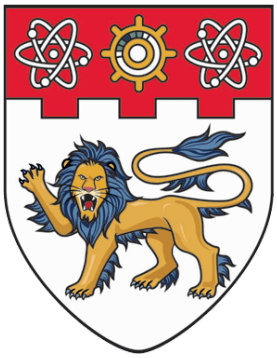


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

DIVERSITY AND DYNAMICS OF RNA VIRUSES IN EQUATORIAL COASTAL WATERS

SANDRA KOLUNDŽIJA

ASIAN SCHOOL OF THE ENVIRONMENT

2022

DIVERSITY AND DYNAMICS OF RNA VIRUSES IN EQUATORIAL COASTAL WATERS

SANDRA KOLUNDŽIJA

Asian School of the Environment

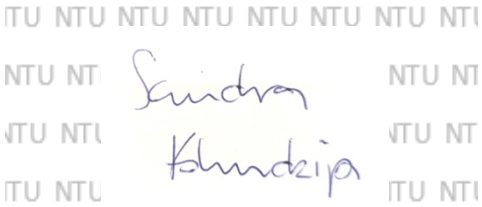
A thesis submitted to Nanyang Technological University in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

2022

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.



10.08.2022

. . . . . .
Date

. . . . . .
Sandra Kolundžija

SUPERVISOR DECLARATION STATEMENT

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

10.08.2022

. . ..
Date

…….…… …….....
Assoc Prof Federico M. Lauro

AUTHORSHIP ATTRIBUTION STATEMENT

This thesis contains material from 2 papers published in the following peer-reviewed journals and from 1 paper accepted at conferences in which I am listed as an author.

Chapter 2 is published with modifications as Kolundžija, S., Cheng, D-Q., Lauro, F.M. RNA Viruses in Aquatic Ecosystems through the Lens of Ecological Genomics and Transcriptomics. *Viruses* **2022**, *14*, 702. https://doi.org/ 10.3390/v14040702

The contributions of the co-authors are as follows:

- Assoc Prof Lauro suggested the topic and edited the manuscript drafts.
- I structured the paper and wrote the drafts of the manuscript. The manuscript was revised together with Dr. Dong-Qiang Cheng.
- I and Dr. Cheng created the figures for the paper.
- All authors read and approved the final manuscript.

Bioinformatic pipeline used for RNA virus discovery in Chapter 3 and Chapter 4 is published as Cheng D-Q., Kolundžija S., Lauro F.M. Global phylogenetic analysis of the RNA-dependent RNA polymerase with OrViT (OrthornaVirae Tree). Front. Virol. 2:981177. doi: 10.3389/fviro.2022.981177

The contributions of the co-authors are as follows:

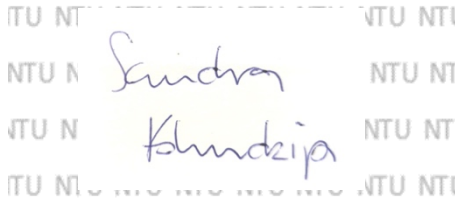
- Assoc Prof Federico Lauro suggested the topic
- Dr. Dong-Qiang Cheng developed the pipeline and wrote the paper
- I and Dr. Cheng interpreted and evaluated the results produced by the OrVit pipeline
- All authors read, edited and approved the final manuscript

Findings from the Chapters 4 and 5 have been presented as a poster presentation at International Society for Microbial Ecology (ISME) conference in August 2020 as: Kolundžija S., Cheng D-Q., Wijaya W., Chua C., Razali H., Suhaimi Z., Sunil R.S., Lindell D., Lauro F.M. Viral control of phytoplankton populations in equatorial coastal waters.

The contributions of the co-authors are as follows:

- Assoc Prof Federico Lauro, Prof Debbie Lindell and I conceived and designed the study
- Rohan Shawn Sunil, Cherlyn Chua, Zahirah Suhaimi, Winona Wijaya and I performed the sampling
- Halimah Razali and I performed all the molecular lab work and long-read sequencing.
- Rohan Shawn Sunil, Cherlyn Chua, Zahirah Suhaimi, Winona Wijaya performed nutrient analysis, flow cytometry and chlorophyll-a extraction
- Dr. Dong-Qiang Cheng and I performed the bioinformatic processing of the data
- Dr. Cheng, Winona Wijaya and I analysed and visualized all the data

We the undersigned agree with the above stated "proportion of work undertaken" for each of the above published (or submitted) peer-reviewed manuscripts contributing to the thesis:



Sandra Kolundžija

Student

Asian School of the Environment

Nanyang Technological University

Date: 10.08.2022

Assoc Prof Federico M. Lauro

Supervisor

Asian School of the Environment

Nanyang Technological University

Date: 10.08.2022

Assoc Prof Benoit Taisne

Associate Chair (Research)

Asian School of the Environment

Nanyang Technological University

Date: 11.08.2022

ACKNOWLEDGEMENTS

"Science knows no country, because knowledge belongs to humanity, and it is the torch which illuminates the world."

This work would not have been possible without the unwavering support and endless patience of my supervisor, Assoc Prof Federico Lauro. He always answered my questions with a smile and constructive suggestions. His unbounded enthusiasm and excitement for science and his thousands of little encouragements along the way have helped to move this ship forward on a long journey. Thank you for masterfully adjusting the sails, Captain Fede! Most of all, thank you for giving me a chance to move to Singapore in 2015 - that chance gave me a whole new life.

I would like to convey my appreciation towards my Thesis Advisory Committee: Assist Prof Hie Lim Kim from the Asian School of the Environment, Assoc Prof Scott Rice from CSIRO, Australia, and Dr. Caroline Chenard from the National Research Council of Canada for their insightful and thoughtful guidance, scientific input and support. I would like to thank Professor Debbie Lindell who opened the door of her lab at Israel Institute of Technology - Technion to learn polony method and Svetlana Goldin, for tirelessly answering my long emails with in-depth explanations and constructive advice that helped me set up polony station in Singapore.

I am truly thankful to have an incredible, supportive community in the Asian School of the Environment and SCELSE. Most of all, I would like to thank the one and only - Halimah Razali - the all-round best intern, best FYP student, and best research assistant I could ever imagine. I will never forget countless hours in the lab where you stood by my side when nothing was working and gave me courage to keep going. Thank you for the late night filtering of tons of seawater and your optimism and perseverance when

our RNA extractions kept failing. I am most grateful to brilliant Dong Qiang Cheng who patiently shared his superb bioinformatic expertise, science rigour and wisdom with me. Thank you building amazing RNA virus discovery pipelines, DQ! I am indebted to Winona Wijaya and George Williams for continuous support with R and data visualization. I will always be grateful for what you have taught me! I am deeply grateful to Zahirah Suhaimi for our joint sampling campaign in the Johor Strait, without which many exciting discoveries could not be made. Christina Tee, thank you for always caring about graduate students' wellbeing and for being so kind and helpful throughout this process. To my professional mentor Assoc Prof Emma Hill and Tamu, your encouragement has meant so much to me.

To all my amazing colleagues at the SLIME lab, THE Annex and ASE, I am fortunate to have you and I truly appreciate all your support. Thanks to everyone who cheered me up with little motivational speeches, especially during the last weeks. You have no idea how much difference it made. Special thanks to Nicoló for being my lunch buddy, my running buddy and my happy hour buddy; to Christaline for our Bali field course (mis)adventures and sharing her knowledge and expertise on beautiful phytoplankton and long-read sequencing; to David Demory for encouraging me to try iron flocculation and for listening about virus concentration (fails) more than anyone ever should.

I was lucky to have wonderful mentors even before starting my PhD! A very special thanks to my first supervisor, Professor Mladen Krajačić from University of Zagreb for sparking my enthusiasm in marine viruses and to Professor Gunnar Bratbak from University of Bergen for comprehensive training in marine viral ecology. This precious experience and joyful days in Norway will always have a special place in my heart.

My deepest gratitude goes my parents and the best sister on the planet for their unconditional love, care, and for supporting me whichever path I chose. To my

wonderful friends Elaine, Sol, Iva, Pauliina, Megan, and Maja, thank you for sharing laughter and happiness with me, and standing by my side in the challenging moments. A wholehearted thank you to my iFly friends and the dynamic flying Saturday group for all the magical moments in the tunnel - without you I would not have made this thesis fly.

TABLE OF CONTENTS

THESIS SUMMARY *1*

CHAPTER 1 INTRODUCTION *2*

1.1. Viruses are the most diverse biological entities in the biosphere **2**

1.2. Ecological and evolutionary roles of viruses in marine ecosystem **5**

1.3. Aims and objectives **9**

CHAPTER 2 RNA VIRUSES IN AQUATIC ECOSYSTEMS THROUGH THE LENS OF ECOLOGICAL GENOMICS AND TRANSCRIPTOMICS *12*

2.1. Abstract **12**

2.2. Introduction **13**

2.3. Methodological challenges in the study of aquatic RNA viral communities **18**

2.4. Environmental RNA metaviromics: lytic positive-sense ssRNA viruses dominate pelagic but not benthic RNA viral assemblages **28**

2.5. Environmental (viral) metatranscriptomics: sampling with size fractionation improves metatranscriptome resolution and uncovers the role of ssRNA viruses **35**

2.6. Environmental dsRNA sequencing: the enrichment of dsRNA from marine samples greatly expands the diversity of dsRNA viruses **42**

2.7. Holobiont metatranscriptomics: marine macroalgae and cultured marine protists reveal the broad distribution of non-lytic strategies in marine RNA viruses **44**

2.8. Recommendations for future studies **46**

2.9. Conclusions **49**

CHAPTER 3 IMPROVING THE SAMPLE-TO-SEQUENCE WORKFLOW FOR SEQUENCING OF AQUATIC METATRANSCRIPTOMES AND RNA VIROMES *50*

3.1. Abstract **50**

3.2. Introduction **51**

3.3.Materials and methods **53**

3.3.1. Sample collection 53

3.3.2. RNA extraction from the cellular fraction (>0.22μm) 54

3.3.3. Concentration of the viral-size (<0.22μm) fraction 56

3.3.4. RNA extraction from the viral-size (<0.22μm) fraction 57

3.3.5. Total RNA sequencing 58

3.3.6. Phylogenetic analysis of the RNA virome 59

3.3.7. Functional annotation of near-complete RNA viral genomes 59

3.4. Results **60**

3.4.1. Quantity and quality of RNA obtained from the cellular (>0.22μm) fraction 60

3.4.2. Quantity and quality of RNA obtained from viral-size (<0.22μm) fraction 63

3.4.3. Effect of physical and in silica rRNA depletion on assembly and recovery of RNA viruses from the viral-size fraction (<0.22μm) fraction 63
3.4.4. Diversity of RNA viruses in the viral-size (<0.22μm) fraction and genomic analysis of novel members of the *Picornavirales*. 66

3.5. Discussion 72
3.5.1. Quantity and quality of RNA obtained from the cellular fraction (>0.22μm) 72
3.5.2. Quantity and quality of RNA obtained from the viral-size fraction (<0.22μm) 73
3.5.3. Effect of physical and in silica rRNA depletion on assembly and recovery of RNA viruses from the viral-size fraction (<0.22μm) fraction 76
3.5.4. Diversity of RNA viruses in the viral-size (<0.22μm) fraction and genomic analysis of novel members of the *Picornavirales* 78

5.6. Conclusions 81

3.7. Acknowledgements 82

APPENDIX 3 83

CHAPTER 4 METATRANSCRIPTOMICS UNCOVERS NOVEL CLADES OF RNA VIRUSES AND GIANT DNA VIRUSES ACTIVELY INFECTING EUKARYOTES IN EUTROPHIC COASTAL WATERS 91

4.1. Abstract 91

4.2. Introduction 92

4.3. Materials and methods 94
4.3.1. Study site and physico-chemical water parameters 94
4.3.2. Microbial sample collection, nucleic acid extraction and clean-up 96
4.3.3. Short-read and long-read DNA metagenome sequencing 97
4.3.4. Metatranscriptome sequencing 97
4.3.5. Taxonomic assignment of the DNA metagenomes 98
4.3.6. Viral discovery in the metatranscriptomic data 98

4.4. Results 101
4.4.1. Eukaryotic phytoplankton community composition in the Johor Strait 101
4.4.2. Phylogenetic diversity of RNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes 102
4.4.3. Phylogenetic diversity of giant DNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes 107

4.5. Discussion 111
4.5.1. Eukaryotic phytoplankton community composition in the Johor Strait 111
4.5.2. Phylogenetic diversity of eukaryote-infecting RNA viruses in the Johor Strait metatranscriptomes 112
4.5.3. Phylogenetic diversity of giant DNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes 115
4.5.4. Host ranges and infection strategies of phytoplankton viruses. 116

4.6. Conclusions 118

4.7. Acknowledgments 119

APPENDIX 4 120

CHAPTER 5 RNA AND GIANT DNA VIRUSES INFECTING EUKARYOTIC PHYTOPLANKTON IN EUTROPHIC COASTAL WATERS DISPLAY DISTINCT LIFE-CYCLE DYNAMICS 125

5.1. Abstract 125

5.2. Introduction 126

5.3. Materials and methods 129
5.3.1. Study site and physical, chemical and biological data collection 129
5.3.2. Concentration of chlorophyll-a and dissolved inorganic nutrients 129
5.3.2. Enumeration of viruses and prokaryotes with flow cytometry 130
5.3.4 Data analysis and visualization 131

5.4. Results 131
5.4.1. Fluctuations of the abiotic and biotic variables in the Johor Strait 131
5.4.2. Temporal dynamics of eukaryote-infecting RNA and giant DNA viruses in the Johor Strait 135
5.4.4. Putative virus - eukaryotic host pairs in the Johor Strait 142

5.5. Discussion 145
5.5.1. Fluctuations of the abiotic and biotic variables in the Johor Strait 145
5.5.2. Temporal dynamics of eukaryote-infecting RNA and giant DNA viruses in the Johor Strait 146
5.5.3. Abiotic and biotic variables influencing the eukaryotic community composition 147
5.4.4. Putative virus - eukaryotic host pairs in the Johor Strait 148

5.6. Conclusions 150

5.7. Acknowledgements 151

APPENDIX 5 152

CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS 162

6.1. Synthesis and significance of this work 162

6.2. Future work 164

BIBLIOGRAPHY 167

LIST OF MAIN FIGURES

Figure 2.1. Phylogenetic analysis of RNA-dependent RNA polymerase (RdRp) of RNA viruses with reverse transcriptase (RT) used as an outgroup to root the tree.. 15
Figure 2.2. Marine viruses simultaneously control two processes in the carbon cycle: (**A**) the viral shuttle and (**B**) the viral shunt.....16
Figure 2.3. Most commonly used meta-omic sequencing approaches in marine virology.....19
Figure 2.4. Geographic locations of aquatic metaviromic (blue), metatranscriptomic (coral), and dsRNA sequencing (yellow) studies focusing on RNA viruses reviewed in this paper.....30
Figure 3.1. Box and whiskers plot of RNA extraction efficiency of different filter-kit combinations.....62
Figure 3.2 Effect of on-column DNase treatment on DNA removal and RNA yield..62
Figure 3.3. Broad taxonomic distribution of the sequencing reads before (A) and after (B) in-silico rRNA depletion. 64
Figure 3.4. Taxonomic assignment of the RNA viruses identified in the viral-size fraction (<0.22μm) on the backbone of a maximum likelihood ortornaviral phylogenetic tree of viral RNA-dependent RNA polymerase.....68
Figure 3.5. Maximum-likelihood phylogenetic tree showing in details distribution of recovered RNA viral contigs over the most abundant order detected in the sample, order Picornavirales of the phylum Pisuviricota. 69
Figure 3.6. (A) Phylogenetic analysis of selected full-length sogarna-like viruses recovered in this study using conserved RNA-dependent RNA polymerase domains (B) Schematic genome organization of selected sogarna-like viruses identified in this study. 70
Figure 3.7. Final wet lab "sample to sequence" integrated workfloe for generation of RNA viromes from the marine viral-size fraction……..75
Figure 4.1. Map showing a) location within the southeast Asia and position relative to the equator. b) magnified view of the region c) study site, located in the Johor Strait, Singapore. 95
Figure 4.2. Stacked bar plot showing the relative abundance of top 10 eukaryotic taxa (genus-level) in the Johor Strait. 102
Figure 4.3. Maximum likelihood phylogenetic tree of viral RNA-dependent RNA polymerase showing phylogenetic position of the RdRp detected in the Johor Strait metatranscriptomes. 103
Figure 4.4. Maximum likelihood phylogenetic tree of RNA-dependent RNA polymerase showing the phylogenetic placement of the the sequences recovered from the Johor Strait metatranscriptomes within the phylum Lenarviricota (A) and Kitrinoviricota (B) 104
Figure 4.5. Maximum likelihood phylogenetic tree of RNA-dependent RNA polymerase showing the phylogenetic placement of the RdRp sequences recovered from the Johor Strait metatranscriptomes within the phylum Pisuviricota:(A) order Picornavirales (B) order Durnavirales. 105
Figure 5.1. Fluctuations of the chlorophyll-a concentrations during the 55-day sampling period in the Johor Strait. 132
Figure 5.2. Fluctuations of the nutrient concentrations during the 55-day sampling period in the Johor Strait.....133

Figure 5.3. Temporal dynamics of distinct viral subpopulations of putative phytoplankton DNA viruses (Phycodnaviridae) distinguished with flow cytometry. A) Vir3 B) Vir4 C) Vir5.. 135
Figure 5.4. Normalized relative expression of major groups of RNA viruses during the 55-day time series in the Johor Strait expressed as Reads Per Kilobase of contig per Million reads 137
Figure 5.5. Non-metric multidimensional scaling (NMDS) ordination plot of RNA viral expression in 55-day time-series in the Johor Strait 137
Figure 5.6. Normalized relative expression of detected giant DNA viruses during the 55-day time-series in the Johor Strait expressed as Reads Per Kilobase of contig per Million reads 139
Figure 5.7. Non-metric multidimensional scaling (NMDS) ordination analysis of giant DNA viral transcription in 55-day time-series in the Johor Strait 139
Figure 5.8. Non-metric multidimensional scaling (NMDS) ordination analysis of phytoplankton community in 55-day time-series in the Johor Strait. 140
Figure 5.9. Canonical correlation analysis (CCA) of phytoplankton community (top 10 most abundant taxa) and environmental variables in the Johor Strait. 141
Figure 5.10. (A) Corrplot showing significant correlations between top 10 phytoplankton taxa and top 20 RNA viruses. (B) Corrplot showing significant correlations between top 10 phytoplankton taxa and top 20 giant DNA viruses. 144

LIST OF APPENDIX FIGURES

Figure A 3.1. Total yields of DNA and RNA obtained from the viral-size fraction (<0.22μm) by Trizol and All Prep Environmental RNA/DNA Kit extraction. 85
Figure A 3.2. Electrophoregrams show the integrity of RNA after different extraction approaches tested with RNA Screen Tape on Tape Station 2200 86
Figure A 3.3. (A) Correlation plot showing negative correlation between filtered biomass expressed as concentration of chlorophyll-a (μg/L) and RNA quality expresses as RNA INtegrity (RIN) scores. (B) Correlation plot showing positive correlation between filtered biomass expressed as concentration of chlorophyll-a (μg/L) and total RNA yield expressed in nanograms (ng). 87
Figure A 3.4. Multiple sequence alignment of protein sequences detected in two analysed genomes and a reference genome of Cheatoceros RNA virus 02.....90
Figure A 4.1. Magnified view of the Johor Strait showing a) depth b) current velocity and direction during spring flood c) current velocity during spring ebb 120
Figure A 4.2. Flowchart of the OrVit pipeline.....121
Figure A 4.3.Stacked bar plot showing the relative abundance of cellular organisms and viruses in the metagenomic reads in the Johor Strait.....122
Figure A 4.4. Stacked bar plot showing the relative abundance eukaryotic taxa (phylum-level) in the Johor Strait.....122
Figure A 4.5.Maximum likelihood phylogenetic tree of the family *Marnaviridae*...123
Figure A 4.6. Close-up of the *Marnaviridae* phylogenetic tree.....124
Figure A 5.1. Flow cytograms of prokaryotes and viruses. 155
Figure A5.2. Temporal dynamics of prokaryotes (A), total virus-like particle (VLP) and Vir1-Vir2 virus subpopulation (B) abundances during the 55-day sampling campaign in the Johor Strait determined with flow cytometry 156
Figure A 5.3. Linear correlations between viral populations captured by FCM, chlorophyll-a concentration and relative expression of giant DNA viruses. 157
Figure A 5.4. Correlation plot showing significant correlations between environmental variables in the Johor Strait. 158
Figure A 5.5. Canonical correlation analysis (CCA) of phytoplankton community (top 10 most abundant taxa) and phytoplankton viruses in the Johor Strait: (A) RNA viruses (B) giant DNA viruses 159
Figure A 5.6. Temporal dynamics of top 20 most abundant RNA viral transcripts in the Johor Strait metatranscriptome. 160
Figure A 5.7. Temporal dynamics of top 20 most abundant giant DNA viral transcripts in the Johor Strait metatranscriptome. 161

LIST OF TABLES

Table 2.1. Advantages and shortcomings of the meta-omic sequencing approaches used in marine viral ecology studies. Adapted from (Roossinck et al., 2015).....21
Table 2.2. Basic geographic, ecosystem, and sampling information about the aquatic viromic studies, metatranscriptomic studies, and dsRNA sequencing studies 24
Table 3.1.Basic sequencing information on a RNA virome sample sequenced with two different RNA sequencing approaches. 65
Table 3.2. Open reading frames, coordinate, length and annotation of a reference sogarnavirus and two putative sogarna-like viruses recovered in this study.
Table 4.1. Taxonomic placement of giant DNA viral transcripts detected in Johor Strait metatranscriptomes.. 109

LIST OF APPENDIX TABLES

Table A 3.1. Initial sampling efforts to test for different combinations of sample volume, concentration methods and RNA extraction kits. 83
Table A 3.2. Filter type and pore size, extraction methods, volume of filtered water for RNA extraction yield tests for metatranscriptome sequencing. 84
Table A 3.3.The extraction efficiency of different RNA extraction approaches of viral concentrates obtained by iron flocculation and sucrose cushion ultracentrifugation....85
Table A 3.4. List of near-complete or complete RNA viral genomes recovered from viral-size fraction. 88
Table A 5.1. Precipitation, chlorophyll-a concentration, temperature and salinity during the 55-day sampling campaign in the Johor Strait. 152
Table A 5.2. Nutrient concentrations and nutrient ratios during the 55-day sampling campaign in the Johor Strait. 153

THESIS SUMMARY

Studies on the ecological effects of marine RNA viruses are critically lacking, despite the fact that viral infections impact health and evolution of individual species, the community structure of populations, and the biogeochemistry of the entire marine ecosystem. In this thesis, using a high-resolution time-series, I uncover a remarkable difference between life-cycles of lytic DNA and RNA phytoplankton viruses with a combination of metagenomics, metatranscriptomics and bioinformatics. The giant DNA viruses, known for lower burst sizes, exhibited low and continuous transcriptional activity, suggesting coexistence with their potential hosts. In contrast, fast-replication RNA viruses, known for high burst sizes, experienced short "bloom and bust" cycles of transcriptional activity which, along with the nutrient limitation, stopped potential bloom formation on two separate instances. Persistent, asymptomatic infections with RNA viruses without an extracellular stage were widely present in the Johor Strait marine ecosystem. Fifty nearly full-length RNA viral genomes and 319 verified RNA viral fragments were discovered using and optimized wet lab protocol and integrated bioinformatic pipeline OrVIT, which through extraction of conserved RdRp domains, produces high-quality phylogenetic trees of RNA viruses. Most recovered sequences clustered within the *Sogarnavirus* genus, which infects diatoms. Both datasets contained a pool of sequences from dsRNA and ssRNA viruses that infect marine animals, suggesting a possible hazard to aquaculture. This thesis underscores the potential of using an integrated multi-omic approach to capture the complex interplay between viruses and their hosts in marine ecosystems and emphasises the critical importance of phytoplankton RNA viruses in top-down control of blooming, fast-growing phytoplankton populations, especially in eutrophic ecosystems.

CHAPTER 1 INTRODUCTION

1.1.Viruses are the most diverse biological entities in the biosphere

Viruses are simple infectious agents consisting of nucleic acids enclosed in a protein shell, and in some cases, a lipid envelope. They lack ribosomes, mitochondria and other organelles and have to rely on the host cellular machinery to multiply, which makes them obligate intracellular parasites (Payne, 2017). Every cellular life form on Earth - bacteria, archaea, unicellular and multicellular eukaryotes can be infected by at least several different viruses. For example, more than different 200 viruses cause infections in humans (Geoghegan and Holmes, 2017). Viruses called virophages can even be parasites of giant viruses and exploit their large viral factories built in the host cell (Gaia et al., 2014; La Scola et al., 2008; Yau et al., 2011). Viromes, viral communities comprising of diverse mixed viral species, are an integral and functionally indispensable part of native microbiomes in different multicellular eukaryotes (Virgin, 2014). From invertebrates like corals to mammals, rich viromes comprise symbiotic, commensal or pathogenic viruses and profoundly influence health and disease of the host organisms (Virgin, 2014; Liang and Bushman, 2021; Shi et al., 2016; Thurber et al., 2017).

Viral infections are categorized as either acute or persistent, and some viruses can switch between different infection types depending on the environmental conditions. In acute, lytic infections, large numbers of produced infective viral particles (virions) are released through the host cell lysis and death (Payne, 2017; Weinbauer, 2004). Persistent infections include two subcategories: chronic and latent. In chronic infections, small numbers of virions are continuously released from host cells (Chappell and Dermody, 2015). In latent infections, viruses do not produce virions, but maintain the viral genome integrated in the host genome or exist like an episome inside the cytoplasm (Chappell

and Dermody, 2015). Latent infection by viruses that infect bacteria (a.k.a. bacteriophages) is often referred as temperate or lysogenic infection (Zimmerman et al., 2020).

Viruses have remarkably versatile genome structures comparing to the cellular life forms which exclusively use double-stranded (ds) DNA as a genetic storage material. Viral genomes can be composed of either DNA or RNA, which may be single-stranded (ss) or double-stranded (ss), circular or linear, and encoded in single (unsegmented) or multiple (segmented) molecules (Krupovic et al., 2019). Based on differences in genome structure and transcription strategies, viruses are classified into 7 groups; I :double-stranded DNA viruses, II: single-stranded DNA viruses, III: double-stranded RNA viruses, IV: (+) single-stranded RNA viruses, V: (-) single-stranded RNA viruses, VI: reverse-transcribing RNA viruses and VII: reverse-transcribing DNA viruses (Baltimore, 1971). This functional, non-hierarchical division, widely known as Baltimore classification of viruses, does not necessarily reflect evolutionary relationships among viruses, but is still commonly used (Koonin et al., 2021). Phylogenetic analysis, based on comparisons of multiple nucleotide or amino acid sequences of conserved genes, is able to reveal shared ancestry and it serves a basis for modern virus classification system, similar to Linnean hierarchical ranking of cellular life (Gorbalenya et al., 2020; Gorbalenya and Lauber, 2022; Koonin et al., 2020). Viruses have polyphyletic origin, meaning they evolved multiple times independently during the course of evolution via different mechanisms. Because they do not share a common ancestor, and therefore lack a universal gene that would be common to all viruses, an universal phylogenetic tree for all viruses cannot be created (Koonin et al., 2020; Krupovic et al., 2019). Genome sizes of viruses are also highly variable and encompass three orders of magnitude. Largest dsDNA viruses like the eukaryote-

infecting Pandoravirus have a genome of 2.5 megabases (Philippe et al., 2013). Genomes of small bacteriophages and jumbo bacteriophages range between 35-100 kilobases and 200-500 kilobases, respectively (Yuan and Gao, 2017). The smallest genomes, those of RNA and ssDNA viruses are typically shorter than 10 kilobases (Rosario et al., 2012; Sadeghi et al., 2021).

Early viral discovery relied on tedious and slow isolation of single viruses in susceptible eukaryotic or prokaryotic cell cultures. This approach can recover only a limited portion of viral diversity, since most of viruses and their hosts may not be culturable at all. (Zhang et al., 2019). The process of viral discovery was radically transformed with 'omic approaches like metagenomics, metatranscriptomics and metaviromics which enable culture-independent detection of viruses, relying only on presence of the viral nucleic acid (Cobbin et al., 2021; Zhang et al., 2019). In theory, they can detect all viruses in the sample, even if they are novel or unculturable (Kolundžija et al., 2022; Roux et al., 2019). Omic - based virus discovery expanded the number of novel viral species by orders of magnitude and described new phyla of viruses (J. Callanan et al., 2020; Edgar et al., 2022; Nayfach et al., 2021; Paez-Espino et al., 2016; Wolf et al., 2020; Zayed et al., 2022). In reality, less than 1% of the global viral diversity has been characterized (Geoghegan and Holmes, 2017). Large number of viral sequences still remains inaccessible with the current bioinformatic methods. Even bioinformatically recovered viral sequences often do not align to any genomes available in the databases and remain uncharacterised. These sequences are collectively known as the "viral dark matter" (Krishnamurthy and Wang, 2017; Zhang et al., 2018). Our current view of viral diversity is likely extremely biased and many surprises await for us in the unexplored corners of the viral universe.

1.2.Ecological and evolutionary roles of viruses in marine ecosystem

The total number of viruses on Earth is estimated to 10 nonillion (10^{31}) (Mushegian 2020; Suttle 2005).Viruses are present across all ecosystems, from 1,000 meters up in the atmosphere to the deepest oceanic trenches, 12,000 meters below the sea surface (Reche et al. 2018; Paez-Espino et al. 2016; Williamson et al. 2017; Li et al. 2021; Zheng et al. 2021; Gregory et al. 2019; Brum et al. 2015; Roux et al. 2016; Whon et al. 2012; Zhao et al. 2013; Zhao et al. 2022). Energy and organic matter flow in the marine ecosystems are mediated by marine microorganisms, consisting of bacteria, archaea and unicellular eukaryotes, protists. Heterotrophic bacteria and archaea drive biogeochemical cycles of major elements and autotrophic bacteria and protists (the latter also known as phytoplankton), fix inorganic CO_2 from the atmosphere into organic carbon via photosynthesis, contributing up to half of the global primary production (Behrenfeld et al., 2006; Falkowski et al., 2008). Every second, 10^{23} microorganisms in the oceans are infected with viruses (Suttle, 2007). Dynamic virus-host interactions hugely influence both microbial and protistan community composition and mortality, as well as organic matter turnover in the oceans (Rohwer and Thurber, 2009; Zimmerman et al., 2020).

Lytic viruses continuously control the abundance of their hosts by following “Kill the winner” dynamics (Thingstad, 2000). They selectively lyse the fastest-growing microorganisms (“the winners”) that could otherwise force out slow-growing, low-abundance microorganisms with diverse ecological roles and maintain the vitally important diversity in the marine ecosystems (Sandaa, 2008) . Extremely rapid increases of phytoplankton biomass, known as phytoplankton blooms, can disrupt diversity and functioning of the ecosystem (Trottet et al., 2021). High viral infection rates of blooming

phytoplankton cells in blooms of *Emiliania huxleyi* (Vincent et al., 2021), *Aureococcus anophagefferens* (Gastrich et al., 2004), *Pheocystis globosa* (Baudoux et al., 2006), and *Heterosigma akashiwo* (Tarutani et al. 2000) demonstrate that viral lysis can substantially contribute to the reduction of the phytoplankton biomass. For some phytoplankton, for example cryptophytes and large diatoms during the Antarctic phytoplankton spring bloom, viral lysis can exceed grazing rates and be the main source of phytoplankton mortality (Biggs et al. 2021).

Viral lysis of the host cell, besides infective viral particles, also releases large amounts of dissolved organic matter (DOM) and inorganic nutrients to seawater and decreases the amount carbon and nutrients transferred to higher trophic levels. This phenomenon, known as the "viral shunt", maintains a constant pool of DOM in the upper ocean that fuels primary and secondary oceanic production (Suttle, 2007; Weitz et al., 2015; Zimmerman et al., 2020). Viral lysis can also direct carbon away from the higher trophic levels and surface pool, to the deep sea. The virally mediated enhancement of the biological carbon pump (a.k.a. the "viral shuttle'), enhances the carbon export through formation of fast-sinking sticky cell aggregates after viral lysis. (Kaneko et al., 2021; Weinbauer, 2004; Zimmerman et al., 2020). Larger phytoplankton cells, like diatoms and *Emiliania huxleyi* are important contributors to the biological carbon pump, especially during blooms when large amounts of carbon are exported to the deep sea (Kaneko et al., 2021; Laber et al., 2018).

Viral infection impacts physiology of the host cell through short-term metabolic reprogramming and/or by affecting long-term host evolutionary processes (Rohwer and Thurber, 2009). The host metabolic reprogramming occurs through the expression of virally-encoded auxiliary metabolic genes during lytic viral infection (Zimmerman et

al., 2020). Viral manipulation of host metabolism boost fitness of their bacterial and phytoplankton host by stimulating carbon metabolism, nutrient uptake, photosynthesis, and nucleotide production, and increases efficient production of viral progeny in the lytic infection. (Lindell et al., 2005; Roux et al., 2016; Thompson et al., 2011). In lysogenic viral infections, integrated bacteriophage genomes (prophages) replicate with host genomes, "piggybacking the winner", a strategy that is especially common in the rich and dense coral microbiome (Knowles et al., 2016). Sometimes prophages provide resistance to superinfection and can alter the phenotype of the host, in a process known as lysogenic conversion. In lysogenic conversion, phage-encoded toxins, adhesion factors promoting biofilm formation, and antibiotic resistance genes benefit host fitness and may confer enhanced virulence traits to bacteria (Argov et al., 2017; Touchon et al., 2017). Viruses exert constant pressure on hosts to develop resistance against viral infection, and vice versa, viruses rapidly evolve virulence to overcome the host's resistance mechanisms with changes transmitted vertically to the viral and host offspring (Argov et al., 2017; Rohwer and Thurber, 2009). In addition, horizontal gene flow facilitated by the viral infection speeds up acquisition of new traits. Extensive horizontal transfer of genetic material has been well documented between all types of viruses and their hosts (Dolja and Koonin, 2018; Irwin et al., 2022; Touchon et al., 2017) and it a widespread process in marine ecosystem, contributing immensely to evolution of marine viral and microbial diversity (Rohwer and Thurber, 2009).

Large scale metagenomic exploration of marine viral communities has been heavily biased/directed towards viruses infecting prokaryotes (Brum et al., 2015; Gregory et al., 2019; Luo et al., 2020). In comparison, diversity and virus – host interactions of marine eukaryote-infecting viruses have been severely overlooked and poorly understood due

to the methodological challenges in sample preparation and bioinformatic analysis (Coy et al., 2018; Kolundžija et al., 2022).

The Nucleocytoplasmic Large DNA Viruses (NCLDVs), also called giant DNA viruses, are eukaryote-infecting DNA viruses with complex virion morphology and large dsDNA genomes that completely or partially replicate in the eukaryotic cytoplasm (Mutsafi et al., 2014). What distinguishes giant DNA viruses from other viruses is that they encode for genes encode numerous genes that were thought to be exclusive to cellular life forms. These genes includes chaperons, transcription factors, and translation-related genes like aminoacyl tRNA synthetases or tRNAs and were acquired from different hosts in multiple independent, horizontal gene transfer events (Koonin and Yutin, 2019; Koonin et al., 2020). Currently, the giant DNA are taxonomically assigned to seven families, and together with the newly discovered, taxonomically unclassified giant viruses (e.g. Pandoravirus, Medusavirus), form a monophyletic group united under the phylum *Nucleocytoviricota,* class *Megaviricetes* (Walker et al., 2022). Viruses from families *Mimiviridae*, *Marseilleviridae* and *Phycodnaviridae* infect both marine phytoplankton and heterotrophic single-celled eukaryotes shape their community structure, and may be especially important during phytoplankton blooms (Moniruzzaman et al., 2016, 2017). Families *Ascoviridae*, *Iridoviridae*, *Poxviridae* and *Asfarviridae* infect broad range of animals and some can cause devastating diseases and high mortality in marine animals, such as shrimp hemocyte iridescent virus (SHIV) or abalone asfa-like virus (AbALV) (Kibenge, 2019; Koonin & Yutin, 2019; Matsuyama et al., 2020).

