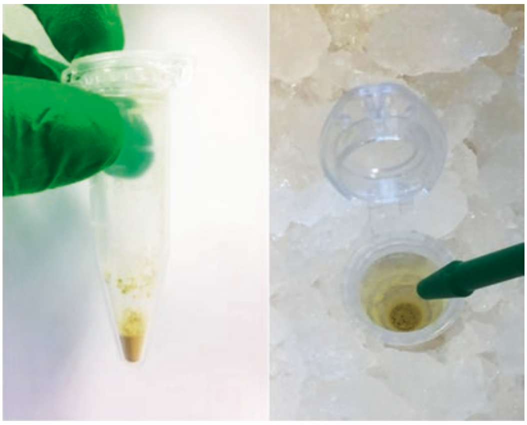


Fig. 1 Side and top view of the sample preparation, grinding the pollen sample in a 5 mL Eppendorf® safe-lock microcentrifuge tube using a sterile polypropylene pellet pestle (Sigma-Aldrich)

extraction protocol. The sample size for the extractions is 100 mg, except for the mirVana protocol where 250 mg of pollen can be processed. The protocol was used on dry pollen (powder) from *Malus*, *Pyrus*, and *Prunus* sp. which was stored at −20 °C until use.

1. Weigh the amount of pollen in an appropriate microcentrifuge tube (5 or 2 mL) and freeze the sample immediately in liquid nitrogen to inactivate RNases.
2. Carefully grind the pollen before starting the extraction protocol (Fig. 1) (*see* **Notes 1** and **2**).

3.2 RNA Extraction

The protocol was evaluated for two different extraction kits. The kit of choice after evaluation is the mirVana kit (Ambion), both when the total RNA strategy or the small RNA strategy is chosen (*see* **Notes 3–6**).

1. Total RNA is obtained by using RNeasy Plant Mini Kit (*see* **Notes 3** and **4**) following the instructions of the manual from “Purification of total RNA from plants cells and tissues and filamentous fungi”.
2. Small RNA isolation (*see* **Note 5**) and total RNA extraction is done by means of the mirVana kit following the manufacturer’s instructions. The most important difference is the amount of ethanol used for binding (Fig. 2).
   (a) After sample preparation, add 10 volumes of lysis/binding solution (provided in the kit) and vortex until all visible clumps are dispersed.
   (b) Go immediately to section E (Organic extraction) and follow the further instructions.

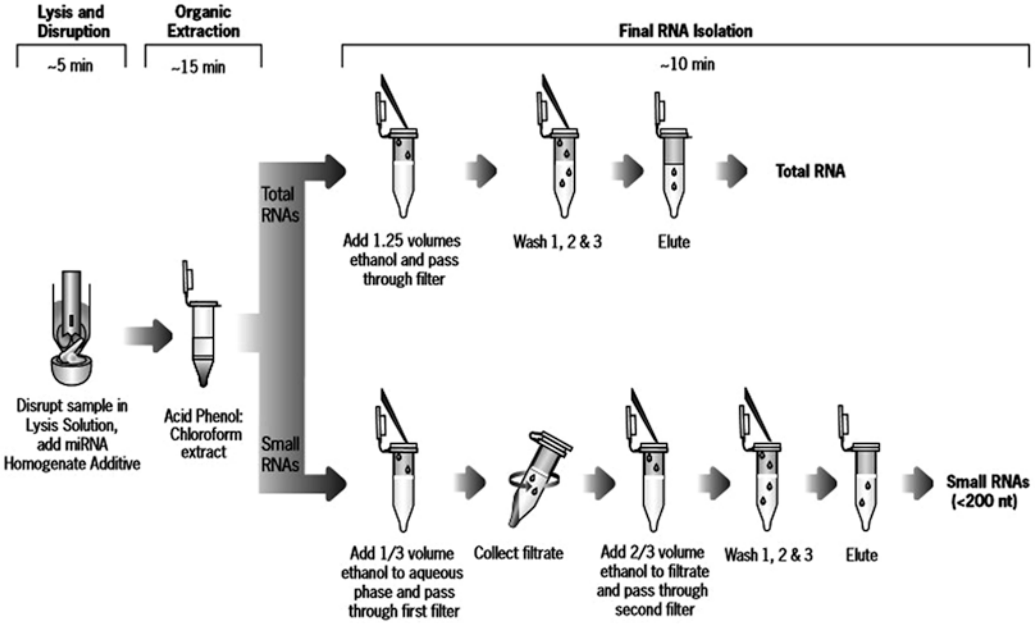

Lysis and Disruption
~5 min
Organic Extraction
~15 min
Final RNA Isolation
~10 min
Total RNAs
Add 1.25 volumes ethanol and pass through filter
Wash 1, 2 & 3
Elute
Total RNA
Disrupt sample in Lysis Solution, add miRNA Homogenate Additive
Acid Phenol: Chloroform extract
Small RNAs
Add 1/3 volume ethanol to aqueous phase and pass through first filter
Collect filtrate
Add 2/3 volume ethanol to filtrate and pass through second filter
Wash 1, 2 & 3
Elute
Small RNAs (<200 nt)

Fig. 2 Overview of the mirVana™ miRNA Isolation Kit Procedure. The sample is first lysed in a denaturing lysis solution which stabilizes RNA and inactivates RNases. The lysate is then extracted once with acid-phenol:chloroform which removes most of the other cellular components, leaving a semi-pure RNA sample. This is further purified by one of the two procedures to yield either total RNA or a size fraction enriched in small RNAs (Figure © Life Technologies Corporation)

(c) Store the RNA at −80 °C and make aliquots for QC control to prevent contamination in each step.

3.3 Ribosomal RNA Depletion

A comparison was made between samples with and without rRNA depletion on the total RNA protocol (*see* **Note 6**). For the rRNA removal, the RiboMinus™ Plant Kit for RNA-Seq was used.

1. After total RNA extraction, an rRNA depletion can be proceeded according to the instruction manual.
2. For further use of downstream applications, it is recommended to concentrate the end solution by concentrating RiboMinus™ RNA using the RiboMinus™ Concentration Module. Follow the manufacturer's instructions.

3.4 Ethanol Precipitation

For concentrating and desalting the RNA, an ethanol precipitation step is recommended.

1. RNA integrity is critical and an additional DNase treatment step is highly recommended. The RQ1 RNase-Free DNase kit was used for this step according to the manufacturer's instructions.
2. Add 2.5 volume of ice-cold 100% ethanol (stored at −20 °C) to 1 volume of RNA.

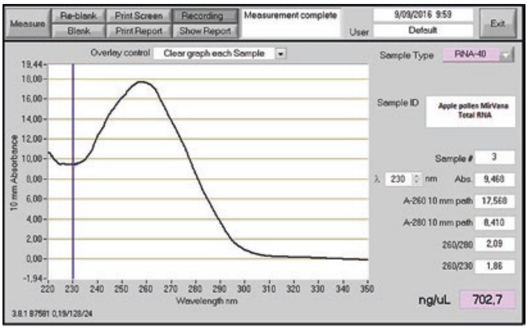
3. Add 1/10 volume of 3 M NaOAc pH 5.2.
4. Incubate the solution for 1 h at −20 °C.
5. Centrifuge (4 °C) at full speed for 20 min and remove the supernatant.
6. Add 100 μL of 70% ice-cold ethanol to the pellet.
7. Centrifuge (4 °C) at full speed for 20 min and remove the supernatant.
8. Air-dry the pellet for 15–30 min.
9. Solve the pellet in nuclease-free water.
10. Store at −20 °C.

3.5 RNA Quality Control, Integrity Check, and RNA Purification

1. Quality control:
   (a) The RNA concentration is measured spectrophotometrically (Nanodrop ND-1000) (*see* **Note 7**).
   (b) Use the elution solution from the corresponding extraction method as blank (*see* Fig. 3).
2. RNA integrity check by means of a "bleach gel":

   A measure of the RNA integrity is RNA integrity number (RIN) (*see* **Note 8**). This protocol describes a quick and easy-to-use RNA quality check method, based on the method of Aranda et al. [11]. The RNA integrity can be quickly analyzed on a simple native agarose gel, by adding a small amount of commercial bleach to the TAE buffer (Fig. 4) (*see* **Note 9**).

   (a) Add 2% of agarose to 0.5× TAE buffer.
   (b) Add 1% of bleach (NaOCl, sodium hypochlorite; 15°) and incubate at room temperature with occasional swirling.
   (c) Heat the suspension to melt the agarose.
   (d) Allow the solution to cool down and add 6 μL (in 100 mL buffer) of Midori Green Advanced DNA stain.



Measurement complete

9/09/2016 9:59

User Default

Overlay control: Clear graph each Sample

Sample Type: RNA-40

Sample ID: Apple pollen MirVana Total RNA

Sample #: 3

λ 230 nm Abs. 9,460

A-260 10 mm path: 17,568

A-280 10 mm path: 8,410

260/280: 2,09

260/230: 1,86

ng/uL 702,7

Fig. 3 Graphic result of a measurement of an apple pollen sample with a yield of 702.7 ng/μL, a 2.09 ratio for 260/280, and a low amount of organic compounds indicated by a high 260/230 ratio (1.86)

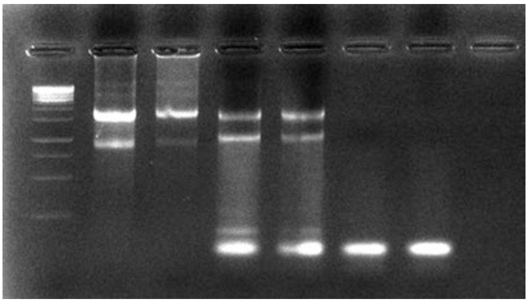


Fig. 4 RNA integrity check on a bleach gel. Lane 1: 1 kb ladder. Lanes 2 and 3: apple and pear pollen extracted by RNeasy, respectively. Lanes 4 and 5: apple and pear pollen, extracted by mirVana total RNA method, respectively. Lanes 6 and 7: apple and pear pollen, extracted by mirVana small RNA method, respectively

(e) Pour the solution into gel mold and allow the "bleach gel" to solidify.

(f) Load the RNA samples mixed with 6× orange DNA loading dye to a final concentration of 1× and add a 1 kB DNA ladder (O'GeneRuler 1 kb ladder).

(g) Electrophorese the gel in 0.5× TAE buffer at 100 V for approximately 35 min and visualize the electrophorese gel with the Azure Gel Imaging Biosystem.

(h) Evaluate the presence of the Eukaryotic 28S and 18S bands, at ±5 kb and ±2 kb, respectively. In addition, rRNA presence can be seen from a third (5.8/5S) band at ±150 bp.

3. RNA purification:

   If the purity of the RNA sample is too low (e.g., 260/230 < 1.8), we continue with an RNA purification protocol by NucleoSpin® RNA Cleanup XS. Follow instructions of the manual provided in the kit.

3.6 Library Preparation and High-Throughput Sequencing

In our case, we outsourced this step. Library preparation was done using the NEBNext Ultra RNA library kit (New England BioLabs) (*see* **Note 10**). Short sequencing reads should be obtained using one of the Illumina platforms (MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq). For sequencing total RNA, we recommend generating at least 20 million read clusters, and a read length of minimum 150 bp (paired-end sequencing). For smallRNA datasets, single-end sequencing of 36 bp is sufficient, with a minimum of three million reads.

3.7 Data Analysis

3.7.1 Quality Control

Data files returned from the sequencing provider have been demultiplexed and are in *fastq* format. In our case, we used paired-end reads of 2 × 150 bp for total RNA datasets and single reads of 36 bp for small RNA datasets. In case of paired-end sequencing of small RNA, we advise to continue working with the forward reads only.

1. Standard quality control is done using the software FastQC. The software can be run with your *fastq* file as input using fastqc input_file.fastq.
2. A quality filtering step is used to remove reads with a low average quality (*see* **Note 11**).
3. Since we recommend to merge the forward and reverse reads, we also recommend not to remove or trim any low-quality sequences at this point. The merging step will automatically select the base with the highest quality in the overlapping region. The merging software can do an additional quality trimming step before the reads are merged (*see* **Note 12**).

3.7.2 Remove Adapter Sequences

1. The reads need to be checked for the presence of Illumina adapter sequences. In the FastQC report the putative adapters are visualized.
2. Using cutadapt [6], the adapters are removed from all sequences. The following command shows how the program can be run for single-read datasets (smallRNA):

   cutadapt -aAGATCGGAAGAGCACACGTCTGAACTC-CAGTCACTACGCT -m 16 -M 28 -o trimmed.fastq input_file.fastq

   The option *-a* is to specify the (beginning of the) 3′ adapter; if the options *-m 16* and *-M 28* are included trimmed fragments between 16 and 28 bp are retained (*see* **Note 13**); the output file name can be specified using *-o*. Finally, the file name of the reads is given as argument. This file can be a file derived from single-read sequencing or in case of paired-end sequencing of total RNA datasets the forward sequencing file can be used. In its default settings, the program allows a maximum of 10% errors compared to the adapter sequence.
3. In case of paired-end sequencing, adapters have to be removed from both sides (total RNA dataset). The following command can be used, where the *-A* option specifies the (beginning of the) 3′ adapter of the reverse reads and the *-p* option the output file name of the second file containing the reverse reads. The command now takes two arguments, corresponding to the names of the input file containing the forward reads and the input file containing the reverse reads:

   cutadapt -aAGATCGGAAGAGCACACGTCTGAACTC-CAGTCACTACGCT -AAGATCGGAAGAGCGTCGTGT AGGGAAAGAGTGTACTCTAG -m 30 -o trimmed-1.fastq -p trimmed-2.fastq input_forward.fastq input_reverse.fastq.

3.7.3 Remove Low-Complexity Sequences (Optional, Small RNA Dataset Only)

Small RNA datasets consist of very short reads; a small portion of these reads can be low-complexity sequences that can be considered as noise. This step is optional and removes low-complexity sequences (e.g., TTTTTTTTT or TATATATATA) using the

DUST approach implemented in PRINSEQ [7]. The software can be run as follows:

```
prinseq-lite -fastq trimmed.fastq -lc_method dust -lc_threshold 10 -out_good trimmed_and_filtered -out_bad low_complexity_seqs
```

The option *-fastq* specifies the input file; *-lc_method* specifies which algorithm you want to use (in this case DUST) with *-lc_threshold* as cutoff (here 10). The name of the output file containing good sequence reads is given with the option *-out_good*, while *-out_bad* contains the name of the file with the filtered low-complexity sequences. Note that we do not add suffixes to the file names since PRINSEQ will add the file extension automatically.

3.7.4 Merging Forward and Reverse Reads (for Total RNA Dataset Only)

This step merges the forward and reverse reads using the program PEAR [8] (*see* **Note 14**). The program adjusts the quality scores in the overlapping zone: if the two reads have the same base in the overlapping region, then the quality score will be excellent, while if the base is different, the base with the highest quality score will be chosen, with its corresponding quality score:

```
pear -f trimmed-1.fastq -r trimmed-2.fastq -o samplename -p 1.0 -v 20 -u 0.1
```

The *-f* and *-r* options specify the forward and reverse input file name, respectively, while the *-o* option can specify the output file name base (PEAR will add extensions to the output files automatically). The option *-v* specifies how many bases in the two reads should overlap; in this case, we choose 20 nt.

We also choose to discard reads with more than 10% Ns by adding the option *-u 0.1*. Using *-p 1.0* you can disable any statistical testing.

Other options are available, for example, to specify the number of cores or the memory the program can use; see PEAR manual for more details. The merged fragments are saved in a file with extension ".assembled". If the fragments cannot be merged, the two reads will remain in separate files for the F and R reads, the ".unassembled.forward" and ".unassembled.reverse" files (*see* **Notes 15** and **16**).

3.7.5 Removal of rRNA (for Total RNA Dataset Only)

Further processing is done using the merged reads only. In our experience, the vast majority of the reads indeed merge, indicating that the RNA fragments in the library were relatively small and overlaps were identified. In the next step, the rRNA sequences (typically mainly derived from the host plant) are removed from the merged reads using the software SortMeRNA [9].

As a reference database of rRNA, we use all eight prepackaged databases that come with the SortMeRNA installation: six based on SILVA [12] (Bacteria 16S, Bacteria 23S, Archaea 16S, Archaea 23S, Eukaryota 18S, Eukaryota 28S) and two based on RFAM [13] (5S, 5.8S). The SortMeRNA command needs the *--ref* option to refer to the reference rRNA databases, with first the reference *fasta* file and then the corresponding index file (.idx) separated by a comma:

```
sortmerna --ref /path/to/sortmerna/rRNA_databases/silva-bac-16s-id90.fasta,/path/to /sortmerna/rRNA_databases/silva-bac-16s-id90.idx:/path/to /sortmerna/rRNA_databases/silva-bac-23s-id98.fasta, /path/to/sortmerna/rRNA_databases/silva-bac-23s-id98.idx: /path/to/sortmerna/rRNA_databases/silva-arc-16s-id95.fasta,/path/to/sortmerna/rRNA_databases/silva-arc-16s-id95.idx: /path/to/sortmerna/rRNA_databases/silva-arc-23s-id98.fasta, /path/to/sortmerna/rRNA_databases/silva-arc-23s-id98.idx: /path/to/sortmerna/rRNA_databases/silva-euk-18s-id95.fasta,/path/to/sortmerna/rRNA_databases/silva-euk-18s-id95.idx: /path/to/sortmerna/rRNA_databases/silva-euk-28s-id98.fasta,/path/to/sortmerna/rRNA_databases/silva-euk-28s-id98.idx: /path/to/sortmerna/rRNA_databases/rfam-5s-database-id98.fasta, /path/to/sortmerna/rRNA_databases/rfam-5s-database-id98.idx: /path/to/sortmerna/rRNA_databases/rfam-5.8s-database-id98.fasta,/path/to/sortmerna/rRNA_databases/rfam-5.8s-database-id98.idx --reads assembled.fastq --fastx --aligned 7_R1_trimmed_and_filtered_rRNA --other non_rRNA
```

Multiple reference databases (in the example below all 8) are separated by a colon. The input file (in this case the merged reads from Subheading 3.7.4) is specified with the option *--reads*. The *--fastx* flag indicates that both rRNA and non-rRNA sequences are outputted in *fasta* format. The output file names can be specified with *--aligned* for the rRNA sequences and with *--other* for the non-rRNA sequences. SortMeRNA will automatically add extensions to these output files.

3.7.6 Virus Detection Using VirusDetect

VirusDetect (http://bioinfo.bti.cornell.edu/cgi-bin/virusdetect/index.cgi) is specifically designed for analyzing small RNA datasets (*see* **Note 16**). The software aligns the small RNA reads to the known virus reference database for reference-guided assembly. In parallel, it uses Velvet for *de novo* assembly of small RNAs and subsequently compares the contigs to the reference database for virus identification. Although there is also an online version available, the following protocol describes the stand-alone version of VirusDetect.

The VirusDetect pipeline runs as follows:

```
perl /path/to/VirusDetect/virus_detect.pl <options> input.fastq.
```

The script takes as argument(s) the input file(s) either in *fastq* or *fasta* format. More than one input file can also be specified, which will then be processed sequentially.

The following options can be used:

--reference: name of the reference virus database, in this case "vrl_plant" (database containing viruses infecting plants) or "vrl_Invertebrates_217_U100" (downloaded database from viruses infecting invertebrates, can be used if interested in the presence of viruses harmful for pollinators residing on the pollen) (*see* **Note 17**).

--host-reference: name of the host reference database used for host sequence subtraction as described in Subheading 2.8.2.

--thread-num: number of processors used for alignments.

Additional parameters can be set for each step (BWA alignment, BLAST, filtering of results). These options are explained well on the VirusDetect website.

Here you can see an example command running on 8 cores: perl /path/to/VirusDetect/virus_detect.pl --reference vrl_plant --host-reference plant_genome.fasta --thread-num 8 trimmed_and_filtered.fastq.

The output is a detailed overview in HTML format showing a list of viral genomes detected in the data, based on Blastn and Blastx searches (*see* **Note 18**). The output shows how many contigs match the reference viral genomes and the coverage, depth, and mean percentage of identity. Each reference sequence can also be graphically viewed, showing the position of the contigs on the virus reference sequence, as well as the alignments.

Based on these results, an experienced plant virologist can make a diagnostic decision on the presence of one or more specific viruses (*see* **Note 19**). However, confirmation using a different diagnostic tool is recommended (*see* **Note 20**).

VirusDetect also delivers an output table containing undetermined contigs, potentially derived from previously unknown viruses (*see* **Note 21**). More details about the output format can be found on the VirusDetect website.

4 Notes

1. As an alternative to the described direct sample preparation, a protocol based on virus enrichment through partial virus purification was evaluated. A general procedure based on filtration, followed by ultracentrifugation, was done. However, subsequent RNA extraction on the purified viral particles from pollen resulted in yields and quality which were insufficient for NGS.
2. A parallel strategy to split the pollen samples into two subsamples, separating the outside contamination of the pollen material from the inside contamination by means of a washing step, is also not recommendable. The aim was to assess the

localization of bee and bumble bee pathogens on the pollen. The washing step resulted in wet and sticky pollen which was difficult to handle during the early steps of the RNA extraction process.

3. As an alternative to the RNeasy and mirVana kit extraction protocols, the TRIZOL nucleic acid extraction method was evaluated on the pollen samples. However, we were unable to retrieve RNA of sufficient purity and integrity (no 28S/18S bands obtained on the bleach gel) to pass the QC for NGS, even after several ethanol precipitation steps and RNA cleanup (Nucleospin® RNA cleanup XS kit).
4. The total RNA prepared with the RNeasy kit passed the QC and was sequenced following the same procedure as for the total RNA prepared with the mirVana kit. The obtained number of reads was higher than for the total RNA from the mirVana kit, and the amount of non-rRNA was comparable (±4%). However, no unique virus contig could be identified from those samples. This was the case for the pollen sample both from apple (presented in Table 1) and from pear, yielding a similar number of non-rRNA reads as for the mirVana kit. No clear explanation for this non-result can be given.
5. NGS sequencing starting from total RNA from the mirVana kit resulted in four times the amount of reads compared to when the small RNA protocol was used (same mirVana kit), as was requested by the sequence provider. The percentage of non-rRNA in the sample that was detected during the data analysis increased from less than 4% to approximately 85% when an rRNA depletion step was done (Table 1). Based on the diagnostic result (*see* **Note 6**), the small RNA protocol is recommended because the protocol is easier (no rRNA depletion step) and less sequencing output is required (hence having a lower cost).
6. The rRNA depletion step applied on the total RNA sample, prepared with the mirVana kit, had a positive effect on the number of unique virus contigs that were detected in the sample, and on the diagnostic result. No viruses were detected in the total RNA sample without this rRNA depletion step, whereas the same viroid was detected in all the protocol variants that were tested. The same four viruses were detected in the sample but more than tenfold of unique virus contigs were generated when started from small RNA, compared to the total RNA (+depletion) (Table 2). In conclusion, if the total RNA protocol is preferred, an rRNA depletion step has to be included.