RNA viruses are another group of eukaryote-infecting viruses widely distributed in marine ecosystems (Kolundžija et al., 2022). Marine RNA viral genomes are small in

size, typically ranging between 2-12kb, and with genes typically organized in two open reading frames (ORFs) or modules (Sadeghi et al., 2021). The replication module consist of genes involved in the replication of the viral genome, such as RNA-dependent RNA polymerase or helicase. The structural module carries proteins needed for virus capsid and particle assembly (Miranda et al., 2016; Moniruzzaman et al., 2017). Higher mutation rates, due to the absence of proof-reading activity of viral RNA polymerases, could be detrimental for larger RNA genomes, and as a consequence RNA viral genomes never exceed 35 kilobases (Holmes, 2003). Highly error-prone replication and susceptibility to mutation of RNA viruses also lead to rapid evolution and great adaptability to new environmental conditions, including new hosts (Moya et al., 2004). The RNA viruses that share a conserved RNA-dependent RNA polymerase gene or reverse transcriptase were recently unified into the realm *Riboviria* (Koonin et al., 2020). Typical marine RNA virus will have an icosahedral capsid symmetry, (+) single-stranded RNA genome with poly-A tail and a lytic cycle with very high burst sizes, releasing up to 10 000 particles per infected cell (Culley, 2018). For a long time, marine RNA viruses were considered to predominantly infect marine microbial eukaryotes, but this is being challenged with recent myriad of discovery of abundant RNA bacteriophages in aquatic 'omic datasets (Callanan et al., 2020; Wolf et al., 2020).

1.3.Aims and objectives

Several technical challenges hinder exploration of RNA viral communities in the marine environment. First, epifluorescence microscopy and flow cytometry, routinely used for enumeration of total viral particles, are not sensitive enough to detect RNA viruses. Viruses with genomes larger than 50kb can be detected above the cut-off for noise in modern flow cytometers after labelling with common nucleic acid fluorescent dyes.

However, RNA viral genomes are smaller than 30kb and are at or below the limit of detection. Consequently, the accurate quantification of RNA virus abundance in the oceans is not possible at the moment (Brussaard et al. 2000; Tomaru and Nagasaki 2007; Kaletta et al. 2020). However, comparisons of DNA/RNA ratio obtained from purified marine DNA and RNA viruses suggest that RNA viruses may constitute more than half of free-floating viruses (Miranda et al., 2016; Steward et al., 2013). Second, despite the high abundance of viral particles in the ocean, their genomes are very small when compared to those of prokaryotic and eukaryotic cells. A major constraint in marine RNA viromics are the low amounts of RNA recovered even after viral enrichment, often with a high host background (Thurber et al. 2009; Duhaime and Sullivan 2012; Kolundžija et al. 2022). Despite the recent increase in metatranscriptomic-based studies of RNA viruses in marine ecosystems, no systematic evaluation of RNA extraction methods for metatranscriptomics has been performed.

This thesis is composed of 6 chapters. In chapter 1, this chapter, I describe the incredible diversity of viral genes, genome structures and size, host ranges, and infections styles that exist and essential roles of viruses in the marine ecosystem. In chapter 2, I review the current knowledge of RNA viruses in aquatic ecosystems, focusing on recent data acquired with high-throughput sequencing approaches. I critically discuss different high-throughput sequencing approaches like viromics, metatranscriptomics and dsRNA sequencing, their usefulness for studies of RNA viruses and their potential biases. I describe global open ocean sampling campaigns, studies of blooms in productive marine coastal environments and lakes, as well as RNA viromes of cultured marine phytoplankton, all offering slightly different, but complementary views of RNA viral diversity present in the marine environment.

In chapter 3, my aim was to improve the current sample preparation methods for two different RNA virus sequencing approaches (namely RNA viromics and metatranscriptomics). I rigorously assesses the quality and quantity of RNA yields in common RNA extraction approaches used in marine viral metatranscriptomics studies. I addresses the technical challenges with RNA extraction from viral concentrates, developing a highly effective and reproducible protocol for extraction and enrichment of RNA viruses. Using this optimized RNA viromics wet lab protocol I was able to describe the diversity of RNA viruses in the viral-size fraction in Singapore coastal waters, identify potentially novel clades of RNA viruses and extract near-complete RNA viral genomes.

The optimized wet lab protocol for metatranscriptomics was used in chapters 4 and 5, where I explore bi-daily temporal dynamics and potential ecological roles of RNA and giant DNA viruses infecting microbial eukaryotes in the Johor Strait, a highly productive coastal location with recurrent phytoplankton blooms. With a high-temporal resolution metatranscriptomic time-series, I describe for the first time the diversity of viruses infecting microbial eukaryotes in Johor Strait waters and demonstrate how lytic viral infection with giant DNA viruses and RNA viruses affects the community composition of their eukaryotic hosts in a different ways. In parallel, a metagenomic study focused on the potential hosts of these viruses reveals how viral suppression of fast-growing phytoplankton can stabilize the community and potentially prevent a bloom. Finally, Chapter 6 provides conclusions, implications, and ideas for future research.

CHAPTER 2 RNA VIRUSES IN AQUATIC ECOSYSTEMS THROUGH THE LENS OF ECOLOGICAL GENOMICS AND TRANSCRIPTOMICS

2.1. Abstract

Massive amounts of data from nucleic acid sequencing have changed our perspective about diversity and dynamics of marine viral communities. Here, we summarize recent metatranscriptomic and metaviromic studies targeting predominantly RNA viral communities. The analysis of RNA viromes reaffirms the abundance of lytic (+) ssRNA viruses of the order *Picornavirales*, but also reveals other (+) ssRNA viruses, including RNA bacteriophages, as important constituents of extracellular RNA viral communities. Sequencing of dsRNA suggests unknown diversity of dsRNA viruses. Environmental metatranscriptomes capture the dynamics of ssDNA, dsDNA, ssRNA, and dsRNA viruses simultaneously, unravelling the full complexity of viral dynamics in the marine environment. RNA viruses are prevalent in large size fractions of environmental metatranscriptomes, actively infect marine unicellular eukaryotes larger than 3 μm, and can outnumber bacteriophages during phytoplankton blooms. DNA and RNA viruses change abundance on hourly timescales, implying viral control on a daily temporal basis. Metatranscriptomes of cultured protists host a diverse community of ssRNA and dsRNA viruses, often with multipartite genomes and possibly persistent intracellular lifestyles. We posit that RNA viral communities might be more diverse and complex than formerly anticipated and that the influence they exert on community composition and global carbon flows in aquatic ecosystems may be underestimated.

Keywords: marine RNA virus; viral metatranscriptomics; metaviromics; dsRNA sequencing; viral ecology; viral diversity

2.2. Introduction

Viruses are vital components of the marine ecosystem (Suttle, 2007). They are major contributors to processes such as carbon and nutrient remineralization (Fuhrman, 1999; Weitz and Wilhelm, 2012; Zimmerman et al., 2020). Viruses also continuously regulate the diversity and abundance of complex microbial communities in the ocean, including of bacteria, archaea, and single-celled eukaryotes (protists) (Suttle, 2007). Plankton protist communities largely consist of single-celled eukaryotic algae (also known as eukaryotic phytoplankton), heterotrophic flagellates and amoebas (Burki et al., 2020). Viruses are incredibly complex when it comes to genome type, replication, infection mode, morphology, and host preference (Payne, 2017). They infect cellular organisms in all three domains of life, as well as other viruses, and a single organism can be infected by multiple viruses (Geoghegan and Holmes, 2017). In aquatic environments, eukaryotic phytoplankton is known to host very diverse viral communities (Coy et al., 2018; Short et al., 2020). Among aquatic bacteriophages, double-stranded (ds) DNA viruses are the most extensively studied and were thought to dominate prokaryotic virosphere (Nasir et al., 2014). Reports of single-stranded (ss) RNA and single-stranded (ss) DNA bacteriophages in environmental viromic and metatranscriptomic datasets (J. Callanan et al., 2020; Labonté & Suttle, 2013; Roux et al., 2012; Roux, Krupovic, et al., 2019; Wolf et al., 2020) challenged this idea and suggested that bacteriophages have equally diverse genome architecture as eukaryotic viruses. On the basis of their genome architecture and replication strategy, viruses were classified in seven groups, now widely known as the Baltimore classes, a classification still used in parallel with the official viral taxonomy (Koonin et al., 2020, 2021). However, the molecular taxonomy has shown that these groups are not monophyletic, and for RNA viruses, the transition

between single-stranded and double- stranded genome types has occurred multiple times during evolutionary history (Wolf et al., 2018).

All RNA viruses are a part of the newly established viral realm, *Riboviria* (Koonin et al., 2020). Within this realm, there are two kingdoms, *Pararnavirae* and *Orthornavirae*. *Pararnavirae* consists of RNA viruses with the reverse transcriptase (RT) while *Orthornavirae* contains RNA viruses with the RNA-dependent RNA polymerase (RdRp). Based on phylogenetic analysis of RNA-dependent RNA polymerase, kingdom *Orthornavirae* is split in into five phyla: *Lenarviricota*, *Pisuviricota*, *Pisuviricota*, *Kitrinoviricota*, and *Negarnaviricota* (Figure 2.1) (Koonin et al., 2020; Wolf et al., 2018). Phylum *Lenarviricota* includes ssRNA with positive (+) genome polarity, such as RNA bacteriophages from family *Leviviridae* and their eukaryote-infecting descendants, *Narnaviridae* and *Mitoviridae*, which are simple RNA viruses without a capsid that are transmitted vertically. Phylum *Pisuviricota* is sometimes called Picornavirus supergroup, and it includes five orders of (+) ssRNA viruses and one order of dsRNA viruses with bipartite genomes, *Durnavirales*. Phylum *Kitrinoviricota* includes only (+) ssRNA viruses infecting eukaryotes, mostly terrestrial plants, and animals. Phylum *Duplornaviricota,* the fourth branch, consists of dsRNA viruses infecting both eukaryotes and prokaryotes. Finally, the fifth branch forms phylum *Negarnaviricota* and includes all known RNA viruses with negative (−) genome polarity (Koonin et al., 2020; Wolf et al., 2018).

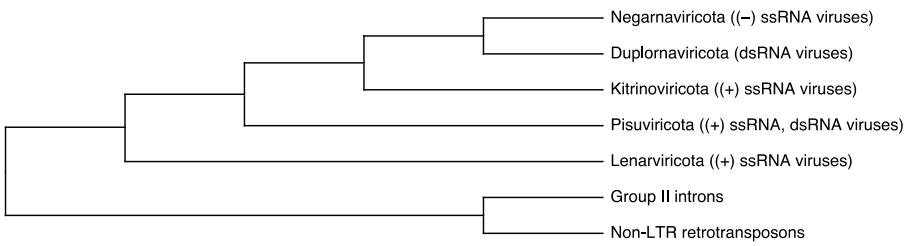


Figure 2.1. Phylogenetic analysis of RNA-dependent RNA polymerase (RdRp) of RNA viruses with reverse transcriptase (RT) used as an outgroup to root the tree. Five major branches have been assigned a phylum rank by Internationational Committee on Taxonomy of Viruses (ICTV): Branch 1 = *Lenarviricota*, Branch 2 = *Pisuviricota*, Branch 3 = *Kitrinoviricota*, Branch 4 = *Duplornaviricota*, and Branch 5 = *Negarnaviricota*. Adapted from (Wolf et al., 2018).

RNA viruses in the aquatic environments may infect a wide range of metazoan hosts (marine mammals, seabirds, fish, crustaceans, and bivalves), resulting with significant economic and ecological consequences in populations of wild and farmed marine animals (Lang et al., 2009; Sivasankar et al., 2017). RNA viruses infecting microbial eukaryotes have important ecological roles in regulating the structure of plankton communities in the ocean (Gustavsen et al., 2014; Miranda et al., 2016; Moniruzzaman et al., 2017) . RNA viral infection of zooplankton and large phytoplankton can alter the efficiency of the biological carbon pump. (Figure 2.2.) (Kaneko et al., 2021; Suttle, 2007; Zimmerman et al., 2020). The analysis of eukaryotic virus composition showed that eukaryotic viral composition can predict carbon sequestration, possibly by affecting the carbon flux through the viral shunt and viral shuttle (Kaneko et al., 2021).

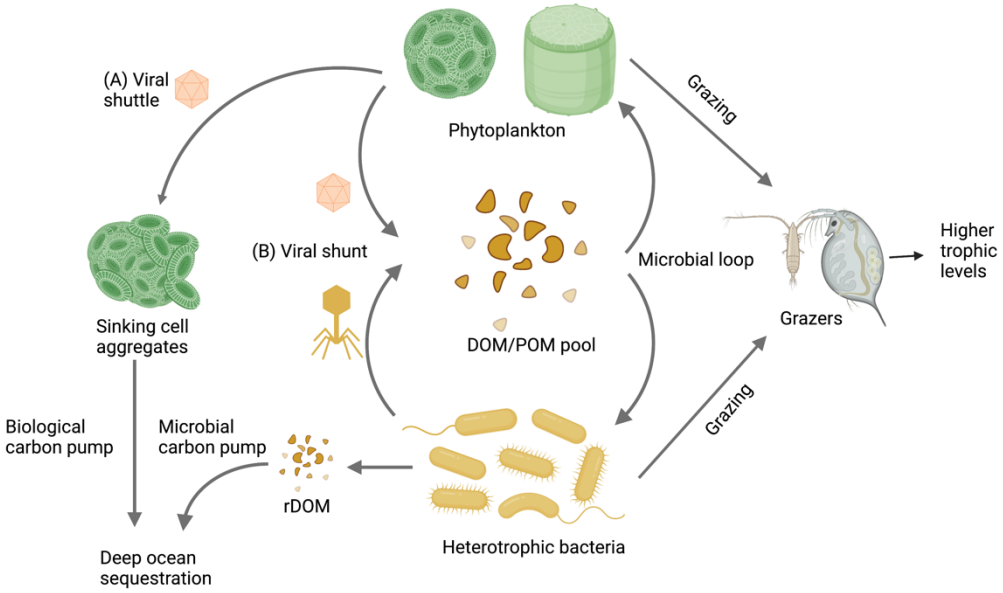


Figure 2.2. Marine viruses simultaneously control two processes in the carbon cycle: (**A**) the viral shuttle and (**B**) the viral shunt. In the viral shuttle, viral lysis of phytoplankton cells produces sticky aggregates with negative buoyancy that enhance the biological carbon pump by sequestrating carbon in the deep ocean. In the viral shunt, viral lysis of cells has the opposite effect—it diverts the organic matter into a dissolved organic matter (DOM) pool that is rapidly and continuously recycled in the surface waters, preventing its sequestration or uptake by higher trophic levels. Adapted from (Zimmerman et al., 2020). Created in BioRender.com.

Yet, despite their fundamental roles, only 12 RNA viruses and 41 DNA viruses have been isolated from protistan hosts due to cultivation challenges (Coy et al., 2018; Sadeghi et al., 2021; Short et al., 2020). Routine methods for assessing viral abundance in aquatic environments, such as epifluorescence microscopy and flow cytometry, currently lack the sensitivity to detect the small genomes of RNA viruses (Brussaard and Payet, 2010; Patel et al., 2007; Tomaru and Nagasaki, 2007). Consequently, the abundance of RNA viruses in the virioplankton is presently unknown. Yet, there is some evidence that RNA viruses may exceed DNA viruses in abundance at times (Miranda et al., 2016; Steward et al., 2013). Marine RNA viruses are small in size, have a lytic cycle, and have very high burst sizes comparing to the DNA viruses, releasing 1000 to 10,000 particles per infected cell, which could affect their abundances (Lang et al. 2009; Culley

2018). Taken together, these results may indicate that abundances and ecological significance of RNA viruses may be underestimated in aquatic ecosystems.

Traditionally, RNA metaviromics has been the key method to study diversity and dynamics of aquatic RNA virus communities. A comprehensive review summarized gene- marker studies targeting the RNA-dependent RNA polymerase gene and the first RNA metaviromics studies revealing consistent dominance of picorna-like RNA viruses in marine environments (Culley 2018). Knowledge about the protist-infecting RNA viruses isolated through culturing was recently summarized by Sadeghi et al. (Sadeghi et al., 2021). Metatranscriptome sequencing greatly expanded known RNA viral diversity associated with invertebrate holobionts, bacteria, plants, and eukaryotes in the soil (Callanan et al., 2020; Roossinck, 2019; Roossinck et al., 2015; Shi et al., 2016; Starr et al., 2019). More recently, a growing number of studies have successfully explored the potential of metatranscriptome mining for RNA virus detection and discovery in aquatic ecosystems. To our knowledge, no reviews have systematically documented this progress from a metatranscriptomic point of view. Here, we describe the most commonly used meta-omic methodological approaches employed in aquatic RNA virology and examine their strengths and limitations. Throughout this review, we refer to meta-omic approaches as viromics, metatranscriptomics, and viral dsRNA sequencing.

We explore in detail 16 recent studies conducted in natural aquatic environments and 4 studies focusing on individual protists and macroalge, and discuss how they have broadened our understanding of RNA viral diversity. The focus of this review is on RNA viruses, but if other viral groups were concurrently detected in a referenced study, they are discussed as well to get a more comprehensive overview of the viruses present

in a specific environment. Finally, we address the remaining challenges and provide suggestions for future explorations of the ecological relevance of RNA viruses in aquatic ecosystems.

2.3. Methodological challenges in the study of aquatic RNA viral communities

Two commonly used approaches for studying viral diversity in aquatic environments are metaviromics (i.e., sequencing of enriched virus particles) and data mining of environmental metagenomes and metatranscriptomes for virus discovery (Figure 2.3). These approaches are so successful that uncultivated viral genomes of DNA viruses now greatly outnumber cultivated viral isolates in the public databases (Roux et al., 2019).

Metaviromics (or viral metagenomics) is the method of sequencing extracellular total virus-like particles of either DNA or RNA viruses (Table 2.1; Figure 2.3). Enrichment of extracellular, free-floating virus-like particles is achieved through four steps: (i) removal of host cells and their nucleic acids by filtration through 0.22 μm filters, (ii) concentration of the viral fraction via tangential flow filtration (Suttle et al., 1991) or iron flocculation (John et al., 2011), (iii) gradient centrifugation and/or nuclease treatment to eliminate contaminating non-encapsidated (host) nucleic acids (Hurwitz et al., 2013), and (iv) nucleic acid extraction from virus-like particles (Thurber et al. 2009). Due to technical challenges, such as filtering large volumes of water and low nucleic acid yields after the extraction, studies using this approach typically target either DNA or RNA viruses. An advantage of the viral enrichment process is an increased proportion of viral reads resulting in improved genome assemblies that enable sensitive detection even of low- abundance viruses (Hurwitz et al. 2013; Thurber et al. 2009). Limitations of this method include cellular contamination of the viral metagenomes, which can hinder the data analysis (Roux et al., 2013).

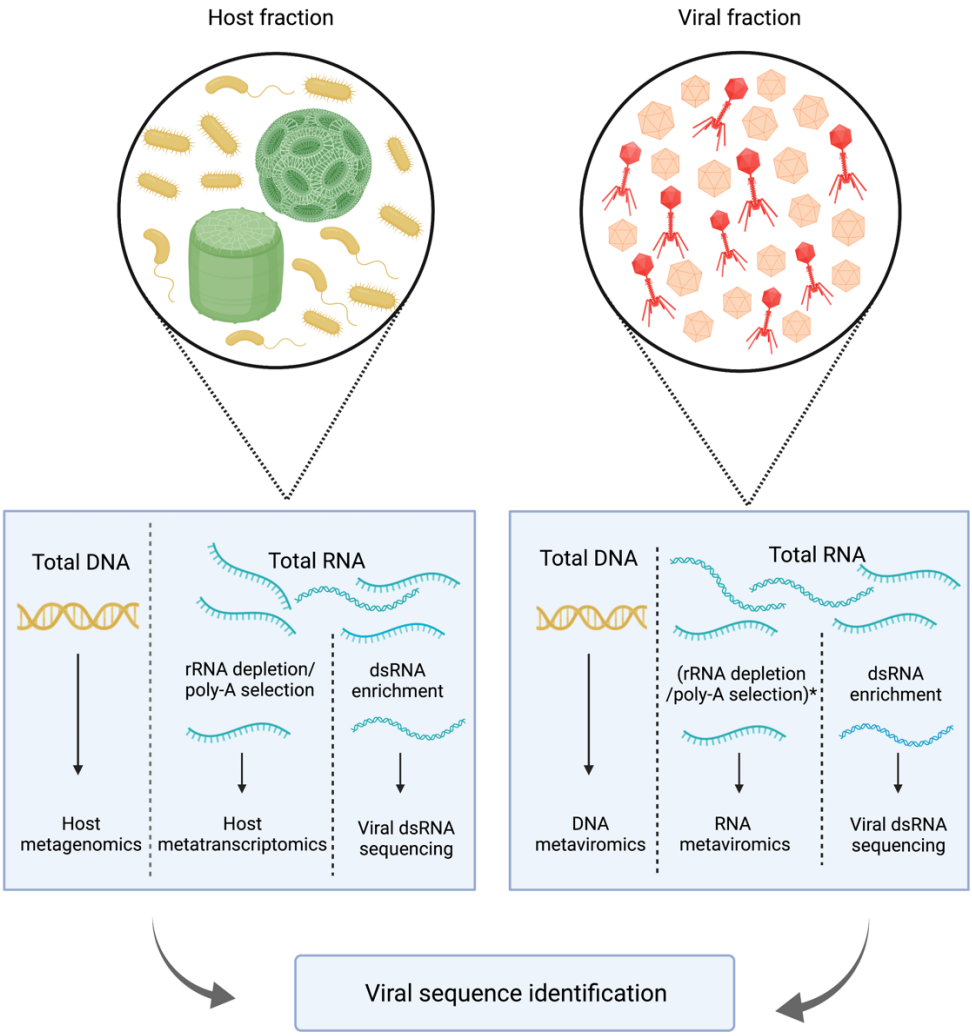


Figure 2.3. Most commonly used meta-omic sequencing approaches in marine virology. Host fraction (>0.22 µm) metagenomes and metatranscriptomes are mined for viruses, or DNA or RNA is extracted from the enriched viral fraction (<0.22 µm). * rRNA depletion or poly-A selection for RNA in the viral fraction is typically not performed due to low yields. An alternative approach that can enrich for replicative forms of RNA viruses inside the cells or for extracellular dsRNA viruses is dsRNA sequencing. Adapted from (Roux, Adriaenssens, et al., 2019). Created with BioRender.com.

Additionally, the filtration process will systematically remove viruses larger than 0.22 µm, such as giant viruses or long filamentous viruses such as archaeal *Lipothrixviridae* or bacterial *Inoviridae* that can be up to 3.5 µm long (Prangishvili et al., 2017; Roux et al., 2019a; Roux al., 2019b). The latter technical limitation should not affect extracellular RNA viruses since the particle size is consistently around 30 nm, but the removal of the cellular fraction will exclude RNA viruses that do not possess a capsid,

do not go through an extracellular state, and are vertically transmitted (Hillman and Cai, 2013; Suzuki et al., 2018; Valverde et al., 2019). Conventional extraction and sequencing library preparation methods may not suit all genome types, limiting the diversity of viruses that can be detected (Labonté and Suttle, 2013; Wilcox et al., 2019). For example, ssDNA viruses require a special extraction method and sequencing approach (Labonté and Suttle, 2013) and are therefore systematically omitted from any study targeting dsDNA viruses, and dsRNA viruses may be underrepresented in the RNA viromes due to inefficient conversion to cDNA during the preparation of the sequencing library (Wilcox et al., 2019).

Table 2.1. Advantages and shortcomings of the meta-omic sequencing approaches used in marine viral ecology studies. Adapted from (Roossinck et al., 2015).

Method	Virus Nucleic Acid Detected	Shortcomings	Advantages
Viromics or viral metagenomics	RNA or DNA viral genomes in the extracellular stage	RNA viruses targeted separately than DNA viruses Needs special enrichment for dsRNA and ssDNA Large DNA viruses are filtered out High-burst-size viruses can be overrepresented	Enriched for viral sequences, better assembly
Metatranscriptomics	Transcripts of (+) and (−) ssRNA, dsRNA, ssDNA, dsDNA	High background of nonviral sequences Potentially fragmented assemblies Misses low-titre viruses Does not distinguish between (+) ssRNA viral genome and transcripts	Captures all types of DNA and RNA viruses simultaneously Captures active infection (for DNA viruses) Can capture RNA viruses without capsids
dsRNA sequencing	ssRNA as replicative intermediate dsRNA genomes	Not as effective for (−) ssRNA and DNA viruses Cellular metatranscriptomes are removed in the enrichment process	Enriched for RNA viruses Can be used for detection of both extracellular (<0.22 µm fractions) and intracellular RNA viruses

For ‘omic analysis, marine microbial cells (single-celled eukaryotes and prokaryotes) are typically collected onto filter membranes with pore sizes ≥0.22 μm (Mueller et al., 2014). Typically, extracellular DNA and RNA viruses will pass through these filters, though viral aggregates or particle-associated extracellular viruses can still be retained (Yamada et al., 2020). Viral sequences can be retrieved from both metagenomes and metatranscriptomes derived from the cellular fraction (Alarcón-Schumacher et al., 2019; Carradec et al., 2018; Endo et al., 2020; Gann et al., 2021; Ha et al., 2021; Hewson et al., 2018; Kaneko et al., 2021; Kolody et al., 2019; Moniruzzaman et al., 2017; Pound et al., 2020; Sieradzki et al., 2019; Zeigler Allen et al., 2017).

Metatranscriptomics, which is the sequencing of RNA transcripts from environmental microbial communities, is well suited to study the active viral infection in diverse viral groups with structurally different genomes (Table 2.1; Figure 2.3). As a part of the infection cycle, all seven Baltimore classes of viruses produce mRNA intermediate inside infected cells and can be simultaneously captured by metatranscriptomics (Depledge et al., 2019; Zhang et al., 2019). The additional benefit of this approach is detection of RNA viruses that lack capsid and extracellular stages and are therefore absent from the viral particle fraction, such as *Narnaviridae*, *Hypoviridae,* and *Endornaviridae* (Hillman and Cai, 2013; Suzuki et al., 2018; Valverde et al., 2019). The main preparations steps include (i) RNA extraction from cells collected on a filter, (ii) DNase treatment to remove the contaminating cellular DNA, and (iii) cellular mRNA enrichment by depletion of highly abundant ribosomal RNAs or by selection of polyadenylated RNA before the library preparation and sequencing. Ribosomal RNA is typically captured by complimentary probes bound to magnetic beads, while polyadenylated RNA is typically selected by binding to oligo-dT probes (Depledge et

al., 2019; Zhang et al., 2019). The removal of ribosomal RNA is a less selective process because it retains all nonpolyadenylated and adenylated cellular RNA. In the context of viruses, this means that with rRNA depletion, RNA viruses that lack a poly-A tail and fragmented or degraded viral RNA genomes can be included in the sample, while during poly-A selection, only RNA viruses (and cellular mRNA) with intact poly-A tail will be retained. This is supported by a recent study demonstrating that rRNA-depleted libraries produced more viral reads, longer contigs, and greater viral richness compared to poly-A-selected libraries (Gann et al., 2021). While these results will require additional experimental confirmation as different filter sizes were used for the two treatments, potentially biasing the data, similar results were observed in metatranscriptomes of marine invertebrates (Rosani et al., 2019). Limitations of this method include a high abundance of rRNA reads in the metatranscriptomes of cellular fractions even after the enrichment steps (Charon et al., 2020; Urayama et al., 2018) and the fact that the majority of non-rRNA sequences in the cellular metatranscriptomes are of host origin, while viral reads typically constitute less than 1% of the sequenced reads (Table 2.2. and references there).

Additional complexity in metatranscriptomics arises from viral genes being transcribed sequentially and in mono- or polycistronic transcription units, which can lead to uneven genome coverage (Depledge et al., 2019). The high background of nonviral sequences, combined with the heterogeneity of viral expression, also leads to more fragmented genome assemblies and is the biggest limitation of this approach (Shi et al., 2018). Furthermore, for RNA viruses whose genome serves as mRNA, it is not possible to distinguish the viral genome from viral transcripts in the metatranscriptomic datasets.

Table 2.2. Basic geographic, ecosystem, and sampling information about the aquatic viromic (turquoise) studies, metatranscriptomic (coral) studies, and dsRNA sequencing (yellow) reviewed in this paper. Asterisk * denotes a temporal study in multiple locations.

Region	Sampling Location	Site Characteristics	Sampling Scheme	Temporal/ Spatial	Host or Viral Fractions Collected	Relative Abundance of Viral Reads in the dataset	Number of RNA Viral Signatures (RdRp) or Genomes Recovered	Reference
Global	Multiple locations	Deep sea sediment	133 locations 133 RNA viromes	Spatial	Only viral fraction (RNA)	Not reported	85,059 RNA viruses (RdRp) 1463 RNA viral genomes	(Zhang et al. 2022)
Polar	West Antarctic Peninsula	Highly productive coastal area	1 location 6 RNA viromes	Temporal	Only viral fraction (RNA)	15%	5 RNA viral genomes	(Miranda et al., 2016)
Polar and temperate	Multiple locations	Multiple site characteristics	11 locations 14 RNA viromes	Spatial	Only viral fraction (RNA)	Not reported	3 RNA viral genomes	(Vlok et al., 2019)
Temperate	Jericho Pier, Georgia Strait, Canada	Highly productive coastal area	2 locations 2 RNA viromes	Spatial	Only viral fraction (RNA)	Not reported	3 RNA viral genomes	(Culley et al., 2007; Culley, et al. 2006)
Temperate	Yangshan Harbour, China	Brackish coastal area	1 location 1 RNA virome 1 DNA virome	Spatial	Only viral fraction (RNA and DNA)	Not reported	4593 RNA viruses (RdRp)	(Wolf et al., 2020)
Subtropical	Kane'ohe Bay, Hawai'i	Coastal area	1 location 2 RNA viromes	Spatial	Only viral fraction (RNA)	54%	6 viral genomes	(Culley et al. 2014)
Subpolar	Honshu, Japan Jamstec cruise	Coastal and pelagic	5 locations 5ssRNA+5dsRNA metatranscriptomes 5ssRNA+5dsRNA virome	Spatial	Host: Unfractionated >0.22 µm dsRNA + ssRNA Viral fraction (RNA) dsRNA + ssRNA	0.1% ssRNA metatranscriptomes 1.3%–36.6% dsRNA metatranscriptomes 5.5% in ssRNA virome 35.2%in dsRNA virome	1270 RNA viruses (RdRp)	(Urayama et al., 2018)
Polar	Chile Bay, Antarctica	Highly productive coastal area	1 location 3 weeks apart 2 metatranscriptomes 2 metagenomes	Temporal	Host: Fractionated 8 µm–0.22 µm fraction	0.04%–0.05% (rRNA depletion)	2 ssRNA viral genomes	(Alarcón-Schumacher et al., 2019)
Global	Multiple locations Tara Ocean Expedition	Multiple site characteristics	68 locations 2 layers (surface, deep) 441 metatranscriptomes	Spatial	Host: Fractionated 2000–180 µm 180–20 µm 20–5 µm 5–0.8 µm	0.0006% to 0.4% (poly-A selection)	29,023 giant virus (PolB) 2753 ss RNA virus (RdRP) 710 ds RNA virus(RdRP)	(Carradec et al., 2018)
Global	Multiple locations Tara Ocean Expeditions	Multiple site characteristics	59 locations	Spatial	Same as above	Same as above	388 giant virus + 3,486 in metagenomes (PolB) 975 RNA viruses (RdRP)	(Kaneko et al., 2021)

Region	Sampling Location	Site Characteristics	Sampling Scheme	Temporal/ Spatial	Host or Viral Fractions Collected	Relative Abundance of Viral Reads in the Metatranscriptome	Number of RNA Viral Signatures (RdRp) and Genomes Recovered	Reference
Global	Multiple locations *Tara* Ocean Expeditions + *Tara* Oceans Polar Circle Expeditions	Multiple site characteristics	121 locations 771 metatranscriptomes (584 eukaryotic and 187 prokaryotic fractions)	Spatial	Same as above	Same as above	44,828 RNA virus (RdRp fragments) 6686 RNA virus (near or full-length RdRp)	(Zayed et al., 2022)
Temperate	Narragansett Bay (NB) Quantuck Bay (QB), USA	Eutrophic (bloom) coastal	2 locations 5 metatranscriptomes in 4 weeks (NB) 3 metatranscriptomes in a week (QB)	Temporal *	Host: Fractionated QB 5–0.22 µm NB >5 µm	0.043%–2.4% (poly-A selection)	18 RNA viral genomes	(Moniruzzaman et al., 2017)
Temperate	Lake Tai, China	Hypereutrophic (bloom) lake	9 locations 1 x monthly for 5 months 35 metatranscriptomes	Temporal *	Host: Unfractionated >0.22 µm	0.02% (rRNA depletion)	457 giant virus (PolB) 193 ssRNA virus (RdRP) 75 dsRNA virus (RdRp) 16 RNA viral genomes	(Pound et al., 2020)
Temperate	Lakes Owasco, Seneca, Cayuga USA	Mezotrophic to eutrophic lakes	3 locations 1 x monthly for 10 months 30 metatranscriptomes	Temporal *	Host: Fractionated 5–0.22 µm	0.6% (rRNA depletion)	35 RNA virus (RdRp)	(Hewson et al. 2018)
Temperate	California Current, USA	Oligotrophic, with upwelling	1 location Every 4 h for 60 h 16 timepoints 29 metatranscriptomes	Temporal	Host: Fractionated >5 µm 5–0.22 µm	Not reported	Not reported	(Kolody et al., 2019)
Temperate	Quantuck Bay (QB), Tiana Beach (TB), USA	Eutrophic (bloom) coastal	2 locations 1x weekly for 10 weeks (QB) 1 x weekly for 8 weeks (TB) 52 metatranscriptomes	Temporal*	Host: Unfractionated >0.22 µm (for rRNA reduction) >1 µm (for poly-A selection)	0.33%–0.53% in rRNA depleted libraries 0.02%–0.023% in poly-A selected libraries	5316 RNA virus (RdRp) in rRNA depleted libraries 1094 RNA virus (RdRp) in poly-A selected libraries	(Gann et al., 2021)
Temperate	Baltic sea, Lake Torneträsk	Eutrophic, mostly brackish	11 location (2 depths) 39 metatranscriptomes 7 DNA viromes	Spatial	Host: Fractionated 200–3.0 µm 3.0–0.8 µm 0.8–0.1 µm + viral fraction (DNA)	3.2% (Separate rRNA depletion and poly-A selection libraries)	Not reported	(Zeigler Allen et al., 2017)

Standard methods for reverse transcription used in RNA sequencing are inefficient in transcribing dsRNA molecules as the presence of complimentary strand will prevent the binding of random primers. Therefore, compared to ssRNA viruses, the dsRNA viral lineages may be underrepresented in both metatranscriptomic and viral RNA metagenomic datasets (Wilcox et al., 2019). An approach that might alleviate this problem is dsRNA sequencing (Table 2.1; Figure 2.3). Presence of long dsRNA in cells is a unique hallmark of viral infection. Double- stranded viral RNA in cells originates from genomes of dsRNA viruses or from replicative intermediates of ssRNA viruses, with the exception of retroviruses. For DNA viruses, bidirectional transcription (sense and anti-sense direction) of overlapping regions can lead to complimentary transcripts that will form a dsRNA duplex. However, only dsRNA and (+) ssRNA viruses produce dsRNA in amounts detectable by immunofluorescence (Weber et al., 2006). Separation of viral dsRNA from total RNA via cellulose chromatography or more recently through immune-based capture, followed by dsRNA sequencing, has been widely used for virus detection and identification in plants and fungi and has led to the discovery of many novel viruses (Blouin et al., 2016; Marais et al., 2018; Roossinck, 2019). Immuno-based capture of dsRNA with anti-dsRNA antibodies was used to pull down multiple (+) ssRNA viral genomes with high recovery (31–74% viral reads) from singular plant samples, including the near-complete genome of an unknown virus (Blouin et al., 2016).

Recently, a novel dsRNA sequencing approach, fragmented and primer-ligated dsRNA sequencing (FLDS), has been developed and applied to marine environmental samples, demonstrating the method's potential to capture RNA viral diversity that previously went undetected (Urayama et al., 2018). After conventional total RNA extraction from either viral concentrates or cells, extracted total RNA is enriched for dsRNA via cellulose chromatography. Primers are ligated to the ends of fragmented dsRNA, which

is then denatured, and reverse transcription into cDNA with complementary primers is initiated from the ends of the molecules. RNA is degraded with RNase and a complimentary DNA strand synthetized by PCR. Using the FLDS method, on average, 11.3–36.6% of reads obtained from the cellular fraction may be identified as viral as opposed to 0.1% using the classic RNA sequencing approach with rRNA depletion. Similar trends were observed for the viral fraction, where viral reads comprised 35% of total reads with the FLDS method, and only 5% with classic RNA sequencing. A very low number of reads were assigned to rRNA (Urayama et al., 2018). Overall, virus enrichment through dsRNA purification might be an interesting avenue to explore for research of aquatic RNA viral communities, followed by either standard dsRNA sequencing used in plant virus detection (Marais et al., 2018) or the FLDS dsRNA sequencing method (Urayama et al., 2018).

In synthesis, there is currently no method that will comprehensively capture all viral diversity. The comparative study of RNA viromes and metatranscriptomes has highlighted tremendous differences between the fractions as well as between different methodological approaches, which will be further discussed below (S. Urayama et al., 2018). This observation is not unique to RNA viruses because, for example, the genomes of uncultivated DNA viruses (>10 kb) in the Tara Oceans dataset recovered from virus-enriched fractions and by metagenome mining exhibit huge differences in diversity with an average of 75% and 90% unique sequences in the microbial and viral fractions, respectively (Roux et al., 2019a). This illustrates that every detection method comes with its biases and using only one approach may show an incomplete or biased snapshot of natural viral communities, underscoring the importance of using multiple approaches.

2.4. Environmental RNA metaviromics: lytic positive-sense ssRNA viruses dominate pelagic but not benthic RNA viral assemblages

The first environmental viral metagenomic surveys consistently showed that most sequences in the RNA viromes were unknown. Among known sequences, typically more than 90% of RNA viral sequences identified as (+) ssRNA viruses from order *Picornavirales* in temperate, tropical, and polar oceans, often clustering with known RNA viruses that infect diatoms. Very few sequences of dsRNA viruses were typically recovered, and RNA phages or negative-sense RNA viruses were absent. Only five RNA metaviromic studies in the water column have been performed to date, predominantly in coastal productive environments, with only one study having an extensive geographical distribution (Wolf et al. 2020; Miranda et al. 2016; Vlok et al. 2019; Culley et al. 2006; Culley et al. 2014). Recently, the first RNA metavirome study focusing on deep sea sediments was published as a preprint (Figure 2.4; Table 2.2) (Zhang et al. 2022).

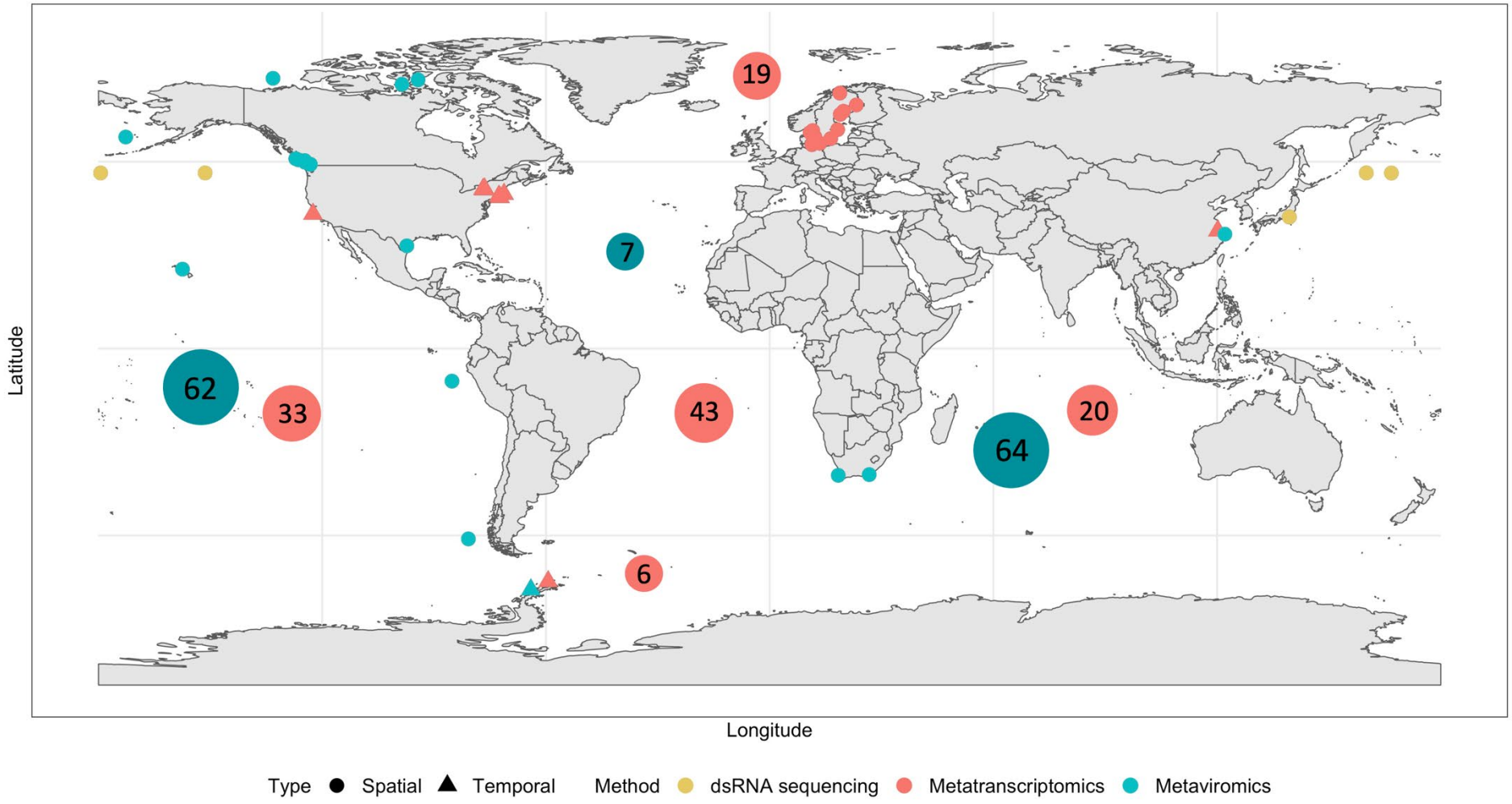

19
7
62
33
43
64
20
6
Latitude
Longitude
Type
Spatial
Temporal
Method
dsRNA sequencing
Metatranscriptomics
Metaviromics

Figure 2.4 (previous page). Geographic locations of aquatic metaviromic (blue), metatranscriptomic (coral), and dsRNA sequencing (yellow) studies focusing on RNA viruses reviewed in this paper. Spatially focused studies are represented with circles, and temporally focused studies are represented with triangles. Two global surveys with extensive geographic coverage, Dayang No. 1 cruises and Tara Oceans Expedition and Tara Oceans Polar Circle Expeditions (Carradec et al., 2018; Zayed et al., 2022; Zhang, 2022), are represented by large bubbles in each of the five oceans. The size of the bubble is proportional to the total number of samples collected from each ocean, with numbers of benthic virome samples from the Dayang No. 1 cruises (Zhang et al. 2022) indicated in teal blue, and number of pelagic metatranscriptome samples from the Tara Oceans Expedition (Zayed et al., 2022) in coral. Samples from the Mediterranean Sea and the Red Sea were included in the counts of the Atlantic Ocean.

In RNA metaviromics study of coastal, temperate waters of British Columbia, 98% of the identifiable reads represented (+) ssRNA viruses (Culley et al. 2006). In the Jericho Pier library, viral sequences from the order *Picornavirales* were most abundant, mainly consisting of families *Marnaviridae* and *Dicistroviridae*. Two complete viral genomes (JP-A and JP-B) were phylogenetically similar to the diatom-infecting *Rhizolenia setigera* RNA virus RsRNAV and could have a protist (diatom) host. In the Strait of Georgia library, the majority of viral sequences fell into *Tombusviridae*, a family of viruses from the phylum *Kitrinoviricota* (Culley et al. 2006). Tropical seawater viral RNA metagenomes of Kaneohe Bay, Hawaii again confirmed the dominance of the order *Picornavirales*, which encompassed 95% of the assignable reads, with a minority of reads being identified as dsRNA viruses (Culley et al. 2014). Similarly, the full genomes recovered from the dataset were phylogenetically related to diatom-infecting viruses, or viruses from families *Marnaviridae* and *Dicistroviridae,* consistent with the previous RNA metaviromic study in temperate waters.

The trend continued with the overwhelming majority of identifiable RNA viral sequences (97.8%) from polar RNA viral metagenomes collected during a spring diatom bloom in West Antarctica Peninsula (Miranda et al., 2016) being classified as (+) ssRNA from the order *Picornavirales*. Out of five full genomes assembled from the dataset, three (PAL128, PAL156, and PAL473) clustered together with diatom-infecting RNA viral isolates or marine RdRp sequences, and two (PAL 438 and PAL E4) contained

highly divergent RdRps. There was a smaller proportion of viral sequences identified as *Marnaviridae* or *Dicistroviridae* compared to the tropical RNA viral metagenomes of Kaneohe Bay. Instead, the relative abundance of diatom-infecting RNA viruses from the genus *Bacilarnavirus* surpassed 90% in three samples or reached around 50%, with unclassified *Picornavirales* taking up the rest. The relative abundance of RNA viruses in the total virioplankton was estimated to range 2–79% in samples on the basis of nucleic acid content of gradient-purified fractions (Miranda et al., 2016).

The largest spatially extensive study of RNA viral metagenomes in the water column to date explored the biogeography of the six marine picorna-like RNA viruses: three complete genomes of picorna-like viruses recovered in this metagenomic study itself (Vlok et al., 2019), two complete picorna-like genomes recovered in a previous metagenomic study (Culley et al. 2006), and one genome originating from an isolate of HaRNAV, which is genetically more divergent and infects a raphidophyte. Out of the three full genomes recovered in this study, two of them have the highest similarity to PAL 156 genome and one has the highest similarity to PAL 128, the full-length genomes recovered in the temporal study of Antarctic diatom bloom (Miranda et al., 2016) that are thought to be infecting diatoms. Reads from global metagenomic datasets comprising marine, freshwater, and reclaimed water RNA viromes were recruited to these genomes, revealing high geographical and temporal variation of RNA viruses. On a temporal scale, samples collected within a year at the same sample site, Jericho Pier in British Columbia, differ remarkably in distribution of six genomes. In 2013, raphidophyte-infecting HaRNAV was the most abundant virus at Jericho Pier, and in 2014, it was diatom-infecting BC-2 and BC-3 viruses, suggesting that the temporal distribution of RNA viruses in one location is controlled by biotic factors such as host availability. On a spatial scale, the diatom- infecting JP-A was the most abundant

genome type off the coast of South Africa, and the raphidophyte-infecting HaRNAV genome dominated the samples from Bering Sea, Laguna Madre in the USA, and samples from Peru and Chile. In the samples from Queen Charlotte Strait and Johnstone Strait in Canada, reads were more equally distributed between diatom- infecting JB-A and JP-B and raphidophyte-infecting HaRNAV genomes. Plenty of viral sequences mapped to the six genomes with low amino acid identity, demonstrating ubiquity and richness of aquatic picorna-like viral communities worldwide. Even if the study did not explore the full diversity of RNA viruses at each sampling location, but instead focused on six selected picorna-like viruses, it demonstrated how dramatically the abundance of certain RNA viral groups can vary between locations (Vlok et al., 2019).

More recently, a study recovered 4593 near-full-length RdRp sequences clustered within 2192 clusters at 75% amino acid identity (AAI), in one water sample from a single location in China (Wolf et al., 2020). Consistent with previous studies, no (+) ssRNA enveloped viruses or (−) ssRNA viruses have been detected and only six RNA viruses were identified as dsRNA viruses, supporting the idea that positive-sense RNA viruses are dominant in the marine environment. In contrast to previous studies, all three phyla of (+) ssRNA viruses were present in the samples, with the order *Picornavirales* constituting only around 26% of RdRp sequences in the sample. Within the phylum *Pisuviricota*, the largest number of RdRp sequences (n = 854) grouped with the aquatic picorna-like viruses (order *Picornavirales*) encompassing *Marnaviridae* and other protist-infecting viruses. *Protopotyviruses*, ancestors of complex plant-infecting *Potyviruses*, formed a new clade that was not previously detected in marine environments. Within the phylum *Kitrinoviricota*, many highly divergent RdRp sequences formed a new, phylogenetically well supported clade with previously "orphan" RdRPs of RNA viruses found in marine invertebrates and wastewater, with

multiple new orders and families such as *Weiviruses*, *Yanviviruses*, *Zhaoviruses,* and *Shangaiviruses,* significantly expanding the presence of this phylum in aquatic environments, previously limited to detection of *Tombusviridae* in one metagenomic library. Within the phylum *Lenarviricota*, a high number of sequences had similarity to the *Leviviridae*, (+) ssRNA bacteriophages, and the simple eukaryote-infecting *Ourmiaviridae*, documenting for the first time the presence of (+) ssRNA bacteriophages as members of marine RNA viral communities. Capsidless endogenous RNA viruses, such as *Mitoviridae* and *Narnaviridae*, were absent from the extracellular RNA virome as they would have been removed together with cells during filtration. The largest RdRp cluster was affiliated with the *Marnaviridae*, viruses of diatoms and stramenopiles and, together with the use of the alternative (ciliate) genetic code, might suggest that a considerable fraction of RNA viruses infect protists (Wolf et al., 2020).

The high numerical abundance of viral particles and the tremendous impacts of viral infection on bacterial and archaeal communities in marine sediments have been well documented (Danovaro et al. 2016; Danovaro et al. 2015). As with the water column studies, virus research in marine sediments has been predominantly focusing on dsDNA viromes (Corinaldesi et al., 2017; De Corte et al., 2019; Z. Li et al., 2021; Yoshida et al., 2013), with reports of ssDNA (Yoshida et al., 2018) and RNA viruses (Zhang et al. 2022) significantly lagging. A recent preprint by Zhang et al. (Zhang et al. 2022) is the only available RNA metaviromic study in marine sediments, and with 133 samples and almost 3 billion of raw reads, is the largest marine RNA virome dataset to date. Sediments were collected between 2010 and 2018 in a wide range of benthic deep seafloor environments around the globe: hydrothermal vents, cold seeps, sea mounts, and ocean basins at depths between 1000 and 6000 m. The sequenced RNA viromes were searched for marker genes for both kingdoms *Orthornavirae* and *Pararnavirae,*

RdRp genes, and RT genes, respectively. Most of the identifiable viral RNA sequences were classified as RT-containing *Retroviridae* and *Metaviridae.* Out of RNA viruses with the RdRp gene, the two most abundant families were surprisingly dsRNA viruses from families *Totiviridae* and *Cystoviridae,* in contrast with water column RNA viromes that are typically dominated with ssRNA viruses. Single-stranded RNA viruses from family *Leviviridae* were the most abundant ssRNA viruses detected, supporting the idea that RNA bacteriophages are important members of marine RNA viral communities in the water column as well as in sediments. However, most of the viral sequences could not be assigned to known viral groups, suggesting that the deep sea might represent a rich reservoir of unexplored viral diversity. Alternatively, this might just reflect the current limitations of the available software programs in assigning and classifying viral sequences from environmental samples (Zhang et al. 2022).

It is still unclear what fraction of the large differences in the estimates of RNA viral diversity between different studies and different habitats is real rather than a result of methodological biases. RNA viruses from order *Picornavirales* have very high burst sizes (Lang et al., 2009), and it has been shown that viruses with higher burst sizes can be overrepresented in enriched samples (Roux et al., 2019). Consistent with this, in the virus-enriched samples from a diatom bloom from polar waters, the diversity was highest in samples with low abundance of RNA viruses (Miranda et al., 2016). Similarly, in the virus-enriched samples from temperate waters, the reads in the two libraries were dominated by a few contigs: in Jericho Pier library, 66% of the library assembled in four contigs, and in Strait of Georgia library, 59% of reads assembled into one viral contig (Culley et al. 2006). It is evident that the study by Wolf et al. (2020) uncovered a staggering diversity of (+) ssRNA viruses in the aquatic environments, broadly expanding the typical range of viral diversity in the first studies. This study had

much deeper sequencing depth, as well as a different library preparation method. In the other pelagic viromic studies (Miranda et al. 2016; Vlok et al. 2019; Culley et al. 2014; Culley et al. 2006), RNA was pre-processed by sequence-independent, single-primer amplification (SISPA) prior to sequencing, which can lead to overrepresentation of the dominant sequences (Karlsson et al., 2013). More recently, the characterization of the benthic RNA virome by Zhang et al. (2020) also had a greater sequencing depth than that of earlier studies, but the results might have been skewed by biases inherent with the use of a Phi29 multiple displacement amplification (MDA) method during library preparation (Marine et al., 2014; Yilmaz et al., 2010). Whether the detected differences in RNA viral diversity are due to potential amplification bias of SISPA libraries and MDA amplification or due to sequencing depth or some other factor will be clarified as more studies become available and published.

2.5. Environmental (viral) metatranscriptomics: sampling with size fractionation improves metatranscriptome resolution and uncovers the role of ssRNA viruses

Metatranscriptomics can capture the active viral infection of diverse viral groups and has been increasingly used in studies of RNA virus community diversity and ecological dynamics in the last few years. Most studies were performed in coastal, productive environments, either on temporal or on spatial resolution (Figure 2.4) and employing different filtration or rRNA removal approaches. Typically, protist and bacterial communities are collected either as “whole” microbial fraction without prefiltration or separated into size fractions using serial filtration, which may allow for higher resolution (Table 2.2. and references there).

One of the first studies that used environmental marine metatranscriptomes followed active infections of diverse viral groups (dsDNA, ssRNA, dsRNA) in two highly

productive sites in the US, Narragansett Bay and Quantuck Bay. Seasonal diatom blooms occur in Narragansett Bay, and harmful brown tide blooms caused by unicellular eukaryotic algae *Aureococcus anophagefferens* take place in Quantuck Bay. Active infections of eukaryotic viruses were tracked during the initiation, peak, and demise of the bloom in Quantuck Bay and in the steady-state, non-bloom conditions in Narraganset Bay using marker gene approach (Moniruzzaman et al., 2017). Contigs were searched for marker genes, including the major capsid protein (MCP) for giant viruses, RdRp for RNA viruses, and the viral replicase for ssDNA viruses. Giant viruses from families *Mimiviridae* and *Phycodnaviridae* were continuously present in both locations, with two distinct dynamics. Some giant viruses were steadily expressed across sampling dates, and others exhibited sharp peaks in abundance, or a "bloom and bust" type of expression. In the non-giant virus community, between 62% and 74% of contigs in both bays were unclassified (+) ssRNA viruses in the order *Picornavirales*, followed by classified families *Picornaviridae*, *Secoviridae,* and *Dicistroviridae* from the same order. Few dsRNA viruses resembling *Totiviridae*, *Partitiviridae,* and *Hypoviridae* were detected. The viral community (excluding giant viruses) in Narraganset Bay was steadily dominated (>90%) by *Picornavirales*. In Quantuck Bay, both ssDNA and unclassified *Picornavirales* were present in the *Aureococcus anophagefferens* bloom samples in equal relative abundances, around 50%. Unclassified *Picornavirales* viruses took over (>90%) when bloom disintegrated, possibly infecting other protists in the bloom succession. Co-occurrence analysis between viruses and host marker genes found that most of the hosts might be phytoplankton or heterotrophic protist, but also fungi (Moniruzzaman et al., 2017). Another temporal study of the RNA viral community of three temperate lakes detected RNA viruses resembling families *Picornaviridae*, *Virgaviridae,* and *Reoviridae* with putative protist or invertebrate hosts. The low number

of contigs retrieved might be due to low sequencing depth of 1–5 million reads per library .

More recently, the active viral community within a *Microcystis*-dominated harmful algal bloom in a temperate Lake Tai in China was characterized using the same viral marker genes approach (Pound et al., 2020). As expected, most of viral transcripts (48%) present in the metatranscriptome assemblies consisted of bacteriophages infecting *Microcystis,* and RNA viruses constituted the remaining 42% of viral transcripts. Though very diverse, giant viruses take up only 8% of the transcripts. Surprisingly, less than 1% of detected viral transcripts was identified as ssDNA bacteriophages or dsDNA bacteriophages infecting heterotrophic bacteria, though *Microcystis* bloom is typically accompanied with an abundant heterotrophic bacterial community. Interestingly, this is one of the first metatranscriptome studies reporting substantial diversity and abundance of dsRNA viruses. Of the total RNA viral contigs, 28% were identified as dsRNA, mainly *Partitiviridae* and *Picobirnaviridae,* and 72% were ssRNA viruses of the *Picornavirales* order, largely *Marnaviridae* and *Discistroviridae* (Pound et al., 2020). This notable increase of detected dsRNA viruses might result from using rRNA depletion rather than poly-A selection (Gann et al., 2021).

Both metagenomes and metatranscriptomes of the microbial community collected during a diatom bloom succession at Chile Bay in the West Antarctic Peninsula were mined for viruses (Alarcón-Schumacher et al., 2019). Viral reads in the cellular metagenomes of the low-chlorophyll sample were dominated by dsDNA viruses: 82% of viral reads belonging to *Myoviridae*, 9% to giant viruses infecting eukaryotic phytoplankton from the family *Phycodnaviridae,* and 8% to the filamentous ssDNA bacteriophages from family *Inoviridae*. In high-chlorophyll cellular metagenome

samples collected during the bloom, the proportion of giant viruses rose to 93% of the DNA reads and previously dominant *Myoviridae* dropped to 3%. In high- chlorophyll metatranscriptomes collected during the bloom, the proportion of (+) ssRNA viruses from family *Picornaviridae* increased to 38%, mostly composed of PAL E4 and PAL 156 viruses, both lytic viruses, and the latter closely related to diatom-infecting RNA viruses from the genus *Bacilarnavirus* (Alarcón-Schumacher et al., 2019). A similar RNA viral community was detected in RNA viromes (<0.22 μm fraction) from Palmer Station at the West Antarctic Peninsula during a summer diatom bloom in 2016 (Miranda et al., 2016). Taken together, these metatranscriptomic studies show that during a bloom, viral communities that are typically dominated by prokaryotic viruses can shift to eukaryotic viruses, further supporting the ecological relevance of lytic DNA and RNA viruses in bloom disintegration. A metatranscriptome study of virus–host dynamics of the coast of California (Kolody et al., 2019) provided evidence of RNA and DNA viruses tightly controlling protist communities even in non-bloom conditions. After serial filtration, the small fraction showed enrichment for bacteriophages of abundant bacterial groups, while a large fraction was enriched for viruses infecting large phytoplankton, the majority (74%) being dsDNA viruses. The large fraction (<5 μm) was also enriched in fungi, demonstrating fungal importance in the surface ocean as well as their potential as hosts, as they are known hosts of mostly dsRNA viruses. Metatranscriptomic reads from a 60 h sampling study with a 4 h resolution (Kolody et al., 2019) were mapped to the reference genomes of selected RNA and DNA viruses and their host to resolve the diel dynamics of virus–host pairs. An interesting observation from this study was that dsDNA transcripts of dsDNA viruses and their putative host peaked at the same time, without a temporal lag. On the contrary, ssRNA viruses followed a typical “Kill the Winner” scenario, with the viral transcripts’ peak

lagging after the putative host peak. Multiple peaks of viral transcripts of three selected ssRNA viruses and one dsRNA virus during the two sampling days indicated that multiple "bloom and bust" cycles of RNA viruses might be happening within short temporal scales. For example, diatom-infecting *Chaetoceros* sp. RNA virus 02 exhibited only one peak, and heterotrophic labyrinthulid protist-infecting *Aurantiochytrium* single-stranded RNA virus 01 peaked five times during the sampling period. This suggests that RNA viruses are important players in regulating the microbial host communities on hourly scales in steady- state systems not perturbed with phytoplankton blooms (Kolody et al., 2019).

A spatial metatranscriptome study in the Baltic Sea at 11 locations spanning a salinity gradient in the brackish Baltic Sea and one freshwater lake identified a wide distribution of ssDNA and RNA populations and high transcriptional activity of fish viruses in different microbial size fractions (Zeigler Allen et al., 2017). Both ssRNA and dsRNA viruses infecting marine protists were detected throughout the dataset, as well as negative-sense RNA viruses from *Mononegavirales* and a high proportion of *Retroviridae*-like sequences. At two high-salinity sites, DNA viruses infecting the protist *Ostreococcus* outnumbered the bacteriophages, amounting to 20%–40% of total viral transcripts, which indicated an ongoing infection. The sample collected in the lake Tornetrask contained virus transcripts mostly belonging to ssDNA bacteriophages from family *Microviridae*, and two samples had a large proportion of ssDNA viruses infecting pigs that could have originated from nearby animal farms. Collectively, these results provide a valuable snapshot into virus diversity at multiple locations and demonstrate that low-abundance viruses can rise to high abundance if conditions allow (Zeigler Allen et al., 2017).

Multiple metatranscriptome size fractions and a metagenome picoplankton size fraction from the Tara Ocean expedition dataset were screened for eukaryotic viruses using marker gene approach: PolB for giant viruses, RdRp for RNA viruses, and RepB for ssDNA viruses (Kaneko et al., 2021). Interestingly, with 975 sequences, RdRp was the most abundant marker gene detected in the metatranscriptomic dataset, followed by 388 PolB sequences of giant viruses and 299 ssDNA viruses. An additional 3846 PolB marker gene sequences of giant viruses were detected in the metagenome picoplankton size fraction. Giant viruses were identified as *Mimiviridae* (mostly algae-infecting), *Phycodnaviridae,* and *Iridoviridae.* Major RNA viral groups in the dataset were *Picornavirales* as well as *Narnaviridae*, *Tombusviridae*, *Virgaviridae,* and dsRNA *Partitiviridae*. Giant viruses and RNA viruses that were positively correlated with carbon export efficiency were abundant in the oceanic regions with high carbon export efficiency, such as the Indian Ocean and the Mediterranean Sea. Hosts of these giant viruses are predicted to be picoplankton (*Mamiellales*) and haptophytes. Hosts of RNA viruses were either copepods or diatoms from the order *Chaetocerales*, both groups being important contributors to the biological carbon pump (Kaneko et al., 2021). The Tara Ocean metatranscriptomic dataset was briefly explored in terms of viral diversity in study by Carradec (2018), reporting 3463 RNA viruses and 29 023 giant viruses using RdRp and PolB as single gene markers. The different size-fractions of the metatranscriptomes largely differed in the viral content, with giant viruses dominating <5µm size fraction, and RNA viruses being enriched in the 5µm -180µm size fraction (Carradec et al., 2018).

Recently, RNA viruses were thoroughly investigated in the Tara Ocean expeditions dataset, complimented with the newly sequenced data from the Tara Oceans Polar Expeditions (Zayed et al., 2022). A total of total 771 size-fractionated

metatranscriptomes was screened using a special pipeline to detect divergent RNA viruses, recovering 6686 viral contigs with near or full-length RdRp gene. The study described potential five new phyla of RNA viruses: *Taraviricota*, *Articviricota*, *Pomiviricota*, *Wamoviricota* and *Paraxenoviricota,* widely present in the marine ecosystem. Interestingly, the phylum *Arcticviricota* seem to be a class of (-)ssRNA viruses and this genome type is typically not detected in the marine viromes. They do not share a common ancestor with the only current established phylum od (-)ssRNA viruses, *Negarnaviricota.* Moreover, the origin of the established phylum *Duplornaviricota* appears to be polyphyletic too, with 3 current classes *(Vidavervircetes, Resentoviricetes* and *Chrymoviricetes)* clustering separately within the global phylogenetic tree (Zayed et al., 2022). The strong support of the polyphyly of the ortornaviral groups and significant expansion of the phylum-level diversity of RNA viruses challenged the stability and robustness of the current five-branch phylum taxonomic structure recognized by the ICTV (Koonin et al., 2020; Wolf et al., 2018).

Studies using the size-fractionated approach show differential enrichment of viral groups in different size fractions (Carradec et al., 2018; Kolody et al., 2019; Zeigler Allen et al., 2017). In the Baltic Sea spatial metatranscriptome dataset and the Tara Ocean Expedition spatial metatranscriptome dataset, RNA viruses were more abundant in libraries originating from the larger cell fractions and are thought to be infecting large unicellular eukaryotes (Carradec et al., 2018; Zeigler Allen et al., 2017). Giant viruses (including *Phycodnaviridae*) and bacteriophages dominated the viral portion of the Baltic Sea metatranscriptome (50–95% viral transcripts) (Zeigler Allen et al., 2017). Similar trends of large DNA viruses dominating viral portions of cellular metatranscriptomes were observed in the Tara Oceans expedition, with 86% of viral genes present originating from giant viruses (Carradec et al., 2018). In conclusion, these

results suggest that size fractionation may improve the resolution and capture a greater diversity of less-abundant viruses. However, care should be taken to avoid using a defined size fraction without a clear understanding of the intended viral targets, because this might lead to more biased results compared to unfractionated samples.

2.6. Environmental dsRNA sequencing: the enrichment of dsRNA from marine samples greatly expands the diversity of dsRNA viruses

The earlier metatranscriptome studies described above recovered mostly ssRNA viruses and very few dsRNA viruses, in agreement with the diversity found in extracellular RNA viromes. A combined metatranscriptomics and RNA metaviromics study compared RNA viruses in the cellular and purified viral particle fractions in four pelagic sampling sites and one coastal sampling site off the coast of Japan (Urayama et al., 2018). For both cellular and viral fractions, the total RNA was separated by chromatography in ssRNA and dsRNA fractions. The double-stranded RNA fraction contains dsRNA genomes and double-stranded replicative intermediates of ssRNA genomes, and it is enriched for viruses compared to the rRNA-depleted ssRNA fraction that contains mostly cellular mRNA and in which viral ssRNA genomes or transcripts are in minority. Single-stranded RNA and dsRNA were sequenced separately, the latter using the newly developed method, fragmented and primer-ligated dsRNA sequencing (FLDS), resulting in four datasets for each sampling station: cellular ssRNA metatranscriptome, cellular dsRNA metatranscriptome, ssRNA virome, and dsRNA virome. A total of 1270 RNA viral contigs were recovered in the study, mostly originating from dsRNA-enriched metatranscriptomes and dsRNA-enriched viromes (>1000), underlining the efficiency of the dsRNA enrichment procedure compared with the traditional sequencing approach for obtaining viral reads. Due to low number of

reads, ssRNA metatranscriptomes were excluded from the analysis. Viral families detected in dsRNA metatranscriptomes were predominantly endogenous ssRNA viruses from family *Narnaviridae*, and dsRNA viruses from families *Picobirnaviridae* and *Reoviridae*. The ssRNA extracellular virome of the coastal site was dominated by (+) ssRNA viruses from the order *Picornavirales,* as reported by other studies (Miranda et al. 2016; Vlok et al. 2019; Culley et al. 2006; Culley et al. 2014). The other sampling sites had too few reads in this fraction for a meaningful comparison. Parallel analysis of the dsRNA virome from the same coastal sampling site revealed an almost equal number of reads belonging to *Picobirnaviridae* and *Reoviridae*, dsRNA viruses with a capsid and an extracellular phase. Many contigs resemble viruses with protist hosts such as diatom colony-associated dsRNA virus-1 (DCADSRV-1) (Urayama, Takaki, and Nunoura 2016) and *Micromonas pussila* reovirus (Attoui et al., 2006). There are two possible explanations for why FDLS sequencing detects an increased number of dsRNA viruses. The first one is that denaturation of dsRNA before transcription allows for efficient transcription of dsRNA genomes, as they are not efficiently amplified during standard metatranscriptomic library preparation (Wilcox et al., 2019). The second is that there is an amplification bias of the FDLS method, such as that shown for SISPA libraries, which are amplified on a similar principle. Whichever the case, these viruses are present in the aquatic environment and had not been previously detected with standard metatranscriptomic methods. This remains the only study that has explored dsRNA sequencing as a way of studying aquatic RNA viral communities, and it demonstrates the potential of the method to uncover novel RNA viruses and contribute to a clearer picture of the aquatic RNA viral diversity.

2.7. Holobiont metatranscriptomics: marine macroalgae and cultured marine protists reveal the broad distribution of non-lytic strategies in marine RNA viruses

There have been several new viral discovery studies in metatranscriptomes of (mostly) axenic protist cultures that have discovered an array of small dsRNA and ssRNA viruses, often with multipartite genomes that coexist with the host and do not cause lysis of the host cell (Charon et al. 2020) . These persistent infections can be chronic or latent (silent). The advantage of viral discovery in cultured metatranscriptomes over broad environmental metatranscriptomics surveys is that the host can be assigned with certainty and low-titre viruses, which are frequently missed by environmental metaviromics and metatranscriptomics, can be picked up.

The first report of RNA viruses in the cultured metatranscriptomes came from six species of phytoplankton belonging to two distant phylogenetic groups, green algae (*Chlorophyta*) and chlorarachniophytes . They recovered 18 novel RNA viruses with very low AAI (27–38%). Most viruses were associated with the green algae *Ostrebium* sp., comprising dsRNA viruses falling into families *Partitiviridae* and (+) ssRNA *Mitoviridae,* non-encapsidated small RNA viruses with tiny genomes. Fewer viruses were detected within chloroarachniophytes, possibly due to the clade being divergent and poorly characterized it- self, and the highly divergent RdRps might have gone undetected. Additionally, segmented (+) ssRNA viruses from family *Virgoviridae* were detected along with the first detection of a (−) ssRNA virus resembling family *Bunyaviridae* in phytoplankton culture, though the confirmation that the chloroarachniophyte algae are the true host requires additional work (Charon et al. 2020). This came as a surprise because both environmental metatranscriptomes and

virus isolates showed evidence that green algae are predominantly infected by giant DNA viruses and 85% of isolated RNA viruses were associated with diatoms (Coy et al. 2018; Charon et al. 2020).