Table 1
Comparison of the key parameters (RNA yield and purity, no. of obtained reads after sequencing, ribosomal RNA information, no. of unique virus contigs, and no. of viruses/viroids identified in a Blastn and Blastx search (*see* Note 10)) for an apple pollen sample

Matrix	Extraction	RNA depletion	RNA conc. (ng/μL)	260/280	260/230	No. of raw reads	No. of non rRNA reads	% Useful reads on total no. of raw reads	No. of unique virus contigs	Contig length (min. to max.)	No. of viruses identified in Blastn
Apple pollen sample	miRNA mirVana (Thermo Fisher)	No	180	2.37	2.17	6,326,669	4,247,384	67.1	543	41–719	4 + 1 viroid
	Total RNA mirVana (Thermo Fisher)	No	438.5	2.44	2.25	38,051,078	1,354,304	3.6	1	363	1 viroid
	Total RNA mirVana (Thermo Fisher)	Yes	438.5	2.44	2.25	25,166,843	21,521,691	85.5	42	107–1375	4 + 1 viroid

Table 2
Diagnostic result of the data analysis following NGS sequencing of the same apple pollen sample as in Table 1

Matrix	Extraction	RNA depletion	No. of identified viruses	ID of the viruses				
Apple pollen sample	miRNA mirVana (Thermo Fisher)	No	4 + 1 viroid	*Pear black necrotic leaf spot virus*	*Apple stem grooving virus*	*Citrus tatter leaf virus* (= ASGV strain)	*Citrus leaf blotch virus*	Apple hammerhead viroid-like circular RNA
	Total RNA mirVana (Thermo Fisher)	No	1 viroid	Apple hammerhead viroid-like circular RNA				
	Total RNA mirVana (Thermo Fisher)	Yes	4 + 1 viroid	*Raspberry bushy dwarf virus*	*Apple stem grooving virus*	*Citrus tatter leaf virus*	*Citrus leaf blotch virus*	Apple hammerhead viroid-like circular RNA

7. For total RNA concentration measurements, the use of the Quantus apparatus, using the QuantiFluor® RNA System kit (Promega), can be used as an alternative to the described Nanodrop method. Concentrations of small RNAs are measured using the Qubit® microRNA Assay Kit (Thermo Fisher Scientific). It is important to work with RNA of high purity, containing as little contaminants as possible. Nucleic acids have absorbance maxima at 260 nm. The ratio of this absorbance maximum to the absorbance at 280 nm has been used as a measure of purity for both DNA and RNA. For pure RNA, the 260/280 ratio should be around 2.0. An additional measure is the 260/230 absorbance ratio, also indicating the presence of organic contaminants, such as phenol, trizol, chaotropic salts, and other aromatic compounds. Samples with 260/230 ratios below 1.8 are considered to have a significant amount of these contaminants.

8. RIN is a tool that calculates the RNA integrity of eukaryotic RNA, based on the entire electrophoretic trace of the RNA sample. RIN values can be obtained using a Bioanalyzer. RIN is automatically assessed by running the RNA sample in an Agilent 2100 Bioanalyzer (Agilent Technologies) with RNA 6000 LabChip® (Caliper Technologies Corporation). The Bioanalyzer can also be used to assess purity and concentration of the RNA (as alternative for the described Nanodrop method).

9. As an alternative to the Bioanalyzer method (*see* **Note 8**), the RNA integrity can also be checked using denaturing agarose gel electrophoresis. Intact RNA will show clear and sharp 28S and 18S rRNA bands (with eukaryotic samples), the 28S rRNA band being approximately twice as intense as the 18S rRNA band (Thermo Fisher Scientific Technotes 8(3) and 11(1), available online: https://www.thermofisher.com/be/en/home/references/ambion-tech-support/rna-isolation/tech-notes.html). However, in this protocol, the "bleach gel method" is recommended over the denaturing type of electrophoresis, since it is time consuming and normally requires toxic reagents.

10. For library preparation, it is important to use a kit that synthesizes cDNA with random primers and not with oligo-dT primers, since viral RNAs usually do not have polyA-tails.

11. In case of an overall low sequencing quality, a quality filtering step can be introduced to remove low-quality reads after steps for small RNA datasets, or for total RNA datasets. Several programs are available to do quality filtering; for example, *fastq_quality_filter* of the FastX toolkit [14] (http://hannonlab.cshl.edu/fastx_toolkit/) allows you to

keep only the reads where a minimum percentage of the bases in the read has a certain quality value. In most cases, the overall quality of the reads is very high (because of the short read size for small RNA datasets), so it is not necessary to include a quality filtering step. For total RNA datasets, the quality tends to decrease toward the end of the read, and overall quality of the reverse read is usually worse than of the forward read.

12. PEAR can also perform quality trimming while merging the reads. You can specify a quality threshold using the option *-q*: if two consecutive bases are below this threshold, the read is trimmed. This option is usually combined with a minimum length that should be retained after trimming, specified by the option *-t*.

13. In the adapter removal step, we specify a minimum (16 bp) and maximum (28 bp) length for the resulting fragment in case of small RNAs. In case of the total RNA dataset, we do not specify a maximum length, since the length of the fragment is variable and depends on the insert size of the library.

14. In the merging step, no maximum length was set for the merged fragment, but if the overlap has to be minimum 20 bp, the resulting merged fragment cannot be longer than 280 bp in the case of 2 × 150 bp sequencing.

15. If the percentage of merged reads is relatively small (and hence you have relatively long inserts in the library), consider to also include the non-merged reads in the subsequent data analysis. In this case, be aware that SortMeRNA can only take one file as input file. Therefore you should first convert the two paired *fastq* files of the non-merged reads to one "interleaved" *fastq* file. SortMeRNA can then be run on the interleaved *fastq* file using the *--paired_in* option to indicate that the input file consists of paired reads. Alternatively, the forward and reverse reads can be analyzed separately in SortMeRNA. By doing this, the information regarding the pairs is lost, but since VirusDetect does not take paired files as input (*see* **Note 16**), this is acceptable.

16. The VirusDetect pipeline was specifically designed for small RNA datasets; hence it cannot handle paired read files. If decided to include the non-merged reads from the total RNA data, it is mandatory to generate a unique input file merging forward and reverse files using the Linux command cat. This means that the paired end information of the remaining read pairs is lost.

17. It is also possible to create own databases to be used as reference viral database (described in Subheading 2.8.1) and add these to the subfolder "databases" within the folder where VirusDetect is installed. These databases still need to be indexed for the mapping tool BWA (using bwa index) and for stand-alone BLAST (using formatdb). Both programs are

included in the VirusDetect install. *See* examples below on how to index the database:

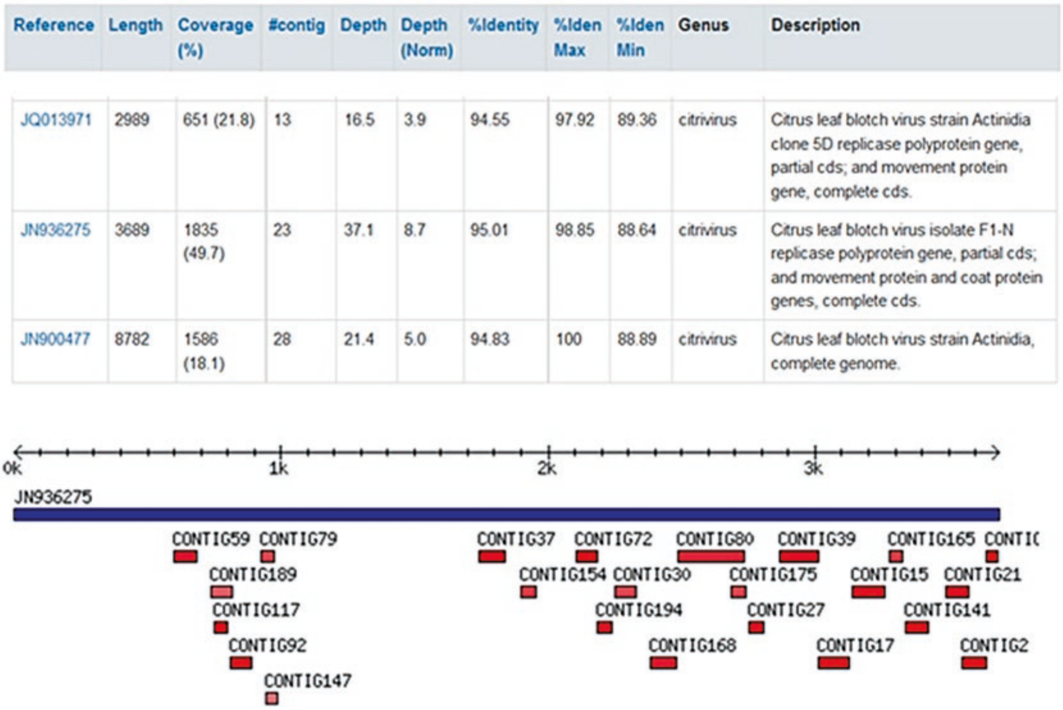
bwa index /path/to/VirusDetect/databases/my_own_database.fasta.

formatdb -i /path/to/VirusDetect/databases/my_own_database.fasta -p F

The *-p F* option tells the *formatdb* program that we are dealing with nucleotide data (protein = false).

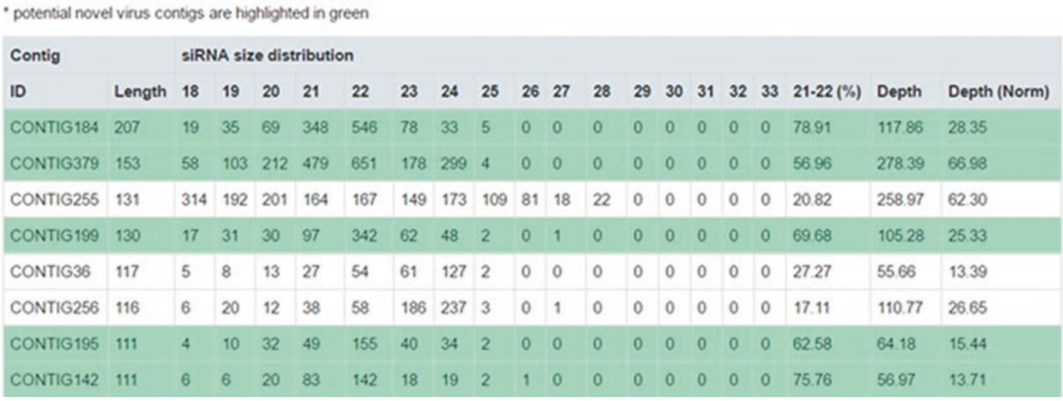
18. It is necessary to execute both Blastn and Blastx searches. However, the results of the pollen samples that were tested in the above-described protocol did not result in the identification of a novel virus, based on the Blastx results.

19. The diagnostic evaluation of the VirusDetect output is not standardized, and expert evaluation is still required. The number of contigs that map with specific reference genomes in GenBank®, their distribution along the genomes, as well as the percentage of identity (the darker red the color the better) of the contig with the corresponding area of the reference isolate are an indication for the presence and identification of a specific virus in the sample (Fig. 5).



Reference	Length	Coverage (%)	#contig	Depth	Depth (Norm)	%Identity	%Iden Max	%Iden Min	Genus	Description
JQ013971	2989	651 (21.8)	13	16.5	3.9	94.55	97.92	89.36	citrivirus	Citrus leaf blotch virus strain Actinidia clone 5D replicase polyprotein gene, partial cds; and movement protein gene, complete cds.
JN936275	3689	1835 (49.7)	23	37.1	8.7	95.01	98.85	88.64	citrivirus	Citrus leaf blotch virus isolate F1-N replicase polyprotein gene, partial cds; and movement protein and coat protein genes, complete cds.
JN900477	8782	1586 (18.1)	28	21.4	5.0	94.83	100	88.89	citrivirus	Citrus leaf blotch virus strain Actinidia, complete genome.

Fig. 5 Example of the specific VirusDetect output, identifying one virus (*Citrus leaf blotch virus*; CLBV) in the apple pollen sample. Three hits with reference genomes, the number of respective contigs, and the contig coverage on genome JN936275 are presented



* potential novel virus contigs are highlighted in green

Contig		siRNA size distribution																		
ID	Length	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	21-22 (%)	Depth	Depth (Norm)
CONTIG184	207	19	35	69	348	546	78	33	5	0	0	0	0	0	0	0	0	78.91	117.86	28.35
CONTIG379	153	58	103	212	479	651	178	299	4	0	0	0	0	0	0	0	0	56.96	278.39	66.98
CONTIG255	131	314	192	201	164	167	149	173	109	81	18	22	0	0	0	0	0	20.82	258.97	62.30
CONTIG199	130	17	31	30	97	342	62	48	2	0	1	0	0	0	0	0	0	69.68	105.28	25.33
CONTIG36	117	5	8	13	27	54	61	127	2	0	0	0	0	0	0	0	0	27.27	55.66	13.39
CONTIG256	116	6	20	12	38	58	186	237	3	0	1	0	0	0	0	0	0	17.11	110.77	26.65
CONTIG195	111	4	10	32	49	155	40	34	2	0	0	0	0	0	0	0	0	62.58	64.18	15.44
CONTIG142	111	6	6	20	83	142	18	19	2	1	0	0	0	0	0	0	0	75.76	56.97	13.71

Fig. 6 Example of a table output with undetermined contigs from the VirusDetect software. The contigs containing a high 21–22 nt fraction potentially indicating the presence of an undescribed virus are highlighted in green

20. It is necessary to confirm the viruses that were identified using the available specific PCR tests and if applicable also virus indexing.
21. The VirusDetect NGS data analysis software also results in a table of undetermined contigs (Fig. 6). These might indicate the presence of new viruses. The table highlights a number of contigs with high (>50%) 21–22 nt fraction of the reads within the small RNA size distribution. To identify potential new viruses, these contigs can be mapped directly against reference genomes of related viruses/virus groups, and/or a read extension protocol can be followed. Based on sequence hot spots, primers can be developed to amplify missing parts of the genome, followed by a biological study.

References

1. Massart S, Olmos O, Jijakli H, Candresse T (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188:90–96. https://doi.org/10.1016/j.virusres.2014.03.029
2. Mumford R, Boonham N, Tomlinson J, Barker I (2006) Advances in molecular phytodiagnostics—new solutions for old problems. Eur J Plant Pathol 116:1–19. https://doi.org/10.1007/s10658-006-9037-0
3. Posthuma-Trumpie GA, Korf J, van Amerongen A (2009) Lateral flow (immuno)assay: its strengths, weaknesses, opportunities and threats. A literature survey. Anal Bioanal Chem 393:569–582. https://doi.org/10.1007/s00216-008-2287-2
4. Boonham N, Kreuze J, Winter S, van der Vlugt R, Bergervoet J, Tomlinson J, Mumford R (2014) Methods in virus diagnostics: from ELISA to next generation sequencing. Virus Res 186:20–31. https://doi.org/10.1016/j.virusres.2013.12.007
5. Andrews S (2015) Babraham bioinformatics—FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
6. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200
7. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864. https://doi.org/10.1093/bioinformatics/btr026
8. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate illumina

paired-end reAd mergeR. Bioinformatics 30:614–620. https://doi.org/10.1093/bioinformatics/btt593

9. Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24):3211–3217. https://doi.org/10.1093/bioinformatics/bts611
10. Zheng Y, Gao S, Padmanabhan C, Li R, Galvez M, Gutierrez D, Fuentes S, Ling KS, Kreuze J, Fei Z (2017) VirusDetect: an automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology 500:130–138. https://doi.org/10.1016/j.virol.2016.10.017
11. Aranda PS, Lajoie DM, Jorcyk CL (2012) Bleach gel: a simple agarose gel for analyzing RNA quality. Electrophoresis 33:366–369. https://doi.org/10.1002/elps.201100335
12. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(D1):D590–D596. https://doi.org/10.1093/nar/gks1219
13. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD (2014) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43(D1):D130–D137. https://doi.org/10.1093/nar/gku1063
14. Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished. http://hannonlab.cshl.edu/fastx_toolkit

Chapter 11

High-Resolution Screening of Viral Communities and Identification of New Pathogens in Fish Using Next-Generation Sequencing

Arnfinn Lodden Økland, Are Nylund, Ali May, Adalberto Costessi, and Walter Pirovano

Abstract

Discovery of viral genomes in fish has historically been based on viral enrichment, random priming, cloning, and Sanger sequencing. However, the development of next-generation sequencing has enabled the possibility to sequence the entire virome of a tissue sample. This has led to an enormous increase in discovery of new viruses. In this chapter, we describe a simple and rapid method for viral discovery in fish. The method is based on Illumina sequencing of total RNA from diseased tissue or cell culture and in silico removal of host RNA.

Key words Fish, Heart and skeletal muscle inflammation, Atlantic salmon, Piscine orthoreovirus, Total RNA sequencing

1 Introduction

Next-generation sequencing has the capacity to produce millions of sequence reads from one sample without the need of any prior knowledge of its genomic composition. Former methods for viral metagenomics usually included random priming, cloning, and Sanger sequencing [1, 2]. Such methods would frequently rely on viral enrichment, using techniques as centrifugation, filtering, gradient centrifugation, and nuclease treatment [3–5]. A similar approach was applied to identify Piscine myocarditis virus, a causative agent of myocarditis in Atlantic salmon (*Salmo salar* L.) [6]. A more recent study utilized a representational difference analysis (RDA) to enrich for viral RNA, followed by cloning of resultant product and Sanger sequencing to identify Atlantic salmon calicivirus [7]. The use of next-generation sequencing to identify viruses in fish has mainly been based on isolation of the virus in a cell culture, centrifugation, and Illumina sequencing of resultant RNA/DNA [8–14]. However, next-generation sequencing of total RNA

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_11, © Springer Science+Business Media, LLC 2018

from host tissue has proved to be successful for identification of viruses that have failed to be isolated in a cell culture. This has been shown for both fish [15] and other organisms [16–18]. The main advantage of using tissue for next-generation sequencing is its ability to sequence the complete virome and microbiome, enabling a greater identification of novel viruses and microorganisms.

Here, we demonstrate a simple and rapid method for identification of new pathogens using next-generation sequencing. This method requires no techniques for enrichment of virus. It is essentially based on next-generation sequencing of total RNA extracted from diseased tissue or cell culture, and in silico removal of the host genome. Using this method we have obtained the near-complete genomes of both RNA and DNA viruses from several fish species (unpublished) and one crustacean [16], but in this chapter we present results from sequencing a known virus, Piscine orthoreovirus (PRV). PRV is a double-stranded RNA virus with a genome consisting of ten segments. The genome consists of three large segments, L1, L2, and L3; three medium segments, M1, M2, and M3; and four small segments, S1, S2, S3, and S4, encoding the proteins λ3, λ2/p11, λ1, μ2, μ1, μNS, σ3/p13, σ 2/p8, σNS, and σ1, respectively. The virus is believed to cause heart and skeletal muscle inflammation (HSMI) in Atlantic salmon (*S. salar*) [8, 19].

2 Materials

2.1 RNA Extraction

1. TRI Reagent® (SIGMA).
2. Chloroform.
3. Isopropanol.
4. Ethanol.
5. Nuclease-free water.
6. Homogenizer.
7. Refrigerated centrifuge with minimum 12,000 × *g*.

2.2 Illumina Library Preparation

1. BioAnalyzer (Agilent).
2. TruSeq RNA library preparation kit (Illumina).
3. 1.7 mL Low-retention microfuge tubes.
4. 0.3 mL PCR tubes.
5. Nuclease-free water.
6. Ethanol.
7. Vortex.
8. Thermomixer.
9. Magnetic stand.

10. SuperScript II Reverse Transcriptase (Life Technologies).
11. AMPure XP Beads (Beckman).
12. Gel electrophoresis equipment.
13. MS-8 Agarose (spearoQ).
14. 1× TAE buffer.
15. Scalpels.
16. Gel DNA recovery kit (Zymo Research).
17. Thermocycler.
18. Applied Biosystems 7500 Real-Time PCR System.
19. KAPA qPCR quantification kit (KAPA Biosystems).

2.3 Illumina Sequencing

1. cBot clustering equipment (Illumina).
2. HiSeq 2500 sequencer (Illumina).
3. TruSeq PE Cluster Kit v3-cBot-HS (Illumina).
4. TruSeq SBS v3-HS sequencing reagents (Illumina).
5. PhiX control library V3 (Illumina).
6. 1 M NaOH.
7. Nuclease-free water.
8. Vortex.
9. Tabletop centrifuge.

2.4 Data Processing and Bioinformatics Analysis

1. Compute server of cloud platform (e.g., Microsoft Azure or Amazon AWS) with at least 48 GB of RAM and sufficient disk space.
2. Bioinformatics analysis software for Illumina data processing (bcl2fastq).
3. Software for sequence trimming, alignment, and assembly (CLC Genomics Workbench).
4. Software for prediction of Open Reading Frames (Prodigal).
5. BLAST analysis software (NCBI, either local or online).

3 Methods

3.1 RNA Extraction

1. Add 1 mL TRI Reagent® (SIGMA) to 50–100 mg tissue.
2. Homogenize the sample using an appropriate homogenizer.
3. Add 0.2 mL chloroform and vortex or shake the sample for 15 s.
4. Incubate at room temperature for 5 min.
5. Centrifuge at 12,000 × *g* at 4 °C for 15 min.
6. Transfer 350–450 μL of the aqueous phase to a new tube containing 0.5 mL isopropanol.

7. Vortex or shake sample for 15 s and incubate at room temperature for 10 min.
8. Centrifuge at 12,000 × *g* at 4 °C for 15 min.
9. Remove supernatant and wash the pellet using 1 mL 70% ethanol.
10. Centrifuge at 9500 × *g* at 4 °C for 5 min.
11. Remove the supernatant and wash the pellet using 1 mL 100% ethanol.
12. Centrifuge at 9500 × *g* at 4 °C for 5 min.
13. Remove the supernatant and let the RNA pellet air-dry for 10 min.
14. Dissolve the RNA pellet in 30–100 μL nuclease-free water heated to 70 °C.
15. Store at −80 °C.

3.2 Illumina Library Preparation

1. Total RNA quality and concentration were measured on a Bioanalyzer 2100 (Agilent)
2. 100 ng of total RNA was used for library preparation using the Illumina TruSeq RNA library preparation kit (Illumina).
3. The total RNA was fragmented and subjected to first-strand cDNA synthesis with random hexamers and second-strand synthesis.
4. Barcoded DNA adapters were ligated to both ends of the double-stranded cDNA.
5. The ligated product was size-selected on agarose gel, whereby the region between 200 and 320 bp was retrieved.
6. PCR amplification was performed for 15 cycles.
7. The resultant sequencing libraries were checked on a Bioanalyzer (Agilent) and quantified by qPCR.

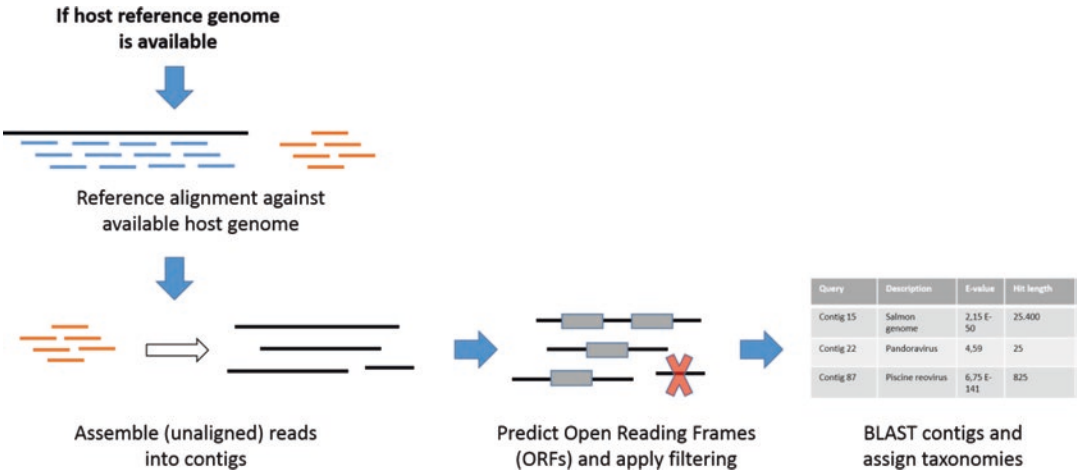
3.3 Illumina Sequencing

1. Each library was clustered on one full lane of a V3 HiSeq flowcell using the cBot (Illumina).
2. Sequencing was performed on the Illumina HiSeq 2500 using a paired-end 50-cycle protocol (*see* **Note 1**).
3. The sequencing run was analyzed with the Illumina bcl2fastq CASAVA pipeline.

3.4 Data Processing and Bioinformatics Analysis

An overview of the bioinformatics analysis procedure is provided in Fig. 1.

1. Enhance the quality of the sequences by trimming off low-quality bases using “Trim sequences” option of the CLC Genomics Workbench.


If host reference genome is available
Reference alignment against available host genome
Assemble (unaligned) reads into contigs
Predict Open Reading Frames (ORFs) and apply filtering
BLAST contigs and assign taxonomies

Query	Description	E-value	Hit length
Contig 15	Salmon genome	2,15 E-50	25.400
Contig 22	Pandoravirus	4,59	25
Contig 87	Piscine reovirus	6,75 E-141	825

Fig. 1 Overview of the bioinformatics viral screening pipeline

2. Align quality-filtered sequence reads to host genome, if available, using the "Map reads to reference" option of the CLC Genomics Workbench (or an alternative alignment tool such as Bowtie [20] or BWA [21]) (*see* **Note 2**).
3. Assemble remaining reads using the "De novo assembly" option of the CLC Genomics Workbench (or an alternative De Bruijn graph-based assembly tool such as Velvet [21] or Spades [21]).
4. Remove all contig sequences below 120 bases.
5. Predict open reading frames (ORFs) using Prodigal software [21] and remove sequences with no ORF (a contig is considered not to have an ORF if it contains less than 40 AA).
6. Remove contigs below 200 bases that contain a stop codon in all six reading frames.
7. Reduce the amount of remaining host sequences by using BLAST+ protocols [21] to remove all sequences that match with the host's phylum and applying a E-value threshold of 0.01 (*see* **Note 2**).

3.5 Results

To illustrate the effectiveness and simplicity of this method we have included the results from an experiment with fish infected with PRV. The material used for this experiment was four samples of total RNA isolated from heart and kidney from five Atlantic salmon. The tissue samples had earlier been used for HSMI transmission trials and were 8 years old. All samples were positive for PRV, with Ct values ranging from 17 to 27. The Illumina sequencing generated 17–27 Gb of data. The lowest yield was obtained for samples 2 and 4, which had the poorest RIN values (*see* **Note 3**).

Table 1
Total yield and aligned reads to Atlantic salmon genome

Sample		Count	Number of bases	Percentage of reads (%)
Sample 1	Total reads	545,169,514	27,630,919,330	100
	Mapped reads	531,314,685	26,928,615,512	97.46
Sample 2	Total reads	393,866,836	20,019,650,896	100
	Mapped reads	389,278,529	19,786,886,435	98.84
Sample 3	Total reads	540,482,438	27,387,712,676	100
	Mapped reads	520,552,234	26,377,196,978	96.31
Sample 4	Total reads	337,905,152	17,196,493,710	100
	Mapped reads	331,095,493	16,850,345,471	97.98

1. Read trimming based on Phred quality scores resulted in clipping of 0.3 bases on average.
2. Alignment of the reads against the Atlantic salmon genome removed approximately 97% of the total reads (Table 1).
3. Assembly of the unmapped reads led to between 70,154 and 192,959 contigs (146,275 on average).
4. Filtering of small sequences (<120 bases) reduced the number of contigs to 31,608 (by ~78%).
5. Filtering of sequences without an ORF reduced the number of contigs to 5427 (by ~83%).
6. Filtering of sequences with stop codons in all six reading frames reduced the number of contigs to 4227 (by ~22%).
7. Through the BLAST+ approach, the number of contigs was reduced to 93–761 sequences per sample. Sequences from all of the ten PRV segments were obtained, and a total of 105 sequences covered 56.7% of the PRV genome (Table 2). The highest number of PRV sequences was obtained from samples 1 and 2. These samples showed the lowest Ct values during preliminary screening. An overview of the alignment of the Illumina-generated PRV contigs to the segments of the PRV reference genome is presented in Fig. 2.

4 Notes

1. Increasing the number of cycles will lengthen the reads. This may facilitate assembly and create longer contigs. Our experience is that using 100 cycles increases the chances of assembling near-complete viral genomes.

Table 2
Coverage of the PRV genome

Segment	Bases	Covered bases	Percent covered (%)
L1	3911	2524	64.5
L2	3935	2838	72.1
L3	3916	974	24.9
M2	2179	1346	61.8
M1	2383	798	33.5
M3	2403	1465	61.0
S2	1329	1045	78.6
S3	1143	849	74.3
S4	1040	567	54.5
S1	1081	806	74.6
SUM	23,320	13,212	56.7

2. Make sure to keep all reads and sequences that are filtered as host genome. Some viral sequences may have been incorrectly annotated as host genome.
3. The recommended RIN value for Illumina sequencing is >8, and using samples with values below this may lower the yield. However, treatment of total RNA with nucleases is also used as a common method for enrichment of viral RNA as viral genome will be protected from RNA degradation within its virion [22]. This was evident in sample 2 which had a RIN value of 2.9. Forty-eight percent of the sequences that remained after the pipeline were from PRV.

Acknowledgments

Research reported in this publication was partially supported by PHARMAQ and PHARMAQ Analytiq, in particular, the DNA sequencing work. The authors are also grateful for their precious advice regarding the bioinformatics pipeline.

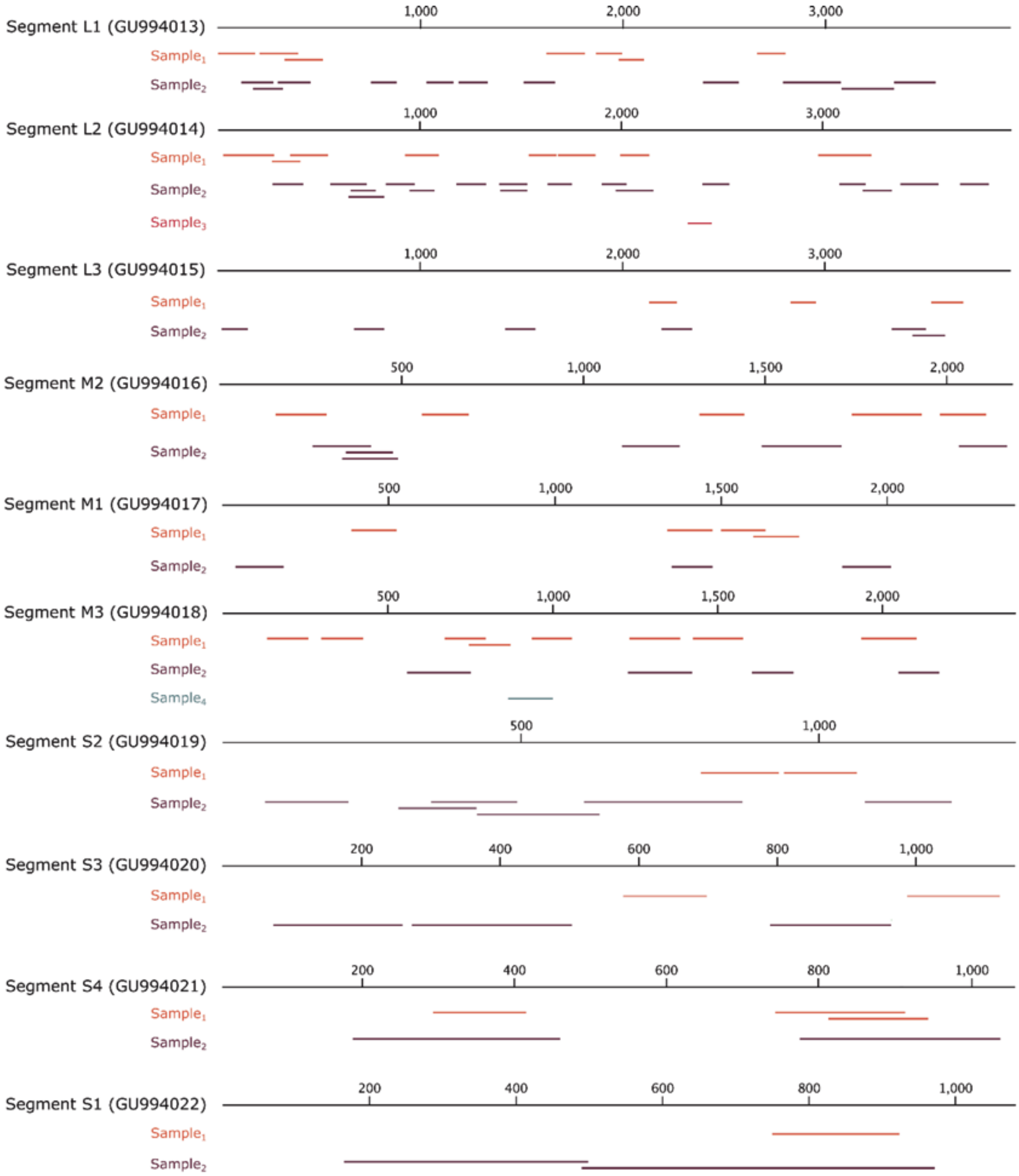


Fig. 2 Schematic representation of the Illumina-generated PRV contigs' coverage of the ten segments of Piscine reovirus (GenBank accession numbers GU994013-GU994022)

References

1. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185(20):6220–6223
2. Djikeng A, Halpin R, Kuzmickas R, DePasse J, Feldblyum J, Sengamalay N, Afonso C, Zhang X, Anderson NG, Ghedin E, Spiro DJ (2008) Viral genome sequencing by random priming methods. BMC Genomics 9(1):1–9

3. Kapoor A, Victoria J, Simmonds P, Wang C, Shafer RW, Nims R, Nielsen O, Delwart E (2008) A highly divergent picornavirus in a marine mammal. J Virol 82(1):311–320. https://doi.org/10.1128/JVI.01240-07
4. Ng TFF, Manire C, Borrowman K, Langer T, Ehrhart L, Breitbart M (2009) Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. J Virol 83(6):2500–2509
5. van Leeuwen M, Williams MMW, Koraka P, Simon JH, Smits SL, Osterhaus ADME (2010) Human picobirnaviruses identified by molecular screening of diarrhea samples. J Clin Microbiol 48(5):1787–1794
6. Haugland O, Mikalsen AB, Nilsen P, Lindmo K, Thu BJ, Eliassen TM, Roos N, Rode M, Evensen O (2011) Cardiomyopathy syndrome of Atlantic salmon (Salmo salar L.) is caused by a double-stranded RNA virus of the totiviridae family. J Virol 85(11):5275–5286. https://doi.org/10.1128/jvi.02154-10
7. Mikalsen AB, Nilsen P, Frøystad-Saugen M, Lindmo K, Eliassen TM, Rode M, Evensen Ø (2014) Characterization of a novel calicivirus causing systemic infection in Atlantic salmon (*Salmo salar*): proposal for a new genus of *Caliciviridae*. PLoS One 9(9):e107132
8. Palacios G, Lovoll M, Tengs T, Hornig M, Hutchison S, Hui J, Kongtorp RT, Savji N, Bussetti AV, Solovyov A, Kristoffersen AB, Celone C, Street C, Trifonov V, Hirschberg DL, Rabadan R, Egholm M, Rimstad E, Lipkin WI (2010) Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus. PLoS One 5(7):e11487
9. Mor SK, Phelps NBD (2016) Molecular detection of a novel totivirus from golden shiner (Notemigonus crysoleucas) baitfish in the USA. Arch Virol 161(8):2227–2234
10. Bacharach E, Mishra N, Briese T, Zody MC, Kembou Tsofack JE, Zamostiano R, Berkowitz A, Ng J, Nitido A, Corvelo A, Toussaint NC, Abel Nielsen SC, Hornig M, Del Pozo J, Bloom T, Ferguson H, Eldar A, Lipkin WI (2016) Characterization of a novel orthomyxo-like virus causing mass die-offs of tilapia. mBio 7(2):e00431
11. Axén C, Hakhverdyan M, Boutrup T, Blomkvist E, Ljunghager F, Alfjorden A, Hagström Å, Olesen N, Juremalm M, Leijon M (2017) Emergence of a new rhabdovirus associated with mass mortalities in eelpout (Zoarces viviparous) in the Baltic Sea. J Fish Dis 40(2):219–229
12. Lange J, Groth M, Fichtner D, Granzow H, Keller B, Walther M, Platzer M, Sauerbrei A, Zell R (2014) Virus isolate from carp: genetic characterization reveals a novel picornavirus with two aphthovirus 2A-like sequences. J Gen Virol 95(1):80–90
13. Naoi Y, Okazaki S, Katayama Y, Omatsu T, Ono S-i, Mizutani T (2015) Complete genome sequences of two Japanese eel endothelial cell-infecting virus strains isolated in Japan. Genome Announc 3(6):e01236–15
14. Subramaniam K, Toffan A, Cappellozza E, Steckler NK, Olesen NJ, Ariel E, Waltzek TB (2016) Genomic sequence of a ranavirus isolated from short-finned eel (Anguilla australis). Genome Announc 4(4):e00843–16
15. Gjessing MC, Yutin N, Tengs T, Senkevich T, Koonin E, Rønning HP, Alarcon M, Ylving S, Lie K-I, Saure B, Tran L, Moss B, Dale OB (2015) Salmon gill poxvirus, the deepest representative of the Chordopoxvirinae. J Virol 89:9348–9367
16. Økland AL, Nylund A, Øvergård A-C, Blindheim S, Watanabe K, Grotmol S, Arnesen C-E, Plarre H (2014) Genomic characterization and phylogenetic position of two new species in *Rhabdoviridae* infecting the parasitic copepod, Salmon louse (*Lepeophtheirus salmonis*). PLoS One 9(11):e112517. https://doi.org/10.1371/journal.pone.0112517
17. Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Chen L-J, Qin X-C, Xu J, Holmes EC, Zhang Y-Z (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife 4:e05378. https://doi.org/10.7554/eLife.05378
18. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, Qin X-C, Li J, Cao J-P, Eden J-S, Buchmann J, Wang W, Xu J, Holmes EC, Zhang Y-Z (2016) Redefining the invertebrate RNA virosphere. Nature 540(7634):539–543
19. Markussen T, Dahle MK, Tengs T, Løvoll M, Finstad ØW, Wiik-Nielsen CR, Grove S, Lauksund S, Robertsen B, Rimstad E (2013) Sequence analysis of the genome of piscine orthoreovirus (PRV) associated with heart and skeletal muscle inflammation (HSMI) in Atlantic salmon (Salmo salar). PLoS One 8(7):e70075
20. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
22. Hall RJ, Wang J, Todd AK, Bissielo AB, Yen S, Strydom H, Moore NE, Ren X, Huang QS, Carter PE, Peacey M (2014) Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. J Virol Methods 195(0):194–204

Chapter 12

Metagenomic Analyses of the Viruses Detected in Mycorrhizal Fungi and Their Host Orchid

Hanako Shimura, Chikara Masuta, and Yasunori Koda

Abstract

In nature, mycorrhizal association with soilborne fungi is indispensable for orchid families. Fungal structures from compatible endo-mycorrhizal fungi in orchid cells are digested in cells to be supplied to orchids as nutrition. Because orchid seeds lack the reserves for germination, they keep receiving nutrition through mycorrhizal formation from seed germination until shoots develop (leaves) and become photoautotrophic. Seeds of all orchid species surely geminate with the help of their own fungal partners, and this specific partnership has been acquired for a long evolutionary history between orchids and fungi.

We have studied the interactions between orchids and mycorrhizal fungi and recently conducted transcriptome analyses (RNAseq) by a next-generation sequencing (NGS) approach. It is possible that orchid RNA isolated form naturally grown plants is contaminated with RNAs derived from mycorrhizal fungi in the orchid cells. To avoid such contamination, we here prepared aseptically germinated orchid plants (i.e., fungus-free plants) together with a pure-cultured fungal isolate and field-growing orchid samples. In the cDNA library prepared from orchid and fungal tissues, we found that partitivirus-like sequences were common in an orchid and its mycorrhizal fungus. These partitivirus-like sequences were closely related by a phylogenetic analysis, suggesting that transmission of an ancestor virus between the two organisms occurred through the specific relation of the orchid and its associated fungus.

Key words In vitro tissue culture, Mycorrhizal fungi, Mycovirus, Orchid, Partitivirus

1 Introduction

Viruses are found throughout the major taxonomic group of fungi. Fungal viruses (mycoviruses) have either double-stranded RNA (dsRNA) or single-stranded RNA (ssRNA) genomes, and are classified into ten families [1]. More recently, a circular ssDNA virus infecting fungi has been reported in the plant pathogenic fungus *Sclerotinia sclerotiorum* [2]. Mycoviruses in pathogenic fungi sometimes cause phenotypic alternations such as debilitation, and the mycovirus-mediated attenuation of fugal virulence is widely known as hypovirulence [3]. Some pathogenic fungi, such as chestnuts blight *Cryphonectria parasitica* and white root rot *Rosellinia necatrix*, have been well studied on the hypovirulence

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_12, © Springer Science+Business Media, LLC 2018

effects by virus infection [4–6]. The hypovirulence of fungi by mycoviruses suggests that viruses change the various traits of fungi and further affect the interaction between fungi and the fungus-infected plants. On the other hand, mycoviruses infecting non-pathogenic fungi such as endophytic fungi and mycorrhizal fungi, which have a mutualistic relationship with plant, have been little studied. Arbuscular mycorrhizal (AM) fungi that belong to the phylum *Glomeromycota* are ubiquitous in terrestrial ecosystems and form mutualistic associations with most land plants [7]. Previously, our research group found that there were four distinct dsRNA viruses in an AM fungus and one of the dsRNA was a biologically active component in the symbiosis [8]. Now, we are studying mycoviruses in the mycorrhizal fungi that associate with the largest plant family, *Orchidaceae*. All orchids need compatible mycorrhizal fungi for seed germination because tiny seeds lack the reserves for germination. The orchid–mycorrhizal fungi interaction is unique in that fungal mycorrhizal structures are digested in orchid cells, and then carbon nutrition derived from fungi is transferred to orchid although beneficial aspect for fungi is not well understood; in a general mycorrhizal interaction like AM interaction, plants give carbohydrates (photosynthesis products) to fungi and take instead water/mineral nutrients from fungi.

To know whether any viruses can infect orchid mycorrhizal fungi, we isolated dsRNA and analyzed the sequences using several fungal isolates that were purified from orchid roots. We found that several mycoviruses were in the mycorrhizal fungi, and that one of the mycoviruses belonged to the family *Partitiviridae* (our unpublished data), which has been shown to infect plants, fungi, and protozoa. In addition, we conducted metagenomic analyses of the viruses in orchid tissues and found that a mycovirus in a compatible mycorrhizal fungus (i.e., capable to induce orchid seed germination) was located in the same clade with the partitivirus-like sequences derived from orchid tissues by a phylogenetic analysis. Horizontal virus transfer of partitivirus beyond kingdom has been previously suggested [9, 10]; however, there is little report on viral transmission between plants and other hosts because possible contamination in the samples is always problematic. To examine our hypothesis that a mycovirus in the orchid mycorrhizal fungi would have the same evolutionary origin with a plant virus in orchids, it is absolutely important to avoid contamination from the samples that we used to isolate viruses. We here isolated RNAs from the orchid plants generated from in vitro tissue culture because the cultured orchid should not have viruses derived from fungus any more, and thus we can analyze the host plant RNAs without contamination of even a trace amount of fungal RNAs.

2 Material

2.1 In Vitro Tissue Culture of Orchid Plants

1. Orchid seed: We use *Cypripedium macranthos* var. *rebunense*, which was derived from a wild population in Rebun Island, Hokkaido, Japan. Mature capsules of *C. macranthos* var. *rebunense* were collected in September (about 90 days after pollination) at Rebun Island with the permission of the Ministry of Environment, Japan.
2. Culture medium for orchid: We used modified T-medium [11]; 545 mg/L KH_2PO_4, 470 mg/L $Ca(NO_3)_2 \cdot 4H_2O$, 245 mg/L $MgSO_4 \cdot 7H_2O$, 200 mg/L KNO_3, 120 mg/L NH_4NO_3, 37.3 mg/L Na_2-EDTA, 27.8 mg/L $FeSO_4 \cdot 7H_2O$, 3 mg/L $MnSO_4 \cdot 4H_2O$, 0.5 mg/L $ZnSO_4 \cdot 7H_2O$, 0.025 mg/L $CuSO_4 \cdot 5H_2O$, 0.025 mg/L $Na_2MoO_4 \cdot 2H_2O$, 0.5 mg/L H_3BO_3, 0.025 mg/L $CoCl_2 \cdot 6H_2O$, 0.025 mg/L KI, 400 mg/L yeast extract, 20 g/L sucrose, 1 μM 6-benzylaminopurine (BA), 6 g/L agar. Adjust pH to 5.5 with KOH. Autoclave for 7 min.
3. 10% Sodium hypochlorite.
4. Tween-20.
5. Sterilized water.
6. Bag (30 × 30 mm) made from nylon mesh (50 μm opening).
7. Scissors (sterilized).
8. Tweezers (sterilized).
9. 9 cm Petri dish (sterilized).

2.2 In Vitro Culture of Mycorrhizal Fungi

1. Orchid mycorrhizal fungi: Those were isolated from roots of *C. macranthos* var. *rebunense* or other orchid species. Some isolates (e.g., WO97 isolate) were confirmed to induce symbiotic germination of *C. macranthos* var. *rebunense* [12, 13]. All isolates are kept on OMA1, 2 g/L fine oatmeal powder (40 mesh), and 15 g/L agar, in the dark at 20 °C.
2. Fungal culture medium for RNA preparation: Oatmeal broth medium; 2 g/L fine oatmeal powder without agar.
3. 9 cm Petri dish (sterilized).

2.3 Preparation of Plant and Fungal RNAs

1. Nucleic acid extraction buffer: 25 mM Tris–HCl (pH 7.5), 25 mM $MgCl_2$, 25 mL KCl, 1% sodium dodecyl sulfate (SDS).
2. DsRNA extraction buffer: 100 mM Tris–HCl (pH 8.0), 200 mM NaCl, 2 mM EDTA, 1% SDS, 0.1% 2-mercaptoethanol. 2-Mercaptoethanol is added to buffer just before use.
3. TE-saturated phenol.
4. TE-saturated phenol-chloroform (w/w = 1:1).

5. Chloroform.
6. 3 M Sodium acetate (pH 5.2).
7. 99.5% Ethanol.
8. 70% Ethanol.
9. DNase I: We use Ambion TURBO DNase (2 U/μL).
10. S1 Nuclease: We use Takara S1 Nuclease (100–200 U/μL).

2.4 Sequence Analysis and Phylogeny

1. DNADynamo (Blue Tractor Software Ltd).
2. Clustal W program.
3. BLAST+ (version 2.2.29+).

   Basic Local Alignment Search Tool (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).
4. Amino acid databases:

   DDBJ viral (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plant/plant.*.protein.faa.gz)

   NCBI RefSeq plant (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dad/ddbjvrl1.DAD.fasta.gz).
5. MEGA 5 software.
6. Microsoft Excel.