Recently, the metatranscriptomes of cultured phytoplankton obtained from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) were screened for the presence of viral RNA-dependent RNA polymerase hallmark gene (Charon et al., 2021). Thirty new RNA viral species with highly divergent RdRps were observed, on average only sharing 35% AAI with known RNA viruses. RNA viruses were observed for the first time in protist families such as *Haptophyceae* and *Chromeraceae*, *Xantophyceae,* and *Bolidophyceae*, which were previously thought to only host dsDNA viruses. One-third of the newly described viruses belong to (+) ssRNA non-encapsidated viruses from families *Narnaviridae* and *Mitoviridae,* known for persistent lifestyles (Hillman and Cai, 2013). The double-stranded RNA viruses discovered were dominated by viruses resembling *Totiviridae* that were previously considered to be fungal pathogens. A member of *Mononegavirales*, a negative-sense RNA virus, was detected in the *Pseudo-Nitzchia* culture, but again, host assignment requires additional experimental work (Charon et al., 2021).

Two studies screened diatom holobionts and marine macroalge for RNA viruses with the FLDS dsRNA sequencing approach (Chiba et al. 2020; Urayama et al. 2016). The first one unveiled 22 and the second six full-length RNA viral genomes, encompassing five families of dsRNA viruses (*Totiviridae*, *Picobirnaviridae*, *Cystoviridae*, *Partitiviridae*, and *Endornaviridae*) and four families of ssRNA viruses (*Flaviviridae*, *Narnaviridae*, *Virgaviridae*, and *Hypoviridae*). Typically, the viral sequences were very divergent, exhibiting less than 50% AAI to known viral RdRp sequences (Chiba et al.,

2020; Urayama et al., 2016). These metatranscriptome studies focused on single protist or macroalgae holobionts have demonstrated that (+) ssRNA, dsRNA, and potentially (−) ssRNA viruses are dis- tributed across many diverse protist groups and went previously undetected in "bulk" environmental approaches. These viruses have small genomes (typically <5 kb), may be segmented, and some lack capsid proteins. Viral families *Amalgaviridae*, *Chrysovirdae*, *Totiviridae, Partitiviridae, Endornaviridae, Mitoviridae,* and *Narnaviridae* are all associated with persistent, asymptomatic infections in plants and fungi (Roossinck, 2019) and likely cause nonlytic, but persistent chronic infection in their marine hosts.

2.8. Recommendations for future studies

Several studies discussed in this review demonstrate the plethora of information available when different sequencing approaches are combined (Alarcón-Schumacher et al., 2019; Urayama et al., 2018; Zeigler Allen et al., 2017), but these studies are still very rare. It is of utmost importance to conduct more studies that integrate metagenomics, metatranscriptomics, and DNA/RNA metaviromics to fully understand the spectrum of the viral diversity and capture the ecological interactions between viruses and their host. Targeted enrichment of dsRNA viral nucleic acid followed by dsRNA sequencing and addition of long-read sequencing could also complement RNA diversity studies by capturing unknown RNA viruses and alleviate the biases of the current methods. To our knowledge, only one study compared RNA viruses in cellular and extracellular viral fractions simultaneously, and it added a novel perspective by demonstrating that both fractions contain a substantial number of dsRNA viruses (Urayama et al. 2018). A valuable addition to the molecular toolshed for the study of aquatic RNA viral communities could be the use of nanopore sequencing (Oxford

Nanopore Technology, Oxford, UK). This technology can be used to sequence native DNA and RNA molecules as they pass through a nanopore; the nucleotide sequence is determined as a change in the current and is specific to the oligomer passing through the pore. Read length is limited only by fragment length, and with the right sample preparation, reads can reach lengths of megabases (Amarasinghe et al., 2020). It was shown that the inclusion of long reads benefits contig length and captures more DNA viral diversity (Beaulaurier et al., 2020; Overholt et al., 2020; Warwick-Dugdale et al., 2019). Incorporating long reads in the studies of RNA viral diversity may improve sequence assemblies and/or enable recovery of full-length RNA genomes from environmental samples in a single long read and therefore circumvent the need for assembly. This may help to identify novel viruses and strains. Sequencing of native RNA molecules could also help to overcome biases associated with reverse transcription and cDNA amplification (Kozarewa et al., 2009). The constraints of the long-read approach are the amount and quality of starting RNA material.

Despite the immense new RNA viral diversity uncovered with multi-omic approaches, it is evident that RNA viruses in aquatic environments are very sparsely sampled (Table 2.2; Figure 2.4.). In addition to conducting diversity surveys, more ecology-focused studies are needed to understand the ecological patterns of RNA viruses in aquatic environments. There was only one RNA metaviromics temporal study conducted during Antarctic diatom bloom (Miranda et al., 2016) and one larger spatial study (Vlok et al., 2019), with no continuous temporal studies on a weekly, monthly, or yearly timescale. Metatranscriptomic datasets are more temporally oriented, typically following a bloom over a few weeks on a weekly sampling basis (Table 2.2). Short time-series, such as that employed by Kolody et al. (2019), demonstrate that RNA viruses can be highly

active and dynamic and have diurnal fluctuations, emphasizing the need for more higher-resolution studies.

Most metatranscriptome studies focus on productive areas during algal blooms (Table 2.2). Efforts to compare across multiple contexts (bloom vs. non-bloom) and in different biomes will undoubtedly lead to a more holistic and comprehensive understanding of RNA and DNA viral dynamics and impacts. This is a worthy endeavour, as the abundance and activity of RNA viruses, and of viruses in general, are intimately connected to the proper functioning of marine food webs, and future climate change will alter the viral-mediated control on biogeochemical cycling (Danovaro et al. 2011), with potentially unpredictable consequences. Another area that deserves attention is standardization and validation of the pre-sequencing and sequencing sample preparation protocols. Enrichment methods, extraction kits, and library preparation can significantly affect the number of viral reads, and consequently the assembly, and skew the relative abundances of particular viral groups and diversity of the viruses present in the sample (Gann et al., 2021; Hjelmsø et al., 2017). To ensure robust comparisons, protocols should be compared and validated with mock viral communities to quantify the efficiency and identify the biases of the methods, as previously done for the human virome (Conceição-Neto et al., 2015; Hjelmsø et al., 2017; Kohl et al., 2015; Rosseel et al., 2015). Lastly, both metaviromics and metatranscriptomics depend on identifying homology with viral sequences present in the database, and highly divergent viral sequences with little or no sequence homology cannot be readily identified. Even in enriched samples, most reads have low similarity to known nucleotide or protein sequences. Therefore, novel bioinformatic approaches for detecting highly divergent RNA viruses are required to further extend the boundaries of our knowledge about RNA viral diversity and evolution (Zhang et al. 2019).

2.9. Conclusions

Multi-omic sequencing approaches have advanced our understanding of RNA viral diversity and expanded our understanding of their ecological roles in aquatic ecosystems. Three common messages have emerged from the synthesis of all these studies: (1) lytic ssRNA viruses from the order *Picornavirales* are a stable component of the aquatic RNA viral communities, but are perhaps not as dominant as previously thought. (2) All three phyla of positive-sense ssRNA viruses are represented in comparable proportions in the aquatic ecosystem as well as the dsRNA viruses. Highly divergent RNA viruses are being discovered frequently, and are infecting protist hosts that were previously thought only to host DNA viruses. (3) Negative-sense ssRNA viruses, which are traditionally associated with multicellular animals and plants, were found for the first time in the metatranscriptomes of some single-celled eukaryotic phytoplankton cultures. The articles discussed above demonstrate significant progress in characterizing aquatic RNA viral communities, but also highlight how undersampled RNA viral communities are. However, as more data accumulate, we expect to paint a more comprehensive picture of RNA viral diversity, abundance, and host range across the aquatic environments.

CHAPTER 3 IMPROVING THE SAMPLE-TO-SEQUENCE WORKFLOW FOR SEQUENCING OF AQUATIC METATRANSCRIPTOMES AND RNA VIROMES

3.1. Abstract

Wet lab sample processing steps is crucial for obtaining good quality sequence data for analysis of marine RNA viruses in the viral-size fraction as well as in the microbial metatranscriptomes. In this chapter I compared different protocols for RNA extractions from the marine microbial fraction (>0.22μm). The combination of an open Sterivex filter and All Prep Environmental RNA/DNA kit gave the best results in terms of quantity and quality of RNA for metatranscriptomic profiling. Excessive biomass negatively affected the RNA integrity and subsequent library preparation. Hence, the filtration volume should be carefully selected depending on the trophic status of the ecosystem under study. Efficient detection of marine RNA viruses in viral-size fraction (<0.22μm) is still hampered by the very small amounts of viral RNA recovered during RNA extraction. Wet lab sample preparation, sequencing, and data analysis of marine RNA viral genomes obtained from the viral-size fraction were optimised to improve yield and remove the contaminating DNA and potential sequencing inhibitors. In just one sample, I recovered highly diverse population of (+) ssRNA and dsRNA viruses (n=319), 4 potentially new clades that did not cluster with known viruses and 50 near-complete RNA genomes. The rRNA depletion step did not affect diversity of the sample or number or length of recovered RNA viral contigs and can potentially be introduced as unbiased pre-step for direct, assembly-free long-read RNA sequencing. Comprehensive assessment of laboratory procedures presented here offers valuable

insights into potential sample processing issues and ease both troubleshooting and experimental design of future ecological studies targeting aquatic RNA viruses.

Keywords: RNA extraction, RNA virus discovery, marine RNA viruses, RNA viromics, RNA viral metatranscriptomics

3.2. Introduction

High-throughput sequencing of viral-size fraction (RNA viromics) and of cellular fraction (metatranscriptomics) revealed astoundingly diverse, often completely unknown, communities of RNA viruses in invertebrates (Shi et al. 2016), wastewater (Callanan et al., 2020; Neri et al., 2022), marine (Wolf et al., 2020; Zayed et al., 2022) and soil ecosystems (Hillary et al., 2022; Starr et al., 2019). Marine RNA viral communities (a.k.a. the RNA virome) span viruses infecting bacteria, microbial eukaryotes (e.g. phytoplankton) and metazoans. RNA viruses, like their DNA counterparts, are ubiquitous and play an important role in the marine ecosystem, influencing the structure of marine phytoplankton and bacterial populations, as well as organic matter fluxes (Kolundžija et al., 2022; Zimmerman et al., 2020). Furthermore, metazoan RNA viruses cause viral disease outbreaks that can inflict huge economic losses to the marine aquaculture industry as well as devastate marine wild animal populations (Chen et al., 2019; Eyngor et al., 2014). Finally, in aquatic environments, human enteric RNA viruses constitute a concern to public health and water quality (Rodríguez-Lázaro et al., 2012).

Sample preparation techniques can have a big impact on the viral diversity recovered (Gann et al., 2021; Hjelmsø et al., 2017; Iker et al., 2013; Li et al., 2015; Rosani et al., 2019). High quality, intact RNA is essential for a successful downstream analysis as degraded RNA will affect the success of rRNA depletion, and bias (viral) transcript or genome abundance (Depledge et al., 2019). Marine studies of RNA viruses rarely focus

on the RNA viruses in the viral size fraction (<0.22μm), with only six studies to date (Kolundžija et al., 2022). The main challenges are the requirement to sample impractical amounts of seawater (often more than 40L) and the difficulty in recovering RNA viruses from the viral-size fraction due to their tiny and highly labile genomes. As little as one microbial cell can contain an amount of DNA and RNA comparable to half a million virions (Greninger, 2018), hence prefiltration is required to eliminate the microorganisms in the seawater. Although filtration effectively removes cells, free-floating ribosomes (20-30nm), which are similar in size to RNA viruses, can pass through a 0.22μm filter (Conceição-Neto et al., 2015; Rosseel et al., 2015). Standard viral concentration procedures (e.g. chemical flocculation; ultrafiltration; purification on gradients; sucrose ultracentrifugation; nuclease treatment (Thurber et al. 2009; Hurwitz et al. 2013)) often co-purify ribosomes with viruses. Therefore, a large fraction of nucleic acids from the viral-size fraction extract will be ribosomal RNA. Low yields of RNA from the marine viral-size fraction frequently require random amplification (e.g. sequence-independent single-primer amplification, RT-SISPA) to achieve appropriate sequencing depth (Culley et al., 2014; Miranda et al., 2016; Vlok et al., 2019). Random amplification is known to preferentially amplify some viral groups, bias viral coverage, and skew the viral taxonomic distribution (Karlsson et al., 2013; Yilmaz et al., 2010).

To aid experimental design for the future metatranscriptome studies, I compared the RNA workflows used in marine metatranscriptomic studies (Table 3.1), and assessed their performance in terms of RNA quality and quality. Next, I designed an optimised and integrated workflow for improved RNA extraction from the viral-size fraction by combining published protocols while optimizing each step of the process. The methodological recommendations on how to achieve high quantity and quality of RNA

extracts could be of use in the experimental design of future ecological studies, supporting discovery of novel RNA viruses as well surveillance of known RNA pathogens of marine animals.

3.3.Materials and methods

3.3.1. Sample collection

More than 2000L of surface seawater was collected at Raffles Marina (1.3445° N, 103.6340° E) from 1 meter depth on multiple dates from 2018 to 2022. More than 2000L of surface seawater was collected at Raffles Marina (1.3445° N, 103.6340° E) from 1 meter depth on multiple dates from 2018 to 2022. Collection of samples of the cellular fraction (>0.22μm) were performed by filtering one litre of seawater (150μm pre-filtered) onto two filters types, with a minimum of two technical replicates. The samples were immediately preserved in RNAlater solution (Thermo Fisher) and frozen until RNA extraction (Table A3.2). The virus-size fraction (<0.22μm) sample preparation is technically demanding and time-consuming. The whole process from collecting, filtering and concentration of big volumes of seawater (40-100L per one sample) and subsequent RNA extraction takes 4-5 days to complete. On each sampling occasion, the seawater was filtered through a 150μm-pore-size nylon filter unit into acid-cleaned polycarbonate carboys (Nalgene) and immediately transported to the lab for further processing. Different combinations of sample volume input (40-100L), virus concentration methods (tangential flow filtration, hollow fibre concentration and iron flocculation) and RNA extraction kits/methods were tested. The first attempts to establish the functional sampling to sequencing workflow consisted of collecting and processing one seawater sample (40-100L). Every sampling effort was repeated in at least 2 biological replicates. Only successful combinations are described in detail in the

methods and results paragraphs. However 6 additional workflows were tested and found unsuitable (data not shown, listed in Table A3.1.). After establishing a functional sample processing workflow, resulting in successful RNA extraction, the sampling was repeated in biological and technical replicates which are summarized in Table A3.3.

3.3.2. RNA extraction from the cellular fraction (>0.22µm)

All Prep PowerViral Environmental DNA/RNA Kit (Qiagen) and DNeasy PowerWater Sterivex Kit (Qiagen) were tested in combination with different filters (Table A3.2). The extraction with DNeasy PowerWater Sterivex Kit was modified and optimized for RNA extraction as described in previously published metatranscriptome studies and published protocols (Davenport et al., 2019; Pound et al., 2020; Stough et al., 2017; Krausfeldt, 2017). First , the RNA later solution was pushed out of the Sterivex cartridge with a syringe and the filter was rinsed with 20mL of sterile PBS buffer (137mM NaCl, 10 mM Na_2HPO_4, 2.7mM KCl, 1.8 mM). Fresh β-mercaptoethanol was added to the lysis solution (STB1) before the start of every extraction. All vortexing steps were extended to 10 minutes and the incubation step was performed at 70°C. All treatments and wash steps were incubated for one minute on the spin column before centrifugation. An additional wash with 100% ethanol to remove residual salts from RNA later solution that could inhibit subsequent steps. All samples were eluted in 100µl of nuclease-free water (Ambion).

The RNA extraction with the All Prep PowerViral Environmental DNA/RNA Kit followed the manufacturer's protocol. Two additional steps were included before the RNA extraction itself. First, the RNA later solution was removed from the Sterivex cartridge and the filter was rinsed with 20mL of sterile PBS buffer in the same way as described above. Second, the casing was cracked opened with sterile pliers and filters

were removed from the frame and cut into small stripes on a sterile Petri dish (sample side up) (Cruaud et al., 2017). To maximize the yield and purity of RNA, the cut filter stripes were divided into two bead-beating tubes and the maximal volume of inhibitor removal solution was used during the inhibitor removal step. Each sample was then loaded onto one spin column, washed and eluted in 100µl of nuclease-free water. The last approach combined polycarbonate filter membrane (Sterlitech) with a pore size of 0.8µm and All Prep PowerViral Environmental DNA/RNA Kit. The extraction procedure followed the manufacturer's protocol. After all extractions, DNA and RNA concentration was quantified with Qubit™ DNA High Sensitivity (HS) Assay Kit and Qubit™ RNA High Sensitivity (HS) Assay Kit (Life Technologies).

To ensure complete removal of DNA before RNA sequencing, on-column and post-extraction DNAse treatment were tested. For on-column DNAse treatment tests, samples were treated with On-column DNAse I digestion set (Sigma-Aldrich) for 15 and 30 minutes, as the later was described as an optimization of the procedure (Davenport et al., 2019). In the second approach, TurboDNAse (Turbo DNA-free kit, Invitrogen) treatment was performed on total nucleic acid extracts after the RNA extraction was completed. Following guidelines for severely contaminated RNA, I tested incubations with different concentrations of the enzyme (2U-8U) at 37°C for 30 minutes. TurboDNAse was inactivated with DNase inactivation reagent to avoid heat degradation of RNA. To remove the degraded DNA and the inactivated enzyme, the RNA was cleaned and concentrated with RNA Clean & Concentrator Kit (Zymo Research). DNA and RNA concentration was quantified after the DNase treatment with Qubit™ DNA High Sensitivity (HS) Assay Kit and Qubit™ RNA High Sensitivity (HS) Assay Kit (Life Technologies). The quality of RNA and the absence of high molecular

weight DNA was assessed with RNA ScreenTape analysis on a Tape Station 2200 instrument (Agilent).

3.3.3. Concentration of the viral-size (<0.22μm) fraction

Seawater was serially filtered through 0.8μm (GFF, Whatman) and 0.22 μm (Sterivex, Millipore) filters to remove the host cells. Following filtration, viruses in the viral-size fraction (<0.22μm) were precipitated with a solution of iron chloride (final concentration of Fe 1mg/L) overnight. Iron flocs with adsorbed virus particles were collected onto 1 μm polycarbonate membrane filters (90mm, Sterlitech) backed by a 0.8μm support membrane filters (Supor® 800, Pall). All filtration was done at a pressure lower than 15psi to avoid damaging virus particles. Iron flocs were resuspended in freshly prepared magnesium-EDTA-ascorbate buffer (0.1M EDTA, 0.2M $MgCl_2$, 0.2M ascorbate, pH 6.0) overnight at 4 °C on a tube revolver rotator (John et al., 2011). Resuspended iron flocs were loaded on a freshly prepared 38% (wt/vol) sucrose cushion and ultracentrifuged at 175 000 x G for 3 hours at 8°C in Type 45 Ti fixed rotor (Optima XPN ultracentrifuge, Beckman Coulter). The pellet beneath the sucrose cushion was resuspended in sterile SM buffer at +4°C overnight with shaking. To remove contaminating host DNA, the pellet was treated with RNAse-free DNAse I (New England Biolabs) in concentration of 100U/mL for 2 hours at room temperature with rotation. The reaction was stopped by inhibiting the DNase I with EDTA and EGTA in final concentration of 100mM to avoid heat degradation of RNA (Hurwitz et al., 2013). Viral concentrates were examined with epifluorescence microscopy to confirm the presence of DNA viruses. The successful concentration of DNA viruses was taken as an indicator that smaller RNA viruses were also likely concentrated, since RNA viruses cannot be enumerated with any of the available enumeration methods.

3.3.4. RNA extraction from the viral-size (<0.22µm) fraction

Each RNA extraction protocol was tested at least on two seawater samples to account for between-sample variability caused by the sampling day. Environmental viral RNA was extracted from 200 - 250µl aliquots of DNAse-treated viral concentrates. An additional concentration step using a Vivaspin column (Sartorius) was included in some of the tests. Every RNA extraction batch included an positive RNA extraction control – a 200µl of MS2 bacteriophage lysate and a negative control -200µl of molecular grade nuclease free water (Ambion). In the phenol/chloroform approach, two hundred fifty microliters aliquots of final viral concentrates were extracted with 1mL of cold Trizol reagent (Invitrogen). The virus concentrates and Trizol reagent were homogenized by pipetting and transferred to pre-centrifuged phase-lock tube (Phasemaker™, Thermo Fisher). After addition of chloroform, the solution was shaken vigorously for 15 seconds, followed by a 15 minute incubation at room temperature. After the centrifugation, the organic phase was locked in the gel and the aquatic phase was transferred to a clean RNAse-free tube (Ambion). Sodium acetate (0.1V) and ice-cold 100% isopropanol (0.8V) to the aqueous phase and mixed by inverting. Optionally, three microliter of nucleic acid co-precipitant Glycoblue (Thermo Fisher) was added in some of the tests, but it was found to be incompatible with column-based purification. Nucleic acids were precipitated for 1h at -80°C and pelleted by centrifugation at 12 000xG at 4°C for 30 minutes. Precipitates were washed with ice-cold, freshly-prepared 75% ethanol. Nucleic acid pellets were collected by centrifugation, supernatant was discarded and the pellets were air dried for 10 minutes and resuspended in 25µl of nuclease free water (Ambion) for 10 minutes at 55°C.

In the glass bead-beating RNA extraction approach, The samples were extracted according to the manufacturer's protocol All Prep PowerViral Environmental

DNA/RNA Kit (Qiagen) with modifications. DNAse-treated viral concentrate was split in multiple aliquots. Each aliquot was separately lysed with lysis buffer with addition of ß-mercaptoethanol (Sigma Aldrich), followed by a bead-beating step and treated with maximum volume of inhibitor removal solution. All aliquots were loaded on one silica column and the extraction was completed following the manufacturer's protocol. Nucleic acids were eluted in 100µl of nuclease-free water (Ambion) and quantified with Qubit fluorometer (High sensitivity RNA and DNA assays, Life Technologies). Only samples with yields higher that >1µg underwent the subsequent steps. Contaminating DNA was removed with rigorous Turbo DNAse treatment, with 18-20U of TurboDNAse at 37°C for 1 hour (Turbo DNA-free kit, Invitrogen). Total RNA was purified with RNA Clean & Concentrator kit (Zymo Research) and eluted in nuclease free water. Extracted total RNA was tested for presence of DNA contamination using qPCR using a 357F/518R primer set (F: 5'CCTACGGGAGGCAGCAG-3', R: 5'ATTACCGCGGGCTGCTGG-3') and KAPA SYBR FAST qPCR Master Mix (2X) (Kapa Biosystems) with the following qPCR conditions: denaturation at 95 °C for 3min, followed by 45 cycles (95 °C for 10 s, 58.3 °C for 60 s) and the final extension at 72 °C for 3 min and a melting and continuous measuring step rising gradually (0.11 °C s−1) to 95 °C.

3.3.5. Total RNA sequencing

Total RNA sequencing libraries were prepared with TruSeq Stranded Total RNA kit (Illumina). For rRNA-depleted libraries, ribosomal RNA was removed from one microgram of total RNA with Ribo-Zero-Plus (Illumina) before the library preparation with TruSeq Stranded Total RNA kit (Illumina). Upon fragmentation, the total or depleted RNA was reverse transcribed with ProtoScript II Reverse Transcriptase and random hexamers with addition of Actinomycin D to prevent spurious DNA-dependent

synthesis. The conditions for the reverse transcription reaction were : 25°C for 10minutes, 42°C for 15 minutes and 70°C for 15minute. Second strand synthesis of cDNA was performed with Second Strand Marking Master Mix. After the end-repair and adenylation, the sequencing adapters were ligated to cDNA. cDNA with adapters was amplified with under with the following cycling conditions: initial denaturation 98°C for 30 seconds, 13 cycles of: denaturation at 98°C for 10 seconds, annealing at 60°C for 30 seconds, extension at 72°C for 30 seconds, with final extension at 72°C for 5 minutes. The quality and concentration of the prepared libraries was analysed with TapeStation D1000 Screen Tape. Paired-end sequencing (2x150bp) with a depth of minimum 60 million reads was performed on the Illumina Novaseq 6000 platform (Macrogen, Singapore).

3.3.6. Phylogenetic analysis of the RNA virome

The quality control of Illumina short sequence reads was performed with FastQC (Andrews 2010). Sequencing adaptors, low quality reads (quality score <20) and reads shorter than 50bp were removed with Cutadapt (Martin, 2011). The remaining host ribosomal rRNA sequences for Bacteria, Archea and Eukaryotes were subtracted *in silico* with SortMeRNA tool (Kopylova et al., 2012). Taxonomic assignment of the reads was performed with Kaiju (Menzel et al., 2016). Cleaned, rRNA-free sequence reads were assembled with de novo sequence assembler MEGAHIT version 1.2.9 with default parameters (Li et al. 2015). RNA viral contigs recovery and taxonomic assignment of the recovered RNA viral contigs were performed with RNA virus discovery pipeline OrViT (Cheng et al. 2022).

3.3.7. Functional annotation of near-complete RNA viral genomes

Contigs were considered to be near-complete RNA viral genomes when both non-structural (RdRp) and structural (capsid) polyproteins were present on the same contig

(Moniruzzaman et al. 2017; Gann et al. 2021). Confirmed RNA contigs (containing RdRp) were extracted from the raw assemblies and open reading frames were predicted with Prodigal v2.6.3. (Hyatt et al., 2010). The predicted open reading frames were annotated searching with HMMSCAN (Prakash et al., 2017) against the PFAM database (v 35.0, November 2021) (Mistry et al., 2021). Multiple alignment of protein sequences was performed with Clustal Omega v1.2.4 (Sievers et al., 2011).

3.4. Results

3.4.1. Quantity and quality of RNA obtained from the cellular (>0.22μm) fraction

To test for efficiency of different kit-filter combinations (listed in Table A3.2), one litre of seawater (prefiltered with 150μm mesh to exclude the mesozooplankton) was collected either on 0.8μm polycarbonate filter membranes or 0.2μm polyethersulfone Sterivex filter units on multiple sampling occasions to compensate for natural variability in samples. All three methods were similar in terms of DNA recovery (Figure 3.1A), with average DNA yields of 2.2μg, 2.1μg and 1.95μg for polycarbonate filter, open Sterivex filter and closed Sterivex filter, respectively. RNA yields for polycarbonate filters and closed Sterivex filter averaged around 1.25μg, and for open Sterivex filter 2.1μg (Figure 3.1).

To ensure complete elimination of DNA, I first tested on-column DNAse treatment (Figure 3.2.). However, on-column DNAse treatment caused inconsistent RNA yields in our experiment, leading to the complete loss of RNA in some of the samples. Overall, RNA recovery in the samples treated with on-column DNAse did not meet the yields required for metatranscriptomic sequencing (>1μg). Hence, I opted for post-extraction TurboDNAse treatment followed by another round of RNA clean-up, which produced

yields that met the requirements for metatranscriptomic sequencing. Next, I investigated the quality of RNA obtained by different filter-kit combinations. RNA integrity number (RIN) algorithm estimates RNA sample integrity after microcapillary RNA electrophoresis. High quality RNA (high RIN number) is crucial for obtaining unbiased gene expression data and detection of low abundance viral transcripts that typically constitute <1% of reads in the sample. All the RNA extracts obtained by different filter-kit combinations have undergone the same TurboDNAse and clean-up treatment. Both closed Sterivex and polycarbonate filter approach had lower quality (RIN 4-5) and lower yields after clean-up comparing to open Sterivex approach (RIN>7). The closed Sterivex approach also showed presence of high molecular weight DNA contamination (Figure A3.2) As expected, the increase in chlorophyll biomass had significant positive effect on RNA yields ($R=0.52$, $p=0.00031$), but negatively impacted RNA quality ($R=-0.58$, $p=4.9e-05$) (Figure A3.3).

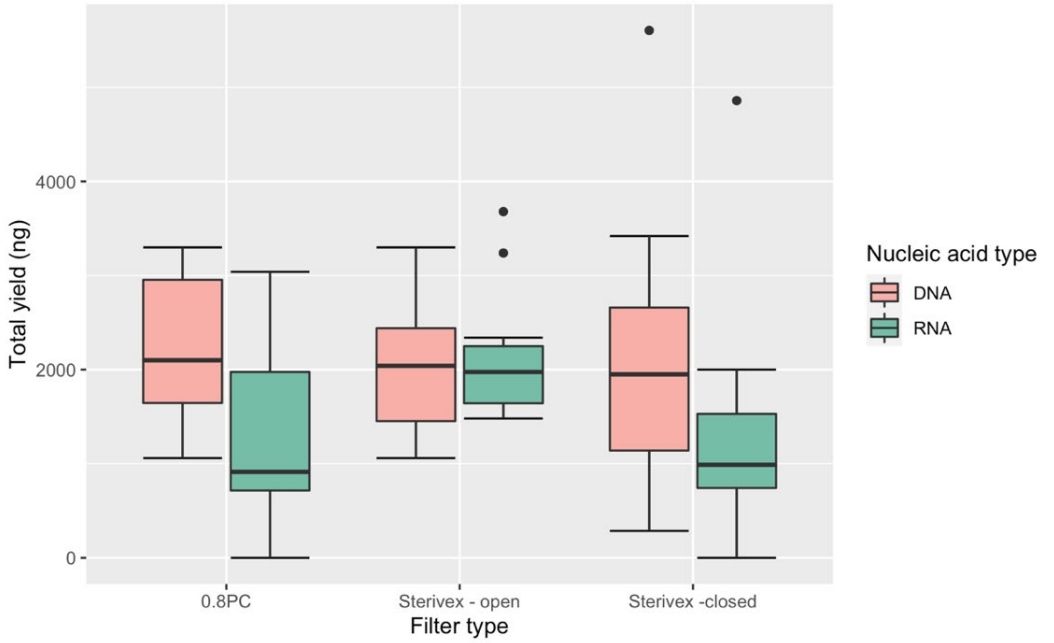


Figure 3.1 Box and whiskers plot of RNA extraction efficiency of different filter-kit combinations*:* RNA extraction from polycarbonate filter with the All Prep PowerViral Environmental DNA/RNA Kit (Qiagen)*,* RNA extraction from the open Sterivex filter with the All Prep PowerViral Environmental DNA/RNA Kit and closed Sterivex filter with Sterivex DNA Easy PowerWater Isolation Kit (Qiagen). Filtered seawater volume was 1L in all experiments*.* Samples were collected on 2-4 different sampling occasions, in minimum 2 technical replicas for each filter-kit combination per sampling occasion. Black dots represent the outliers.

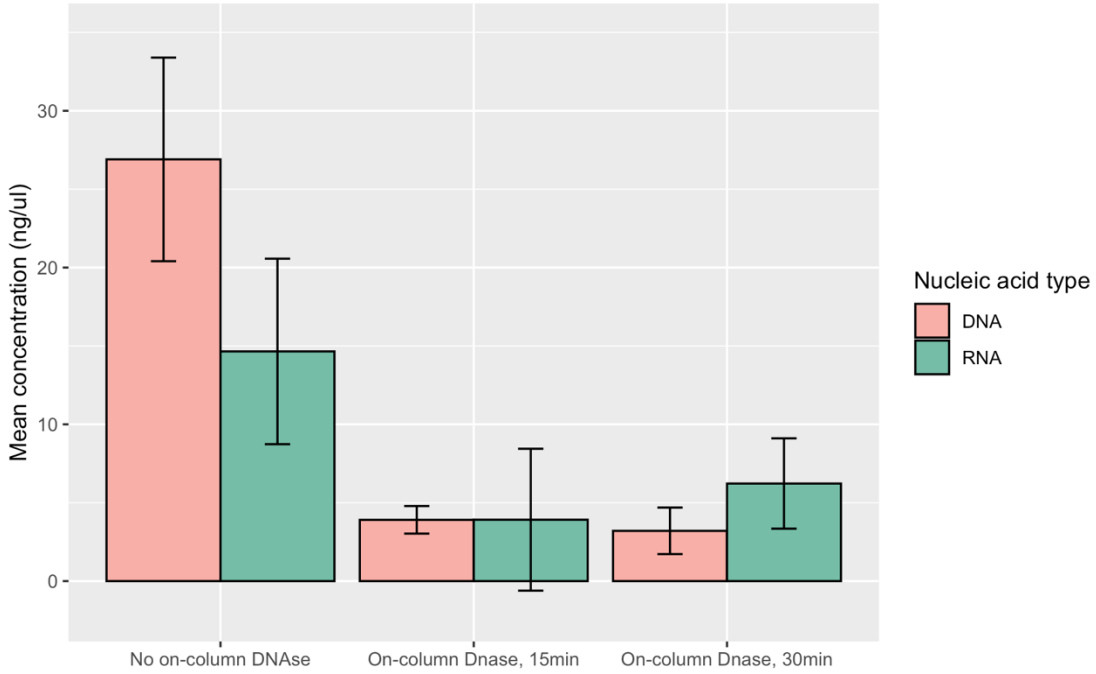


Figure 3.2. Effect of on-column DNase treatment on DNA removal and RNA yield. All extractions for this experiment were performed with All Prep Environmental RNA/DNA kit and 0.8µm polycarbonate filter and eluted in 100µL. All filters are technical replicates (n=16) collected on the same sampling day.

3.4.2. Quantity and quality of RNA obtained from viral-size (<0.22μm) fraction

Trizol RNA extraction yielded between 247-610ng of total RNA from 40L of seawater (Table A3.3., Figure A3.1.). Adding Glycoblue precipitant increased the yield of RNA, but it caused problems in the subsequent clean-up and it was not further investigated as a method of choice. Bead-beating based RNA extraction with Qiagen All Prep Environmental DNA/RNA kit, resulted in a significantly higher yields of both RNA and DNA the viral fraction, with an average of 1.868μg of RNA and 9.8μg of DNA recovered from 40L of seawater (Table A3.3., Figure A.3.1.). Due to overwhelming amount of coextracted viral DNA, another rigorous DNase treatment (90U/mL) was introduced as described in the material and methods. A clean-up step was performed to remove the enzyme, degraded nucleotides and potential residual inhibitors from iron flocculation or seawater that may interfere with sequencing. RNA yield after this step was on average 1.202μg (Table A.3.3). If desired, an aliquot of sample for viral DNA sequencing can be taken prior to DNase treatment to explore DNA and RNA viruses in the viral size fraction simultaneously.

3.4.3. Effect of physical and in silica rRNA depletion on assembly and recovery of RNA viruses from the viral-size fraction (<0.22μm) fraction

Sequencing of the total RNA sample produced 86 pair-end million reads (Table 3.1) and rRNA depleted sample produced 76 million pair-end reads. After additional *in silico* removal of rRNA reads, 50 and 68 million reads remained for total RNA virome and depleted RNA virome, respectively. As expected, around 70% of the reads could not be classified for both samples. Physical depletion of rRNA the RNA virome reduced the host background (Figure 3.3.) and increased the relative abundance of viral reads. Total

RNA samples contained 10% of absolute viral reads (equals to 8.6 millions viral reads), while depleted sample contained 17% of absolute viral reads (equals to 12.75 million reads). *In silico* rRNA depletion step further decreased the proportions of host-derived RNA sequences, but the absolute number of viral reads remained the same (data not shown). Assembly of rRNA- depleted reads resulted in 5.5-fold more recovered contigs than the assembly of total RNA reads with MEGAHIT (Table 3.1). However, I observed no difference in N50, maximum contig length or number of recovered RNA viral contigs (Table 3.1).