3 Methods

In a natural field, there are numerous soilborne fungi including mycorrhizal fungi and other microorganisms in the soil where orchids grow. When we collect wild orchid samples and isolate RNAs, exhaustive sequence analyses using NGS would detect any RNAs other than orchid derived. It is known that orchid seeds can germinate in vitro without a compatible mycorrhizal fungus if they are cultured in a suitable nutritious medium. We previously developed in vitro propagation for *Cypripedium macranthos* var. *rebunense*, which is a Japanese wild orchid, and maintained several mycorrhizal fungi isolated from *C. macranthos* and other orchids by pure culture [12, 14]. To detect any viruses from an orchid and mycorrhizal fungi, we isolated orchid RNAs derived from both wild and in vitro-cultured plants in addition to cultured fungal mycelia. This approach is indeed useful to avoid contamination between orchid and fungi, and would give informative results on the viruses in each of orchid and fungi.

3.1 In Vitro Tissue Culture of Orchid Plants

1. Dissect mature capsule (Fig. 1a) longitudinally and scrape off seeds from capsule (Fig. 1b). Transfer seeds (ca. 5 mg) into a bag of nylon mesh. Sterilize the nylon mesh bags with 10% sodium hypochlorite supplemented with a few drops of

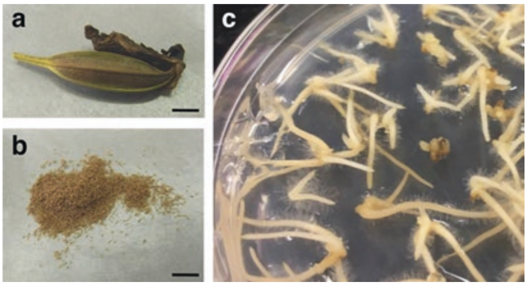


Fig. 1 Orchid seedlings propagated by in vitro tissue culture. (**a**) A mature capsule of *Cypripedium macranthos* var*. rebunense*. (**b**) Mature seeds pulled out from a capsule. (**c**) Aseptically cultured seedlings of *Cypripedium macranthos* var*. rebunense* in 9 cm Petri dish. Bars in (**a**) and (**b**) are 1 cm

Tween-20 for 30 min, and then wash thoroughly with sterilized water.

2. Nylon mesh bags are transferred on sterilized Petri dish. Open the bag by scissors and tweezers. Sow seeds on 20 ml culture medium in 9 cm Petri dish.
3. Keep the plates at 4 °C in the dark to break seed dormancy for 3 months, and then incubate at 20 °C in the dark.
4. After 2–3-month incubation at 20 °C, seeds usually germinate and grow into protocorm (protocorm is an orchid-specific, young seedling form before organ development). When seedlings have roots and grow to a large size (Fig. 1c), transfer seedlings to the same fresh medium in 100 mL conical flask or 300 mL culture bottle (*see* **Note 1**).

3.2 In Vitro Culture of Mycorrhizal Fungi

1. To prepare an inoculum, incubate and maintain fungi isolated from roots of *C. macranthos* var. *rebunense* (Fig. 2a, b) or other orchid roots in OMA1. For RNA extraction, we used WO97 isolate (Fig. 2c), which is a compatible mycorrhizal fungus and has an ability to induce efficient seed germination of *C. macranthos* var. *rebunense* [12].
2. When extracting RNAs from fungal mycelium, incubate fungi in a liquid oatmeal medium for 2 months (20 °C, in the dark). After incubation, remove excess medium by filtration. Freeze the harvested mycelium with liquid N_2 and keep at −80 °C until use (*see* **Note 2**).

3.3 Preparation of Single-Stranded RNA

1. Total nucleic acids are extracted from orchid or fungal tissues as follows: homogenize 0.1 g tissues in liquid N_2 using a mortar and pestle, and immediately transfer the frozen powder into a 1.5 mL microtube containing 500 μL of TE-saturated phenol and 500 μL of the nucleic acid extraction buffer (*see* **Note 3**).

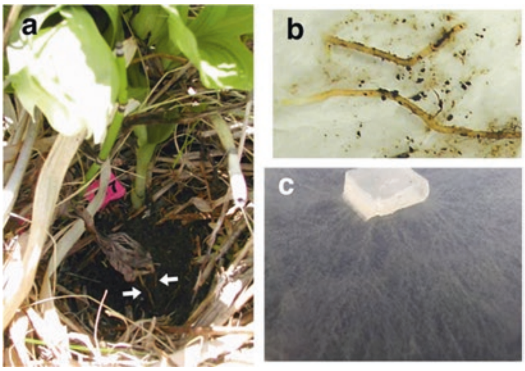


Fig. 2 Sampling of the field-growing orchid and isolation of mycorrhizal fungi. (**a**) Sampling of roots from a wild *Cypripedium macranthos* var*. rebunense* plant. Arrows indicate exposed roots. (**b**) Roots collected from a wild *C. macranthos* var*. rebunense* plant. (**c**) A representative image of mycorrhizal fungi isolated from roots of *C. macranthos* var*. rebunense.* A square disc on the plate is a fungal inoculum. Linear, radially distributed structures from an inoculum are growing hyphae. The hyphal colonies of the isolate are hyaline or white on an oatmeal agar medium

2. Vortex for 15 s and centrifuge for 3 min at 10,000 × *g* at room temperature.
3. Transfer the aqueous phase to a new tube and add an equal volume of TE-saturated phenol-chloroform. Vortex the mixture for 15 s and then centrifuge for 5 min at 12,000 × *g* at room temperature (approximately 450 μL of aqueous phase can be recovered).
4. Repeat TE-saturated phenol-chloroform extraction (*see* **Note 4**).
5. Collect the aqueous phase in a new tube and add 1/10 volume of 3 M sodium acetate (pH 5.2) and 3 volumes of 99.5% ethanol, and then mix and turn over the tube gently.
6. Centrifuge for 15 min at 12,000 × *g* at 4 °C. Carefully remove the supernatant and wash the pellet with 500 μL of 70% ethanol followed by a spin at 12,000 × *g* for 5 min at 4 °C.
7. Discard the supernatant, and spin again at 12,000 × *g* for 1 min to remove the residual supernatant.
8. Vacuum-dry or let it stand until it dries (for 5 min at room temperature). Then, resuspend the pellet in 30–50 μL of RNase-free water.
9. Quantify the extracted nucleic acids by spectrophotometer, and check a normal band pattern by 1.2% agarose gel electrophoresis.
10. Process total nucleic acids to DNase I treatment as follows: incubate reaction mixture (5 μL of 10× buffer, 5 μg of total nucleic acids, 2 μL of TURBO DNase and RNase-free water up to 50 μL) at 37 °C for 30 min.

11. Add RNase-free water up to 100 μL and extract RNAs by phenol-chloroform.
12. Transfer aqueous phase (ca. 100 μL) to a new tube, add 1/10 volume of 3 M sodium acetate (pH 5.2) and 3 volumes of 99.5% ethanol, and then mix and turn over gently. Keep the mixtures at −30 °C overnight.
13. For precipitation of RNAs, centrifuge a tube at 12,000 × *g* for 20 min (4 °C) and resuspend the pellet in 20–30 μL of RNase-free water.
14. Quantify and check RNAs by spectrophotometer and 1.2% agarose gel electrophoresis.

3.4 Preparation of Double-Stranded RNA

1. For preparation of dsRNA, extract total nucleic acids in the same way as in Subheading 3.3 except for use of dsRNA extraction buffer.
2. DNase I treatment is also conducted as in Subheading 3.3 and then treat the extracted RNA samples by S1 nuclease as follows: incubate a reaction mixture (5 μL of 10× buffer, 5 μg of RNA, 2 μL of S1 nuclease, and RNase-free water up to 50 μL) at 37 °C for 30 min.
3. Add RNase-free water up to 100 μL and extract RNAs by phenol-chloroform (*see* **Note 5**).
4. Transfer the aqueous phase (ca. 100 μL) to a new tube and add 1/10 volume of 3 M sodium acetate (pH 5.2) and 3 volumes of 99.5% ethanol. Mix and turn over the tube gently.
5. After cooling at −30 °C overnight, precipitate dsRNAs by centrifugation (12,000 × *g*, 20 min, 4 °C), and resuspend the pellet in RNase-free water.
6. Quantify and check dsRNAs by spectrophotometer and 1.2% agarose gel electrophoresis (*see* **Note 6**). A flow of preparation of ssRNA and dsRNA for each orchid and fungi sample is summarized in Fig. 3.

3.5 Next-Generation Sequencing of RNA Samples from Orchid and Fungus (Our Example)

1. For next-generation sequencing (NGS) analyses, we prepared four kinds of RNA samples from orchid plant and fungus: (i) ssRNA extracted from orchid tissue from a wild plant (orchid [wild] ssRNA), (ii) ssRNA extracted from in vitro-cultured orchid tissue (orchid [in vitro] ssRNA), (iii) dsRNA extracted from in vitro-cultured orchid tissue (orchid [in vitro] dsRNA), and (iv) dsRNA extracted from cultured fungal mycelium (fungi [cultured] dsRNA) (Fig. 3). For dsRNA sample from fungi, we used dsRNA extracted from WO97 isolate. These RNAs are processed for RNA-seq analysis by a standard protocol depending on which NGS provider we use (e.g., Hokkaido System Science [HSS], Japan).

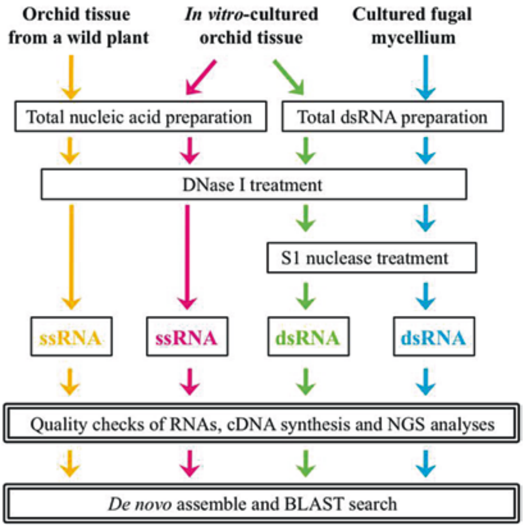

Orchid tissue from a wild plant
In vitro-cultured orchid tissue
Cultured fugal mycellium
Total nucleic acid preparation
Total dsRNA preparation
DNase I treatment
S1 nuclease treatment
ssRNA
ssRNA
dsRNA
dsRNA
Quality checks of RNAs, cDNA synthesis and NGS analyses
De novo assemble and BLAST search

Fig. 3 A flow of sample preparation for NGS analyses from the orchid and fungi

2. Process RNA samples for library preparation using TruSeq RNA sample Prep Kit (Illumina) according to standard protocols. For ssRNA samples, conduct poly(A)-RNA purification using oligo-dT beads. Fractionate RNAs by divalent cation treatments and then conduct reverse transcription, adaptor ligation, and PCR amplification (*see* **Note 7**). After purification and elimination of small molecules (<200 bp) by AMPure XP beads, apply the prepared library samples to 100 bp paired-end sequencing using Illumina Hiseq 2000 (*see* **Note 8**).
3. Select the sequenced raw data by filtering and sort only high-quality data by the index tag sequence in the adaptor primer of each sample. In our example, we obtained the numbers of sequence reads between 40 and 50 million for each library: orchid (wild) ssRNA, 47,016,804 reads; orchid (in vitro) ssRNA, 42,302,340 reads; orchid (in vitro) dsRNA, 40,486,758 reads; and fungi (cultured) dsRNA, 46,798,910 reads.
4. Process the obtained reads in each sample for adapter trimming, and then execute de novo assembly by Trinity [15]. In our example, we obtained 117,408 contigs for orchid (wild) ssRNA, 103,118 contigs for orchid (in vitro) ssRNA, 21,468 contigs for orchid (in vitro) dsRNA, and 4474 contigs for fungi (cultured) dsRNA.
5. For annotations of the obtained contig, run Blastx + (version 2.2.29+) program against the amino acid databases of DDBJ viral and NCBI RefSeq plant.

3.6 Extraction of Viral Sequences and Phylogenetic Analysis

1. On the basis of the annotation information, sort the contigs of the viral sequences and the others in an Excel program. In our example, the contigs of the virus sequences included not only plant viruses but also many animal and protist viruses (Fig. 4). The numbers of virus species were 81 for orchid (wild) ssRNA, 17 for orchid (in vitro) ssRNA, 36 for orchid (in vitro) dsRNA, and 14 for fungi (cultured) dsRNA. We found that various virus species were detected in each sample, and that some virus species were common among different samples (Fig. 4).
2. Because it is normally difficult to find significant homology at the nucleic acid level among the viral sequences unless they are

a

Origin	RNA type	Host species of detected viruses					Total virus species
		Plant	Animal	Fungi	Protist	Other	
Orchid (wild)	ssRNA	30	26	7	17	1	81
Orchid (*in vitro*)	ssRNA	9	5	2	1	0	17
Orchid (*in vitro*)	dsRNA	18	9	1	8	0	36
Fungi (cultured)	dsRNA	1	5	3	4	1	14

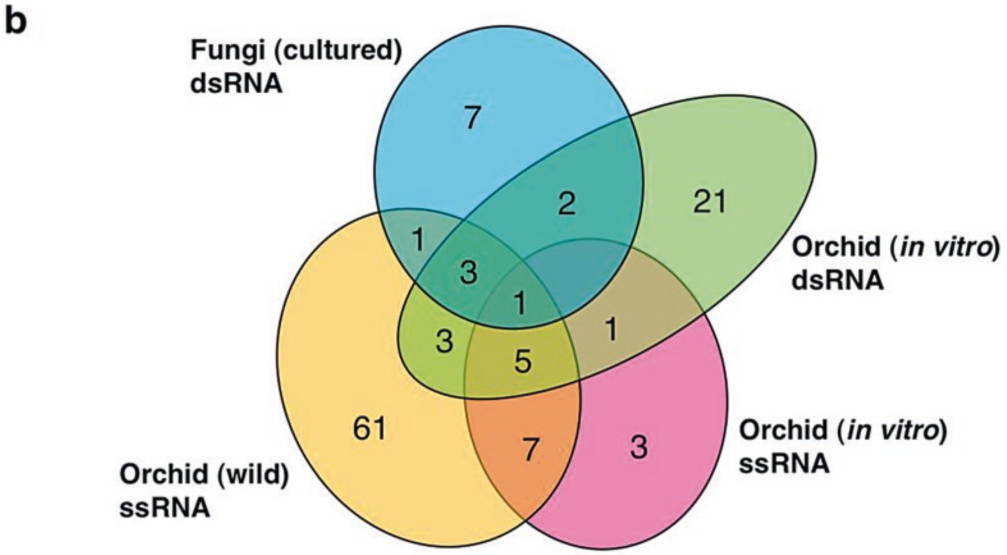


Fig. 4 Example of viral species detected by NGS. (**a**) The table shows the numbers of viral sequences detected from each sample. (**b**) The Venn diagram shows common sequences among the samples

closely related, we should compare viral amino acid sequences. We thus extracted the viral nucleic acid sequences from the Excel file and converted them into amino acid sequences after confirming the direction. We used a conventional DNA analysis tool such as DNADynamo (Blue Tractor Software Ltd). For phylogenetic analysis, multiple alignments of amino acid sequences using Clustal W [16] and phylogenetic analyses using a neighbor-joining program in MEGA 5 [17, 18] were conducted.

3. We show an example of our phylogenic analysis of partitivirus-like RNA sequences from an orchid plant (*C. macranthos* var. *rebunense*) and its mycorrhizal fungi (Fig. 5). In our preliminary experiments, we identified a partitivirus as a dsRNA band (in a gel) extracted from WO97 by a conventional cloning method. This time, we could confirm the real presence of the partitivirus sequence in WO97 (i.e., in the sample of fungi [cultured] dsRNA) by NGS. We here isolated both ssRNA and dsRNA from the orchid plants of two different origins, one is field isolated, and the other is in vitro cultured, and examined virus-like sequences by NGS. Interestingly, we found several partitivirus sequences in the cDNA libraries from both in vitro-cultured and filed-isolated orchid tissues. We then created a phylogenetic tree using the partitivirus-like sequences together with more than 20 known partitiviruses (fungal viruses) and cryptic viruses (plant viruses); the partitiviruses and cryptic viruses are phylogenetically related [9].

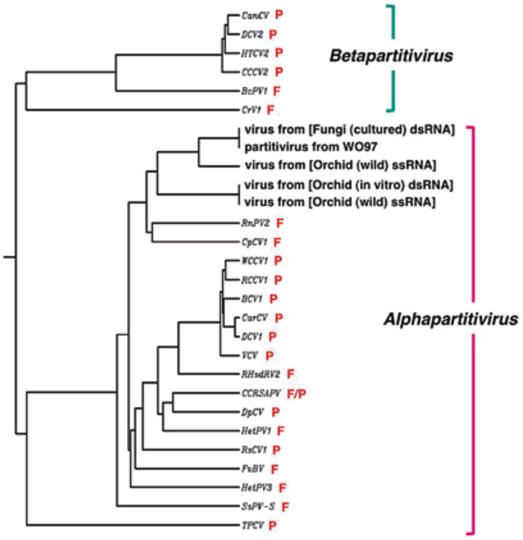


Fig. 5 Example of a phylogenetic analysis of partitivirus-like RNA sequences from an orchid and its mycorrhizal fungi. The classification of *Alphapartitivirus* and *Betapartitivirus* is based on Nibert et al. (2014). Putative host of each partitivirus is shown on the right: F, fungus; P, plant

Unexpectedly, a partitivirus from WO97 and the partitivirus-like sequences derived from orchid tissues were located in the same clade, suggesting that these sequences may have evolved from a common ancestor. Taken together, we can now hypothesize horizontal transmission of a partitivirus (or cryptic virus) between an orchid and fungus because we could find a similar sequence also in the in vitro-cultured orchid tissue, in which any contamination of mycorrhizal or endophytic fungi should not occur (*see* **Note 9**).

4 Notes

1. To avoid excessive browning that causes negative effects to normal tissue development, it is better that protocorm and seedling are transferred to fresh medium every 3 months.
2. Depending on the fugal isolate, it takes more than 4 months to obtain enough volume of mycelium for RNA preparation. However, a prolonged culture (about 8 months) may cause lower efficiency for RNA extraction.
3. As appropriate, the buffer volume should be scaled up.
4. TE-saturated phenol-chloroform extraction should be repeated until the middle phase almost disappears.
5. Instead of TE-saturated phenol, use of acid phenol is better.
6. When mycoviruses abundantly accumulate in a fungal isolate, we can see some bands of dsRNA (derived from mycoviruses) in this electrophoresis.
7. This PCR amplification step is conducted in all samples to increase the library concentration to an appropriate level for sequencing. The number of the amplification cycle is generally less than 4.
8. About construction of cDNA libraries: first-strand cDNA synthesis is performed using random hexamers as a primer to ensure that we capture all the viral RNAs in the sample. We can give RNAs directly to a provider, but recommend that first-strand cDNA synthesis should be performed with our own hands because dsRNAs is normally very small in quantity and must be well denatured before first-strand synthesis. The RNA-DNA hybrids can be given to a provider to eventually generate dsDNA libraries; the provider adds an adapter to the dsDNAs.
9. To verify that the viral sequences found by NGS actually existed in the plants and fungi, we should perform RT-PCR to detect the same sequences as those in NGS using the primers that are designed based on the NGS results.

Acknowledgments

We thank Dr. Tatsuhiro Ezawa (Hokkaido University, Japan) for his helpful advice for the experiments. We also thank Ms. Chieko Hirata for her technical help. We appreciate Botanical Garden, Hokkaido University, for giving us orchid materials. This work was supported in part by grants from the Ministry of the Environment, Government of Japan (Global Environment Research Fund).

References

1. Ghabrial SA, Suzuki N (2009) Viruses of plant pathogenic fungi. Annu Rev Phytopathol 47:353–384
2. Yu X, Li B, Fu Y, Jiang D, Ghabrial SA, Li G, Peng Y, Xie J, Cheng J, Huang J, Yi X (2010) A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc Natl Acad Sci U S A 107:8387–8392
3. Nuss DL (2005) Hypovirulence: mycoviruses at the fungal-plant interface. Nat Rev Microbiol 8:632–642
4. Anagnostakis SL (1982) Biological control of chestnut blight. Science 215:466–471
5. Wei CZ, Osaki H, Iwanami T, Matsumoto N, Ohtsu Y (2004) Complete nucleotide sequences of genome segments 1 and 3 of Rosellinia antirot virus in the family Reoviridae. Arch Virol 149:773–777
6. Chiba S, Salaipeth L, Lin YH, Sasaki A, Kanematsu S, Suzuki N (2009) A novel bipartite double-stranded RNA Mycovirus from the white root rot Fungus Rosellinia necatrix: molecular and biological characterization, taxonomic considerations, and potential for biological control. J Virol 83:12801–12812
7. Smith SE, Read DJ (2008) Mycorrhizal symbiosis, 3rd edn. Academic Press, London, p 800
8. Ikeda Y, Shimura H, Kitahara R, Masuta C, Ezawa T (2012) A novel virus-like double-stranded RNA in an obligate biotroph arbuscular mycorrhizal fungus: a hidden player in mycorrhizal symbiosis. Mol Plant-Microbe Interact 25:1005–1012
9. Nibert ML, Ghabrial SA, Maiss E, Lesker T, Vainio EJ, Jiang D, Suzuki N (2014) Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research. Virus Res 188:128–141
10. Szego A, Enünlü N, Deshmukh SD, Veliceasa D, Hunyadi-Gulyás E, Kühne T, Ilyés P, Potyondi L, Medzihradszky K, Lukács N (2010) The genome of Beet cryptic virus 1 shows high homology to certain cryptoviruses present in phylogenetically distant hosts. Virus Genes 40:267–276
11. Tsutsui K, Tomita M (1990) Suitability of several carbohydrates as the carbon sources for symbiotic seedling growth of two orchid species. Lindleyana 5:134–139
12. Shimura H, Koda Y (2005) Enhanced symbiotic seed germination of Cypripedium macranthos var. rebunense following inoculation after cold treatment. Physiol Plant 123:281–287
13. Shimura H, Sadamoto M, Matsuura M, Kawahara T, Naito S, Koda Y (2009) Characterization of mycorrhizal fungi isolated from the threatened Cypripedium macranthos in a northern island of Japan: two phylogenetically distinct fungi associated with the orchid. Mycorrhiza 19:525–534
14. Shimura H, Koda Y (2004) Micropropagation of Cypripedium macranthos var. rebunense through protocorm-like bodies derived from mature seeds. Plant Cell Tissue Organ Cult 78:273–276
15. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
16. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
17. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
18. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Chapter 13

DNA Multiple Sequence Alignment Guided by Protein Domains: The MSA-PAD 2.0 Method

Bachir Balech, Alfonso Monaco, Michele Perniola, Monica Santamaria, Giacinto Donvito, Saverio Vicario, Giorgio Maggi, and Graziano Pesole

Abstract

Multiple sequence alignment (MSA) is a fundamental component in many DNA sequence analyses including metagenomics studies and phylogeny inference. When guided by protein profiles, DNA multiple alignments assume a higher precision and robustness. Here we present details of the use of the upgraded version of MSA-PAD (2.0), which is a DNA multiple sequence alignment framework able to align DNA sequences coding for single/multiple protein domains guided by PFAM or user-defined annotations. MSA-PAD has two alignment strategies, called "Gene" and "Genome," accounting for coding domains order and genomic rearrangements, respectively. Novel options were added to the present version, where the MSA can be guided by protein profiles provided by the user. This allows MSA-PAD 2.0 to run faster and to add custom protein profiles sometimes not present in PFAM database according to the user's interest. MSA-PAD 2.0 is currently freely available as a Web application at https://recasgateway.cloud.ba.infn.it/.

Key words Genomic rearrangement, Multiple sequence alignment, Conserved protein domains, Phylogeny, Sequence assignment

1 Introduction

Multiple sequence alignment (MSA) is central to bioinformatics analyses as it constitutes an integral component in many sequence analysis applications, such as phylogeny inference, variant calling, and taxonomic and sequence assignment in metagenomics analysis workbenches (i.e., Mothur [1], pplacer [2]). A refined and accurate MSA can radically improve the quality and precision of biological conclusions drawn out of a DNA dataset. In this context, MSA-PAD [3] represents a multiple sequence alignment framework with outstanding advantages. In details, MSA-PAD (1) translates given DNA sequences using a user-defined genetic code and open reading frame/s (ORFs), (2) identifies the protein domains they code for, (3) aligns all amino acid sequence fragments coding for the same domain, (4) deals with multiple and repeated domain

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_13, © Springer Science+Business Media, LLC 2018

copies, (5) back-aligns amino acid alignments into DNA ones, and (6) provides a merged DNA alignment of all coding domains detected in the input sequences. Moreover, when applied to genomic sequences, MSA-PAD considers either intron occurrence and gene order variations (e.g., viral [4, 5] and mitochondrial genomes [6, 7]) resulting in apparently truncated protein domains or variations in their arrangement along the genome regions under investigation. Several algorithms, such as Muscle [8], Mafft [9], ClustalO [10], PROMALS [11], and PRANK [12], focus on a single final target, which is an MSA output without accounting for the above-mentioned characteristics especially for protein domain information embedded in the input DNA sequences. Other algorithms, such as TranslatorX [13] and tranalign (EMBOSS package) [14], take into consideration protein domain information present in the input DNA but do not account for intron occurrence or genomic arrangements. The final MSA produced by MSA-PAD can be generated following two different modes: (i) Gene or (ii) Genome. "Gene" mode alignment respects domain order organization from 5′ to 3′, and resolves the alignment of repetitive domains even when they are repeated in tandem, while "Genome" mode alignment provides a super-gene-like alignment ignoring domain order constraints.

Here we present MSA-PAD 2.0 with additional options compared to its older version. While the first version maps the translated DNA sequence fragments only against PFAM [15] conserved domains in order to assign each chunk to the relevant domain, MSA-PAD 2.0 offers the possibility to use custom user profiles either alone or together with PFAM. In other words, the user can upload his or her own trusted protein profile/s and use them alone or merged with PFAM to assign the amino acid sequences to a known domain. This novelty has many advantages, namely it reduces the computational time when user profiles are used alone, it trims the input sequences (i.e., genomic sequences) at the target protein domain positions, and it allows to include domains not present in PFAM database.

2 Materials

2.1 Preparing Input File/s

1. Prepare the input DNA sequences in FASTA format. This format consists of sequence identifier (seqID) preceded by greater-than symbol (">") followed by the DNA sequence itself on a new line. The DNA sequence can be on one or on several lines, where whole IUPAC code [16] is accepted. Input sequences can be downloaded from public primary (NCBI [17], ENA: https://www.ebi.ac.uk/ and others) and/or specialized databases (BOLD [18], ViPR [19]) or from newly

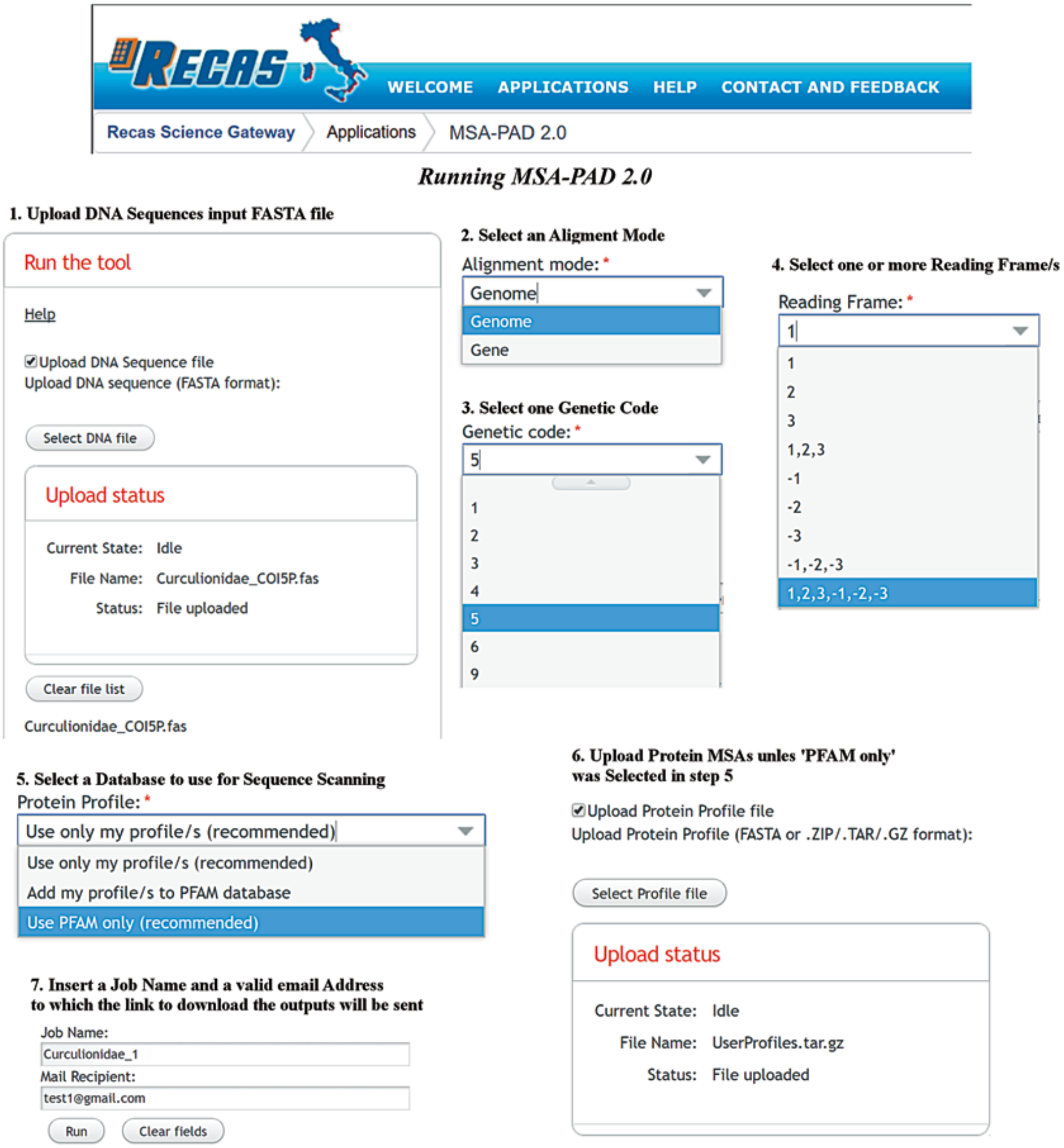
produced sequences by Sanger method. A text editor of your choice (i.e., Notepad++, TextPad, gedit, TextWrangler) should be used to prepare and save the DNA input file with ".fasta" or ".fas" extension (*see* **Note 1**).

2. Be sure that the sequences are protein coding. Note that MSA-PAD will detect intrinsically the ORF, exon, and intron positions and the coding strand (*see* Subheading 3.1).
3. To guide your MSA by your own protein profile/s instead of PFAM's ones, prepare your trusted protein profiles in FASTA format and align them singularly using your favorite multiple protein alignment algorithms such as Muscle, ClustlaO, PRANK, or others. In case of more than one protein profile, paste all profiles files into one single folder and compress it with your system archiver algorithm (*see* **Note 2**).

3 Methods

3.1 Running MSA-PAD

1. To run MSA-PAD, go to https://recasgateway.cloud.ba.infn.it/, and register and/or log in with your personal or your favorite social network account. Once logged in, from the "APPLICATIONS" menu tab choose MSA-PAD 2.0. This page (Fig. 1) will allow running the tool when the required inputs (marked with asterisk) are supplied as described below. Example input files can be downloaded from the web page with their corresponding outputs. Consult also the help page if you need additional details on MSA-PAD parameters.
2. Upload your DNA input sequences in FASTA-formatted text file by clicking the button "Select DNA File."
3. Choose an "Alignment mode." There are two different modes, namely "Gene" or "Genome" (*see* **Note 3**).
4. Choose one "Genetic code." The genetic codes refer to NCBI genetic codes (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), also available in MSA-PAD help page.
5. Choose one or more reading frame/s from "Reading Frame" drop-down menu. This refers to the translation starting position in each DNA sequence present in the input file. Choose forward (1 or 1, 2 or 1, 2, 3) and/or reverse (−1 or −1, −2 or −1, −2, −3) according to the input DNA sequences. Note that the choice of all six frames leaves the possibility to the algorithm to select across all six frames the correct one, which is a strength point in this MSA framework (*see* **Note 4**).
6. Choose a database, from "Protein profile" drop-down menu, against which the input DNA sequences will be searched. At this point, you can consider PFAM database only, your own



Running MSA-PAD 2.0

1. Upload DNA Sequences input FASTA file

Run the tool

☑Upload DNA Sequence file
Upload DNA sequence (FASTA format):

Upload status

Current State: Idle
File Name: Curculionidae_COI5P.fas
Status: File uploaded

Curculionidae_COI5P.fas

2. Select an Aligment Mode

Alignment mode:* Genome

3. Select one Genetic Code

Genetic code:* 5

4. Select one or more Reading Frame/s

Reading Frame:* 1

5. Select a Database to use for Sequence Scanning

Protein Profile:* Use only my profile/s (recommended)

Use only my profile/s (recommended)
Add my profile/s to PFAM database
Use PFAM only (recommended)

6. Upload Protein MSAs unles 'PFAM only' was Selected in step 5

☑Upload Protein Profile file
Upload Protein Profile (FASTA or .ZIP/.TAR/.GZ format):

Upload status

Current State: Idle
File Name: UserProfiles.tar.gz
Status: File uploaded

7. Insert a Job Name and a valid email Address to which the link to download the outputs will be sent

Job Name: Curculionidae_1
Mail Recipient: test1@gmail.com

Fig. 1 Snapshots of MSA-PAD web application. All steps needed to successfully run MSA-PAD are numbered from one to seven

trusted protein profile/s alone (prepared as stated in **item 3** of Subheading 2), or a joined database of both PFAM and user profile/s (*see* **Note 5**).

7. Add a "Job Name" by which you can recognize your own MSA-PAD run.
8. Add a valid e-mail address in "Mail Recipient" field to which the program outputs will be sent (*see* **Note 6**).

3.2 Output Retrieval and Interpretation

1. The MainOutput folder includes three files:
   (a) The final MSA in FASTA format.
   (b) "AlignmentDomainsPartitions" file, which specifies the coordinates of each protein domain in the final MSA. For instance, "domain_1 = 0–100" indicates that positions 0–100 in the final multiple alignment belong to domain 1 (*see* **Note 7**).
   (c) "ExcludedSequencesIDs" file containing sequence identifiers (separated by comma) not present in the final MSA, because they did not satisfy the criteria of the most frequent domain pattern in "Gene mode" or they did not have any unique domain in "Genome mode" (*see* **Note 8**).
2. The AdditionalFiles folder includes three file types:
   (a) File/s with *hmmAligned* suffices: These files consist of the alignments (in STOCKHOLM format) of each protein sequence block. The prefix indicates the name of PFAM or user profile the sequences were aligned against. Sequence identifiers included in this file contain information about the original position in the DNA sequence belonging to the corresponding domain. For example, "Sequence1;code5;frame-1;S8169;E9603@8478@9279" indicates that "Sequence1" was translated from position 8169 to position 9603 using the genetic code number 5 on the reverse frame 1 and had a significant profile match at the DNA sites 8478–9279. Another example representing the presence of intron can be as follows: "Sequence2;code1;frame1;S0;E172;code1;frame1;S456;E666@3@100@489@595"; this means that the first exon in sequence2 spans the positions 3–100 while the second one is located at positions 489–595 (*see* **Note 9**).
   (b) File/s with *Backaligned.fasta* suffices consisting of the DNA alignments of each back-aligned amino acid sequence block. In their prefix, PFAM domain names to which DNA sequences belong are mentioned (*see* **Note 10**).
   (c) The "MissingSites_Report" file containing the original DNA site position missing from the final multiple alignment (*see* **Note 11**).

4 Notes

1. Only underscore ("_") is allowed to use in the seqID. Do not use special characters in your input sequences especially "@, comma or semicolon." It is best to use only accession numbers or your own seqIDs. In addition, DNA sequences longer than

80,000 or shorter than 30 nucleotides (10 amino acids) are not acceptable as the underneath algorithms will not be able to analyze them. Mitochondrial and many viral genomes are compatible with sequence length limit of MSA-PAD.

2. We recommend the use of "7zip" on all main operating systems (Windows, Linux, and Mac) or "tar" or "zip" in Linux and Mac OS environments. These algorithms are able to create ".zip," ".tar," and ".gz" archives compatible with MSA-PAD.
3. Gene mode conserves genomic organization from 5′ to 3′ including repeated domains. For that, if gene/domain order is important in the case study, be careful to use "Gene mode" alignment; otherwise "Genome mode" will eliminate all redundant domains and keeps unique ones to avoid paralogy mapping.
4. Using more than one ORF in forward, reverse, or both directions lets MSA-PAD to detect automatically intron/s positions and frameshifts. This will prevent you to manage manually your input DNA sequences in order to report them to the right ORF and to splice the introns out.
5. Use the database you trust more. We recommend the choice of PFAM database or the user profile/s alone, especially if PFAM may contain the same domain as the user's one. In this latter case, if the user profile was added to PFAM, MSA-PAD will choose one of them (sorted by alphabetical order) and it will seem as false negative or unexpected match.
6. Upon job completion, an e-mail will be sent to the address previously provided with a link to download the job zipped-folder output. If the run was not successful, the output will contain a single file with an error message; otherwise, two main output folders called "MainOutput" and "AdditionalFiles" will be generated.
7. The "AlignmentDomainsPartitions" file is useful not only for DNA sequence fragment annotations but also for phylogeny inference using partitioned models where each partition of the MSA can be analyzed with different evolutionary model.
8. In Gene mode alignment, MSA-PAD will determine the occurrence of the most common domains respecting their orders from 5′ to 3′. For this reason, it is important to use homologous sequences; otherwise, sequences that do not map to the most common domains in the majority of sequences in the input dataset will be eliminated. On the other hand, Genome mode alignment will ignore repeated domains but it deals with genomic rearrangement as it concatenates the final sequence mapping on different domains arbitrarily (often sorted by alphabetical order). Check out the domain partitions

file in the "MainOutput" folder (*see* Subheading 3.2) to localize your initial sequence coordinates.

9. The protein alignment blocks correspond to each single domain embedded in the input DNA sequences. These can be used in a second time as user profile domain/s for another case study. If used, MSA-PAD will select only the fragment/s of sequences mapping on this specific domain and will trim out all DNA sites not belonging to it. This operation is useful in preparing a consistent MSA for phylogenetic analyses or a consistent reference dataset in sequence assignment analyses such as Metagenomics studies.
10. The back-aligned files are useful if you wish to analyze each domain singularly at nucleotide level.
11. MSA-PAD will eliminate introns as it works definitely on coding sequences. In addition, PFAM database contains conserved protein domains often missing few sites at both profile/s ends. All the above will be reported as excluded DNA sites in the "MissingSites" file.

Acknowledgment

This work was supported by the Italian nodes of Lifewatch and ELIXIR Research Infrastructures.

References

1. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537–7541. https://doi.org/10.1128/AEM.01541-09
2. Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538. https://doi.org/10.1186/1471-2105-11-538
3. Balech B, Vicario S, Donvito G, Monaco A, Notarangelo P, Pesole G (2015) MSA-PAD: DNA multiple sequence alignment framework based on PFAM accessed domain information. Bioinformatics 31(15):2571–2573. https://doi.org/10.1093/bioinformatics/btv141
4. Yang XF, Peng JJ, Liang HR, Yang YT, Wang YF, Wu XW, Pan JJ, Luo YW, Guo XF (2014) Gene order rearrangement of the M gene in the rabies virus leads to slower replication. Virusdisease 25(3):365–371. https://doi.org/10.1007/s13337-014-0220-1
5. Flanagan EB, Zamparo JM, Ball LA, Rodriguez LL, Wertz GW (2001) Rearrangement of the genes of vesicular stomatitis virus eliminates clinical disease in the natural host: new strategy for vaccine development. J Virol 75(13):6107–6114. https://doi.org/10.1128/JVI.75.13.6107-6114.2001
6. D'Onorio de Meo P, D'Antonio M, Griggio F, Lupi R, Borsani M, Pavesi G, Castrignano T, Pesole G, Gissi C (2012) MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. Nucleic Acids Res 40(Database issue):D1168–D1172. https://doi.org/10.1093/nar/gkr1144
7. Gai Y, Song D, Sun H, Yang Q, Zhou K (2008) The complete mitochondrial genome of

Symphylella sp. (Myriapoda: Symphyla): extensive gene order rearrangement and evidence in favor of Progoneata. Mol Phylogenet Evol 49(2):574–585. https://doi.org/10.1016/j.ympev.2008.08.010

8. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113
9. Katoh K, Standley DM (2016) A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32(13):1933–1942. https://doi.org/10.1093/bioinformatics/btw108
10. Sievers F, Higgins DG (2014) Clustal omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 1079:105–116. https://doi.org/10.1007/978-1-62703-646-7_6
11. Pei J, Grishin NV (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23(7):802–808. https://doi.org/10.1093/bioinformatics/btm017
12. Loytynoja A (2014) Phylogeny-aware alignment with PRANK. Methods Mol Biol 1079:155–170. https://doi.org/10.1007/978-1-62703-646-7_10
13. Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38(Web Server issue):W7–13. https://doi.org/10.1093/nar/gkq291
14. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16(6):276–277
15. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230. https://doi.org/10.1093/nar/gkt1223
16. Johnson AD (2010) An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics 26(10):1386–1389. https://doi.org/10.1093/bioinformatics/btq098
17. Coordinators NR (2017) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 45 (D1):D12-D17. doi:https://doi.org/10.1093/nar/gkw1071
18. Ratnasingham S, Hebert PD (2007) Bold: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7(3):355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
19. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN, Jen W, Klem EB, Scheuermann RH (2012) Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Virus 4(11):3209–3226. https://doi.org/10.3390/v4113209

Chapter 14

From Whole-Genome Shotgun Sequencing to Viral Community Profiling: The ViromeScan Tool

Simone Rampelli and Silvia Turroni

Abstract

ViromeScan is an innovative metagenomic analysis tool that allows the viral community characterization in terms of taxonomy from raw data of metagenomics sequencing. It efficiently denoises samples from reads of other microorganisms. Users can adopt the same shotgun metagenomic sequencing data to fully characterize complex microbial ecosystems, including bacteria and viruses. Here we apply ViromeScan pipeline to some examples, thus illustrating the processes computed from raw data to the final output.

Key words Virome, Shotgun metagenomics, Sequencing, Bioinformatics, Pipeline

1 Introduction

The most advanced experimental procedures for profiling the virome include isolation and extraction of the encapsidated viral fraction [1–3] and only at a later stage sequencing and characterization of the viral community by either assembled or read mapping strategies [4–7]. An emerging possibility is to detect the viral reads directly from shotgun metagenomic sequence data, without the need of preparative procedures. In particular, the assignment of the taxonomic ID to sequencing unprocessed samples allows a faster and more reliable profiling of the virome. In the context of the entire microbiome, it eliminates the risk of missing information during the extraction step, as already demonstrated for giant viruses [8]. However, metagenomic reads contain nucleic acid fragments from several microorganisms, including bacteria, archaebacteria, eukaryotes, phages, and eukaryotic viruses, moving the experimental challenge on how to discriminate viral reads when mixed with sequences belonging to other organisms. ViromeScan is an innovative bioinformatics tool that accurately profiles viral communities directly from raw data of metagenomic sequencing, responding to the previous issue through a set of consecutive bioinformatic filters, which select and discard the metagenomic nonviral reads such

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_14, © Springer Science+Business Media, LLC 2018

as those of human or bacterial origin [9]. ViromeScan works with shotgun reads and detects the presence of DNA and/or RNA viruses, depending on the input sequences to be processed. By default, ViromeScan profiles the eukaryotic viral community within the microbiome, but it could be implemented and/or personalized by supplying a customized hierarchical reference database. ViromeScan is available at the website http://sourceforge.net/projects/viromescan/.

2 Materials

ViromeScan needs a PC or workstation with Linux or OS 10.X as operating system and Bash version release 4.1.2 or later. The machine has also to read the programming languages R (version >3.0.0), Perl v5.10.1 or later, and Java v1.6 or other more recent version. Other required software are Bowtie2 [10], BMTagger [11], and Picard tools [12].

3 Methods

3.1 Input Files and Database

1. Input files should be single-end or paired-end reads in Fastq format (for paired-end reads compressed files in gzip, bzip2, and zip formats are also accepted) retrieved from shotgun sequencing or RNA-seq.
2. Based on the research strategy, ViromeScan offers to users the ability to choose from different in-house-built reference databases, including human DNA virus database, human DNA/RNA virus database, eukaryotic DNA virus database, and eukaryotic DNA/RNA virus database. The human virus databases are made up only of viruses that have the human being as a natural host; on the other hand, the eukaryotic virus databases also include viruses for vertebrates, invertebrates, fungi, algae, and plants while excluding bacteriophages. All databases are built using the complete viral genomes available on the NCBI website [13] (*see* **Note 1**).

3.2 Workflow of ViromeScan

1. The first step consists of an accurate screening of the reads, in order to select candidate viral sequences. Performing this procedure before filtering steps allows a remarkable gain of time in the other pipeline steps, due to the reduction of the dataset to less than 1% of the total amount of metagenomic reads (*see* Fig. 1 to visualize the workflow of ViromeScan).
2. The subsequent quality filtering step of the candidate viral reads has been implemented as described in the processing procedure of the Human Microbiome Project (HMP) [14].