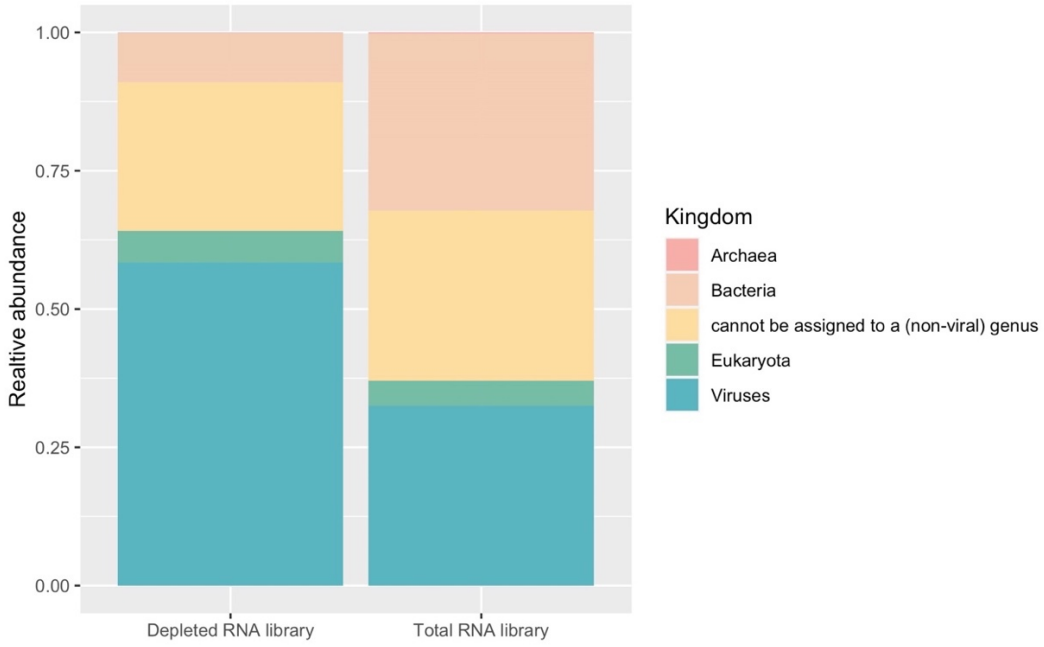


Figure 3.3. Broad taxonomic distribution of the sequencing reads before (A) and after (B) in-silico rRNA depletion.

Table 3.1. Basic sequencing information on a RNA virome sample sequenced with two different RNA sequencing approaches.

	Total RNA seq	rRNA depletion (Ribo-Zero Plus)
Total reads	86 211 226	76 050 048
Reads after filtering and trimming	86 133 668	75 989 764
Reads after *in silico* rRNA depletion (%)	50 762 588 (58.88%)	68 943 676 (90.65%)
Number of total assembled contigs /N50/longest contig	13 827 N50 640 Max 16923 bp	70 454 N50 643 Max 14 933 bp
Number of confirmed RNA viral contigs	158	161
Average length of confirmed RNA viral contigs	3665bp	2576bp
Near-complete RNA viral genomes recovered	26	24

3.4.4. Diversity of RNA viruses in the viral-size (<0.22μm) fraction and genomic analysis of novel members of the *Picornavirales*.

The majority of 319 confirmed RNA viral contigs in the Raffles Marina samples was assigned to the order of (+) ssRNA *Picornavirales* in the phylum *Pisuviricota* (n=244) (Figure 3.4.). Among Picornavirales, most abundant groups in Raffles Marina virome were protist-infecting *Marnaviridae* (n=107), unclassified *Picornavirales* (n=85) and invertebrate-infecting *Discistroviridae* (n=46) (Figure 3.5.). Most of the marna-like sequences detected in Raffles Marina dataset clustered with diatom-infecting viruses of the genus *Sogarnavirus.* Second most abundant group of RNA contigs in Raffles Marina dataset was assigned to (+) ssRNA phylum *Kitrinoviricota* (n=22). Third most abundant group of RNA viral contigs clustered under order of double-stranded RNA order *Durnavirales* of the phylum *Pisuviricota*. Among these sequences, four clades did not cluster with any sequences from the RefSeq database used to make the global ortornaviral tree, and were named unknown Raffles clade 1 - 4 (Figure 3.4). Interestingly, the clades were very diverse and were phylogenetically related to 4 different phyla. Raffles clade 1 (n=5) was most closely related to dsRNA bacteriophages from the family *Cystoviridae*, phylum *Duplovericota*. Raffles clade 2 (n=7) was related to order *Nidovirales (*phylum *Pisuviricota)*, enveloped (+) ssRNA virus order that includes nidoviruses, common pathogens of aquatic animals as well as coronaviruses. Novel Raffles clade 3 (n=5) clustered with one unclassified "orphan" viruses resembling *Weiviruses* and could likely represent a Yangshan clade uncovered by Wolf et al (2020). Finally novel Raffles clade 4 with only one representative was related to (-) ssRNA *Negarnaviricota*. Out of 319 RNA viral contigs detected, 50 contigs had both non-structural protein (RdRP) and structural capsid protein present, meeting the criteria to be categorized as near-complete genome (Gann et al., 2021; Moniruzzaman et al., 2017).

The basic information about all 50 near-complete or complete RNA viral genomes is presented in table A3.4. Two full - length genomes (contig 3188 and contig 13583), and one selected reference genome (Cheatoceros RNA virus 02, NC_055125.1), all clustered within the genus *Sogarnavirus* (Figure 3.6), were analysed in more details. The assembled RNA genomes were 9645bp and 11325bp in length, for contig 3188 and contig 13583 respectively. The reference genome of Cheatoceros RNA virus 02 is 9417bp in length. Comparative analysis revealed that all three genomes (NC_055125, Contig 3188 and contig 13583) contain two large open reading frames (ORF) encoding non-structural polyprotein (ORF1) and structural polyprotein (ORF2). Non-structural open reading frame (ORF1) encoded two replication-related domains: a helicase domain and RdRp domain. No protease domain was detected in ORF1 of any of the 3 analysed genomes (Figure 3.6B, Table 3.2). The second ORF encoded for structural capsid proteins in all three analysed genomes. One conserved structural domain, most closely related to Cricket paralysis virus VP4 protein (Dicistro_VP4, PF11492), was detected in the reference genome of Cheatoceros RNA virus 02. Two conserved structural domains, similar to Cricket paralysis virus VP4 protein (Dicistro_VP4, PF11492) and Cricket Paralysis Virus capsid-like protein (CRPV, PFAM 08762.13) were detected in the contig 3188. Finally, in the contig 13583 structural ORF, I detected three structural domains, two with similarity to the Cricket paralysis virus VP4 protein (Dicistro_VP4, PF11492) and one with similarity to the Cricket Paralysis Virus capsid-like protein (CRPV, PFAM 08762.13). Multiple sequence alignment of the detected open reading frames revealed low protein sequence similarity to the reference genome. Structural ORFs of the investigated genomes had 58.5% and 55.6% of similarity to the structural ORF of the Chaetoceros RNA virus 02 genome for the contig 3188 and contig 13583, respectively. Similarly, non-structural ORFs of the investigated genomes had 65.5% and

59.1% of similarity to the non-structural ORF of the Chaetoceros RNA virus 02 genome for the contig 3188 and contig 13583, respectively (Figure A3.4.)

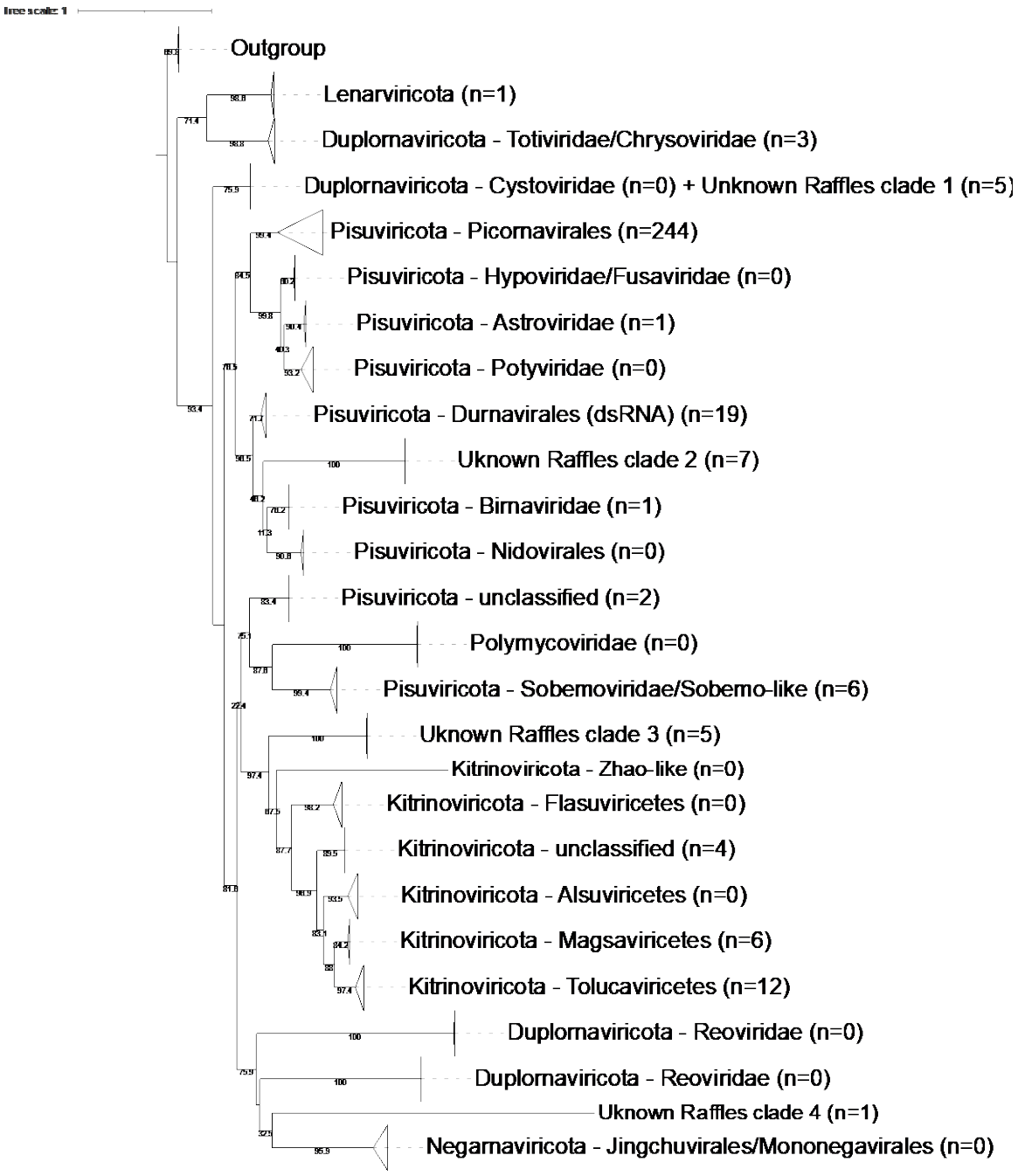


Figure 3.4. Taxonomic assignment of the RNA viruses identified in the viral-size fraction (<0.22μm) on the backbone of a maximum likelihood ortornaviral phylogenetic tree of viral RNA-dependent RNA polymerase. Five major ortornaviral phyla are currently recognized by ICTV, *Lenarviricota*, *Pisuviricota, Kitrinoviricota, Duplornaviricota* and *Negarnaviricota*. They backbone tree was built by using viral sequences from the RefSeq database, with *Artverviricota* and *Uroviricota* as an outgroup. Size of the triangle is proportional to the number of sequence used for tree-building. The number of recovered contigs per order/family/unclassified lineage in shown in brackets.

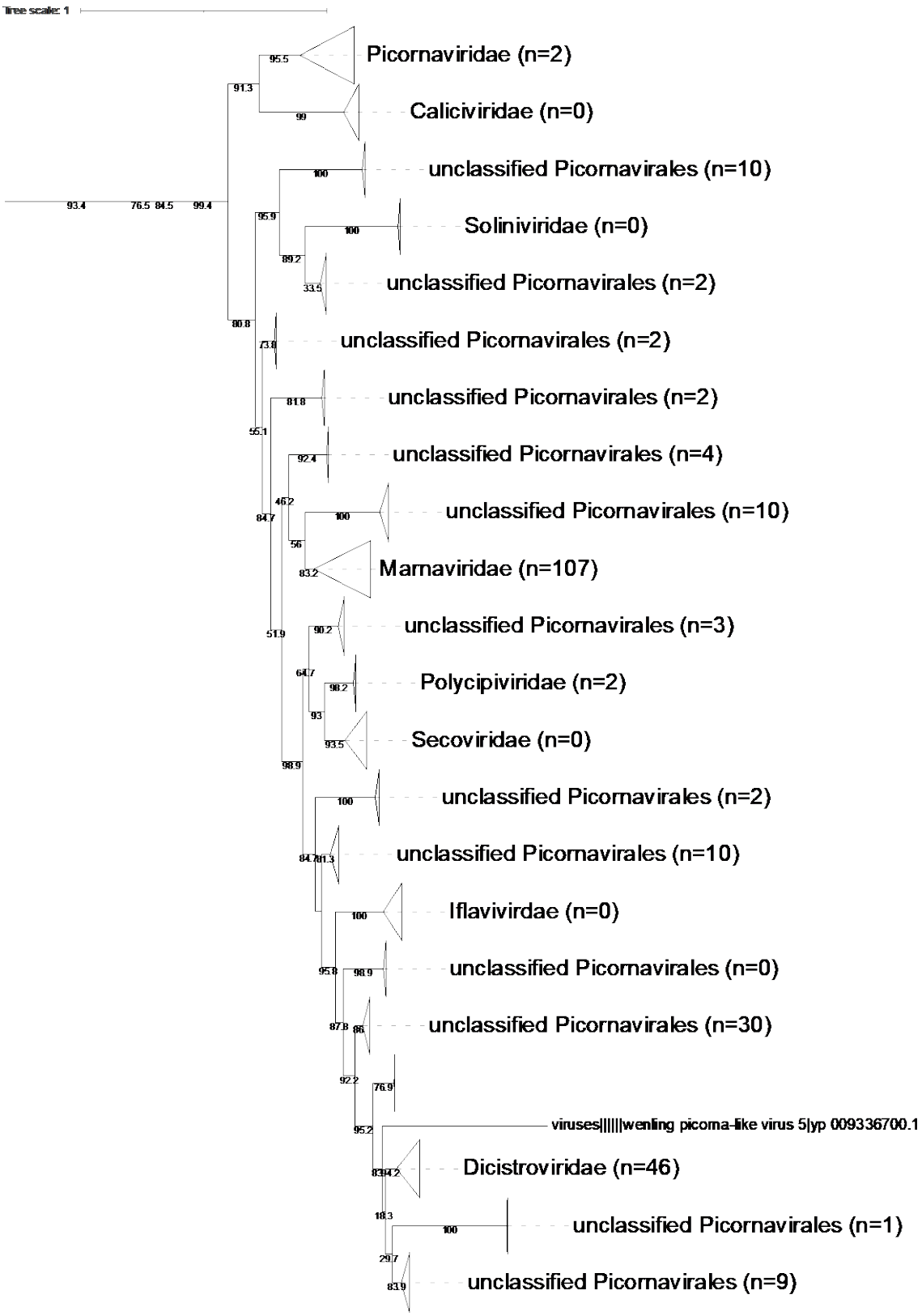


Figure 3.5. Maximum-likelihood phylogenetic tree showing in details distribution of recovered RNA viral contigs over the most abundant order detected in the sample, order *Picornavirales* of the phylum *Pisuviricota*. The number of recovered contigs per order/family/unclassified lineage in shown in brackets.

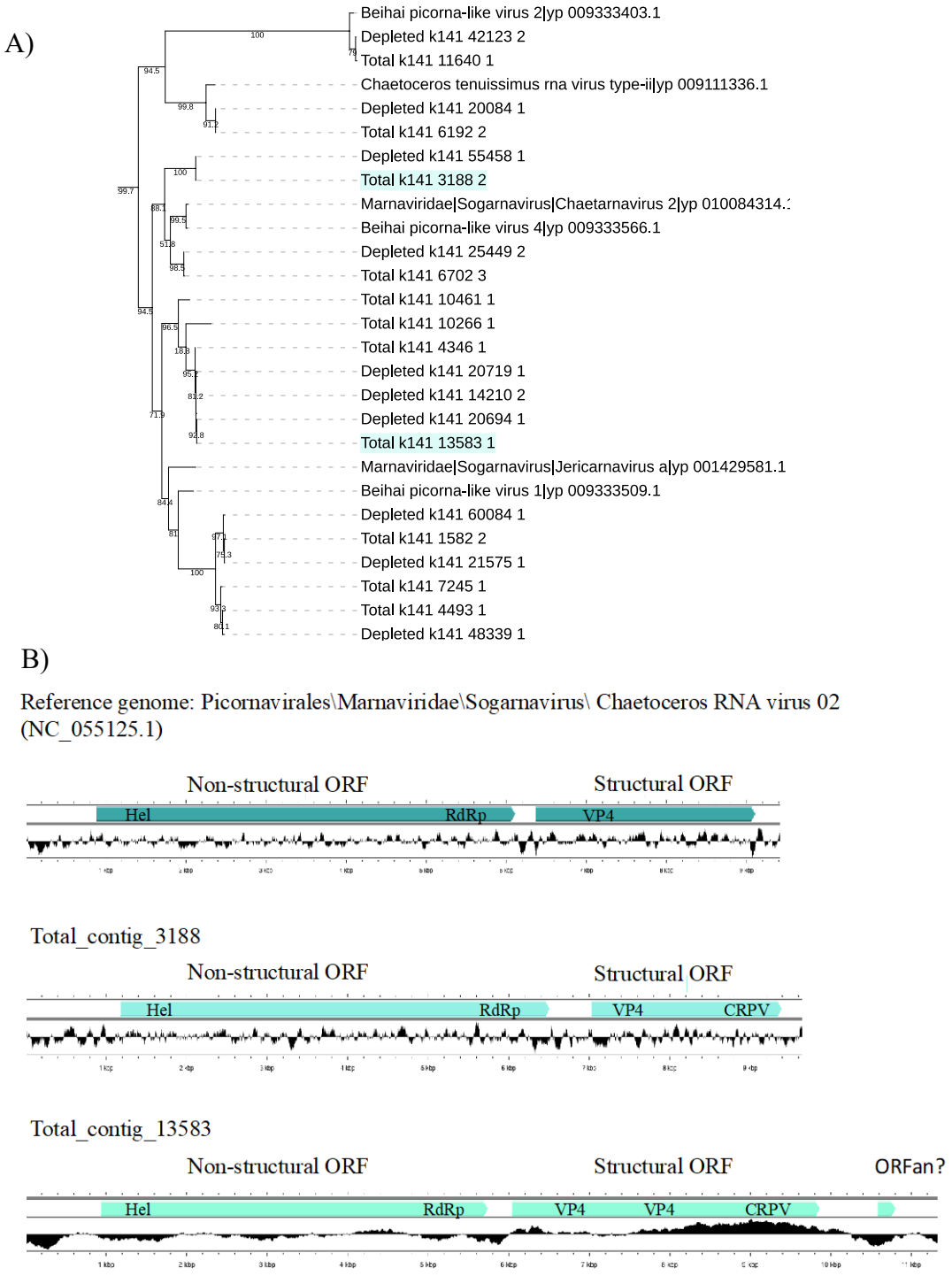


Figure 3.6 (A) Phylogenetic analysis of selected full-length sogarna-like viruses recovered in this study using conserved RNA-dependent RNA polymerase domains. Sogarnaviruses fall under the family *Marnaviridae*, order *Picornavirales*, phylum *Pisuviricota*. (B) Schematic genome organization of selected sogarna-like viruses identified in this study. The open reading frames were predicted using the standard genetic code. Protein sequences were predicted with Prodigal and annotated against the PFAM database. Hel=helicase; RdRp=Rna-dependent RNA polymerase; VP4=capsid protein VP4; CRPV = cricket paralysis virus capsid-like protein.

Table A 3.2. Open reading frames, coordinate, length and annotation of a reference sogarnavirus and two putative sogarna-like viruses recovered in this study.

Contig ID	Full Contig Length (bp)	ORF	ORF Coordinates (bp)	ORF Length (bp | aa)	Gene Annotation | Region (aa) | (PFAM)
Chaetoceros sp. RNA virus 2 NCBI Reference genome: NC_055125.1	9417	ORF1	883 - 6011	5229 | 1708	RNA_helicase |427-537aa| (PF00910.25) RdRp_1 |1234-1649aa| (PF00680.22)
		ORF2	6365 - 9106	2742 | 790	Dicistro_VP4 |160-215| (PF11492)
Total_3188	9645	ORF1	1,182 - 6,506	5325 | 1774 2685	RNA_helicase |401-511aa| (PF00910.25) RdRp_1 |12-1606aa| (PF00680.22)
		ORF2	7,027 - 9,390	2634 | 809	Dicistro_VP4 | 215-271 aa| (PF11492) CRPV_capsid | 567-805 aa | (PFAM 08762.13)
Total_13583	11325	ORF1	935- 5743	4809 |1602	RNA_helicase |391-501aa| (PF00910.25) RdRp_1 |1153-1565aa| (PF00680.22)
		ORF2	6039 - 9863	3825 | 1289	Dicistro_VP4 | 333-391aa | (PF11492) Dicistro_VP4 | 692-748 aa| (PF11492) CRPV_capsid | 1048-1284aa | (PFAM 08762.13)
		ORF3	10579 - 10806	228 | 76	No match (Potential ORFan)

3.5. Discussion

3.5.1. Quantity and quality of RNA obtained from the cellular fraction (>0.22μm)

In this chapter I directly compared different RNA extraction and filter types typically used in the marine metatranscriptomics studies that investigated viruses present in the cellular fraction (>0.22μm). Best yields and quality of RNA among compared kits were achieved with the open Sterivex extraction approach, which was then was selected as a extraction method of choice for the ecological study in the Johor Strait. It was suggested earlier that lysis and subsequent washes may be less effective inside the limited space of closed Sterivex casing (Cruaud et al., 2017), which may explain better yields obtained with the open Sterivex approach. Another advantage of Sterivex filters over filter membrane type of filters is that the closed cylindrical casing reduces the risk of contamination in the field (Cruaud et al., 2017). I also speculate that in the closed Sterivex protocol and the polycarbonate filter protocol tested the RNA Later was not fully removed. The carryover probably affected both the efficiency of extraction and the clean-up procedures as these salts would precipitate during clean-up step, block the column and reduce the final, clean RNA yields. This also may explain the genomic DNA contamination observed, as carryover salts will cause non-optimal conditions for TurboDNase enzyme and cause incomplete DNA degradation. However, we were not able to identify accurately which variables (kit, filter type, pore size, RNALater carryover) in the extraction process accounted the observed differences in RNA quality (RIN).

Besides the choice of RNA extraction methods, both filtration process and phytoplankton biomass will alter the quality of RNA. Due to natural degradation of cells during filtering, obtaining highly intact RNA (RIN 9-10) from environmental samples may not be possible. In samples with higher phytoplankton biomass, filter overload and

clogging may lead to increased pressure and cell lysis, and ultimately to a poor RIN score. In my experiments, filtering 1L of seawater led to good RIN scores (>7) in samples with low chlorophyll concentrations (<5μg/L). However, for chlorophyll concentrations >5μg/L, filtering 1L of seawater decreased the RIN score to 4-5. In ecosystems with highly variable chlorophyll biomass, care should be taken to experimentally test the filtration volume at the specific sampling site to ensure both high yield and high quality of RNA.

3.5.2. Quantity and quality of RNA obtained from the viral-size fraction (<0.22μm)

Marine RNA virome studies are hampered by sampling large volumes of water for viral concentrations and small yield of viral RNA after nucleic acid extraction. Generally, protocols optimized for DNA viruses are often used for RNA viruses, with varying degrees of success. Currently, there are no published protocol or evaluation specifically focuses on extraction of RNA from viral-size fraction (<0.22μm). The RNA virome sample-to-sequence protocol that I have established, summarized in Figure 3.7, builds and improves on existing wet lab steps and integrates parts of previously published protocols. The main steps include viral concentration (via TFF of iron flocculation), ultracentrifugation, and extraction and clean-up of nucleic acid (Thurber et al., 2009).

In my tests, RNA extractions following the viral concentration with TFF were not successful. Tangential flow filtration is well-established, commonly used approach for viral concentration (Hurwitz et al., 2013; Thurber et al., 2009; Wolf et al., 2020; Wommack et al., 2010). I suspect that omitting the ultracentrifugation step and DNAse treatment in my initial experiments, and the use of ultrafiltration devices for the final concentration instead, resulted in unsuccessful RNA extraction. Firstly, the ultrafiltration devices would often get clogged due to viscosity of the viral concentrate.

Similarly, the extraction column was overloaded presumably by free-floating cellular DNA, presumably due to lack of DNAse pre-treatment.

The final sample-to-sequence workflow first used iron flocculation as a method of viral concentration (John et al., 2011). In the following step of the workflow, purification of viruses by ultracentrifugation on a sucrose cushion combined with DNase treatment, has been shown to minimize microbial contamination of viral concentrates (Hurwitz et al., 2013). Small technical modifications introduced mainly in RNA extraction and DNAse steps in our protocol led to a robust, reproducible, and efficient protocol for successful RNA extraction from marine viral concentrates, recovering 1.868μg of RNA from just 40L of seawater. In comparison, other published RNA virome study focused on the viral fraction reported yields of 15-30ng per 50L of seawater (Miranda et al., 2016). Environmental samples, like viscous viral concentrates may require harsher procedure to release the viral nucleic acid content, like phenol chloroform, bead-beating of combination of both. I suspect this is the reason why some kits that are meant for serum samples, though routinely used for environmental RNA viromes (Culley et al., 2014; Culley et al., 2006), performed sub-optimally with the viral concentrates in our tests. Besides high yields, an additional benefit of All Prep Environmental DNA/RNA kit is the efficient inhibitor removal in environmental samples (Iker et al., 2013) and higher viral richness compared to other extraction kits (Hjelmsø et al., 2017). The viral concentration process can concentrate organic inhibitors like humic acids, that originate from terrestrial runoff and are present in high amounts in productive coastal environments (Wnuk et al., 2020). Also, any leftover iron from the iron flocculation process must be removed as it can inhibit library preparation enzymes (Erin Eggleston, personal communication). Most of the current protocols include one brief DNAse digestion step of 1-2U for 10min (Wolf et al. 2020; Culley et al. 2006), the protocol

presented here has two DNAse digestion steps. The first one is used to remove extracellular DNA before extraction (100U/mL, 2h) as suggested by Hurwitz et al (2013), and the second one (90U/mL, 1h) to remove the coextracted viral DNA. This "sample-to-sequence" workflow can also be applied to RNA viromes obtained from the viral-size fraction of other aquatic environments such as lakes, wastewater, recreational or drinking water supplies.

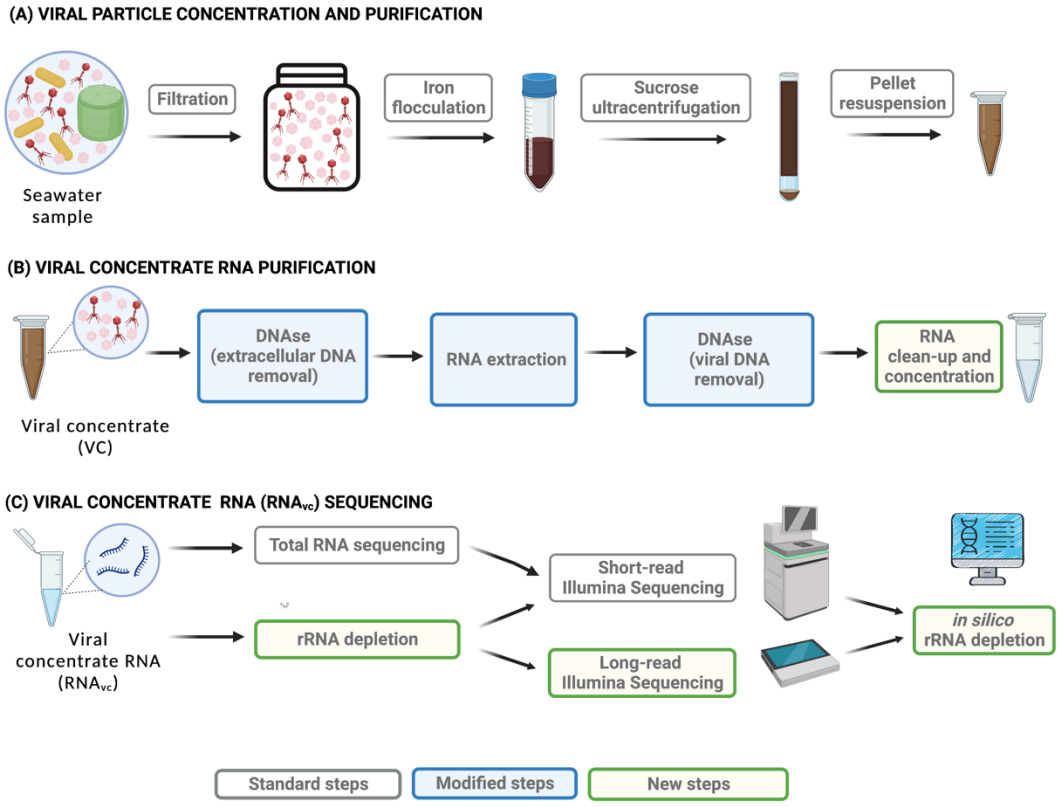


Figure 3.7. Final wet lab "sample to sequence" integrated workflow for generation of RNA viromes from the marine viral-size fraction: (A) viral particle concentration and purification (based on Hurwitz et al. 2013), (B) modified viral concentrate RNA purification and (C) viral concentrate RNA (RNAvc) sequencing. Standard steps are labelled in grey, the parts of protocol that include modifications are labelled in blue and completely new steps introduced in the protocol are labelled in green.

3.5.3. Effect of physical and in silica rRNA depletion on assembly and recovery of RNA viruses from the viral-size fraction (<0.22μm) fraction

The same concentration approach, combining iron flocculation and sucrose ultracentrifugation resulted in 35% of reads being assigned as viruses in the investigation of dsDNA viromes (Hurwitz et al., 2013), while we recovered 10-17% reads assigned as RNA viruses. Since RNA viral genomes are substantially smaller than dsDNA viral genomes, a lower recovery is to be expected. Other studies focusing on RNA viruses in the viral-size fraction reported the recovery of a maximum of 5% of reads that could be assigned to RNA viruses (Urayama et al. 2018; Solomon and Hewson 2022). In comparison, my workflow significantly increased the enrichment of RNA viruses.

Up to 40% of taxonomically assignable readings from the DNA viral metagenomes in the viral fraction contain non-viral sequences, even after physical enrichment of viral particles with tangential flow filtering or iron flocculation (Roux et al. 2013; Hurwitz et al. 2013; Rosseel et al. 2015). Only 14% of the total RNA sample from the Raffles Marina sample and 5% of the RNA from the depleted sample were determined to be of cellular origin. This may be due to the fact that the majority of cellular mRNA are quickly damaged during this multi-step, multi-day sample preparation technique and therefore the number of host-derived sequences is lower for RNA viromes than for DNA viromes.

Still, ribosomes are frequently coenriched with viral particles (Rosseel et al., 2015) and intact ribosomes are resistant to activity of RNAses (Gerashchenko & Gladyshev, 2017). Therefore, RNA extracted from the viral-enriched fraction will unavoidably contain a sizable amount of rRNA. Here, I demonstrated that rRNA depletion before sequencing can almost double the number of viral reads that can be retrieved from viral-size fraction

samples. Hence, more viral reads can be recovered with less sequencing depth, reducing sequencing costs. Additionally, *in silico* rRNA minimized the number of host-derived sequences (2-5%) and should be incorporated as a standard step to "clean" the virome data from host sequences. The removal of host-derived sequences reduces complexity and leads to better assembly and higher contig number in DNase treated soil DNA viromes (Sorensen et al., 2021). The removal of the host background could explain the 6-fold increase of total assembled contigs in rRNA depleted sample as opposed to untreated sample (Table 3.1). Recovery of the same number of RNA viral contigs was unexpected, since Gann et al. (2021) demonstrated 5-fold increase RNA viral contigs detected when comparing poly-A selected and rRNA-depleted metatranscriptomic libraries.

Introduction of the rRNA depletion step may also be advantageous for direct long-read RNA sequencing (Oxford Nanopore Technologies) of marine viral-size fraction (<0.22μm). Extended sample preparation times (up to 4 days from sampling to clean RNA sample) can lead to degradation and lower quality RNA. The removal of rRNA may be used in place of the suggested poly-A selection since it does not require a presence of intact poly-A tail on RNA molecules, hence will have a higher recovery (Kolundžija et al., 2022). The benefits of direct long-read RNA sequencing (i.e. without a reverse transcription step) may prevent many of the biases inherent to short read sequencing protocols that rely on cDNA generation (Liu & Graber, 2006). Native RNA strands up 26kb long have been successfully sequenced (Jain et al., 2022), suggesting that complete RNA viral genomes, which are often shorter than 15 kb (Holmes, 2003), might be recovered. Generating longer RNA viral reads may circumvent the assembly of misoriented and/or chimeric sequences and reduce the computationally challenging process of short-read assembly (Garalde et al., 2018; Jain et al., 2022).

3.5.4. Diversity of RNA viruses in the viral-size (<0.22μm) fraction and genomic analysis of novel members of the *Picornavirales*

Recent RNA virome study high diversity (>4500) of RNA viruses in just one sample in the Yangshan harbor, reporting multiple clades encompassing 3 phyla of (+)ssRNA viruses: *Lenarviricota*, *Pisuviricota* and *Kitrinoviricota* (Wolf et al., 2020). Earlier RNA virome studies include random amplification of RNA viral genomes with RT-SISPA before the library preparation and RNA virome sequencing (reviewed in Chapter 2, Kolundzija et al., 2022), and report narrower diversity range, typically >90% belonging to the order *Picornavirales* of the phylum *Pisuviricota.* Random amplification methods are known to cause biases in the taxonomic representation of viruses and negatively influence viral diversity (Karlsson et al., 2013; Yilmaz et al., 2010). The diversity of RNA viral community composition in the Raffles Marina viral-size fraction (<0.22 μm) was explored with the integrated OrVit workflow (Cheng et al., 2022). Seventy-five percent of my RNA viral contigs were identified as *Picornavirales*, compared to 26% in the Yangshan harbour study (Wolf et al., 2020), placing the results I obtained in the middle of the phylogenetic range reported by different RNA virome studies. Viral families like *Marnaviridae*, *Dicistroviridae* and unclassified *Picornavirales* are abundant in other ocean RNA viromes studied (Culley et al., 2014; Culley et al. 2006; Miranda et al., 2016) and are commonly detected in the metatranscriptomes of marine microbial eukaryotes as well (Moniruzzaman et al. 2017, Chapter 3). The high representation in my dataset could be a result of high burst size viruses like *Picornavirales* taking up disproportionally high of free-floating virus "pool" and/or high abundance of their hosts and reflects true diversity in the investigated ecosystem. Alternatively, it's possible that my approach is limited in its ability to detect the full range of diversity that exists in the environment. Direct comparison with the sample

preparation method used in the Yangshan harbour study (Wolf et al., 2020) would be needed to clarify this.