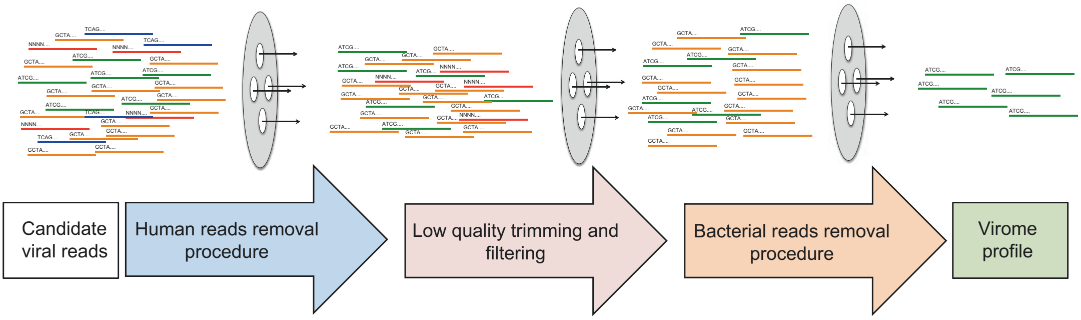
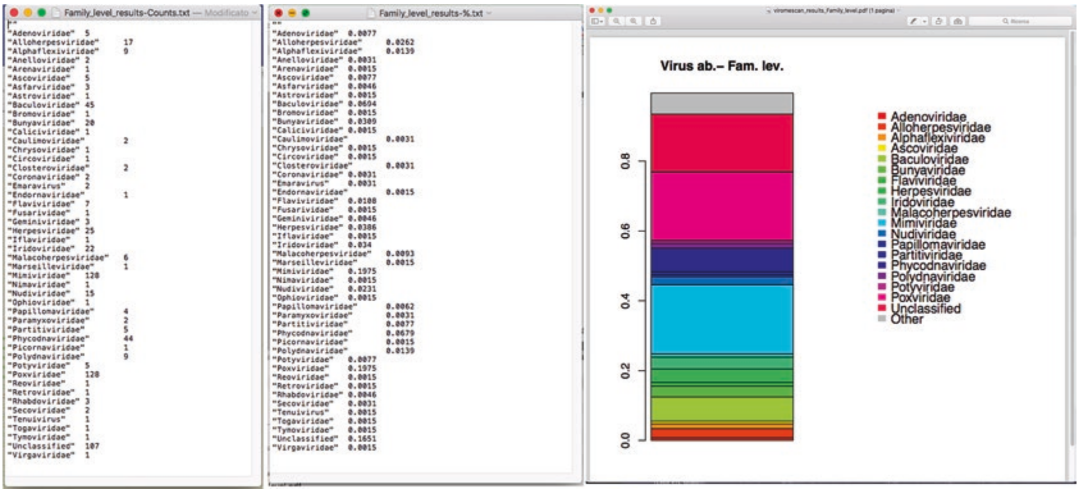

Candidate viral reads
Human reads removal procedure
Low quality trimming and filtering
Bacterial reads removal procedure
Virome profile

Fig. 1 Workflow of ViromeScan. Candidate viral reads are identified by mapping the sequences to a reference database and then filtered using three subsequent steps to trim low-quality reads and completely remove any human and bacterial contaminants

In brief, sequences are trimmed for a low-quality score using the script trimBWAstyle.pl [15]. The script is specifically utilized to trim bases off the ends of sequences, which show a quality value of two (or lower). This threshold is taken to delete all the bases with uncertain quality as defined by Illumina's EAMMS filter (End Anchored Max Scoring Segments). In addition, reads trimmed to less than 60 bp are also discarded.

3. In the light of the fact that the sequences to be analyzed derive from whole genome or RNA sequencing, it is possible that the candidate viral reads contain a small percentage of human sequences. In order to eliminate these possible contaminants, an additional screening step for human contamination has been implemented. As reported in the HMP procedures [16], Human Best Match Tagger (BMTagger) [11] is an efficient tool that allows discriminating among human, microbial, and viral reads. First, BMTagger attempts to discriminate between human reads and the other reads by comparing the 18mers produced from the input file with those contained in the reference human database. If this first attempt fails, an additional alignment step is performed to ensure the detection of all possible matches with up to two errors.
4. It is plausible that the human-filtered reads also contain a certain amount of bacterial sequences. For this reason, it is advisable to check the sequences for bacterial contamination. Similar to the human sequence removal procedure, bacterial reads are identified and removed using BMTagger [11]. By default, human-filtered reads are screened against the genomic DNA of a representative group of bacterial taxa that are known to be common in human body niches, but this bacterial database can be implemented and/or customized with sequences of interest (*see* **Note 2**).
5. Filtered reads are finally compared to the viral genomes of the selected hierarchical viral database using Bowtie2 [10].



```
""
"Adenoviridae"  5
"Alloherpesviridae"     17
"Alphaflexiviridae"     9
"Anelloviridae" 2
"Arenaviridae"  1
"Ascoviridae"   5
"Asfarviridae"  3
"Astroviridae"  1
"Baculoviridae" 45
"Bromoviridae"  1
"Bunyaviridae"  20
"Caliciviridae" 1
"Caulimoviridae"        2
"Chrysoviridae" 1
"Circoviridae"  1
"Closteroviridae"       2
"Coronaviridae" 2
"Emaravirus"    2
"Endornaviridae"        1
"Flaviviridae"  7
"Fusariviridae" 1
"Geminiviridae" 3
"Herpesviridae" 25
"Iflaviridae"   1
"Iridoviridae"  22
"Malacoherpesviridae"   6
"Marseilleviridae"      1
"Mimiviridae"   128
"Nimaviridae"   1
"Nudiviridae"   15
"Ophioviridae"  1
"Papillomaviridae"      4
"Paramyxoviridae"       2
"Partitiviridae"        5
"Phycodnaviridae"       44
"Picornaviridae"        1
"Polydnaviridae"        9
"Potyviridae"   5
"Poxviridae"    128
"Reoviridae"    1
"Retroviridae"  1
"Rhabdoviridae" 3
"Secoviridae"   2
"Tenuivirus"    1
"Togaviridae"   1
"Tymoviridae"   1
"Unclassified"  107
"Virgaviridae"  1
```

```
""
"Adenoviridae"  0.0077
"Alloherpesviridae"     0.0262
"Alphaflexiviridae"     0.0139
"Anelloviridae" 0.0031
"Arenaviridae"  0.0015
"Ascoviridae"   0.0077
"Asfarviridae"  0.0046
"Astroviridae"  0.0015
"Baculoviridae" 0.0694
"Bromoviridae"  0.0015
"Bunyaviridae"  0.0309
"Caliciviridae" 0.0015
"Caulimoviridae"        0.0031
"Chrysoviridae" 0.0015
"Circoviridae"  0.0015
"Closteroviridae"       0.0031
"Coronaviridae" 0.0031
"Emaravirus"    0.0031
"Endornaviridae"        0.0015
"Flaviviridae"  0.0108
"Fusariviridae" 0.0015
"Geminiviridae" 0.0046
"Herpesviridae" 0.0386
"Iflaviridae"   0.0015
"Iridoviridae"  0.034
"Malacoherpesviridae"   0.0093
"Marseilleviridae"      0.0015
"Mimiviridae"   0.1975
"Nimaviridae"   0.0015
"Nudiviridae"   0.0231
"Ophioviridae"  0.0015
"Papillomaviridae"      0.0062
"Paramyxoviridae"       0.0031
"Partitiviridae"        0.0077
"Phycodnaviridae"       0.0679
"Picornaviridae"        0.0015
"Polydnaviridae"        0.0139
"Potyviridae"   0.0077
"Poxviridae"    0.1975
"Reoviridae"    0.0015
"Retroviridae"  0.0015
"Rhabdoviridae" 0.0046
"Secoviridae"   0.0031
"Tenuivirus"    0.0015
"Togaviridae"   0.0015
"Tymoviridae"   0.0015
"Unclassified"  0.1651
"Virgaviridae"  0.0015
```

Fig. 2 Standard output of the ViromeScan software. The results of the ViromeScan pipeline are tabulated as both read count and relative abundance, and visualized as histograms at different phylogenetic levels. An example at the family level is shown

This step performs the definitive association of each virome read to a viral genome.

6. The final output consists of a table with the total amount of reads for each detected viral taxon, expressed as number of hits and relative abundance, and additional graphs showing the abundances at family, genus, and species level. These graphs are provided using the "graphics" and "base" R packages (*see* Fig. 2 and **Note 3**).

4 Notes

1. In addition to the reference databases built within ViromeScan, users can build a customized nucleotide database with the preferred genomes or genes. First, they have to type the command "*bowtie2-build database.fa database*" for obtaining a bowtie2 database to be integrated with the viral sequences of interest. This database should be put in the directory "*$PWD/viromescan/database/bowtie2/*" and used as input in the command line after the "-d" option.
2. Users can customize the filtering procedure by implementing or replacing the bacterial database built within the ViromeScan folder ("*$PWD/viromescan/database/Bacteria_custom.fa*") with the bacterial sequences of interest, associated with environments other than the human body (e.g., animal, soil, or water microbiome). Please note that the files containing the

bacterial sequences of interest must have the same names as the original files. To perform this procedure, users have to create the new indexed database for running BMTagger [11]. Please also note that BMTagger and the other bmtools necessary to run ViromeScan are already present in the "*$PWD/viromescan/tools*" folder. The procedure consists of three steps: (1) to create the indexes for bmfilter with the command "*bmtool –d Bacterial_custom.fa –o Bacterial_custom.bitmask –A 0 –w 18*"; (2) to create the indexes for srprism with the command "*srprism mkindex –i Bacterial_custom.fa –o Bacterial_custom.srprism –M 7168*"; and (3) to create the blast database with the command "*makeblastdb –in Bacterial_custom.fa –dbtype nucl*".

3. In order to get a clean output, the original script foresees an "rm" procedure at the end of the computation. This prevents the elaboration of intermediate data, such as the retrieving of the hit sequences from the .sam output. Users can choose to remove this step of the script to use data for further analysis.

References

1. Thurber RV, Haynes M, Breitbart M et al (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4(4):470–483
2. Duhaime MB, Sullivan MB (2012) Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology 434(2):181–186
3. Willner D, Hugenholtz P (2013) From deep sequencing to viral tagging: recent advances in viral metagenomics. BioEssays 35(5):436–442
4. Lorenzi HA, Hoover J, Inman J et al (2011) The Viral MetaGenome Annotation Pipeline (VMGAP): an automated tool for the functional annotation of viral metagenomic shotgun sequencing data. Stand Genomic Sci 4(3):418–429
5. Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434(2):162–174
6. Wommack KE, Bhavsar J, Polson SW et al (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci 6(3):427–439
7. Roux S, Tournayre J, Mahul A et al (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15:76
8. Colson P, Fancello L, Gimenez G et al (2013) Evidence of the megavirome in humans. J Clin Virol 57(3):191–200
9. Rampelli S, Soverini M, Turroni S et al (2016) ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics 17:165
10. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
11. BMTagger (2011) ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/. Accessed 30 Aug 2012
12. Picard tools website. https://broadinstitute.github.io/picard/. Accessed 30 Aug 2012
13. The NCBI viral genome database. http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?opt=virus&taxid=10239. Accessed 2 June 2015
14. Turnbaugh PJ, Ley RE, Hamady M et al (2007) The human microbiome project. Nature 449(7164):804–810
15. TrimBWAstyle.usingBam.pl (2010) https://github.com/genome/genome/blob/master/lib/perl/Genome/Site/TGI/Hmp/HmpSraProcess/trimBWAstyle.usingBam.pl. Accessed 9 Sept 2012
16. NIH Human Microbiome Project website. http://www.hmpdacc.org. Accessed 17 June 2015

Chapter 15

Shannon Entropy to Evaluate Substitution Rate Variation Among Viral Nucleotide Positions in Datasets of Viral siRNAs

Aysan Ghasemzadeh, Marta Małgorzata ter Haar,
Masoud Shams-bakhsh, Walter Pirovano, and Vitantonio Pantaleo

Abstract

Next-generation sequencing has opened the door to the reconstruction of viral populations and examination of the composition of mutant spectra in infected cells, tissues, and host organisms. In this chapter we present details on the use of the Shannon entropy method to estimate the site-specific nucleotide relative variability of turnip crinkle virus, a positive (+) stranded RNA plant virus, in a large dataset of short RNAs of *Cicer arietinum* L., a natural reservoir of the virus. We propose this method as a viral metagenomics tool to provide a more detailed description of the viral quasispecies in infected plant tissue. Viral replicative fitness relates to an optimal composition of variants that provide the molecular basis of virus behavior in the complex environment of natural infections. A complete description of viral quasispecies may have implications in determining fitness landscapes for host-virus coexistence and help to design specific diagnostic protocols and antiviral strategies.

Key words Shannon entropy, Viral quasispecies, Mutant clouds, *Turnip crinkle virus*, Viral siRNAs, *Cicer arietinum*

1 Introduction

Viral quasispecies (also known as "mutant clouds") are aggregates of closely related viral genomes generated by replication of the RNA virus in the host [1]. Viral mutant clouds may arise as a result of the numerous replication rounds that take place during intracellular amplification associated with high mutation rates of the viral RNA-dependent RNA polymerase, given their lack of proofreading activity [2]. Thus, variant clouds comprise viral variants that deviate from a consensus master genome by one or more single-nucleotide polymorphisms (SNPs) and/or insertions/deletions. Mutant clouds could play a biological role in fitness and virulence, since they can rapidly evolve in changing environments,

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_15, © Springer Science+Business Media, LLC 2018

i.e., leading to the emergence of resistance-breaking strains [3]. During the earlier stages of infection by plant viruses, mutant clouds are established within single infected cells. The spreading of the viral population within a host begins from the primary infected cells to the nearest neighbors in a process known as cell-to-cell movement. Several studies on different virus/host systems have shown that systemic movement implicates population bottlenecks and generation of heterogeneous viral subpopulations in different organs of the same plant [4–8].

Virus-infected plants accumulate 21–24 nucleotide (nt) viral (v) short interfering (si) RNAs, which are generated by the evolutionarily conserved RNA silencing machinery responsible for regulating gene expression and which is also involved in plant immunity against invading nucleic acids [9]. Classic methods of viral diagnostics using antibodies and polymerase chain reaction (PCR), including the more sensitive real-time PCR, may fail to identify new pathogenic and virulent strains; the mutant cloud could mask their presence in analyzed tissues. Moreover, conventional methods are not applicable for emerging viruses with unknown genomes. v-siRNAs can be used (1) to cover known viral genomes by aligning reads to the reference sequences (ref_seq) (i.e., a simple and cost-effective method for the detection of known viruses) [10–12]; (2) for de novo reconstruction of the complete genome of a known plant RNA virus from multiple contigs of v-siRNAs [13–15]; and (3) for the nonhomologous discovery of novel plant infectious entities [16]. Deep sequencing technology could ensure an elevated coverage of every single nucleotide of the viral genome, thus allowing the representation of the viral mutant cloud in the sample. This is particularly true in the case of a high viral titer in tissues that could be further strengthened by the adoption of protocols for enriching short RNAs of viral origin [10].

We conducted a recent survey of viral entities associated with ancient varieties of *Cicer arietinum* L. The plants were infected with Turnip crinkle virus (TCV), a (+) stranded RNA plant virus belonging to the genus Carmovirus (*Tombusviridae* family) [17] used as a model for molecular biology studies on replication and recombination [18]. Herein, we present details of the procedure adopted for estimating the variation of the TCV ref_seq at the single-nucleotide position in a large v-siRNA dataset. The leaves and flowers of *C. arietinum* in an open-field cultivation were used for providing the data. The variation rate was calculated at a single-nucleotide position by Shannon's entropy value [19], a sensitive tool to estimate the diversity of a system and which is commonly employed in multiple DNA alignment using nucleotide frequencies.

2 Material

2.1 RNA Extraction

1. TRIzol® Reagent (Thermo Fisher Scientific).
2. Chloroform.
3. Isopropanol.
4. 70% Ethanol (in DEPC-treated water).
5. RNase-free water.
6. Centrifuge and rotor capable of reaching up to 12,000 × *g*.

2.2 Small RNA Library Preparation

1. NEXTflex Small RNA-Seq Kit v3 (Bio Scientific).
2. 1–10 μg Total RNA in up to 10.5 μL nuclease-free water.
3. Isopropanol.
4. 80% Ethanol.
5. 2, 10, 20, 200, and 1000 μL pipettes.
6. RNase-free pipette tips.
7. Centrifuge and rotor capable of reaching up to 12,000 × *g*.
8. Thin-wall nuclease-free PCR tubes.
9. Nuclease-free 1.5 mL microcentrifuge tubes.
10. Thermocycler.
11. Heat block.
12. Vortex.
13. Magnetic stand.
14. 8% Acrylamide:*bis*-acrylamide (19:1), 1× TBE PAGE gels.
15. 1× TBE buffer.
16. Scalpel.
17. SYBR Gold.
18. Gel documentation instrument (i.e., ChemiDoc Touch Imaging System, Biorad).

2.3 Small RNA Library Sequencing

1. HiSeq 2500 sequencer (Illumina).
2. Bioanalyzer (Agilent).

2.4 Bioinformatics Analysis

1. Workstation, operative system GNU/Linux 14.04.
2. FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) for dataset preparation.
3. https://usegalaxy.org/ for managing data (convert format of data, group and count of sRNAs).
4. Bowtie or Burrows-Wheeler aligner software for alignments [20, 21].
5. Python libraries: Numpy, Collections, Random.

6. SAMTOOLS [22] version 1.3.1 for generation of BAM and PILEUP files from the SAM files generated by Bowtie or Burrows-Wheeler aligners.
7. Custom python script to extract nucleotide frequencies per reference position for the TCV reference sequence from the PILEUP files.

Pseudocode custom script 1

```
Index the reference sequence
For each SAM file:
    –Convert the SAM file to an indexed BAM file
    –Create a file containing base-pair information at each reference position
     –Convert the file with base-pair information to raw counts per reference position
```

8. Custom python script to calculate the Shannon's entropy values for the extracted alignment positions.

Pseudocode custom script 2

```
For each nucleotide counts file:
    For each reference position in nucleotide file:
        Calculate coverage per position with coverage >10
        Calculate entropy per position with coverage >10
        Preform bootstrapping for positions with coverage >20
            Randomly shuffle nucleotides
            Repeat the following 10 times:
                Select 20 nucleotides
                Calculate entropy based on the selected 20 nucleotides
                Remember the obtained entropy value
              Calculate the average entropy of the 10 bootstrapped values
```

2.5 Detection of Turnip Crinkle Virus by PCR Method

1. Thermocycler.
2. PCR primer set: *Forward primer* 5′-ATCCTGAACGAATTCCCTACAAC-3′ from 2362–2376 bp of genome region and *reverse primer* 5′-CCCGTGACTAGCAGAACCT-3′ from 2501–2515 bp of genome region.
3. 10 μM dNTPs.
4. 4 U/μL Taq polymerase.
5. 10× Taq polymerase buffer.

3 Methods

3.1 RNA Extraction

1. Prepare ~80 mg of leaf and flower samples (*see* **Note 1**).
2. Homogenize each sample in a separate ice-cold and sterilized mortar.
3. Add 300 μL Trizol to each sample.
4. Homogenize very soon by grinding.
5. Transfer 200 μL of homogenized sample in a tube containing 800 μL Trizol.
6. Vortex a little and keep for 5 min at room temperature.
7. Add 200 μL chloroform, invert ten times, and keep for 8 min at room temperature.
8. Centrifuge for 15 min at 4 °C at 12,000 × *g*.
9. Transfer 600 μL of supernatant to a tube containing 500 μL isopropanol.
10. Invert ten times, keep for 8 min at room temperature, and centrifuge for 10 min at 4 °C at 12,000 × *g*.
11. Remove supernatant and wash two times with 70% ethanol.
12. Dissolve the RNA pellet in 30–100 μL nuclease-free water.
13. Check the quality of RNA by gel electrophoresis and spectrophotometer.

3.2 Small RNA Library Preparation

1. 10 μg of total RNA was used for library preparation using the NEXTflex Small RNA-Seq Kit v3.
2. Barcoded RNA adapters were ligated to both ends of the sRNAs.
3. The sample was subjected to first-strand cDNA synthesis with RT primer.
4. PCR amplification with NEXTflex™ Universal Primer and NEXTflex™ Barcoded Primer was performed for 25 cycles.
5. The PCR product was size-selected on 8% PAGE gel, and the band 150 bp in size was retrieved.
6. Resultant libraries were cleaned up and prepared for sequencing as indicated by the NEXTflex manual's instructions.

3.3 Small RNA Library Sequencing

1. Biological quality of libraries was controlled by Bioanalyzer.
2. Libraries with high quality were sequenced by HiSeq 2500 sequencer platform.

3.4 Preparation of the Libraries and Alignments

1. Convert fastQ to fastA format with FASTX-toolkit.
2. Clip the 3′ adapter sequence by FASTX-toolkit.
3. Filter sequence size in 16–26 bp range by FASTX-toolkit.

4. Align selected sRNAs with the genomic Ref_Seq of turnip crinkle virus (*see* **Note 2)** using Bowtie aligner and save alignment outputs as SAM format.
5. Deposit alignments in a data repository such as BioProject (https://www.ncbi.nlm.nih.gov/bioproject) (*see* **Note 3**).

3.5 Evaluation of Shannon Entropy

1. Generation of BAM files using SAMTOOLS "view" using SAM files as input.
2. Generation of sorted BAM file using SAMTOOLS "sort" using BAM files as input.
3. Index the sorted BAM files using SAMTOOLS "index."
4. Conversion of sorted/indexed BAM files to PILEUP format using SAMTOOLS "mpileup."
5. Conversion of PILEUP format to nucleotide counts and frequencies per reference position in alignments using a custom python script (pseudocode custom script 1).
6. Generation of entropy values for each sample individually, for each reference position with at least 10× coverage, using a custom python script (pseudocode custom script 2). Equation used to calculate per-base Shannon's entropy: $H_i = - \sum f_{a,i} \times \log_2 f_{a,\,i}$ where $f_{a,i}$ represents the relative frequency of base a at position i. The python script also includes an additional control that uses bootstrapping (*see* **Note 4**).

3.6 Output Retrieval, Representation, and Interpretation

1. The values of the Shannon entropy for the estimated substitution rate variation among viral nucleotide positions are summarized as in Table 1.
2. Each value of the Shannon entropy can be plotted for a visual representation (we have used Microsoft Excel) as in Fig. 1 (*see* **Note 5**).

4 Notes

1. An old local variety of *Cicer arietinum* with black seeds denoted as "Gioia black" has been used for this study.
2. NCBI Reference Sequence: NC_003821.3. Alignment parameters should allow a number of three mismatches for each sRNA.
3. In the case of the present analysis data are stored under the BioProject PRJNA386437 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA386437).
4. The entropy per alignment position with 20 times or more coverage is calculated after randomly selecting 20 aligned nucleotides, and repeating this process ten times. The final sequence entropy is the average of the 10 bootstrap values.

Table 1
Summary of the Shannon entropy analysis for substitution rate variations

				Coverage				Entropy			
Sample	**Reference**	**No. of positions with coverage**	**No. of positions[a]**	**Minimum**	**Maximum**	**Mean**	**Standard**	**Minimum**	**Maximum**	**Mean**	**Standard**
Flowers	gi|635546938|ref|NC_003821.3|	4018	1576	10	105	31.787	18.328	0	0.863	0.022	0.086
Leaves	gi|635546938|ref|NC_003821.3|	4053	1576	11	267	82.251	46.381	0	0.915	0.016	0.060

[a] With coverage ≥20, and where the entropy was >0 for at least one sample after bootstrapping (10×)

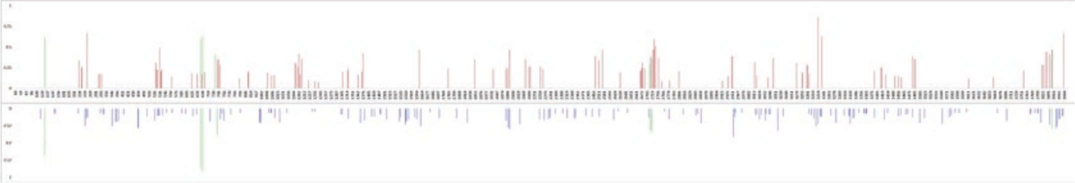


Fig. 1 Shannon entropy values of each position of turnip crinkle virus (TCV) in tissues of *C. arietinum*. Single-nucleotide entropy of TCV genome in flowers (upper panel) and leaves (bottom panel). In green, highly variable nucleotide sites for both viral populations, in red and in blue highly variable nucleotide sites for either flowers' or leaves' viral clouds. X-axis represents the nucleotide position for the Ref seq and Y-axis indicates Shannon entropy values from each covered position

Acknowledgments

This work was supported by the project SaVeGrainINPuglia—"Progetti integrati per la biodiversità" PSR 2014-2020 Reg. (CE) 1698/2005. A.G. and M.S.B were supported by the Ministry of Science, Research and Technology of Iran.