The main similarity between published datasets and my dataset seems to be the omnipresence of *Marnaviridae*, that is as abundant in coastal waters as their diatom hosts (Malviya et al., 2016). However, a major difference between our dataset and other published datasets is the significantly higher proportion of dsRNA viruses (27 out of total of 319). For example the Yangshan harbour study detected only 6 dsRNA viruses as compared to 4593 (+) ssRNA viruses (Wolf et al., 2020), while early studies detected none (reviewed in Chapter 2, Kolundžija, Cheng, and Lauro 2022). Double-stranded RNA viruses I detected resemble *Chrysoviridae*, *Totiviridae, Partitiviridae,* and *Picobirnaviridae* families. Interestingly, though these viruses are not detected other studies of the viral size fraction(<0.22μm) with classic RNA viromic approach, they constituted majority of viral-size fraction in study that used dsRNA sequencing approach (Urayama et al., 2018). With exception of *Picobirnaviridae*, the other dsRNA viruses detected are known to cause nonlytic, asymptomatic persistent infections in plant and fungi (Roossinck, 2019). There is evidence that *Picobirnaviridae* might infect bacteria though they were originally considered to be vertebrate-infecting viruses (Cobbin et al., 2021; Krishnamurthy and Wang, 2018; Neri et al., 2022). Given the high number of levi-like sequences resembling (+)ssRNA bacteriophages *Leviviridae* in the Yangshan harbour RNA virome (Wolf et al., 2020), their absence in my dataset was unexpected. Instead, I detected dsRNA viruses resembling dsRNA bacteriophages *Cystoviridae,* which are also abundant in the marine sediments (Zhang et al., 2022) as well as dsRNA *Picobirnaviridae*, that are also potentially bacteria-infecting. Lastly, the Raffles clade 2, a set of completely new sequences that did not cluster with any viruses from the RefSeq database seem to be related either to birnaviruses or nidoviruses which

are both known to be pathogens of fish and crustaceans. (Delmas et al., 2019; Monalisha et al., 2020) Some members of *Dicistroviridae*, which were also abundantly represented in my dataset, are potentially pathogenic to marine invertebrates (Lightner, 2011).This suggests the existence of a stable pool of animal viruses present in Singapore coastal waters which may interfere, and potentially be detrimental to aquaculture activities.

From a single sample, I recovered 50 near-complete genomes of RNA viruses, surpassing the total count of genomes retrieved from the viral-size fraction in all previous studies combined. Typically, only 2-6 full length genomes are recovered even with multiple virome samples studied (reviewed in Chapter 2, Kolundžija et al, 2022), revealing the capacity of this approach for recovery of novel RNA genomes from environmental dataset. The functional annotation of selected sogarna-like full length genomes revealed that their genome is organized into two open reading frames (ORFs) or modules. The two modules, namely the non-structural (replication) module and the structural (capsid) module, are a standard for the family *Marnaviridae* and the entire order *Picornavirales*, as indicated by Sadeghi et al. (2021). Another notable characteristic of *Picornavirales* is the organization of the replication polyprotein ORF into three domains arranged in a conserved order: superfamily III helicase, protease, and superfamily I RNA-dependent polymerase (RdRP), which is commonly known as the Hel-Pro-Pol module. (Le Gall et al., 2008).

The first ORF in the two newly recovered genome as well as in the reference genome of Chaetoceros RNA virus 02, encoded only a putative helicase and RdRp domains. Similarly, the protease domain was not detected in the replication ORF of Guinardia delicatula RNA virus (GdelRNAV) isolate (Arsenieff et al., 2019) and in majority of metavirome-assembled genomes recovered from tropical waters (Culley et al., 2014). The proteolytic activity of virally encoded proteases is essential to cleave the viral

polyproteins and host proteases do not participate in that process (Le Gall et al., 2008). Considering that the gene is required for replication of all (+) ssRNA viruses and that the ORFs are similar in size to ORF that contain protease, I hypothesize that the replication ORF of the two new viruses encodes proteases that are too divergent from the known proteases to be picked-up with current bioinformatic approaches. The structural ORF was encoded downstream of the Hel-Pro-Pol module in the analysed genomes, which is a configuration is conserved in the entire order *Picornavirales,* including family *Marnaviridae* (Le Gall et al., 2008; Zell et al., 2017). The structural ORF of *Marnaviridae* genomes typically encodes 3 to 4 capsid proteins (Arsenieff et al., 2019; Culley et al., 2014; Lang et al., 2021). The lower number of structural protein domains per structural ORF in the analysed genomes could also be explained the divergence from the known structural proteins. On a genome-level, these novel sogarna-like viruses are highly divergent even they all share the same RNA-dependent RNA polymerase domain with similarity to superfamily I (RdRP_1, PF00680). Divergence of structural (capsid) proteins may especially be related to host range and specificity, indicating these viruses infect different species (Munke et al., 2020; Sadeghi et al., 2021). Despite the presence of untranslatable regions (UTR) at the both ends of each analysed genome, I did not detect poly-A tails in the 3'UTR, which might indicate that incomplete assemblies were obtained. This is not surprising given the limitations of assembly softwares when dealing with low-complexity but high-coverage regions such as poly-A tails.

5.6. Conclusions

The sensitivity of virus detection critically depends on successful extraction of viral nucleic acid. A low ratio of viral nucleic acid compared to the overwhelming amount of cellular nucleic acid poses a challenge in sequencing-based studies in marine viral

ecology. Here, I evaluated commonly used RNA extraction methods and optimized the sample preparation to achieve high yields (>1μg) and high quality (RIN>7) of RNA, suitable for metatranscriptome sequencing the marine cellular fraction (150μm-0.22μm). Additionally, I present an optimized sample-to-sequence method to recover > 1μg of RNA from 40L of seawater for sequencing of the virus-enriched fraction (<0.22μm). Improving the yield and quality of RNA makes it possible to avoid random amplification procedures and include an rRNA depletion step to further enrich for viral RNA. Consequently, a large number of full-length genomes of new RNA viral species could be recovered using my approach. In future studies, optimized protocols such as these can be utilized to discover new marine RNA viruses, identify potential aquaculture pathogens, and study the virus-host dynamics in marine ecosystems from an ecological perspective.

3.7. Acknowledgements

I would like to thank Dr. David Demory for our fruitful discussions about virus concentration, as well as Dr. Lauren Krausfeldt and Dr. Natalie Solonenko for tips on how to use their protocols. I am grateful to Halimah Razali, Winona Wijaya, Grace Tay, Roy Tan, Dawn Ng and Avneet Kaur for their help with sampling at Raffles Marina. I thank Assist Prof Adriana Lopes dos Santos for lending us the filtration equipment for the iron flocculation experiments. This study was funded by the Intra-CREATE Seed Collaboration Grant (Award NRF2018-ITS004-0001) and Competitive Research Programme (Award CRP21-2018-0005) awarded by the National Research Foundation (NRF) of Singapore. Sandra Kolundžija is supported by a Singapore International Graduate Award (SINGA) of the Agency for Science, Technology & Research (A*STAR).

APPENDIX 3

Table A 3.1. Initial sampling efforts to test for different combinations of sample volume, concentration methods and RNA extraction kits. TFF =Tangential flow filtration, IF= iron flocculation, HF=hollow fibre concentration.

Date	Volume collected (L)	Concentration method	Extraction method/kit	Successful (Y/N)
31.01.2018	40	TFF + Amicon+DNAse	QIAamp MinElute Virus (Qiagen)	N
05.06.2018	40	TFF + Amicon+DNAse	QIAamp MinElute Virus (Qiagen)	N
05.03.2019	100	TFF + Amicon	QIAamp MinElute Virus (Qiagen)	N
07.03.2019	100	TFF + Amicon	QIAamp MinElute Virus (Qiagen)	N
04.04.2019	100	TFF + Amicon+DNAse	QIAamp MinElute Virus (Qiagen)	N
10.04.2019	40	TFF + Amicon	QIAamp MinElute Virus (Qiagen)	N
10.04.2019	100	TFF + Amicon	QIAamp MinElute Virus (Qiagen)	N
24.06.2019	100	TFF + Amicon	Split RNA Extraction Kit (Lexogen)	N
11.07.2019	100	TFF+Amicon	Split RNA Extraction Kit (Lexogen)	N
02.09.2019	40	TFF + Amicon	Split RNA Extraction Kit (Lexogen)	N
03.09.2019	60	TFF + Amicon	Split RNA Extraction Kit (Lexogen)	N
09.09.2019	40	TFF + Amicon	TRIzol (Thermo Fisher)	N
10.09.2019	60	TFF + Amicon	TRIzol (Thermo Fisher)	N
30.10.2019	60	HF + Amicon	Split RNA Extraction Kit (Lexogen)	N
03.12.2019	60	HF + Amicon	Split RNA Extraction Kit (Lexogen)	N
10.12.2019	60	HF + Amicon	TRIzol (Thermo Fisher)	N
13.12.2019	60	HF + Amicon	TRIzol (Thermo Fisher)	N
17.02.2020	65	HF + Ultracentrifugation	Split RNA Extraction Kit (Lexogen)	N
24.02.2020	65	HF + Ultracentrifugation	Split RNA Extraction Kit (Lexogen)	N
10.03.2020	2x60	HF + Ultracentrifugation	TRIzol (Thermo Fisher)	N
22.06.2020	60	IF + Sucrose ultracentrifugation +DNase	TRIzol (Thermo Fisher)	Y
07.07.2020	40	IF + Sucrose ultracentrifugation +DNase	TRIzol (Thermo Fisher)	Y
15.07.2020	40	IF + Sucrose ultracentrifugation +DNase	TRIzol (Thermo Fisher)	Y
03.08.2020	40	IF + Sucrose ultracentrifugation +DNase	TRIzol (Thermo Fisher)	Y

Date	Volume collected (L)	Concentration method	Extraction method/kit	Successful (Y/N)
03.08.2020	40	IF + Sucrose ultracentrifugation +DNase	All Prep Environmental DNA/RNA kit (Qiagen)	Y
07.09.2020	2x 40 (duplicate)	IF + Sucrose ultracentrifugation +DNase	All Prep Environmental DNA/RNA kit (Qiagen)	Y
29.09.2020	40	IF + Sucrose ultracentrifugation +DNase	All Prep Environmental DNA/RNA kit (Qiagen)	Y
17.02.2021	2x40 (duplicate)	IF + Sucrose ultracentrifugation +DNase	All Prep Environmental DNA/RNA kit (Qiagen)	Y
21.01.2022	2x40 (duplicate)	IF + Sucrose ultracentrifugation +DNase	All Prep Environmental DNA/RNA kit (Qiagen)	Y

Table A 3.2. Filter type and pore size, extraction methods, volume of filtered water for RNA extraction yield tests for metatranscriptome sequencing.

Filter type	Extraction method	Volume filtered
Sterivex - closed	Sterivex DNA Easy PowerWater Isolation Kit (Qiagen) modified for RNA extraction.	1L
Sterivex -open	All Prep PowerViral Environmental DNA/RNA Kit (Qiagen)	1L
Polycarbonate filter, 0.8µm	All Prep PowerViral Environmental DNA/RNA Kit (Qiagen)	1L

Table A 3.3. The extraction efficiency of different RNA extraction approaches of viral concentrates obtained by iron flocculation and sucrose cushion ultracentrifugation. The unconcentrated seawater sample volume was 40L in all the experiments except in the experiment denoted with an asterisk, where the volume was 20L.

Extraction aliquots	Extraction method	Total RNA yield (ng)	Total DNA yield (ng)	Total RNA yield post clean -up (ng)	Notes
3x250µL	(Vivospin)Trizol +Glycoblue	414	611.5	Unsuccessful	Incompatible with column-based clean-up
3x250µL	(Vivospin)Trizol +Glycoblue	247.75	513.75	Unsuccessful	Incompatible with column-based clean-up
2x250µl	(Vivospin) Trizol -No glycoblue	610	722.5	Not attempted	Very dirty
3x200µl	(Vivospin) All Prep	420*	2140*	Not attempted	
8x200µL	All Prep	2255	17820	1858ng	
8x200µL	All Prep	2220	6720	1023ng	
8x200µL	All Prep	2280	6900	1257ng	
8x200µL	All Prep	1290	1180	673ng	
8x200µL	All Prep	1296	6040	Not attempted	

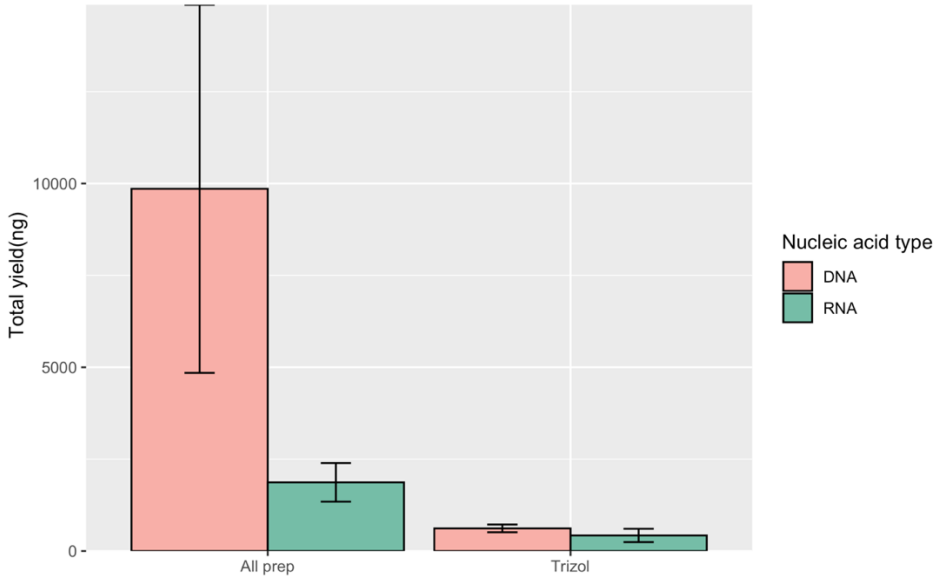


Figure A 3.1. Total yields of DNA and RNA obtained from the viral-size fraction (<0.22µm) by Trizol and All Prep Environmental RNA/DNA Kit extraction.

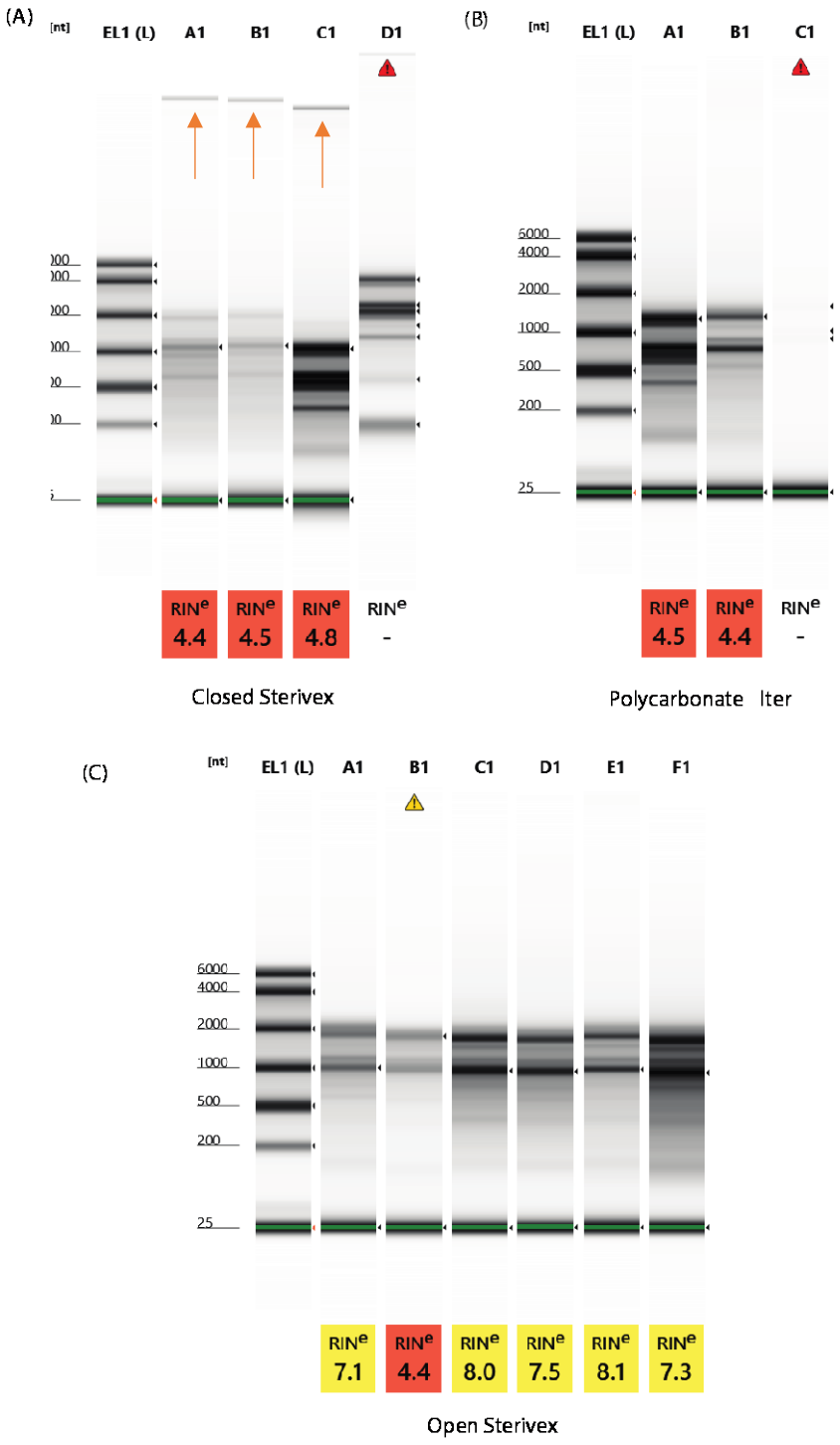


Figure A 3.2. Electrophoregrams show the integrity of RNA after different extraction approaches tested with RNA Screen Tape on Tape Station 2200 (Agilent). RNA integrity number (RIN) higher than 7 is required for metatranscriptomic sequencing. Lane EL1 represent the electronic molecular ladder with molecular weights ranging from 6000 to 25 nucleotides. (A) lanes A1- D1 represent technical replicates of closed Sterivex filter with Sterivex DNA Easy PowerWater Isolation Kit (Qiagen) modified for RNA extraction from 1L seawater samples collected at Raffles Marina. The orange arrows indicate presence of high-molecular DNA contamination even after the DNAse treatment. (B) lanes A1- C1 represent technical replicates of RNA extraction from polycarbonate filter with the All Prep PowerViral Environmental DNA/RNA Kit (Qiagen) from 1L seawater samples collected at Raffles Marina. (C) technical replicates of RNA extraction from the open Sterivex filter with the All Prep PowerViral Environmental DNA/RNA Kit (Qiagen) from 1L seawater samples collected at Raffles Marina (lanes A1-B1) and Johor Strait (C1-F1). After the extraction, all samples underwent identical TurboDNase treatment and RNA clean-up.

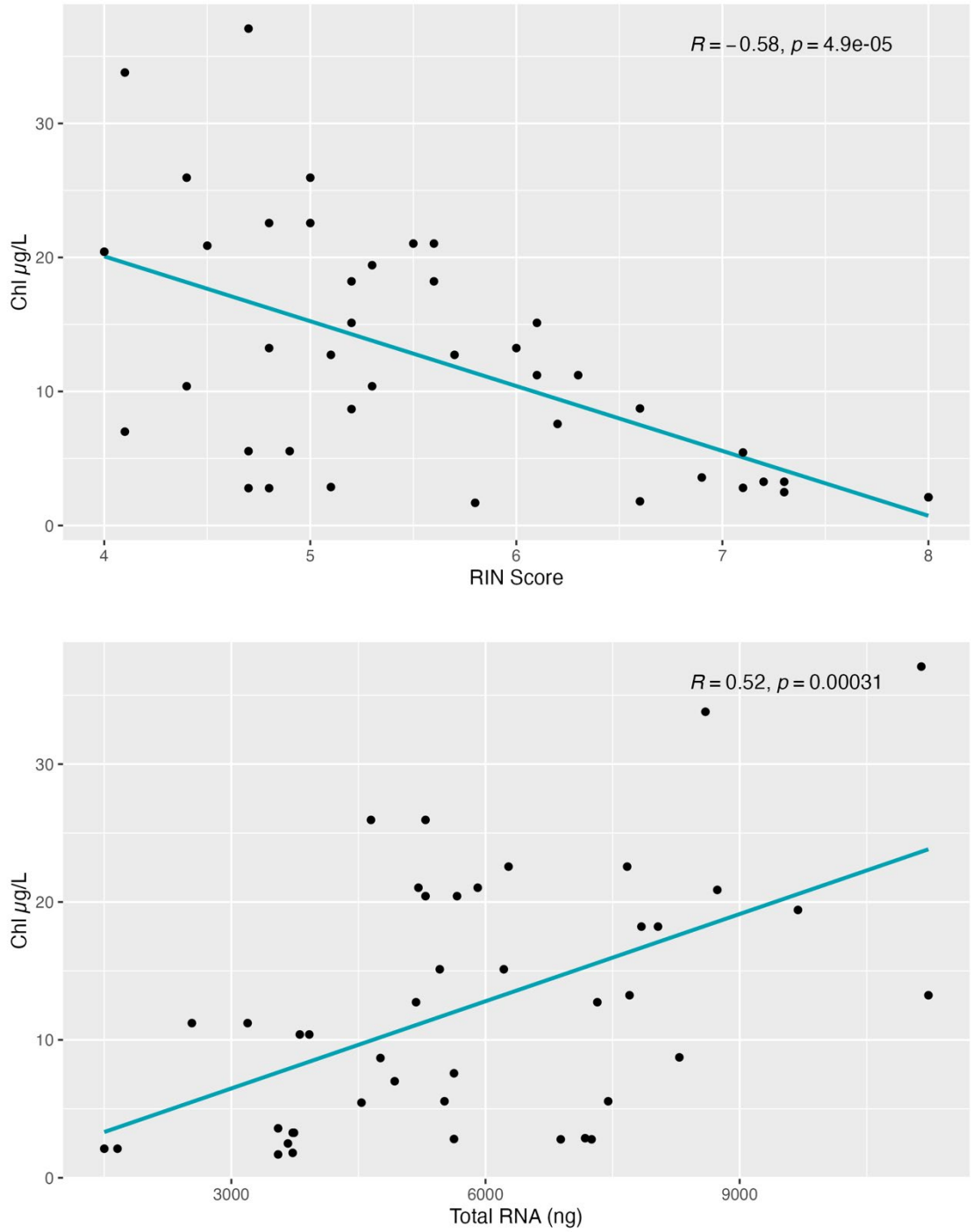


Figure A 3.3. (A) Correlation plot showing negative correlation between filtered biomass expressed as concentration of chlorophyll-a (µg/L) and RNA quality expresses as RNA INtegrity (RIN) scores. (B) Correlation plot showing positive correlation between filtered biomass expressed as concentration of chlorophyll-a (µg/L) and total RNA yield expressed in nanograms (ng). All RNA extractions(n=43) were performed with All Prep Environmental RNA/DNA kit and open Sterivex filter, with consistent sampling volume (1L) on multiple occasions (n=30) to capture different chlorophyll-a concentrations.

Table A 3.4. List of near-complete or complete RNA viral genomes recovered from viral-size fraction.

Contig ID	Contig length (bp)	Structural ORF/module	Replication ORF/module
k141_49626	2060	Rhv picornavirus capsid protein	RdRP_1 Viral RNA-dependent RNA polymerase
k141_37217	9606	CRPV_capsid CRPV capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_12486	2981	Dicistro_VP4 Cricket paralysis virus	RDRP-1 Viral RNA-dependent RNA polymerase
k141_64033	9145	Dicistro_VP4 Cricket paralysis virus	RDRP-1 Viral RNA-dependent RNA polymerase
k141_65926	9407	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_21575	7430	Dicistro_VP4 Cricket paralysis virus VP4	RDRP-1 Viral RNA-dependent RNA polymerase
k141_69492	3129	Dicistro_VP4 Cricket paralysis virus VP4	RdRp_1 Viral RNA-dependent RNA polymerase
k141_42972	1954	Rhv picornavirus capsid protein	RdRP_1 Viral RNA-dependent RNA polymerase
k141_32411	4196	CRPV capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_4141	3505	CRPV capsid protein like/ picornavirus capsid protein	RdRP_1 Viral RNA-dependent RNA polymerase
k141_25449	8181	Dicistro_VP4 Cricket paralysis virus VP4	RdRP_1 Viral RNA-dependent RNA polymerase
k141_20084	9481	Dicistro_VP4 Cricket paralysis virus VP4	RdRP_1 Viral RNA-dependent RNA polymerase
k141_16632	1417	Calicivirus coat protein	RdRP_1 Viral RNA-dependent RNA polymerase
k141_50723	3841	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_52547	7243	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_36645	4872	Calicivirus coat protein	RdRp_1 Viral RNA-dependent RNA polymerase
k141_70263	5415	CRPV capsid protein like	RdRp_1 Viral RNA-dependent RNA polymerase
k141_1225	5228	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_42123	6651	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_37005	5729	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_24636	3974	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_8689	7055	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_17630	2289	Viral coat protein (S domain)	RdRP_3 Viral RNA dependent RNA polymerase
k141_13583	11325	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase

Contig ID	Contig length (bp)	Structural	RdRp
k141_13622	6813	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_8652	9814	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_10793	5149	Viral coat protein (S domain)	RdRP_3 Viral RNA dependent RNA polymerase
k141_11518	4995	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_3619	9398	CRPV capsid protein like	RdRp_1 Viral RNA-dependent RNA polymerase
k141_7962	4734	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_6192	9592	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_3346	8584	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_8736	9011	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_5177	9400	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_3032	9156	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_11640	9718	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_13081	6575	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_7358	6620	Calicivirus coat protein	RdRp_1 Viral RNA-dependent RNA polymerase
k141_10279	4594	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_4949	1869	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_9203	4089	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_7803	9475	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_4227	5241	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_6702	9184	Dicistro_VP4 Cricket paralysis virus	RdRp_1 Viral RNA-dependent RNA polymerase
k141_685	9238	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_9598	4331	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_13167	9758	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_3164	3372	Rhv picornavirus capsid protein	RDRP-1 Viral RNA-dependent RNA polymerase
k141_13197	9755	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase
k141_3188	9645	CRPV_capsid protein like	RdRP_1 Viral RNA-dependent RNA polymerase

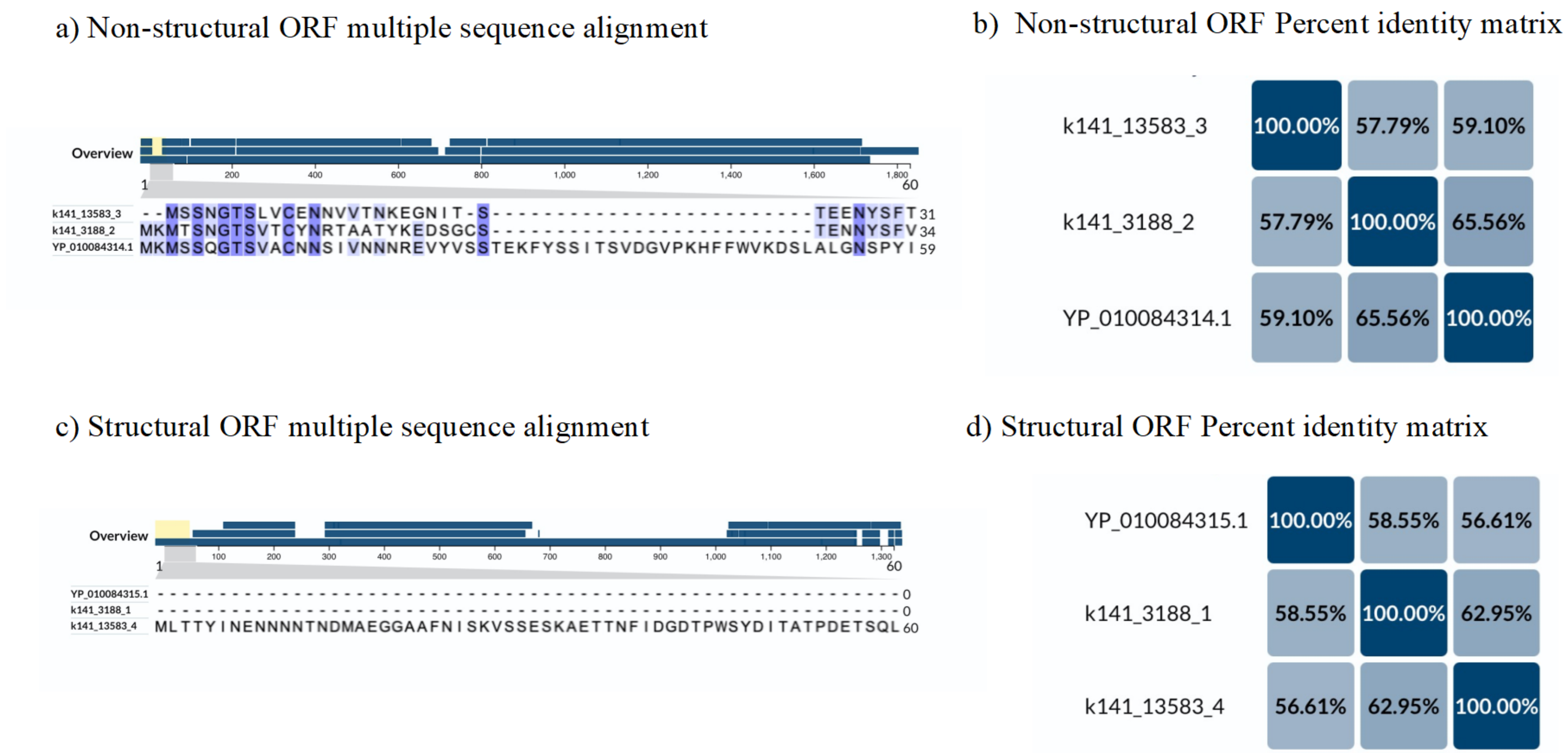


Figure A 3.4. Multiple sequence alignment of protein sequences detected in two analysed genomes (contig 3188 and contig 13583) and a reference genome of Cheatoceros RNA virus 02 (reference protein sequences YP_010084314.1, YP_010084315.1) Sequence alignment of ORF1 encoding for non- structural proteins (a) and the corresponding percent identity matrix (b). Sequence alignment of ORF 2, encoding for structural proteins (c) and the corresponding identity matrix (d).

CHAPTER 4 METATRANSCRIPTOMICS UNCOVERS NOVEL CLADES OF RNA VIRUSES AND GIANT DNA VIRUSES ACTIVELY INFECTING EUKARYOTES IN EUTROPHIC COASTAL WATERS

4.1. Abstract

Viruses infecting microbial eukaryotes play critical roles in marine ecosystem functioning. Yet, they are severely understudied relative to marine prokaryotic viruses. In a high-resolution 55-day time-series I used metagenomics to explore the diversity of phytoplankton and other marine microbial eukaryotes in the Johor Strait, an equatorial ecosystem known for excessive nutrient input and sporadic phytoplankton blooms. Bloom-forming diatoms *Chaetoceros*, *Thalassiosira* and *Skeletonema* had consistently high relative abundances, only surpassed by the prasinophyte *Ostreococcus* at the beginning and the end of the time-series. In parallel, I used metatranscriptomics to profile the phylogenetic extent of RNA and giant DNA viruses actively infecting eukaryotes, identifying 483 RNA viruses and 48 giant DNA viruses. Among the RNA viruses, the lytic (+) ssRNA picorna-like viruses infecting phytoplankton were the most diverse and abundant group. Surprisingly, capsidless (+) ssRNA and dsRNA viruses that exhibit exclusively intracellular lifestyle were the second most diverse group, suggesting the possible widespread existence of latent infections in marine microbial eukaryotes. Additionally, I detected invertebrate and fish RNA viruses that could potentially impact aquaculture activities in the Strait. Among giant DNA viruses, prasinoviruses had highest diversity in the Johor Strait. This study provides new insights into diversity of phytoplankton and their viruses and provides a baseline for

understanding the responses of marine ecosystem diversity to unremitting anthropogenic pressures

Keywords: metagenomics, metatranscriptomics, diversity, RNA virus, NCLDV, giant DNA virus, phytoplankton

4.2. Introduction

The marine photosynthetic microbial eukaryotes (or phytoplankton) support a large fraction of the worldwide primary production (Field et al., 1998), constitute the basis of the marine food webs (Simon et al., 2009) and play an important role in the biological carbon pump (Behrenfeld, 2014; Turner, 2015). Coastal ecosystems are highly variable and dynamic, with continuous natural or anthropogenically-induced fluctuations in physical and chemical conditions, which cause rapid changes in biomass, diversity and physiological activity of coastal phytoplankton communities (Beman et al., 2005; Spatharis et al., 2007).

The Johor Strait is a narrow passage separating the island of Singapore and peninsular Malaysia, one degree north of the equator. This coastal ecosystem supports intense port activities and important aquaculture sites for both countries, exposed to high anthropogenic pressures and excessive nutrient loads (Tan et al., 2016). The phytoplankton community of the Johor Strait undergoes large monthly changes in diversity and extreme variations in biomass, with chlorophyll-a concentration ranging from 1-200 μg/L (Chénard et al., 2019; Hii et al., 2021; Wijaya et al., 2022). Outbreaks of the toxic dinoflagellate *Karlodinium australe* and high-biomass diatoms blooms in the Johor Strait have been associated with massive fish kills, severe aquaculture losses and water treatment costs (Gin et al., 2000; Kok & Leong, 2019; Lim et al., 2014; Mohd-Din et al., 2020; Trottet et al., 2021).

Increased activity or abundance of phytoplankton viruses has been consistently observed during the decline of phytoplankton blooms (Alarcón-Schumacher et al., 2019; Baudoux et al., 2006; Biggs et al., 2021; Jacquet et al., 2002; Moniruzzaman et al., 2016, 2017; Tarutani et al., 2000), consistent with the classic "Kill the Winner" model (Thingstad, 2000). Electron microscopy and single-molecule in situ hybridization of RNA viral transcripts, the only methods that can detect virus particles of transcripts inside the phytoplankton cells, show that 27-37% of cells in a bloom can be infected by viruses (Gastrich et al., 2004; Vincent et al., 2021). Enormous diversity of the marine viruses infecting phytoplankton and other eukaryotes has been discovered with modern high-throughput sequencing approaches like metagenomics, metatranscriptomics and viromics (Charon et al., 2021; Endo et al., 2020; Ha et al., 2021; Kolundžija et al., 2022; Moniruzzaman et al., 2017; Schulz et al., 2020; Shi et al., 2016; Vlok et al., 2019; Wolf et al., 2020; Zayed et al., 2022). These include nucleocytoplasmic large dsDNA viruses (NCLDVs), also referred to as giant DNA viruses, with huge genomes with up to 2.5 Mb and 1000–2500 genes. (Claverie & Abergel, 2018; Ha et al., 2021; Philippe et al., 2013). RNA and ssDNA viruses have small genomes average size of 10kb, and carry a small number of essential genes (Belshaw et al., 2007; Holmes, 2003). Cellular fraction (>0.22μm) metagenomics can only detect only dsDNA viruses (including giant DNA viruses) that did not pass through the filter during the virus enrichment process. Viromics (viral-size (<0.22μm) fraction metagenomics) normally focuses either on DNA or RNA viruses, and largely excludes giant DNA viruses because of their size (Chapter 2, Kolundžija et al., 2022). In contrast, metatranscriptomics of the cellular fraction (>0.22μm), enables the simultaneous, unbiased detection of expressed RNA of all viral types inside a host cell (Cobbin et al., 2021; Zhang et al., 2019), allowing for exploration of full diversity of phytoplankton viruses and potential virus-phytoplankton

interactions in the ecosystem. The metatranscriptomic approach is able to identify subtler changes in viral populations because the intracellular virus population changes more rapidly than the free virus population ("bank of viruses") (Aylward et al., 2017; Hevroni et al., 2020). Major limitation of previous viral metatranscriptomic studies is the low sampling (e.g. weekly or monthly) frequency and the focus on temperate or polar environments (described in Chapter 2, Chapter 5 and (Kolundžija et al., 2022).

I conducted a high frequency sampling (every 1-2days), 55-day long monitoring study to understand the importance of certain viral groups in controlling the phytoplankton community in a dynamic coastal tropical environment. Using metagenomics, I describe diversity of major phytoplankton groups and investigate the presence of bloom-forming phytoplankton in the Johor Strait. Using metatranscriptomics, I unravel phylogenetically diverse community of eukaryote-infecting RNA and giant DNA viruses. The dynamics of virus - phytoplankton interactions will be further explored in detail in Chapter 5 of this thesis.

4.3. Materials and methods

4.3.1. Study site and sample collection

The study site was located in the Johor Strait, Singapore (1.46°N,103.84°E). The area (Figure 4.1) is impacted by anthropogenic activities (e.g. port activities, shipyards), and a thriving aquaculture industry. Short-term disturbances (e.g. rainfall and nutrient discharge from nearby rivers and reservoirs) have been implicated in changes in phytoplankton community composition, even triggering phytoplankton blooms (Hii et al., 2021; Mohd-Din et al., 2020; Trottet et al., 2021). The causeway connecting peninsular Malaysia and island of Singapore increases water residence times and intensifies nutrient retention (Tan et al., 2016). The field sampling campaign was conducted during the early Northeast monsoon, from November 4th to December 28th

2020. November and December are the rainiest months in Singapore, with the mean monthly rainfall (>250mm) (Hassim & Timbal, 2019). High rainfall events and increased freshwater inflow may lead to additional nutrient loading into Johor Strait (Palani et al., 2012). Seawater samples for analysis of chlorophyll-a biomass, chemical analysis of inorganic nutrients and microbial community composition were collected in technical duplicate over a 55-day sampling period, resulting in total of 30 timepoints. Physical parameters of water column - temperature, salinity, and turbidity (Secchi depth) of the water column, were analysed in the field (Extech Instruments; United Scientific Supplies Secchi Disk). The rainfall data for Johor Strait was obtained from the National Environmental Agency (NEA) of Singapore and the tidal measurements from the Maritime and Port Authority of Singapore. The bathymetry data, the current direction and velocity for the Johor strait during Spring Flood and Spring Ebb (Figure A4.1) were provided by the DHI Group.

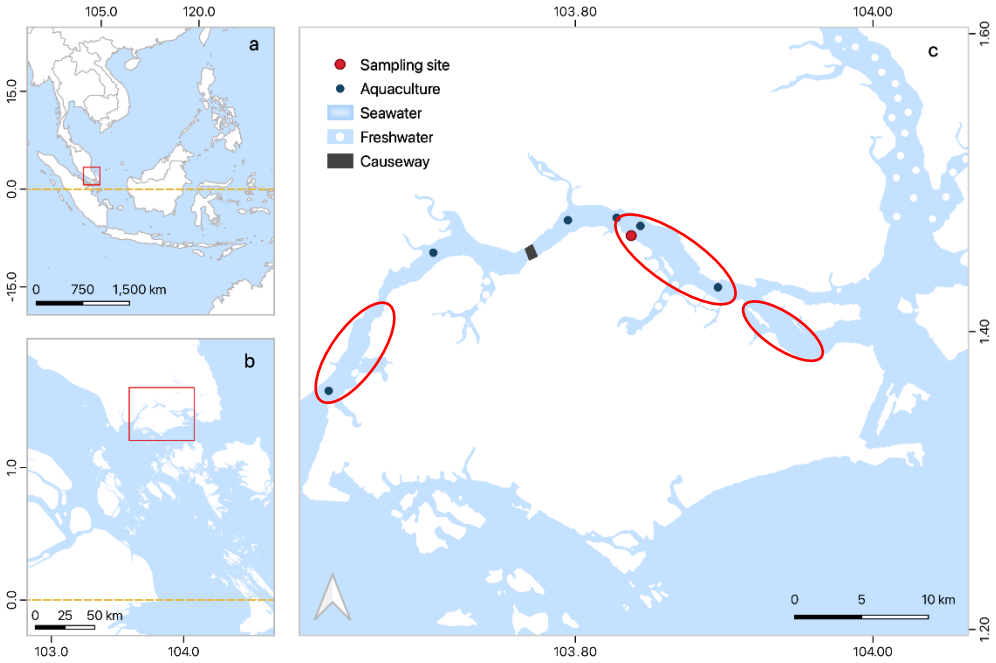


Figure 4.1. Map showing a) location within the southeast Asia and position relative to the equator. b) magnified view of the region c) study site, located in the Johor Strait, Singapore. Yellow dashed line represents the equator. Red dot denotes the sampling site. White circles indicate freshwater inputs. Dark grey rectangle represents the Johor-Singapore causeway. Dark blue dots illustrate the location of the aquaculture farms (Mohd-Din et al., 2022). Red circles represent algal bloom hotspots (Trottet et al., 2021).

4.3.2. Microbial sample collection, nucleic acid extraction and clean-up

For microbial community samples, one litre of seawater was prefiltered with 150µm mesh to exclude mesozooplankton (typically larger than 200µm). Seawater was pumped directly in the field using a custom-made autosampler device OSMO (Ocean Sailing Microbiome Observatory) (Lauro et al., 2014). Biomass was collected onto a 0.22µm nominal-pore Sterivex GP filter cartridge (EMD Milipore) for 15 minutes under minimal pressure to avoid degradation. All seawater samples for microbial community analysis were collected at 1m depth between 9am and 10am to avoid the influence of diel signals. The Sterivex filters were immediately preserved in 2mL RNA Later (ThermoFisher Scientific) and flash frozen in liquid nitrogen in the portable dry shipper. Frozen samples were transported in the lab within 2h and kept at -80°C until nucleic acid extraction. Samples were randomized before the extraction to minimize the batch effects. Total RNA and DNA were co-extracted with AllPrep Power Viral DNA/RNA kit (Qiagen) following protocol optimized in the Chapter 3. First, RNALater was removed from the Sterivex filter cartridge by washing with sterile phosphate buffered saline (137mM NaCl, 10 mM Na_2HPO_4, 2.7mM KCl, 1.8 mM KH_2PO_4) prepared with nuclease-free water (Ambion). The filter cartridge was cracked open, and the filter was cut in pieces with a sterile razor blade and placed in two bead-beating tubes (Cruaud et al. 2017, Chapter 3). The nucleic acids were eluted in total of 120µl of nuclease-free water and quantified on Qubit Fluorometer 2.0 with high sensitivity DNA and RNA Qubit assays (Life Technologies). The nucleic acid extract was split into two aliquots, one for DNA metagenome sequencing and the other for RNA metatranscriptome sequencing. Due to mistakes in sample collection and processing, two samples (sampling day 22 and 52) were excluded from metatranscriptome and metagenome sequencing to avoid skewing the results.

4.3.3. Short-read and long-read DNA metagenome sequencing

Short-read DNA sequencing was performed with TruSeq Nano DNA kit (Illumina) on NovaSseq 6000 sequencing platform by Macrogen (Singapore), generating approximately 60 million paired-end reads per sample. Long-read DNA libraries were prepared with PCR barcoding kit (SQK-PBK004, Oxford Nanopore Technologies). One hundred nanograms of DNA was sheared with Covaris G-tubes (Agilent). Fragmented DNA was end-repaired and A-tailed with NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs). PCR adapters were ligated to the repaired DNA using a NEB Blunt/TA Ligase Master Mix (New England Biolabs) for 10 minutes at room temperature. Adapter-ligated DNA was amplified with LongAmp Hot Start Taq 2x Master Mix (New England Biolabs) and whole genome primers in the following cycling conditions: initial denaturation for at 95°C for 3 minutes, 14 cycles of denaturation at 95°C for 15 seconds, annealing at 56°C for 15 seconds and extension at 65°C for 6 minutes. Final PCR extension was performed at 65°C for 6 minutes. Barcoded and amplified samples were mixed in equimolar ratios and ligated with the sequencing adapters. The libraries were sequenced on Nanopore flow cells with 9.4.1 chemistry and produced around 1 million raw reads per sample.

4.3.4. Metatranscriptome sequencing

For metatranscriptome sequencing, the nucleic acid extract was further treated with TurboDNase-free Kit (Ambion) under rigorous conditions to remove the contaminating DNA and cleaned with RNA Clean & Concentrator Kit (Zymo Research) as established in Chapter 3. Nucleic acid concentrations in the samples were quantified on Qubit Fluorometer 2.0 with a Qubit dsDNA HS assay kit and Qubit RNA HS assay kit (Life Technologies). RNA Screen Tape assay (Agilent) was used to ensure that genomic DNA was successfully removed, and that RNA has not been degraded. One microgram of

total RNA was depleted of rRNA using Ribo-Zero Plus rRNA removal kit (Illumina). The rRNA-depleted RNA went through library preparation with Stranded Total RNA kit (Illumina). Library preparation and sequencing was performed on the NovaSeq 6000 sequencing platform (Macrogen, Singapore) to produce a minimum of 60 million, 150-bp, paired end reads per sample.

4.3.5. Taxonomic assignment of the DNA metagenomes

Short read Illumina DNA sequence quality was checked with FastQC software. Sequencing adapters were removed with Cutadapt (Martin, 2011) and the reads were quality filtered with a custom Phyton script in two steps with following conditions: 1) only reads with 90% of bases with quality score above 20 were kept 2) bases with quality score <25 were trimmed from the 3' end and only reads longer >30bp were retained to ensure protein information can be retrieved. Quality-filtered reads were assembled with MEGAHIT version 1.2.9 (Li et al., 2015) and produced 19 066 313 contigs with N50 of 757bp. Raw fast5 Nanopore DNA files were base called, demultiplexed and quality-filtered with Guppy basecaller (version 5.0.7). Reads were taxonomically assigned with Kaiju (Menzel et al., 2016). Only eukaryotic reads were analysed further.

4.3.6. Viral discovery in the metatranscriptomic data

Raw RNA sequence data quality was visualized using FastQC (Andrews, 2010). The sequencing adapters were trimmed with Cutadapt (Martin, 2011). Low quality sequences were filtered with a custom Perl script. Filtering step retained only reads where 95% of the read length has quality score >20. The remaining reads were trimmed at 3' to remove bases with quality score <25 and only reads longer than 30bp were retained. Archaeal, bacterial and eukaryotic ribosomal RNA (rRNA) was removed *in silico* with SortMeRNA software package (Kopylova et al. 2012). Quality-filtered

metatranscriptome reads were co-assembled using rnaSPAdes (Bushmanova et al., 2019).

To search for RNA viruses/RNA viral transcripts, the metatranscriptome assemblies were further processed with the OrViT pipeline which is shown in Figure A4.2 (Cheng et al. 2022). Briefly, protein sequences were predicted with Prodigal v 2.6.3 (Hyatt et al., 2010). Potential RdRp homologs in metatranscriptome assemblies were identified using Hidden Markov Models (HMMs) using hmmsearch in HMMMER v3.3.2. Following publicly available HMMS for RdRp were used: Mononeg_RNA_pol [PF00946], RdRP_5 [PF07925], Flavi_NS5 [PF00972], Bunya_RdRp [PF04196], Mitovir_RNA_pol [PF05919], RdRP_1 [PF00680], RdRP_2 [PF00978], RdRP_3 [PF00998], RdRP_4 [PF02123], RVT_1 [PF00078], RVT_2 [PF07727], Viral_RdRp_C [PF17501], and Birna_RdRp [PF04197], with an additional HMM profile of Cystoviridae (Starr et al., 2019). To identify RNA viruses, RdRp palmprint, a highly conserved catalytic core domains (A,B,C), were excised from the full length RdRp. The palmprint subsequence was progressively aligned to the global ortornaviral phylogenetic tree constructed from RNA viral protein sequences from the RefSeq database (Cheng at al. 2022). Phylogenetic trees were visualized and edited using iTOL (Letunic & Bork, 2019).

Potential giant DNA viral transcripts were recovered by transcript mapping to a custom-built giant DNA virus database following previously described method (Aylward et al., 2021; Ha et al., 2021). Briefly, I downloaded giant DNA viral genomes of cultured isolates and viral metagenome-assembled genomes (vMAGs) from RefSeq database (147 genomes), GenBank database (1600 genomes) and total of 2 623 giant DNA virus genomes from the recent publications (Aylward et al., 2021; Ha et al., 2021; Matsuyama et al., 2020; Moniruzzaman et al., 2020; Needham et al., 2019; Needham, Yoshizawa,

et al., 2019; Rozenberg et al., 2020; Schulz et al., 2020) , resulting in acquisition of 4370 giant DNA viral genomes. To dereplicate, compiled giant DNA viral genomes were compared using MASH v2.3 (Ondov et al., 2016) with parameters -k 16 and -s 300 and clustered together via single linkage clustering distance using a Mash distance of 0.05, which corresponds to 95% average nucleotide identity (ANI). Viral genomes sharing ANI >95% are considered to be the same one viral species. The longest contig (highest N50) was selected from each species cluster was selected as a representative giant DNA genome. Dereplicated giant DNA viral genomes were screened with ViralRecall v2.0, bioinformatic tool that can identify and remove the genomes contaminated with sequences of cellular organisms or bacteriophages before the downstream analyses. The genomes with a negative score, indicative of presence of contaminating sequences not originating for giant DNA viruses, were removed (Aylward & Moniruzzaman, 2021). Decontamination resulted in final database of 2870 giant DNA viral genomes. The protein were predicted with Prodigal v2.6.3. (Hyatt et al., 2010) and masked with Tantan v.22 to eliminate false homologies (Frith, 2011). Protein reference database was prepared with lastdb and Johor Strait metatranscriptome reads were mapped with LAST v.1060 (Kiełbasa et al., 2011). The giant DNA viral genomes that had reads mapped to >10 % genes were treated as confirmed giant DNA viruses.

4.4. Results

4.4.1. Eukaryotic phytoplankton community composition in the Johor Strait

The raw environmental metagenomic dataset of the Johor Strait encompassed 2.2 billion reads distributed in 28 samples. Only 10 to 24.44% of the sequence data could be taxonomically assigned. Between 6.9% and 15.7% of sequences was assigned to Bacteria and 1%-9.1% of sequences was assigned to Eukaryotes (Figure A4.3, Table A4.1). Only eukaryotic reads were further analysed. Phylum-resolution taxonomic profiling of the eukaryotic community revealed high relative abundance of diatoms (*Bacillariophyta*), as they encompassed more that 75% of assigned eukaryotic reads during more than half of the time-series (Figure A4.4). Chlorophytes (*Chlorophyta*) had significant relative contribution (up to 40% of eukaryotic reads) during the early sampling period/low chlorophyll days. Dinoflagellates (*Dinophyta*) and ciliates (*Ciliophora*) constituted on average 3-4% of eukaryotic reads, with higher relative contribution on low chlorophyll days. Opisthokonts (*Opisthokonta*) were exclusively represented by fungal phyla were contributing to on average 6%, and maximum 11.7% of eukaryotic reads. Heterotrophic microbial eukaryotes (e.g. *Discoba*, *Amoebozoa*, *Rhizaria*) were continuously present throughout the dataset in relative abundances <1%. Top taxa on a genus level were diatoms *Chaetoceros, Cyclotella, Fragilariopsis, Skeletonema and Thalassiosira,* picoeukaryotic chlorophytes *Ostreococcus and Micromonas,* dinoflagellate *Symbodinium* and ciliate *Stylonychia* (Figure 4.2). Compositional variability over time and possible biotic and abiotic factors influencing it will be discussed in Chapter 5.

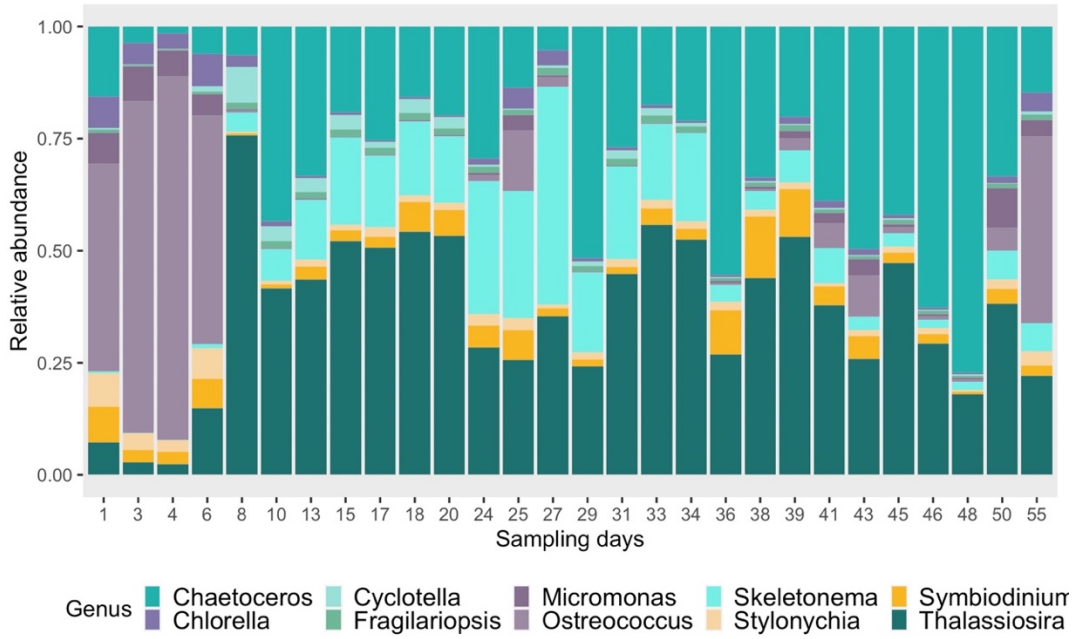


Figure 4.2. Stacked bar plot showing the relative abundance of top 10 eukaryotic taxa (genus-level) in the Johor Strait. Chlorophytes are shown in purple, diatoms in blue-green and dinoflagellates and ciliates in yellow.

4.4.2. Phylogenetic diversity of RNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes

The raw environmental metatranscriptomic dataset of the Johor Strait encompassed 2.2 billion reads distributed in 28 samples and resulted 3.95 million assembled transcripts. Using RNA-dependent RNA polymerase as a marker gene, I recovered 483 RNA viral transcripts from the Johor Strait metatranscriptomes that were taxonomically classified by mapping to a previously constructed global reference tree of *Orthornavirae* (Figure 4.3). The majority of the recovered RdRp sequences mapped within two phyla, *Lenarviricota* and *Pisuviricota*.

4.4.2.1. Phylum *Lenarviricota*

Total of 120 of all RdRp detected (25%) clustered within the phylum *Lenarviricota*, out of which the majority (n=99) resembled narna-like and mito-like viruses, and the remaining 21 RdRp sequences were identified as levi-like viruses (Figure 4.3, Figure 4.4.A).

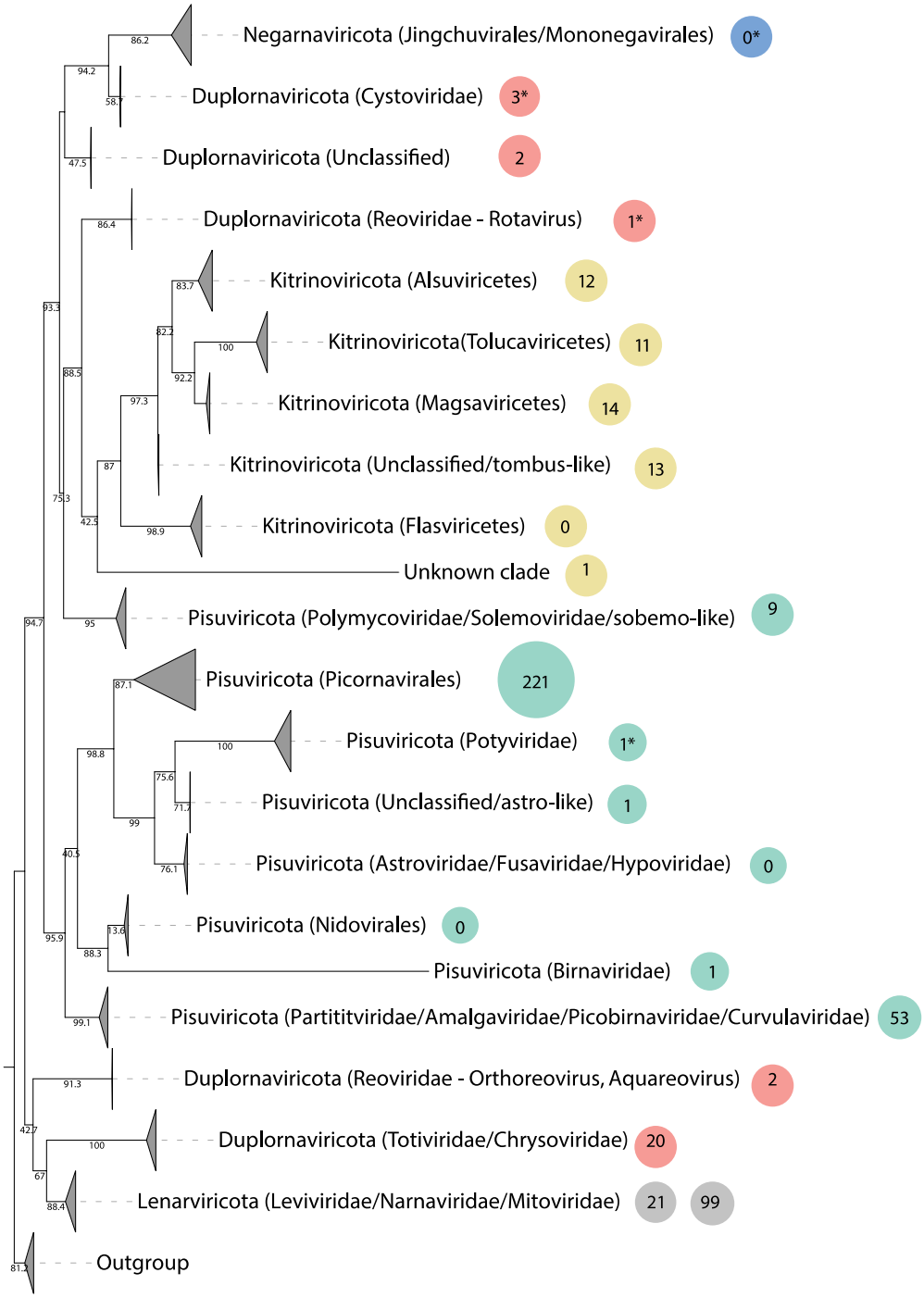


Figure 4.3. Maximum likelihood phylogenetic tree of viral RNA-dependent RNA polymerase showing phylogenetic position of the RdRp detected in the Johor Strait metatranscriptomes (in bubbles). Major RNA viral phyla: *Lenarviricota* (grey), *Pisuviricota* (green), *Kitrinoviricota* (yellow), *Duplornaviricota* (coral) and *Negarnaviricota* (blue). *Artverviricota* and *Uroviricota* were used as an outgroup. The orders/families and unclassified lineages falling under the specific branch are indicated in the brackets. The number in the bubble indicates the number of recovered transcripts per taxonomic group. Asterisk* denotes that an orphan RdRp (n=1-3) sequence was collapsed with the mentioned clade. Bootstrap values are shown. The size of the triangle is proportional to number of sequences that were used to build the tree.

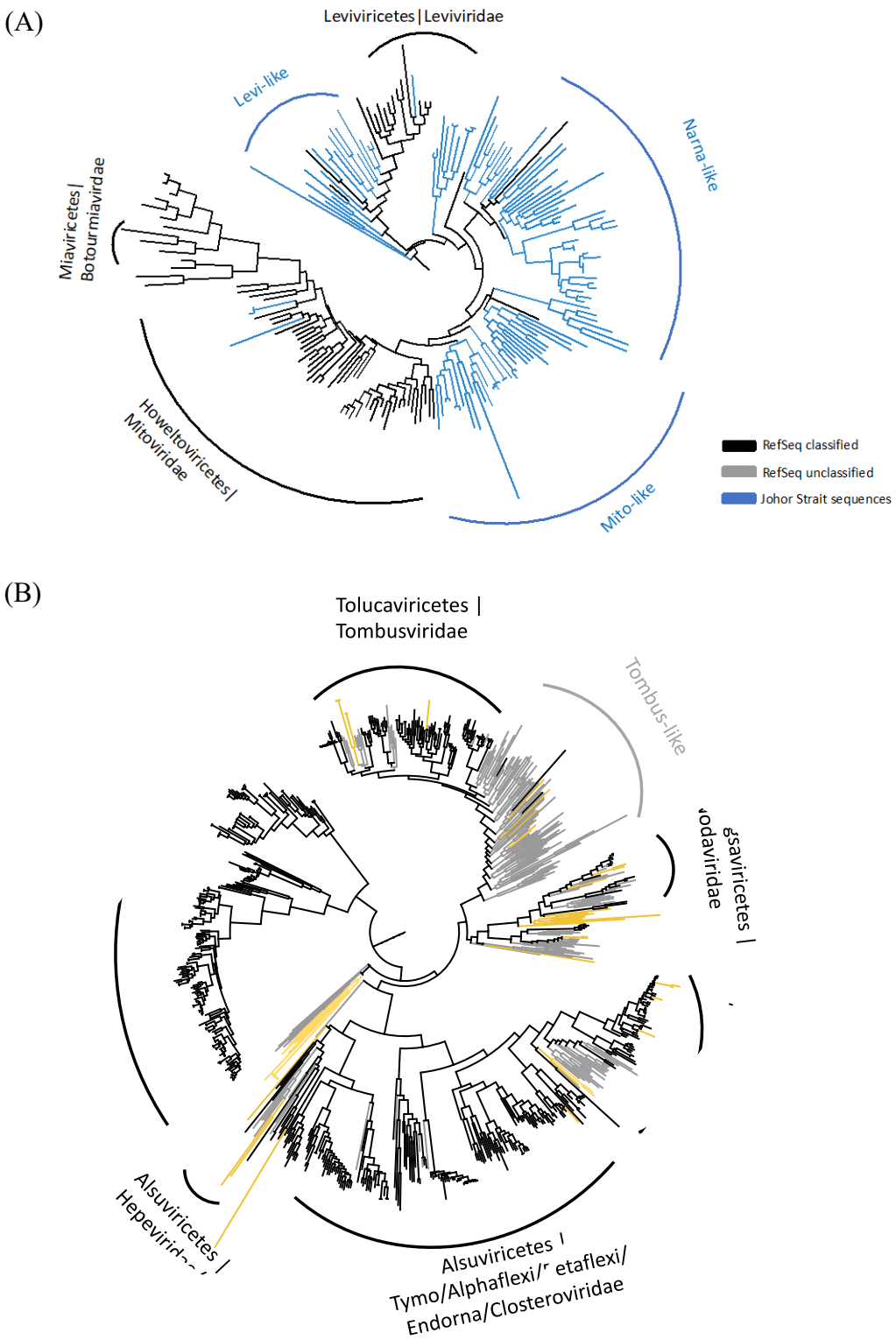


Figure 4.4. Maximum likelihood phylogenetic tree of RNA-dependent RNA polymerase showing the phylogenetic placement of the the sequences recovered from the Johor Strait metatranscriptomes within the phylum *Lenarviricota* (A) and *Kitrinoviricota* (B). Grey branches represent taxonomically unclassified viruses from the Ref Seq database. Black branches represent virus sequences from the RefSeq that have been formally classified by the ICTV and their classification is shown on two taxonomic levels (class | family). Newly recovered sequences within the phylum *Lenarviricota* are shown in blue and newly recovered sequence within the phylum *Kitrinoviricota* are shown in yellow.

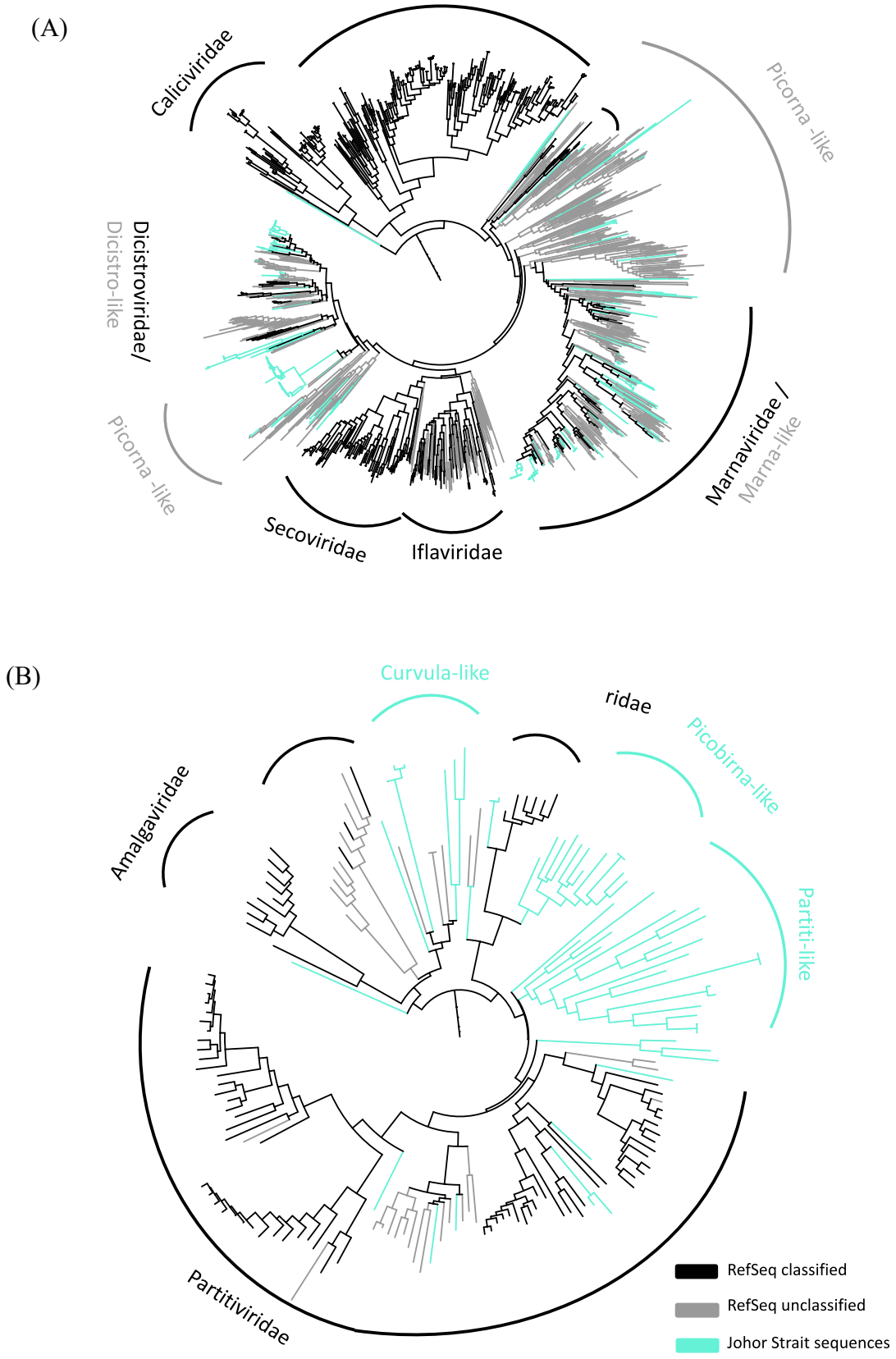


Figure 4.5. Maximum likelihood phylogenetic tree of RNA-dependent RNA polymerase showing the phylogenetic placement of the RdRp sequences recovered from the Johor Strait metatranscriptomes within the phylum *Pisuviricota:*(A) order *Picornavirales* (B) order *Durnavirales.* Grey branches represent taxonomically unclassified viruses from the Ref Seq database. Black branches represent virus sequences from the RefSeq that have been formally classified by the ICTV and their family-level classification is shown. Newly recovered sequences within the phylum *Pisuviricota* are shown in turquoise.

4.4.2.2. Phylum *Pisuviricota*

The largest number of recovered RdRp (n=286) was assigned to the phylum *Pisuviricota* (Figure 4.3). Within *Pisuviricota*, the order *Picornavirales* hosted the majority of recovered RdRp sequences (n=221). Animal nidoviruses *(*order *Nidovirales)* and plant potyviruses (order *Patatavirales*) were absent from the Johor Strait metatranscriptome. Families *Solemoviridae* and *Polymycoviridae*, including nine sobemo-like RdRp sequences recovered from the Johor Strait metatranscriptome formed a separate clade from the rest of viruses classified as *Pisuviricota* (Figure 4.3). Within the order *Picornavirales*, recovered RdRp sequences clustered with the families *Marnaviridae* (n=83) and *Dicistroviridae* (n=53) or formed clusters with unclassified picorna-like sequences (n=85) (Figure 4.5). The majority of marna-like sequences grouped within the genus *Sogarnavirus* (n=45), including the most expressed transcripts detected during the time-series (Figure A4.5,Figure A4.6). No representative of animal virus families like *Caliciviridae*, *Picornaviridae or Iflaviridae* or plant *Secoviridae* were detected either (Figure 4.5A). Order *Durnavirales*, that hosts double-stranded viruses of the phylum *Pisuviricota*, accommodated 53 viral RdRp, encompassing curvula-like (n=7) picobirna-like viruses (n=14), partiti-like viruses (n=30) and one representative of both birna-like and amalga-like viruses. Both picobirna-like and partiti-like sequences from the Johor Strait formed novel clusters (Figure 4.5B).

4.4.2.3 Phylum *Kitrinoviricota*

A total of 50 RdRp sequences from the Johor Strait metatranscriptome was assigned to the phylum *Kitrinoviricota* (Figure 4.3., Figure 4.4B). The class *Tolucaviricetes*, encompassing family *Tombusviridae* as well as unclassified tombus-like sequences, hosted 11 of Johor Strait RdRp sequences. Within the class *Magsaviricetes*, a cluster large of noda-like sequences (n=10) formed a sister clade to *Nodaviridae*. Within the

class *Alsuviricetes*, bromo-like and virga-like sequences (n=12), as well as hepe-like (n=4) were detected. A large number of Johor Strait sequences (n=13) formed a separate cluster with a number of unclassified *Kitrinoviricota* sequences. No RNA viruses with similarity to class *Flaviviricetes* were detected.

4.4.2.4 Phylum *Duplornaviricota*

The viruses classified under the phylum *Duplornaviricota* formed 3 distinct clusters in our analysis: vertebrate-infecting class *Resentoviricetes*, protist and fungi-infecting class *Chrysmotiviricetes* and bacteria-infecting class *Vidaverviricetes.* A total of twenty RdRp sequences resembling toti-like viruses clustered within class *Chrysmotiviricetes* that encompasses families *Totiviridae, Megabirnaviridae, Chrysoviridae* as well as unclassified toti-like viruses (Figure 4.3). Three RdRp sequences formed a sister clade to double-stranded RNA bacteriophages from the family *Cystoviridae* (class *Vidaverviricetes*) but were clustered together in this analysis (Figure 4.3) . Likewise, 2 "orphan" viral RdRp detected in our dataset that did not fall into any known clade were collapsed within the phylogenetically closest clade and not analysed further due to insufficient information (Figure 4.3, denoted with asterisk). No negative sense single-stranded RNA from the phylum *Negarnaviricota* were detected.