References

1. Domingo E, Sheldon J, Perales C (2012) Viral quasispecies evolution. Microbiol Mol Biol Rev 76(2):159–216. https://doi.org/10.1128/MMBR.05023-11. pii: 76/2/159
2. Drake JW, Holland JJ (1999) Mutation rates among RNA viruses. Proc Natl Acad Sci U S A 96(24):13910–13913
3. Sardanyes J, Elena SF (2011) Quasispecies spatial models for RNA viruses with different replication modes and infection strategies. PLoS One 6(9):e24884. https://doi.org/10.1371/journal.pone.0024884
4. Jridi C, Martin JF, Marie-Jeanne V, Labonne G, Blanc S (2006) Distinct viral populations differentiate and evolve independently in a single perennial host plant. J Virol 80(5):2349–2357. https://doi.org/10.1128/JVI.80.5.2349-2357.2006
5. Sacristan S, Malpica JM, Fraile A, Garcia-Arenal F (2003) Estimation of population bottlenecks during systemic movement of tobacco mosaic virus in tobacco plants. J Virol 77(18):9906–9911
6. French R, Stenger DC (2003) Evolution of Wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. Annu Rev Phytopathol 41:199–214. https://doi.org/10.1146/annurev.phyto.41.052002.095559
7. Li H, Roossinck MJ (2004) Genetic bottlenecks reduce population variation in an experimental RNA virus population. J Virol 78(19):10582–10587. https://doi.org/10.1128/JVI.78.19.10582-10587.2004
8. Monsion B, Froissart R, Michalakis Y, Blanc S (2008) Large bottleneck size in Cauliflower Mosaic Virus populations during host plant colonization. PLoS Pathog 4(10):e1000174. https://doi.org/10.1371/journal.ppat.1000174
9. Pantaleo V (2011) Plant RNA silencing in viral defence. Adv Exp Med Biol 722:39–58. https://doi.org/10.1007/978-1-4614-0332-6_3
10. Pirovano W, Miozzi L, Boetzer M, Pantaleo V (2014) Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to a Cicer arietinum population. Front Microbiol 5:790. https://doi.org/10.3389/fmicb.2014.00790
11. Morelli M, Giampetruzzi A, Laghezza L, Catalano L, Savino VN, Saldarelli P (2017) Identification and characterization of an isolate of apple green crinkle associated virus involved in a severe disease of quince (Cydonia oblonga, Mill.) Arch Virol 162(1):299–306. https://doi.org/10.1007/s00705-016-3074-6
12. Miozzi L, Pantaleo V (2015) Drawing siRNAs of viral origin out from plant siRNAs libraries. Methods Mol Biol 1236:111–123. https://doi.org/10.1007/978-1-4939-1743-3_10
13. Zhang Z, Qi S, Tang N, Zhang X, Chen S, Zhu P, Ma L, Cheng J, Xu Y, Lu M, Wang H, Ding SW, Li S, Wu Q (2014) Discovery of replicating circular RNAs by RNA-seq and computational algorithms. PLoS Pathog 10(12):e1004553. https://doi.org/10.1371/journal.ppat.1004553

14. Giampetruzzi A, Roumi V, Roberto R, Malossini U, Yoshikawa N, La Notte P, Terlizzi F, Credi R, Saldarelli P (2012) A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus Res 163(1):262–268. https://doi.org/10.1016/j.virusres.2011.10.010

15. Seguin J, Rajeswaran R, Malpica-Lopez N, Martin RR, Kasschau K, Dolja VV, Otten P, Farinelli L, Pooggin MM (2014) De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. PLoS One 9(2):e88513. https://doi.org/10.1371/journal.pone.0088513. PONE-D-13-40747 [pii]

16. Wu Q, Wang Y, Cao M, Pantaleo V, Burgyan J, Li WX, Ding SW (2012) Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. Proc Natl Acad Sci U S A 109(10):3938–3943. https://doi.org/10.1073/pnas.1117815109. 1117815109 [pii]

17. Rochon D, Lommel S, Martelli GP, Rubino L, Russo M (2012) Tombusviridae. In: AMQ K, Adams MJ, Carstens EB, Lefkowitz EJ (eds) Virus taxonomy: classification and nomenclature of viruses, vol IX. Elsevier, London, pp 1111–1138

18. Simon AE (1999) Replication, recombination, and symptom-modulation properties of the satellite RNAs of turnip crinkle virus. Curr Top Microbiol Immunol 239:19–36

19. Shannon CE (1997) The mathematical theory of communication. 1963. MD Comput 14(4):306–317

20. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

21. 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(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25. gb-2009-10-3-r25 [pii]

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) Genome Project Data Processing Subgroup. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Chapter 16

Insect Virus Discovery by Metagenomic and Cell Culture-Based Approaches

Finny S. Varghese and Ronald P. van Rij

Abstract

Insects are the most abundant and diverse group of animals on earth, but our knowledge of their viruses is biased toward insect-borne viruses that cause disease in plants, animals, or humans. Recent metagenomic studies and systematic surveys of viruses in wild-caught insects have identified an unanticipated large repertoire of novel viruses and viral sequences. These include new members of existing clades, new clades, and even entirely new virus families. These studies greatly expand the known virosphere in insects, provide opportunities to study virus-host interactions, and generate new insights into virus evolution. In this chapter, we discuss the methods used to identify novel viruses in insects and highlight some notable surprises arising from these studies.

Key words Metagenomics, Transcriptomics, Arbovirus, Insect virus, Small interfering RNA, RNA interference

1 Introduction

Viruses are obligate parasites that infect organisms from every domain of life, ranging from unicellular bacteria to large, multicellular vertebrates including whales [1–3]. Recognized as filterable infectious agents that are smaller than bacteria, viruses were isolated and grown in cell culture for the first time in the early 1900s [2]. With the advent of the electron microscope and increased understanding of the adaptive immune system, microscopy- and serology-based techniques came to be used for virus detection and discovery [4]. Later, techniques became available that detect specific viral nucleic acid sequences, including southern and northern blots, PCR, RT-PCR, and microarrays [4]. Most of these techniques have their own limitations. For example, not all viruses can be grown in cell culture, and for those viruses that can be cultured, specific cell lines may be required. Nucleic acid-based methods can be used to detect non-culturable viruses, but require prior knowledge of the nature of the infecting virus [2, 4]. Now, more than a

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6_16, © Springer Science+Business Media, LLC 2018

century after the word "virus" was first coined by Willem Beijerinck in 1898 [5], new viruses are being discovered at an unanticipated pace through the use of next-generation deep sequencing. This technology overcomes some of the limitations of traditional methods for virus discovery and opens opportunities for cost-effective, unbiased metagenomic surveys for new viruses in environmental samples and from non-symptomatic organisms, including species that have hitherto received little attention [6].

Among the different routes through which virus transmission occurs, insect-borne transmission is arguably the most complex one. Insects are responsible for spreading a multitude of viruses not only to humans, such as the arthropod-borne (arbo-)viruses chikungunya, dengue, and Zika, but also to plants and animals (e.g., tomato spotted wilt virus, bluetongue virus, Schmallenberg virus). In addition, insects carry a large number of viruses that only replicate in their invertebrate hosts (insect-specific viruses). Some of these are known for the pathology they cause in the insect host, such as chronic bee paralysis virus and nuclear polyhedrosis virus [7, 8]. Other insect-specific viruses belong to classical arbovirus families and it has been suggested that they are ancestral to arboviruses [9–11].

Although insects are the most abundant group of animals on the planet and of major ecological, agricultural, and medical importance, our knowledge of insect viruses is limited. There are several incentives to chart the insect virosphere. First, to prepare for future arboviral threats, we need to be aware of the reservoir of viruses in insect vectors that have the potential to switch hosts and infect vertebrates. Second, pathogenic insect viruses can be used for biological control of insect pests [12] and the identification of novel insect viruses may enhance the arsenal of potential biocontrol agents. Third, insect-specific viruses may affect transmission of arboviruses [13] and persistent infections may thus contribute to the poorly understood observation that local mosquito populations differ in vector competence for specific arboviruses [14]. Fourth, insect-specific viruses can be used as a platform for vaccine development and diagnostics, its safety enhanced by the inability to replicate in mammalian cells [15, 16]. For example, a recombinant insect-specific alphavirus, Eilat virus (EILV), expressing the structural proteins of chikungunya virus (CHIKV) elicited a strong and long-lasting neutralizing antibody response to CHIKV [16]. Fifth, and perhaps most importantly, charting the invertebrate virosphere will provide crucial insights into the ecology and evolution of viruses.

Over the last decade, several metagenomic and systematic cell culture-based studies have explored the insect virosphere. These studies discovered novel viruses in existing clades and families, but also identified previously unknown viruses in novel clades and even entirely new virus families. In this chapter, we provide an overview of the commonly used metagenomic and culture-based methods for identifying novel insect viruses. We point out advantages and disadvantages of these methods and highlight some of the key findings of recent virus discovery studies. Detailed discussion of the experimental methods and bioinformatic analyses is beyond the scope of this review; the interested reader is referred to other chapters in this volume.

2 Methods for Virus Discovery

2.1 RNA Sequencing-Based Virus Discovery

Metagenomic studies rely on the unbiased sequencing of (viral) nucleic acids in samples of interest. Already before the wide and affordable availability of next-generation sequencing, transcriptome data have been used for virus discovery, for example through the analyses of expressed sequence tagged (EST) libraries. In this approach, viral RNA from purified virions is reverse-transcribed into cDNA, cloned into EST libraries, sequenced by conventional Sanger sequencing, assembled into contiguous sequences (contigs), followed by bioinformatic analyses based on sequence similarity searches [17]. Next-generation sequencing has greatly improved the efficiency, sensitivity, and throughput of this process by eliminating the rate-limiting cloning step and by increasing sequencing depth, resulting in the rapid generation of millions of sequence reads [17, 18].

The following is a general workflow for next-generation RNA-sequencing (RNA-Seq)-based metagenomic surveys to uncover novel viral sequences in insects; different studies have modified their protocols to best suit their research questions. Field-collected insects are first identified based on morphological traits and pooled into convenient units for further processing, for example, based on (related) species or numbers [19]. In some studies, host species identities were retrospectively confirmed by comparing the sequence of the mitochondrial cytochrome c oxidase subunit I (COI) gene against the NCBI nucleotide and BOLD databases (Barcode of Life Data Systems) [19–23].

Ribosomal RNA and transfer RNA account for up to 90–95% of the total RNA in a cell, also in virus-infected cells [24]. RNA samples, therefore, need to be depleted of rRNA or enriched for mRNA or viral RNA for metagenomic studies. In some studies, samples have been enriched for viral material by eliminating host cell debris and bacterial contamination by centrifugation and filtration, followed by cesium chloride density gradient ultracentrifugation [25]. This approach, however, is labor intensive and can lead to contamination of the samples [26]. Another approach to enrich for viral nucleic acids is to treat the supernatant of insect homogenates with nuclease to degrade nucleic acids that are not protected by viral capsids [27]. As an alternative, total RNA may be subjected to poly(A) RNA selection, using commercial reagents such as the polyA Spin mRNA Isolation kit (New England Biolabs) or the NucleoTrap mRNA kit (Macherey Nagel). Using this approach, only polyadenylated sequences are selected and many viruses that do not produce poly(A)-tailed RNA will be missed. Therefore, the current method of choice for unbiased virus discovery is ribosomal RNA depletion, for which several commercial reagents are available such as the Ribo-Zero rRNA Removal Kit, the Ribo-Zero Gold Epidemiology kit (Illumina), and the RiboMinus kit (Thermo Fisher) [19, 21, 28, 29]. Although RNA-based metagenomic studies have the potential to identify novel DNA viruses [28], studies aimed specifically at discovering DNA viruses may omit the RNA isolation and reverse-transcription steps and generate sequencing libraries from purified DNA directly [25].

Sequencing libraries are prepared from the purified RNA, using commercial reagents such as the TruSeq mRNA Library Prep kit (Illumina) or ScriptSeq RNA-Seq Library Preparation kit (Epicentre). Library preparation for next-generation sequencing is reviewed in ref. [30]. Briefly, the procedure involves the following steps: physical or chemical fragmentation of total RNA, reverse-transcription using random primers for cDNA synthesis, end-repairing, A-tailing, 5′ adaptor ligation, PCR amplification of the adapter-bound sequences, purification and quantification of the library, followed by high-throughput sequencing. Currently, the Illumina (Hi-Seq or Mi-Seq) sequencing platform seems to be most often used, due to its relatively low cost and high sequencing depth [31]. However, the Roche 454 pyrosequencing platform has also been used [25, 27, 32]. Recently, third-generation sequencing technologies like the single-molecule real-time sequencing (Pacific Biosciences) [33] and Oxford Nanopore [34] have been developed. These platforms produce longer read lengths than the 50–100 nt reads of Illumina (Roche, ~700–1000 bp; Pacific Biosciences, ~10 kb), which may provide higher consensus accuracy and more uniform coverage, thereby reducing variation in the obtained sequences and facilitating assembly of viral genomes.

Prior to bioinformatics analyses, the raw sequence reads are processed using freely or commercially available packages for base calling (e.g., Bustard, BayesCall, Seraphim [35]), removal of adaptor sequences (e.g., CutAdapt, Skewer, AdapterRemoval [36–38]), and trimming of low-quality reads (e.g., Sickle, Trim Galore, Trimmomatic [39]). The trimmed and quality-controlled reads are then *de novo* assembled into contigs using assembly packages, such as Trinity or Velvet [40–42].

Assembled contigs are analyzed for similarity to known viruses by BLAST searches (blastx, tblastx, blastn) against the NCBI virus genome database or user-curated databases of specific virus families of interests (e.g., [19, 29]). Contigs that align to viral sequences with a desired E-value are extracted (usually a threshold $<1 \times 10^{-5}$

is used). To rule out contaminating reads from host or bacterial sources, the contigs are once again BLASTed against the NCBI nonredundant (nr) database. Confirmed viral contigs are then merged by identifying unassembled overlaps between neighboring contigs using the SeqMan program (part of the Lasergene software package). Alternatively, the reads can first be BLASTed against a database of viral sequences and high-scoring reads can then be used for assembly of viral genomes. Viral genome assemblies often have missing gaps, which are then filled by PCR, RT-PCR, and/or RACE (random amplification of cDNA ends) analyses. The obtained viral sequences are subjected to phylogenetic analyses to establish their relationship to known viruses. This framework has been successful in identifying novel viruses from a range of insects and other invertebrate species, including *Drosophila melanogaster*, mosquitoes, and honeybees (for examples, see Table 1).

Although next-generation sequencing has emerged as a powerful tool for insect virus discovery, there are several limitations to this approach. First, prior knowledge of related viruses is required and it is not possible to identify viruses with no homology to known viruses. Yet, all RNA virus genomes code for an RNA-dependent RNA polymerase (RdRP) and protein blasts should be able to detect relatively distant viruses using this conserved domain. For example, this strategy led to the discovery of a novel group of negative-sense RNA arthropod viruses from a proposed new family *Chuviridae* [19]. Second, the majority of the assembled genomes are partial sequences, which may complicate downstream phylogenetic analyses and assignment to a taxonomic class. Third, there is an inevitable bias toward detecting viruses that are present at high titers in the sample and other, low-abundance viruses may escape detection (although this is also true for other virus discovery methods). Fourth, in some cases, trimming of low-quality reads can deplete them of viral sequences. For example, the full-length genome of the Aphis glycines virus from the soybean aphid could only be assembled from low-quality reads prior to standard trimming procedures [18]. Finally, RNA samples from insects are often contaminated with RNA of associated bacteria, fungi, unrecognized parasite infections, gut contents, or surface contamination, which makes it difficult to ascertain that a newly identified virus constitutes an active infection of the insect host. Therefore, conclusions about host associations based solely on viral metagenomic data should be interpreted with caution and are ideally complemented with additional experimental support. Detection of virus-derived small RNAs is one approach to address this concern [28]; this approach can also be used as a strategy for virus discovery and is discussed in the next section.

2.2 Small RNA-Based Virus Discovery

RNA interference (RNAi) is a crucial antiviral defense system that depends on the production of 21-nt small interfering RNAs (siRNAs) from viral double-stranded RNA (dsRNA) by the host

Table 1
Examples of recent metagenomic studies for insect virus discovery

Authors [reference] Method	Host (common name)	Host (Linnaean name)	Material	Sequencing platform	Enrichment for viral nucleic acids	Viruses detected/ identified[a]	Classification (alphabetic order)
Small RNAs							
Webster et al. [28]	Fruit fly	*Drosophila melanogaster*	Adult flies	Illumina	Periodate oxidation/ beta elimination	Chaq virus, Galbut virus	Unclassified
Wu et al. [45]	Mosquito	*Aedes aegypti*	Adult mosquito (lab colony)	Illumina	–	Mosquito nodavirus	Nodaviridae
Ma et al. [54]	Mosquito	*Culex tritaeniorhynchus*	Adult mosquito	Illumina GA-II	–	Culex tritaeniorhynchus densovirus	Parvoviridae
Carissimo et al. [84]	Mosquito	*Anopheles coluzzii*	Adult mosquito (lab colony)	Illumina	–	Anopheles C virus, Anopheles cypovirus	Dicistroviridae, Reoviridae
Aguiar et al. [56]	Fruit fly, mosquito, sand fly	*Drosophila melanogaster, Aedes aegypti, Lutzomyia longipalpis*	Adult insects (lab strain for Drosophila)	Illumina	–	2 new viruses in fruit flies, 3 in sand flies, 2 in mosquitoes (dsRNA, (+) and (−)RNA viruses)	Unclassified, Bunyavirales, Nodaviridae, Reoviridae

RNA-Seq							
Webster et al. [28]	Fruit fly	*Drosophila melanogaster*	Adult flies	Illumina	rRNA depletion	4 previously known RNA viruses; >20 novel RNA viruses; 1 novel DNA virus	Unclassified, Flaviviridae, Iflaviridae, Negevirus, Nodaviridae, Nudiviridae, Partitiviridae, Permutotetraviridae, Picornavirales, Phasmaviridae, Reoviridae
Cox-Foster et al. [32]	Honeybee	*Apis mellifera*	Adult bees, royal jelly	Roche 454	–	7 previously known (+)RNA viruses	Unclassified, Dicistroviridae, Iflavividae
Runckel et al. [85]	Honeybee	*Apis mellifera*	Adult bees	Illumina GA-II	Poly(A) selection or no enrichment	3 previously known (+)RNA viruses; 4 novel (+)RNA viruses	Unclassified, Dicistroviridae, Iflaviviridae
Remnant et al. [86]	Honeybee	*Apis mellifera*	Adult bees (thoraxes)	Illumina HiSeq	rRNA depletion	5 previously known (+)RNA viruses; 7 novel (+) and (−) RNA viruses	Unclassified, Bunyaviridae, Dicistroviridae, Iflavividae, Rhabdoviridae
Ng et al. [25]	Mosquito	Mosquitoes (undefined)	Adult mosquitoes	Roche 454	Virion purification by CsCl gradient	Broad range of animal, plant, insect, and bacterial species[b]	

(continued)

Table 1 (continued)

Authors [reference] Method	Host (common name)	Host (Linnaean name)	Material	Sequencing platform	Enrichment for viral nucleic acids	Viruses detected/ identified[a]	Classification (alphabetic order)
Shi et al. [27]	Mosquito	*Culex tritaeniorhynchus, Anopheles sinensis, Armigeres subalbatus, Culex quinquefasciatus*	Adult mosquitoes	Roche 454	Nuclease treatment of mosquito lysate	12 novel viruses (ssDNA, (+)RNA, (−)RNA)	Anneloviridae, Dicistroviridae, Parvoviridae, Rhabdoviridae
Ballinger et al. [87]	Phantom midge	*Chaoborus trivittatus; Chaoborus cf. flavivans*	Larvae	Illumina HiSeq	rRNA depletion	Kigluaik phantom virus, Nome phantom virus	Phasmaviridae
Shi et al. [29]	Mixed pools of invertebrate species	–	–	Illumina HiSeq	rRNA depletion	5 novel segmented Jingmen viruses; 12 novel flavi-like viruses	Unclassified (Jingmenviruses and Flavivirus-like viruses)
Li et al. [19]	Mixed pools of 70 arthropod species	–	–	Illumina HiSeq	rRNA depletion	112 novel viruses (−) RNA viruses	Arenaviridae, Bunyavirales, Chuviridae, Mononegavirales, Orthomyxoviridae
Shi et al. [21]	Mixed pools of 220 invertebrate species	–	–	Illumina HiSeq	rRNA depletion or poly(A)-selection	1445 novel RNA viruses	All major clades of RNA viruses

[a] (+)RNA, positive-sense RNA virus; (–)RNA, negative-sense RNA virus. For most of the identified viruses, their formal taxonomic status, e.g., whether they represent new virus species or strains of existing species, awaits to be established

[b] The authors suggest that the identified animal and plant viruses may have been obtained from vertebrates during blood feeding and feeding on plant nectar

ribonuclease Dicer-2. It is well established that RNAi targets the major clades of RNA viruses in insects, including positive- and negative-sense single-stranded RNA viruses and dsRNA viruses [43]. Viral siRNAs (vsiRNA) generally map across the entire length of the viral genome [43], thus providing the opportunity to reconstitute viral genome sequences from vsiRNAs [44, 45]. Indeed, initial analyses of small RNA deep sequencing datasets of *Drosophila* cell lines resulted in the recovery of previously known viruses (Drosophila A virus, Drosophila C virus, Drosophila X virus) as cell culture contaminants, as well as the identification of novel viruses such as American nodavirus, Drosophila totivirus, and Drosophila birnavirus [44–46]. In line with the observations that vsiRNAs map across viral RNA genomes in experimental infections, de novo reconstitution of viral genomes from vsiRNAs in some cases resulted in the recovery of entire viral genomes [47], making siRNA profiling a promising tool for insect virus discovery.

An advantage of vsiRNA profiling is that it taps into an antiviral system of the host, and thus naturally enriches for sequences of foreign origin. After their production, vsiRNAs are incorporated into Argonaute 2 (AGO2) in the RNA-induced silencing complex (RISC), after which the 2′OH of the 3′ terminal nucleotide is methylated by the methyltransferase Hen-1 [48]. 2′O methylation renders small RNAs resistant to periodate oxidation/β-elimination. Thus, AGO2-associated siRNAs are protected from β-elimination, whereas microRNAs, which typically associate with Argonaute 1 in a miRISC, are not [48]. As β-eliminated small RNAs are unclonable by standard methods for small RNA library preparation, this provides an opportunity to enrich small RNA libraries for AGO2-associated, putative viral siRNAs [28]. Thus, small RNAs that do not map to the host genome, but are resistant to β-elimination, are strong candidates for being of viral origin.

It has been shown that vsiRNAs are also produced during DNA virus infection of insects [49] in several infection models: the baculoviruses Helicoverpa armigera single nucleopolyhedrovirus (HaSNPV) and Autographa californica multiple nucleopolyhedrosis virus (AcMNPV) in the cotton bollworm moth (*Helicoverpa armigera*) and the fall army worm (*Spodoptera frugiperda*) respectively, invertebrate iridovirus-6 in *Drosophila melanogaster* [50–52], and vaccinia virus in a *Drosophila* cell line [53]. Typically, vsiRNA profiles for DNA viruses are less uniformly distributed than those of RNA viruses, and it has been proposed that dsRNA from overlapping transcripts from both genomic strands or RNA structures in viral transcripts give rise to vsiRNA production [49, 50, 52, 53]. Thus, whereas small RNA-based metagenomics has the potential to uncover DNA viruses, it is unlikely to uncover complete genomes. Yet, it could pave the way for PCR-based approaches to recover the complete viral genome, as was recently used for the identification of Culex tritaeniorhynchus densovirus from wild-caught mosquitoes [54].

The workflow for small RNA-based recovery of viral sequences is similar to the workflow for RNA-Seq, with several modifications. Library preparation for small RNA sequencing requires size selection of RNAs in the size range of ~17–30 nt and an optional periodate oxidation/β-elimination treatment. Adaptors are ligated to the 5′ and 3′ ends, followed by PCR amplification of the adapter-ligated sequences and gel purification of the library. Several commercial reagents are available to prepare small RNA libraries, such as the TruSeq small RNA Library Preparation kit (Illumina). Small RNA sequence reads may be assembled using programs like Velvet, which was specifically designed for shorter reads, and then analyzed for sequence similarity to reference databases as described for RNA-Seq data (Subheading 2.1).

Small RNA profiling of insects has been successfully used not only to identify novel viral entities, but also to verify the insect-host association of newly identified viruses and to deduce a viral origin of sequences detected by RNA-Seq that lack similarity to reference sequences [28]. For a more detailed list, please refer to Table 1. Recently, Aguiar et al. proposed a novel, sequence-independent method to infer a viral origin of contigs without similarity to known viruses [55, 56]. These authors noted that small RNAs of different viruses in their datasets show specific size profiles, which were also different from the profiles of fungi or bacterial derived contigs. Humaita-Tubiacanga virus, for example, showed a strong peak at 21 nt (reflecting Dicer-2-dependent siRNAs), whereas the bunyavirus Phasi Charoen-like virus showed a size profile reminiscent of processing by both the siRNA- and PIWI-associated RNA (piRNA) pathways [57]. Other viruses did not show enrichment of specific size classes, perhaps reflecting random degradation of viral RNA or the activity of virus-encoded suppressors of siRNA production. Size profiles were presented as heatmaps and subjected to hierarchical clustering, in which contigs derived from the same virus appeared in single clusters [55, 56]. The authors thus proposed that unknown contigs appearing in viral clusters are likely of viral origin and, thus, that small RNA signatures can be used to identify novel viruses lacking homologous sequences in reference databases.

It has recently become clear that both vertebrates and invertebrates carry in their genomes sequences derived from non-reverse-transcribing RNA viruses, called endogenous viral elements (EVEs) or non-retroviral integrated RNA viruses (NIRVs) [58–60]. Moreover, it seems that mosquito genomes encode an extraordinary amount of such EVEs, some of which seem to be transcribed and give rise to small RNAs [61–65]. Consequently, both RNA-Seq and small RNA-based metagenomic approaches have the potential to uncover such EVE sequences, which obviously do not reflect actively replicating viruses. However, it seems that EVEs rarely cover entire viral genomes and if sequence reads are first eliminated by mapping to the host genome (if available), EVE-derived reads will be discarded. EVEs may, however, also be useful for virus discovery, following the approach of Shi et al. [21], who used sequence similarity to EVEs to infer host tropism of newly identified viral sequences in samples containing a mix of host species.

2.3 Cell Culture-Based Methods for Virus Discovery

Even though metagenomic studies are producing new viral sequences at an unanticipated pace, classical cell culture-based methods for virus detection/identification continue to play an important role in virus discovery in insects. Virus culture remains the only way to obtain the infectious virus isolate and to analyze its properties, including—but not limited to—host and tissue tropism, pathogenicity, and other aspects of virus-host interactions that cannot be predicted from sequence alone.

Recently, several groups have performed large cell culture-based surveys for viruses in field-collected mosquitoes (e.g., [22, 66–75]). In this approach, mosquitoes were collected, identified to the species level based on morphological traits (sometimes in combination with sequence analysis of the COI gene), and pooled into convenient group sizes (usually 10–100). Adult female mosquitoes have most often been surveyed, likely because of their importance for arbovirus transmission, but adult males or larvae have also been collected (e.g., [76]). The mosquito samples are then homogenized, filtered, and inoculated onto *Aedes albopictus* C6/36 cells, which are then monitored for cytopathic effects (CPE). These cells lack a functional RNAi response, rendering them highly sensitive to virus infection [77]. Indeed, many viruses induce CPE in C6/36 cells, but not in the RNAi-competent U4.4 cell line derived from the same mosquito species (e.g., [71]). Thus, the RNAi defective phenotype of the C6/36 cell line is critical for the success of cell culture-based virus discovery, as many viruses would escape detection if CPE in an RNAi-competent cell line had been used as read-out. After CPE-inducing mosquito pools have been identified, these can be prioritized for further study based on the presence or absence of members of specific virus families (e.g., by immunofluorescence for flavivirus antigen [66] or by RT-PCR [73]). The nature of the infecting virus can subsequently be established by virion morphology using electron microscopy and by (next-generation) sequencing.

A major limitation is that this approach for virus discovery is labor intensive and requires a susceptible cell line. Whereas the first concern can be met by hard work and persistence, the latter is more problematic. In fact, it is inevitable that cell culture-based methods under-appreciate virus diversity. First, viruses can be highly cell type and host specific, and the cell line used for screening may simply not support replication of specific viruses, especially when they naturally infect evolutionarily distant hosts. Second, for many insect species no or only a limited number of cell lines are available. Moreover, if they are available, they are unlikely to be RNAi defective, which, as outlined earlier, may be essential for virus discovery (of note, it should now be feasible to generate RNAi-defective cells using CRISPR/Cas9 technology). Third, viruses that establish a persistent, non-cytopathic infection will not be detected in a CPE-based readout.

Despite these limitations, cell culture-based screening has proven to be a powerful tool to identify new viruses, especially from mosquitoes (for reviews, see [78–80]). Some notable findings include the discovery of a clade of insect-specific flaviviruses that cluster with mosquito-borne flaviviruses [67, 81], identification of insect-specific viruses in families that were thought to only infect vertebrates (Cavally virus and Nam Dinh virus, founding members of the family *Mesoniviridae* in the order *Nidovirales* [73, 75, 82]), and identification of novel bunyaviruses, Herbert virus, Gouléako virus, Jonchet virus, and Ferak virus [70–72] that are now type species of newly established genera (*Herbevirus*, *Goukovirus*) and families (*Jonviridae*, *Feraviridae*) in the order *Bunyavirales*.

3 Conclusion: New Insights into Virus Diversity

Systematic virus discovery programs have revolutionized our understanding of virus diversity and evolution. For example, in a metagenomic survey for negative-sense RNA viruses in invertebrates, Li et al. discovered a monophyletic group of viruses they named chuviruses, in a proposed family *Chuviridae*. Phylogenetic analyses of sequences of the L-segment (RdRp) revealed that *Chuviridae* occupy a position intermediate to that of segmented and non-segmented negative-sense RNA viruses. Moreover, chuviruses seem to display a variety of genome organizations, including non-segmented, bi-segmented, and even circular RNA genomes. These results suggest that *Chuviridae* might be an evolutionary link between virus taxa with different genome organizations [19]. In another study with a similar setup, Shi et al. identified close to 1500 new viruses in a wide range of invertebrate species and, although their formal taxonomic status has not been established, phylogenetic analyses of the RdRP genes suggest that many of these could represent new virus families [21].

Recent hypothesis-free virus discovery studies show that invertebrates carry an unprecedented diversity of viruses. These studies have filled phylogenetic gaps in current taxonomies by identifying novel virus lineages. Moreover, these studies provide support for an invertebrate origin for several clades of vertebrate-pathogenic viruses by showing that invertebrate viruses are in basal phylogenetic relationship to vertebrate viruses and that their diversity engulfs that of vertebrate virus clades [9, 19, 21, 29, 71, 78–80].

Both cell culture-based and deep sequencing-based approaches will remain invaluable in attempts to comprehensively chart the insect virosphere. They should be considered complementary approaches, the power of which relies on their ability to yield infectious virus isolates and their high sensitivity and broad applicability. Culture-based approaches have thus far mostly been used in mosquitoes. This is due to their importance as vectors of animal and human diseases, but likely also due to the lack of reagents for many other insect species. A challenge for the future is the development of robust virus isolation tools for other insects, analogous to the frequently used and highly susceptible C6/36 cell line. Yet, it is unrealistic to expect that experimental infection models can be generated for the wide variety of invertebrate species that can be assessed by metagenomic approaches. In this respect, it is an important step forward that the International Committee on Taxonomy of Viruses (ICTV) has endorsed the proposal to include viruses that are known only from metagenomic sequences in the official virus taxonomy [83]. The ongoing application of these powerful techniques across and beyond the class *Insecta* will surely yield many more surprises.

Acknowledgements

We thank members of the Van Rij laboratory for fruitful discussions. This work was financially supported by a Consolidator Grant from the European Research Council under the European Union's Seventh Framework Programme (grant no. ERC CoG 615680).

References

1. Flint J, Racaniello VR, Rall GF, Skalka AM (2015) Principles of virology, 4th edn. American Society of Microbiology
2. Kapoor A, Lipkin WI (2001) Virus discovery in the 21st century. In: eLS. Wiley. https://doi.org/10.1002/9780470015902.a0023621
3. Hinshaw VS, Bean WJ, Geraci J, Fiorelli P, Early G, Webster RG (1986) Characterization of two influenza a viruses from a pilot whale. J Virol 58:655–656
4. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77
5. Wilkinson L (2001) Beijerinck, Martinus Willem. In: eLS. Wiley. https://doi.org/10.1038/npg.els.0002363
6. Lipkin WI, Firth C (2013) Viral surveillance and discovery. Curr Opin Virol 3:199–204
7. Ribiere M, Olivier V, Blanchard P (2010) Chronic bee paralysis: a disease and a virus like no other? J Invertebr Pathol 103(Suppl 1):S120–S131
8. Myers JH, Cory JS (2016) Ecology and evolution of pathogens in natural populations of Lepidoptera. Evol Appl 9:231–247
9. Cook S, Moureau G, Kitchen A, Gould EA, de Lamballerie X, Holmes EC, Harbach RE

(2012) Molecular evolution of the insect-specific flaviviruses. J Gen Virol 93:223–234

10. Halbach R, Junglen S, van Rij RP (2017) Mosquito-specific and mosquito-borne viruses: evolution, infection, and host defense. Curr Opin Insect Sci 22:16–27
11. Junglen S (2016) Evolutionary origin of pathogenic arthropod-borne viruses-a case study in the family Bunyaviridae. Curr Opin Insect Sci 16:81–86
12. Lacey LA, Grzywacz D, Shapiro-Ilan DI, Frutos R, Brownbridge M, Goettel MS (2015) Insect pathogens as biological control agents: back to the future. J Invertebr Pathol 132:1–41
13. Nasar F, Erasmus JH, Haddow AD, Tesh RB, Weaver SC (2015) Eilat virus induces both homologous and heterologous interference. Virology 484:51–58
14. Chouin-Carneiro T, Vega-Rua A, Vazeille M, Yebakima A, Girod R, Goindin D, Dupont-Rouzeyrol M, Lourenco-de-Oliveira R, Failloux AB (2016) Differential susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika virus. PLoS Negl Trop Dis 10:e0004543
15. Erasmus JH, Needham J, Raychaudhuri S, Diamond MS, Beasley DW, Morkowski S, Salje H, Fernandez Salas I, Kim DY, Frolov I, Nasar F, Weaver SC (2015) Utilization of an Eilat virus-based chimera for serological detection of chikungunya infection. PLoS Negl Trop Dis 9:e0004119
16. Erasmus JH, Auguste AJ, Kaelber JT, Luo H, Rossi SL, Fenton K, Leal G, Kim DY, Chiu W, Wang T, Frolov I, Nasar F, Weaver SC (2017) A chikungunya fever vaccine utilizing an insect-specific virus platform. Nat Med 23:192–199
17. Liu S, Vijayendran D, Bonning BC (2011) Next generation sequencing technologies for insect virus discovery. Virus 3:1849–1869
18. Liu S, Chen Y, Bonning BC (2015) RNA virus discovery in insects. Curr Opin Insect Sci 8:54–61
19. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife 4. https://doi.org/10.7554/eLife.05378
20. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3:294–299
21. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ (2016) Redefining the invertebrate RNA virosphere. Nature 540:539–543
22. Huhtamo E, Moureau G, Cook S, Julkunen O, Putkuri N, Kurkela S, Uzcategui NY, Harbach RE, Gould EA, Vapalahti O, de Lamballerie X (2012) Novel insect-specific flavivirus isolated from northern Europe. Virology 433:471–478
23. Ratnasingham S, Hebert PD (2007) Bold: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364
24. Lodish H, Berk A, Zipursky SL, Matsudaira P, Darnell J (2002) Molecular cell biology, 4th edn. W.H. Freeman, New York
25. Ng TF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, Anthony S, Ruan Y, Rohwer F, Breitbart M (2011) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6:e20579
26. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J Jr, Delwart EL, Chiu CY (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol 87:11966–11977
27. Shi C, Liu Y, Hu X, Xiong J, Zhang B, Yuan Z (2015) A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. PLoS One 10:e0129845
28. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui JM, Bayne EH, Longdon B, Buck AH, Lazzaro BP, Akorli J, Haddrill PR, Obbard DJ (2015) The discovery, distribution, and evolution of viruses associated with Drosophila melanogaster. PLoS Biol 13:e1002210
29. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X, Qin XC, Widen SG, Wood TG, Tesh RB, Xu J, Holmes EC, Zhang YZ (2015) Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. J Virol 90:659–669
30. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56:61–64, 66, 68, passim
31. Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M (2016) Challenges in the analysis of viral metagenomes. Virus Evol 2:vew022. https://doi.org/10.1093/ve/vew022

32. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui L, Hutchison SK, Simons JF, Egholm M, Pettis JS, Lipkin WI (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318:283–287

33. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

34. Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30:326–328

35. Ledergerber C, Dessimoz C (2011) Base-calling for next-generation sequencing platforms. Brief Bioinform 12:489–497

36. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12

37. Jiang HS, Lei R, Ding SW, Zhu SF (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinform 15:182

38. Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 9:88

39. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

40. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

41. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512

42. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

43. Bronkhorst AW, van Rij RP (2014) The long and short of antiviral defense: small RNA-based immunity in insects. Curr Opin Virol 7:19–28

44. Vodovar N, Goic B, Blanc H, Saleh MC (2011) In silico reconstruction of viral genomes from small RNAs improves virus-derived small interfering RNA profiling. J Virol 85:11016–11021

45. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li WX, Ding SW (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc Natl Acad Sci U S A 107:1606–1611

46. van Mierlo JT, van Cleef KW, van Rij RP (2010) Small silencing RNAs: piecing together a viral genome. Cell Host Microbe 7:87–89

47. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7

48. Horwich MD, Li C, Matranga C, Vagin V, Farley G, Wang P, Zamore PD (2007) The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. Curr Biol 17:1265–1272

49. Bronkhorst AW, Miesen P, van Rij RP (2013) Small RNAs tackle large viruses: RNA interference-based antiviral defense against DNA viruses in insects. Fly (Austin) 7:216–223

50. Bronkhorst AW, van Cleef KW, Vodovar N, Ince IA, Blanc H, Vlak JM, Saleh MC, van Rij RP (2012) The DNA virus invertebrate iridescent virus 6 is a target of the Drosophila RNAi machinery. Proc Natl Acad Sci U S A 109:E3604–E3613

51. Jayachandran B, Hussain M, Asgari S (2012) RNA interference as a cellular defense mechanism against the DNA virus baculovirus. J Virol 86:13729–13734

52. Kemp C, Mueller S, Goto A, Barbier V, Paro S, Bonnay F, Dostert C, Troxler L, Hetru C, Meignin C, Pfeffer S, Hoffmann JA, Imler JL (2013) Broad RNA interference-mediated antiviral immunity and virus-specific inducible responses in Drosophila. J Immunol 190:650–658

53. Sabin LR, Zheng Q, Thekkat P, Yang J, Hannon GJ, Gregory BD, Tudor M, Cherry S

(2013) Dicer-2 processes diverse viral RNA species. PLoS One 8:e55458

54. Ma M, Huang Y, Gong Z, Zhuang L, Li C, Yang H, Tong Y, Liu W, Cao W (2011) Discovery of DNA viruses in wild-caught mosquitoes using small RNA high throughput sequencing. PLoS One 6:e24758

55. Aguiar ER, Olmo RP, Marques JT (2016) Virus-derived small RNAs: molecular footprints of host-pathogen interactions. Wiley Interdiscip Rev RNA 7:824–837

56. Aguiar ER, Olmo RP, Paro S, Ferreira FV, de Faria IJ, Todjro YM, Lobo FP, Kroon EG, Meignin C, Gatherer D, Imler JL, Marques JT (2015) Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. Nucleic Acids Res 43:6191–6206

57. Miesen P, Joosten J, van Rij RP (2016) PIWIs go viral: Arbovirus-derived piRNAs in vector mosquitoes. PLoS Pathog 12:e1006017

58. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6:e1001191

59. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM, Tomonaga K (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463:84–87

60. Taylor DJ, Bruenn J (2009) The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol 7:88

61. Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Tu Z, Van Rij RP, Bonizzoni M (2017) Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. BMC Genomics 18:512

62. Chen XG, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J, Armbruster P, Huang X, Yang Y, Zhang H, He W, Peng H, Liu Y, Wu K, Chen J, Lirakis M, Topalis P, Van Leeuwen T, Hall AB, Jiang X, Thorpe C, Mueller RL, Sun C, Waterhouse RM, Yan G, Tu ZJ, Fang X, James AA (2015) Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci U S A 112:E5907–E5915

63. Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, Belhouchet M, Lemasson JJ, de Micco P, de Lamballerie X (2004) Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J Gen Virol 85:1971–1980

64. Suzuki Y, Frangeul L, Dickson LB, Blanc H, Verdier Y, Vinh J, Lambrechts L, Saleh MC (2017) Uncovering the repertoire of endogenous flaviviral elements in Aedes mosquito genomes. J Virol 91:e00571

65. Andino R, Zach W, Dolan PT, Kunitomi M, Tassato M (2017) Long-read assembly of the Aedes aegypti genome reveals the nature of heritable adaptive immunity sequences. Curr Biol 27:3511–3519

66. Huhtamo E, Cook S, Moureau G, Uzcategui NY, Sironen T, Kuivanen S, Putkuri N, Kurkela S, Harbach RE, Firth AE, Vapalahti O, Gould EA, de Lamballerie X (2014) Novel flaviviruses from mosquitoes: mosquito-specific evolutionary lineages within the phylogenetic group of mosquito-borne flaviviruses. Virology 464–465:320–329

67. Junglen S, Kopp A, Kurth A, Pauli G, Ellerbrok H, Leendertz FH (2009) A new flavivirus and a new vector: characterization of a novel flavivirus isolated from uranotaenia mosquitoes from a tropical rain forest. J Virol 83:4462–4468

68. Junglen S, Kurth A, Kuehl H, Quan PL, Ellerbrok H, Pauli G, Nitsche A, Nunn C, Rich SM, Lipkin WI, Briese T, Leendertz FH (2009) Examining landscape factors influencing relative distribution of mosquito genera and frequency of virus infection. EcoHealth 6:239–249

69. Lv X, Mohd Jaafar F, Sun X, Belhouchet M, Fu S, Zhang S, Tong SX, Lv Z, Mertens PP, Liang G, Attoui H (2012) Isolates of Liao ning virus from wild-caught mosquitoes in the Xinjiang province of China in 2005. PLoS One 7:e37732

70. Marklewitz M, Handrick S, Grasse W, Kurth A, Lukashev A, Drosten C, Ellerbrok H, Leendertz FH, Pauli G, Junglen S (2011) Gouleako virus isolated from West African mosquitoes constitutes a proposed novel genus in the family Bunyaviridae. J Virol 85:9227–9234

71. Marklewitz M, Zirkel F, Kurth A, Drosten C, Junglen S (2015) Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. Proc Natl Acad Sci U S A 112:7536–7541

72. Marklewitz M, Zirkel F, Rwego IB, Heidemann H, Trippner P, Kurth A, Kallies R, Briese T, Lipkin WI, Drosten C, Gillespie TR, Junglen S (2013) Discovery of a unique novel clade of mosquito-associated bunyaviruses. J Virol 87:12850–12865

73. Nga PT, Parquet Mdel C, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K, Ichinose A, Snijder EJ, Morita K, Gorbalenya AE (2011) Discovery of the first insect nidovirus, a missing evolutionary link in

the emergence of the largest RNA virus genomes. PLoS Pathog 7:e1002215

74. Quan PL, Junglen S, Tashmukhamedova A, Conlan S, Hutchison SK, Kurth A, Ellerbrok H, Egholm M, Briese T, Leendertz FH, Lipkin WI (2010) Moussa virus: a new member of the Rhabdoviridae family isolated from Culex decens mosquitoes in Cote d'Ivoire. Virus Res 147:17–24
75. Zirkel F, Kurth A, Quan PL, Briese T, Ellerbrok H, Pauli G, Leendertz FH, Lipkin WI, Ziebuhr J, Drosten C, Junglen S (2011) An insect nidovirus emerging from a primary tropical rainforest. MBio 2:e00077-00011
76. Haddow AD, Guzman H, Popov VL, Wood TG, Widen SG, Haddow AD, Tesh RB, Weaver SC (2013) First isolation of Aedes flavivirus in the Western Hemisphere and evidence of vertical transmission in the mosquito Aedes (Stegomyia) albopictus (Diptera: Culicidae). Virology 440:134–139
77. Brackney DE, Scott JC, Sagawa F, Woodward JE, Miller NA, Schilkey FD, Mudge J, Wilusz J, Olson KE, Blair CD, Ebel GD (2010) C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. PLoS Negl Trop Dis 4:e856
78. Blitvich BJ, Firth AE (2015) Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. Viruses 7:1927–1959
79. Bolling BG, Weaver SC, Tesh RB, Vasilakis N (2015) Insect-specific virus discovery: significance for the Arbovirus community. Viruses 7:4911–4928
80. Junglen S, Drosten C (2013) Virus discovery and recent insights into virus diversity in arthropods. Curr Opin Microbiol 16:507–513
81. Huhtamo E, Putkuri N, Kurkela S, Manni T, Vaheri A, Vapalahti O, Uzcategui NY (2009) Characterization of a novel flavivirus from mosquitoes in northern europe that is related to mosquito-borne flaviviruses of the tropics. J Virol 83:9532–9540
82. Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE (2012) Mesoniviridae: a proposed new family in the order Nidovirales formed by a single species of mosquito-borne viruses. Arch Virol 157:1623–1628
83. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM (2017) Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev. Microbiol 15:161–168

84 Carissimo G, Eiglmeier K, Reveillaud J, Holm I, Diallo M, Diallo D, Vantaux A, Kim S, Menard D, Siv S, Belda E, Bischoff E, Antoniewski C, Vernick KD (2016) Identification and characterization of two novel RNA viruses from Anopheles gambiae species complex mosquitoes. PLoS One 11:e0153881

85 Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, Andino R, DeRisi JL (2011) Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. PLoS One 6:e20656

86 Remnant EJ, Shi M, Buchmann G, Blacquiere T, Holmes EC, Beekman M, Ashe A (2017) A diverse range of novel RNA viruses in geographically distinct honey bee populations. J Virol 91:e00158

87 Ballinger MJ, Bruenn JA, Hay J, Czechowski D, Taylor DJ (2014) Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. J Virol 88:8783–8794

Index

Vitantonio Pantaleo and Michela Chiumenti (eds.), *Viral Metagenomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1746, https://doi.org/10.1007/978-1-4939-7683-6, © Springer Science+Business Media, LLC 2018