4.4.3. Phylogenetic diversity of giant DNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes

I further analysed the microbial metatranscriptomes for presence of active ongoing infection with giant DNA viruses and detected total of 48 confirmed nucleocytoplasmic large dsDNA viruses (NCLDVs) of the phylum *Nucleocytoviricota* (Figure 4.6, Table 4.1). All of giant DNA viral transcripts detected in the Johor Strait metatranscriptome was placed within 2 orders, *Imiteravirales* (n=24) and *Algavirales* (n=24). Phylogenetic analysis by Schulz et al. (2020) split the phylum *Nucleocytoviricota* into 10 superclades

(SC1-10). Based on this division, our sequences were placed within 3 clades, SC3, SC9 and SC10. SC3 clades includes prasinoviruses, chloroviruses, raphidoviruses and other viruses with typically algal hosts, while clades SC9 and SC10 have both phytoplankton and heterotrophic microeukaryotes as hosts (Schulz et al., 2020). Almost half of the giant DNA viruses detected was identified as prasinoviruses or unclassified prasinovirus-like genomes from the SC3 clade. *Tetraselmisviruses* and unclassified genomes from clade SC10, and *Chrysochromulinaviruses* and unclassified genomes from the clade SC9 counted to the remaining giant DNA viruses in our dataset. Potential virus-host associations will be discussed in Chapter 5. As expected, no fish or marine invertebrate associated giant DNA viruses like *Poxviridae*, *Iridoviridae* or *Asfarviridae* were detected since our experimental design captures viruses actively replicating in the cellular fraction 150µm-0.22µm.

Table 4.1. Taxonomic placement of giant DNA viral transcripts detected in Johor Strait metatranscriptomes. All the identified giant DNA viral transcripts were classified within class *Megaviricetes* and within two orders*: Imitervirales* and *Algavirales.*

Order	Family	SeqID (Giant DNA viral transcripts)	SeqID (Giant DNA virus database)
Algavirales	Prasinoviruses (SC3 superclade)	n813	GVMAG-M-3300001460-12.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2719	GVMAG-S-ERX556003-78.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2271	GVMAG-S-1074330-31.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2300	GVMAG-S-1092455-40.fna
Algavirales	Prasinoviruses (SC3 superclade)	n1933	GVMAG-M-3300027206-1.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2654	GVMAG-S-3300013195-8.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2218	GVMAG-S-1037383-112.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2275	GVMAG-S-1074346-73.fna
Algavirales	Prasinoviruses (AG-1)	n89	ERX552295.44.dc.fa
Algavirales	Micromonas pusilla virus 12T strain=12T (AG-01)	n741	GCF_000906035.1_ViralProj195482 _genomic.fna
Algavirales	Ostreococcus tauri virus RT-2011 OtV6 (AG-01)	n247	GCA_003051685.1_ASM305168v1_ genomic.fna
Algavirales	Micromonas sp. RCC1109 virus MpV1 (AG-01)	n737	GCF_000890375.1_ViralProj61013_ genomic.fna
Algavirales	Late Phycodnaviridae (Late Phyco clade 1)	n166	ERX556003.9.dc.fa
Algavirales	Prasinoviruses (SC3 superclade)	n2297	GVMAG-S-1092455-30.fna
Algavirales	Ostreococcus lucimarinus virus 2 Olv2 (AG-01)	n758	GCF_001399285.1_ViralProj298920 _genomic.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2298	GVMAG-S-1092455-32.fna
Algavirales	Ostreococcus lucimarinus virus 1	n734	GCF_000888835.1_ViralProj61011_ genomic.fna
Algavirales	Ostreococcus mediterraneus virus 1 OmV1 (AG-01)	n757	GCF_001399265.1_ViralProj298990 _genomic.fna
Algavirales	Prasinovirus (SC3 superclade)	n928	GVMAG-M-3300008012-2.fna
Algavirales	Prasinoviruses (AG-1)	n214	ERX633348.16.dc.fa
Algavirales	Prasinoviruses (SC3 superclade)	n2213	GVMAG-S-1037371-165.fna
Algavirales	Prasinoviruses (SC3 superclade)	n2555	GVMAG-S-3300010299-67.fna
Algavirales	Prasinovirus (SC3 superclade)	n2737	GVMAG-S-ERX556034-54.fna
Imitervirales	Chrysomulinaviruses (IM-01 clade)	n70	ERX552270.86.dc.fa
Imitervirales	MGVL67 (SC10 superclade)	n1510	GVMAG-M-3300021961-10.fna
Imitervirales	MGVL62 (SC10 superclade)	n1301	GVMAG-M-3300017956-2.fna
Imitervirales	MGVL62 (SC10 superclade)	n2241	GVMAG-S-1055481-12.fna
Imitervirales	MGVL62 (SC10 superclade)	n2212	GVMAG-S-1037362-78.fna
Imitervirales	MGLV42 (SC9 superclade)	n1191	GVMAG-M-3300012936-4.fna
Imitervirales	MGVL58 (SC10 superclade)	n867	GVMAG-M-3300005613-6.fna
Imitervirales	Tetraselmisviruses (IM-12)	n30	ERX552261.23.dc.fa

Imitervirales	MGVL42 (SC9 superclade)	n1188	GVMAG-M-3300012936-34.fna
Imitervirales	Chrysomulinaviruses (SC9 superclade)	n1856	GVMAG-M-3300024294-5.fna
Imitervirales	MGVL62 (SC10 superclade)	n1294	GVMAG-M-3300017824-14.fna
Imitervirales	Tetraselmisviruses (clade IM-12)	n1897	GVMAG-M-3300025695-2.fna
Imitervirales	Chrysomulinaviruses (SC9 superclade)	n1822	GVMAG-M-3300023210-19.fna
Imitervirales	Chrysomulinaviruses (SC9 superclade)	n1912	GVMAG-M-3300025849-19.fna
Imitervirales	MGVL57 (SC9 superclade)	n2554	GVMAG-S-3300010299-45.fna
Imitervirales	MGVL57 (SC9 superclade)	n1850	GVMAG-M-3300024261-8.fna
Imitervirales	Chrysomulinaviruses (IM-01 clade)	n2839	SRX802143.26.dc.fa
Imitervirales	MGVL62 (SC10 superclade)	n1302	GVMAG-M-3300017956-25.fna
Imitervirales	Mimiviridae (Mimiviridae clade 1)	n2828	SRX802077.122.dc.fa
Imitervirales	MGVL62 (SC10 superclade)	n903	GVMAG-M-3300007236-6.fna
Imitervirales	MGVL57 (SC9 superclade)	n1297	GVMAG-M-3300017950-28.fna
Imitervirales	Tetraselmisviruses (SC10 superclade)	n1311	GVMAG-M-3300017967-22.fna
Imitervirales	Tetraselmisviruses (SC10 superclade)	n2757	GVMAG-S-ERX556109-29.fna
Imitervirales	MGVL62 (SC10 superclade)	n2475	GVMAG-S-3300001941-2.fna
Imitervirales	MGVL57 (SC9 superclade)	n1343	GVMAG-M-3300018420-30.fna

4.5. Discussion

4.5.1. Eukaryotic phytoplankton community composition in the Johor Strait

Nutrient pulses have been shown to have disruptive effects on phytoplankton communities, triggering sporadic phytoplankton blooms in the Johor Strait (Kok and Leong, 2019; Lim et al., 2014; Trottet et al., 2021). Proliferation of phytoplankton species capable of high biomass growth or toxin production harmful algal blooms (HABs) can harm marine life by poisoning fish and shellfish, clogging fish gills or causing oxygen depletion (Trottet et al., 2021). Several molecular barcoding or microscopy-based studies investigated the diversity of phytoplankton in the Johor Strait. These studies reported dominance of centric diatoms like *Cerataulina*, *Thalassiosira, Rhizolenia* and *Cheatoceros* and dinoflagellates including *Gyrodinium*, *Gonyaulax, Amphodinium* and *Alexandrium* (Chénard et al., 2019; Gin et al., 2000; Hii et al., 2021; Kok & Leong, 2019; Mohd-Din et al., 2022). Toxic bloom-forming diatoms *Pseudo-nitzchia*, dinoflagellates *Alexandrium* and *Karenia,* and raphidophytes *Heterosigma* and *Fibrocapsa* are widely distributed and abundant in the Johor Strait (Hii et al., 2021; Mohd-Din et al., 2022). Consistent with previous microscopy-based reports of phytoplankton diversity in the Johor Strait reporting high absolute abundances of *Chaetoceros*, *Skeletonema, Rhizolenia* and *Thalassiosira* (Mohd-Din et al., 2020, 2022), our dataset reports the highest relative abundance of the same diatom taxa. In the metabarcoding study performed in the Johor Strait, the dinoflagellate sequences dominated the relative abundances, particularly the *Dinophyceae – Gyrodinium* and *Woloszynskia* (Chénard et al., 2019). In contrast, in my metagenomic dataset dinoflagellates had low relative abundance throughout the sampling period. Species with high 18S rDNA copy number like dinoflagellates can to be overrepresented in metabarcoding studies based on 18S amplification (Zhu et al., 2005) which could

explain the difference observed between metabarcoding and metagenomic studies. Unexpectedly, I detected high relative abundances of prasinophytes like *Ostreococcus* and *Micromonas* during early sampling period, while they were absent from both metabarcoding and microscopy dataset collected previously in the Johor Strait (Chénard et al., 2019; Mohd-Din et al., 2022). One plausible explanation could be that small naked flagellates like prasinophytes may be overlooked by microscopy.

4.5.2. Phylogenetic diversity of eukaryote-infecting RNA viruses in the Johor Strait metatranscriptomes

4.5.2.1. Phylum *Lenarviricota*

Phylum *Lenarviricota* encompasses 4 classes of RNA viruses*:* (+) ssRNA bacteriophages (class *Leviviricetes*), capsidless eukaryote-infecting families *Narnaviridae* and *Mitoviridae* (class *Amabiliviricetes* and *Howeltoviricetes*, respectively) and capsidated eukaryote-infecting family *Botourmiaviridae* (class *Miaviricetes*) (Callanan et al., 2021; Neri et al., 2022; Sadiq et al., 2022). Recently, single-stranded RNA bacteriophages have been reported in marine environment (Wolf et al., 2020). Soil metatranscriptomes are rich in capsidless viruses (Starr et al., 2019), but these viruses have not been reported in marine metatranscriptomes (Kolundžija et al., 2022). High number of RdRp sequences resembling mitoviruses and narnaviruses in our samples may be related to technical factors such as library preparation (Gann et al., 2021) or very high sequencing depth of my samples or to environmental factors such as host diversity and abundance.

4.5.2.2. Phylum *Pisuviricota*

The aquatic picorna-like RdRp sequences, comprising of marna-like, dicistro-like and picorna-like sequences, were most abundant taxa detected in the Johor Strait metatranscriptome, consistent with their reported abundance in RNA viromes (<0.22µm

fraction) and metatranscriptomes of productive coastal areas around the world (Culley, 2018; Miranda et al., 2016; Moniruzzaman et al., 2017; Vlok et al., 2019; Wolf et al., 2020). Diatom-infecting RNA viruses from the family *Marnaviridae* have been implicated in the dynamics of Antarctic diatom blooms (Alarcón-Schumacher et al., 2019; Miranda et al., 2016), An overwhelming amount of unclassified picorna-like viruses dominated the metatranscriptome in the aftermath of harmful brown tide bloom (Moniruzzaman et al., 2017). This suggests that marna-like and picorna-like viruses detected may also have similar ecological roles in the diatom-rich ecosystem of the Johor Strait. Large number of dicistro-like viruses present implies that the Johor Strait hosts high diversity of invertebrate viruses, some of which could be dangerous pathogen like Taura syndrome virus (Bujarski et al., 2019; Eynde et al., 2020). Within *Pisuviricota*, double-stranded order *Durnavirales* accommodated RdRp sequences resembling *Picobirnaviridae* and *Partitiviridae.* While occasionally reported in marine aquatic samples (Pound et al., 2020; Urayama et al., 2018) *Picobirnaviruses* were mainly detected in stool and wastewater samples and presumed to infect animals (Adriaenssens et al., 2018; Delmas et al., 2019b). New lines of evidence suggest that some *Picobirnaviridae*, alongside (+) ssRNA *Leviviridae* and dsRNA *Cystoviridae* also can infect prokaryotes. In fact, the true host of picobirnaviruses may be the microorganisms associated with animals, not animals themselves (Krishnamurthy and Wang, 2018; Neri et al., 2022). One birna-like RdRP sequence that was present in the dataset had most similarity to aquabirnaviruses, whose natural host are salmonid fish (Delmas et al., 2019a).

4.5.2.3. Phylum *Kitrinoviricota*

Phylum *Kitrinoviricota* hosts (+) ssRNA viruses with broad range of host including terrestrial fungi, plants and animals as well as microbial eukaryotes as natural hosts (Shi

et al., 2016; Wolf et al., 2018). Nodaviruses (class Magsaviricetes) infect insects, crustaceans and fish (Hameed et al., 2019; Shi et al., 2016; Zhang et al., 2017). A recent study (viral fraction, <0.22μm) in the Yanghan harbour reported high diversity of noda-like viruses, grouping with oomycete-infecting viruses (Wolf et al., 2020). Noda-like cluster of the Johor Strait of the Johor Strait metatranscriptome was phylogenetically more similar to nematode-infecting nodaviruses, suggesting that may have animal host as well.The Johor Strait hepe-like sequences most closely resembled hepe-like viruses detected in viromes of marine invertebrates (Shi et al., 2016), suggesting that they may be pathogens of crustacean zooplankton (Dong et al., 2020). The absence of the Flavivirus supergroup (class *Flasuviricetes)* and negative sense single-stranded RNA of the phylum *Negarnaviricota* is in agreement with previously published work (Wolf et al., 2020)

4.5.2.4. Phylum *Duplornaviricota*

In previous published analyses, all of the phylum *Duplornaviricota* branched off the (+) single stranded RNA viruses (Wolf et al., 2018, 2020). Our analysis showed that the three classes (*Chrysmotiviricetes, Resentoviricetes, Videviricetes)* evolved independently and likely represent 3 different phyla. Similar tree topology and polyphyletic origin of *Duplornaviricota* was suggested by Zayed et al. (2022) in the recent Science paper. Hence, current 5 phyla taxonomic division will likely need to be expanded.

Within the *Duploviricota*, 2 of the unknown "orphan" RdRp sequences from the Johor Strait metatranscriptome clustered next to the family *Reoviridae*, class *Resentoviricetes*, containing rotaviruses, aquatic orthoreoviruses and aquareovires that infect fish and shell fish (Blindheim et al., 2015; Kibenge, 2019; Polinski et al., 2020) and likely represent a sister clade. Class *Chrysmotiviricetes* that encompasses families *Totiviridae,*

unclassified toti-like viruses and *Chrysoviridae*, and are known to have fungal and protistan host (Charon et al., 2020, 2021; Urayama et al., 2016). Almost half of recovered toti-like sequence clustered very closely to toti-like diatom colony associated viruses (Charon et al., 2021; Urayama et al., 2016), suggesting that the hosts of these viruses may be diatoms. Totiviruses can also infect heterotrophic protist like *Leishmania* and *Giardia*, crustaceans and fish (Sandlund et al., 2021; Sunarto and Naim, 2016). A few of detected toti-like sequences resembled fish viruses like piscine myocarditis virus (Sandlund et al., 2021).

The observations of viruses associated with animal diseases or belonging to native viromes of fish, crustaceans, molluscs and other invertebrates (noda-like, hepe-like, reo-like, birna-like, dicistro-like picorna-like and toti-like viruses described above) highlights a potential reservoir of aquaculture pathogens in the waters of the Johor Strait. Since the sampled cellular fraction was prefiltered to eliminate the mesozooplankton (>200μm), these viruses could be originating from viruses attached to particles or animal detritus.

4.5.3. Phylogenetic diversity of giant DNA viruses infecting eukaryotes in the Johor Strait metatranscriptomes

Nucleocytoplasmic large DNA viruses from the orders *Imitervirales* and *Algavirales* are tightly associated with diverse microbial eukaryotes in the marine ecosystem (Endo et al., 2020; Hingamp et al., 2013; Xia et al., 2021). In our dataset, giant DNA viral genomes of unclassified and cultured prasinoviruses (order *Algavirales*) were the most diverse group of giant DNA viruses detected. Prasinoviruses are the smallest giant DNA viruses and infect picoplankton prasinophytes *Ostreococcus* and *Micromonas* (Coy et al., 2018; Weynberg et al., 2017). Prasinoviruses have never been reported in other

metatranscriptome studies (Ha et al., 2021; Moniruzzaman et al., 2017). I attribute this to experimental design: I collected the 150μm - 0.22μm fraction, while other studies used a fractionated approach and analysed only the fraction larger than 5μm, filtering out the picoeukaryotic phytoplankton.

The remaining detected diversity of giant DNA viruses grouped within the clades of unclassified giant DNA viruses as well as Chrysochromulinaviruses and Tetraselmisviruses from the order *Imitervirales,* that are known to infect phytoplankton and other microbial eukaryotes (Aylward et al., 2021; Ha et al., 2021; Sandaa et al., 2022). Greater diversity of giant DNA viruses might be recovered from the DNA metagenomes of the same cellular fraction, as large giant DNA virus particles are often retained by the 0.22μm filter (Kolundžija et al., 2022).

4.5.4. Host ranges and infection strategies of phytoplankton viruses.

Both RNA and giant DNA viruses infecting phytoplankton are generally considered to have a high host specificity and lytic lifestyles (Brussaard, 2004b; A. Culley, 2018; Nagasaki & Bratbak, 2010). The burst sizes of lytic DNA and RNA phytoplankton viruses can differ by more than two orders of magnitude. Burst size of phytoplankton DNA viruses typically range between 300-1000 virus particles (Nagasaki & Bratbak, 2010; Sandaa et al., 2022) and for prasinoviruses can be as low as 25 virus particle (Derelle et al., 2018), while RNA viruses can reach burst sizes up to 6×10^4 (Lang et al., 2009; Takao et al., 2005). The difference in burst sizes can be partially explained by virus genome size as well as genome size of the host and usage of host resources (Edwards et al., 2021).

Recently, based on virus-host dynamics of cultured haptophytes, it was proposed that host density could have a major effect on the viral infection of phytoplankton in the ocean (Sandaa et al., 2022). Fast-growing, high-abundance phytoplankton ("opportunists") are accompanied with highly virulent viruses with acute lytic infection style that are transmitted horizontally. Low abundance phytoplankton ("gleaners") are accompanied with low virulence viruses that do not kill the host cell immediately. Instead, the virus particles are continuously released from a live cell for a longer period of time, and both horizontal and vertical transmission can occur (Sandaa et al., 2022). Few scarce reports of the latter infection style in the marine environment include the coexistence of haptophytes and haptophyte-infecting viruses through different seasons (Johannessen et al., 2017). Similarly, continuous expression of some giant DNA and occasional "bloom and bust" peaks of other giant DNA were observed during a brown tide bloom (Moniruzzaman et al., 2017). Potential hosts of some viruses in the Johor Strait dataset, specifically Chrysochromulinaviruses and Tetraselmisviruses, include low abundance, slow growing haptophytes and chlorophytes, that could follow the slow infection dynamics proposed by Sandaa et al. (2022).

The Johor Strait metatranscriptome included a diverse groups of both dsRNA and (+) ssRNA viruses with potentially very broad host ranges and alternative infection strategies. Positive-sense single-stranded (+) ssRNA viruses with acute lytic lifestyles are the dominant viruses reported in most marine environmental metatranscriptome studies (reviewed in Chapter 2, Kolundžija et al., 2022) and many similar viruses (e.g picorna-like, marna-like viruses) were present in the Johor Strait. However, his study also captured multiple (+) ssRNA and dsRNA viruses with persistent, non-lytic, lifestyles that are transmitted vertically, that are rarely reported in marine environmental studies. These include *Narnaviridae* and *Mitoviridae*, capsidless (+) RNA viruses from

the phylum *Lenarviricota* (Hillman & Cai, 2013) and capsidated dsRNA viruses families like *Partitiviridae, Amalgaviridae, Totiviridae* and *Chyrsoviridae*, known for their persistent lifestyles (Roossinck, 2019; Vainio et al., 2018). All these viral families were originally considered fungal and plant viruses (Ghabrial & Suzuki, 2009; Roossinck, 2019; Roossinck et al., 2015). Recently, similar viruses have been detected in metatranscriptomes of a natural diatom colony and cultured phytoplankton, expanding potential host ranges of these viral groups (Charon et al., 2020, 2021; Chiba et al., 2020; Urayama et al., 2016). In complex environmental samples it may not be possible to elucidate if the true hosts are phytoplankton or associated fungal parasites, like chytrid fungi and oomycetes. However, non-lytic infections like the ones suggested by my study appear to be prevalent in the marine ecosystem. These findings indicate that the widely accepted idea that marine RNA viruses are mostly lytic may require some re-evaluation.

4.6. Conclusions

The Johor Strait is a coastal ecosystem exposed to short-term nutrient pulses and economically important area for the surrounding countries. This is the first study that tackles the diversity of phytoplankton and their viruses in the Johor Strait using a combination of metagenomics and metatranscriptomics. The high abundance of fast-growing phytoplankton species with the potential to reach high cell concentrations suggests the recurrent likelihood of phytoplankton blooms in the area. The metatranscriptomic analysis of the seawater samples uncovered a multitude of phytoplankton DNA and RNA viruses, some belonging to previously unknown clades. The dynamic life-cycles of these viruses suggest that they could have a profound impact on the phytoplankton community structure. Additionally, the presence of RNA animal viruses in the samples underlines the value of metatranscriptomics as a surveillance tool

for identifying potentially harmful RNA viruses that could pose a threat to the marine ecosystem, including aquaculture and wildlife.

4.7. Acknowledgments

I am grateful to George Williams for the help in QGIS with Figure 4.1 and Figure A4.1. I am incredibly thankful to Dong-Qiang Cheng, Winona Wijaya and George Williams for their help with R and data visualization. I would like to acknowledge Robert Nichols Scott from the DHI group for providing the bathymetry and current data for the Johor Straits. Funding for this project has been provided by the Intra-CREATE Seed Collaboration Grant (Award NRF2018-ITS004-0001) and Competitive Research Programme (Award CRP21-2018-0005) awarded by the National Research Foundation (NRF) of Singapore.

APPENDIX 4

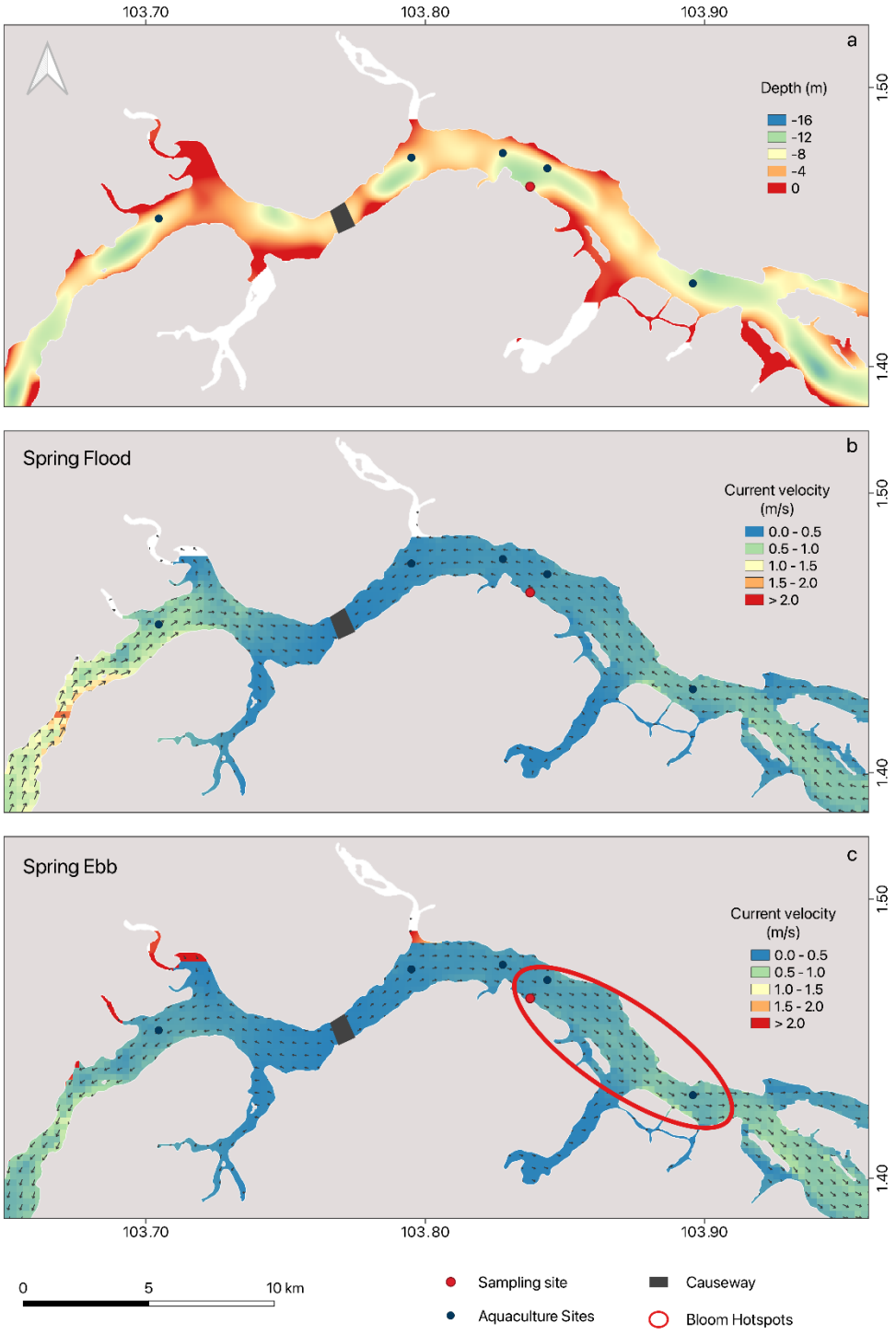


Figure A 4.1. Magnified view of the Johor Strait showing a) depth b) current velocity and direction during spring flood c) current velocity and direction during spring ebb. Sampling site is indicated with a red dot. The Johor-Singapore Causeway is shown with black rectangles. Aquaculture farms are shown in dark blue dots. The bloom hotspot area closest to the sampling location is shown only on figure c) for clarity.

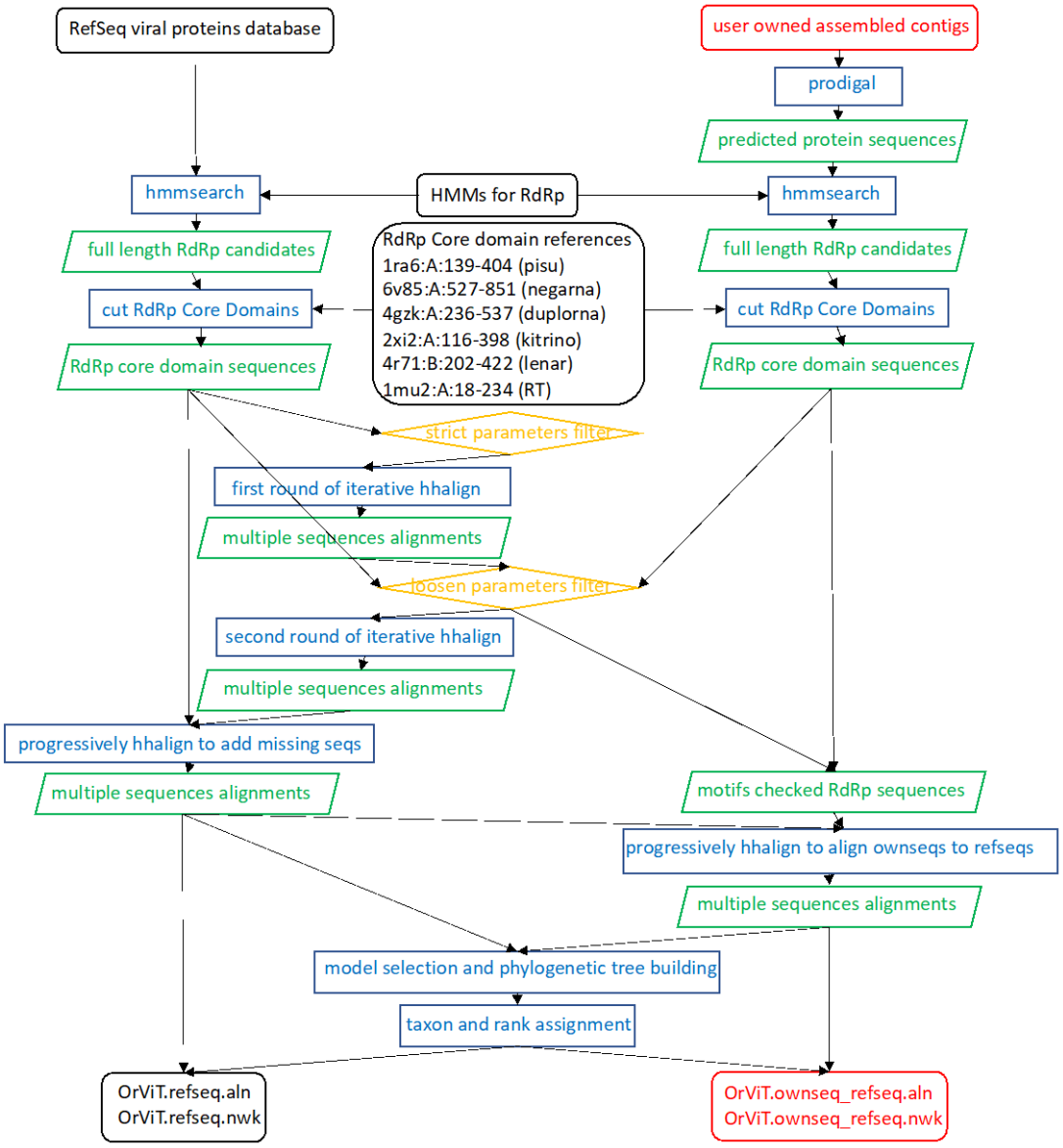


Figure A 4.2. Flowchart of the OrVit pipeline (Cheng et al., 2022)

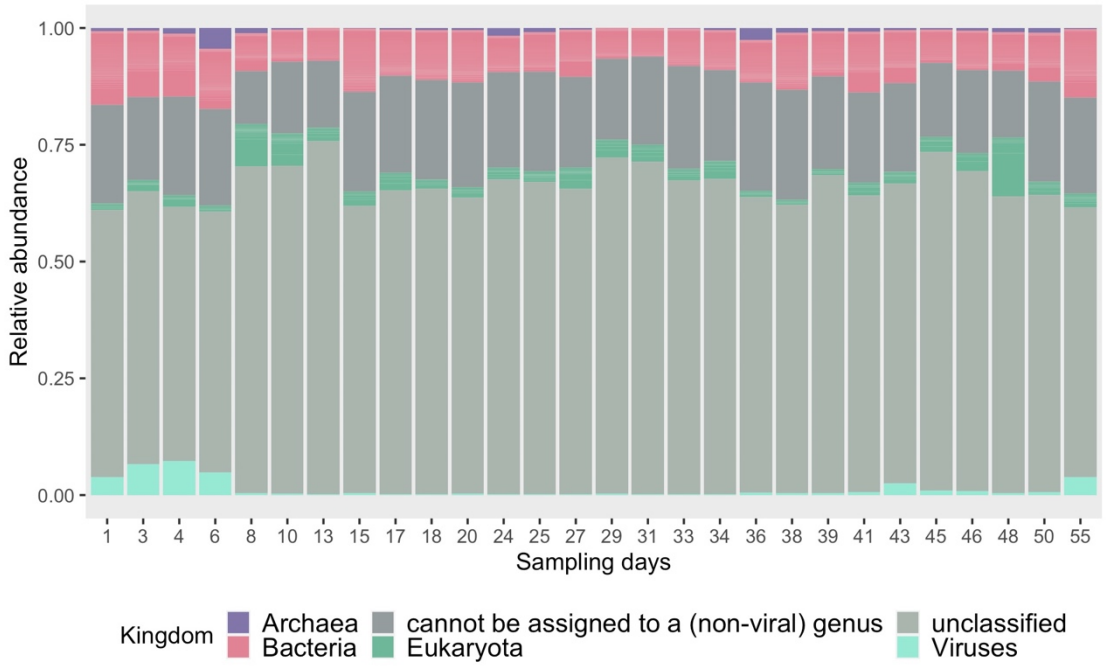


Figure A 4.3. Stacked bar plot showing the relative abundance of cellular organisms (*Archea*, *Bacteria* and *Eukaryota*) and viruses in the metagenomic reads in the Johor Strait.

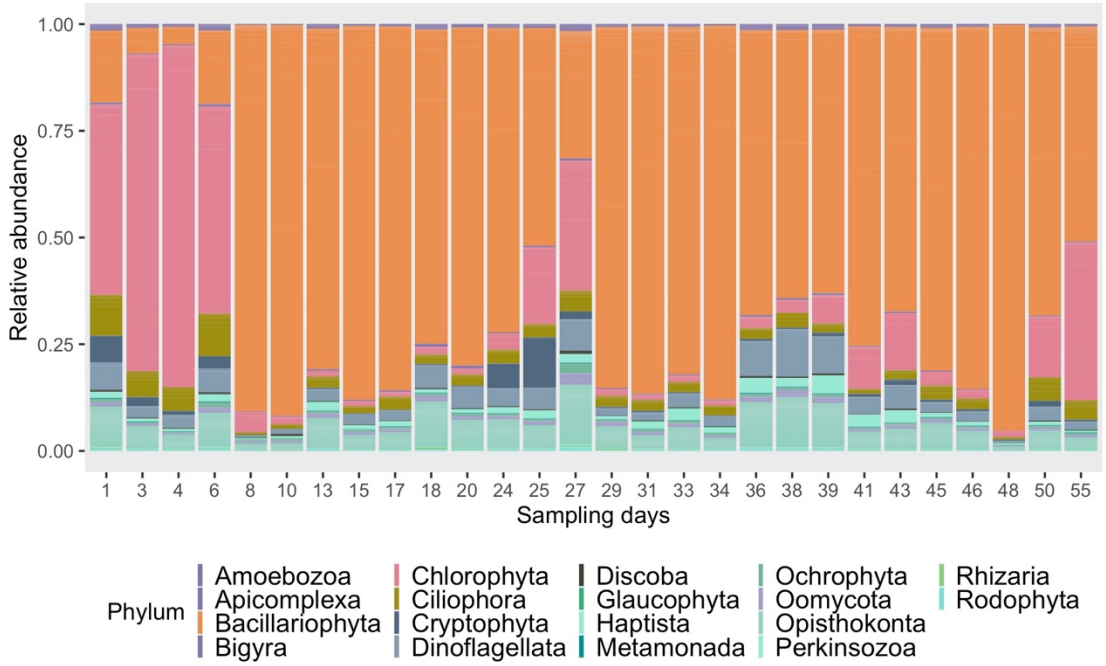


Figure A 4.4. Stacked bar plot showing the relative abundance eukaryotic taxa (phylum-level) in the Johor Strait. Chlorophytes are shown in pink, diatoms in orange, ciliates in olive green, dinoflagellate and cryptophytes in grey and dark grey. The heterotrophic phyla are shown in light blue palette.

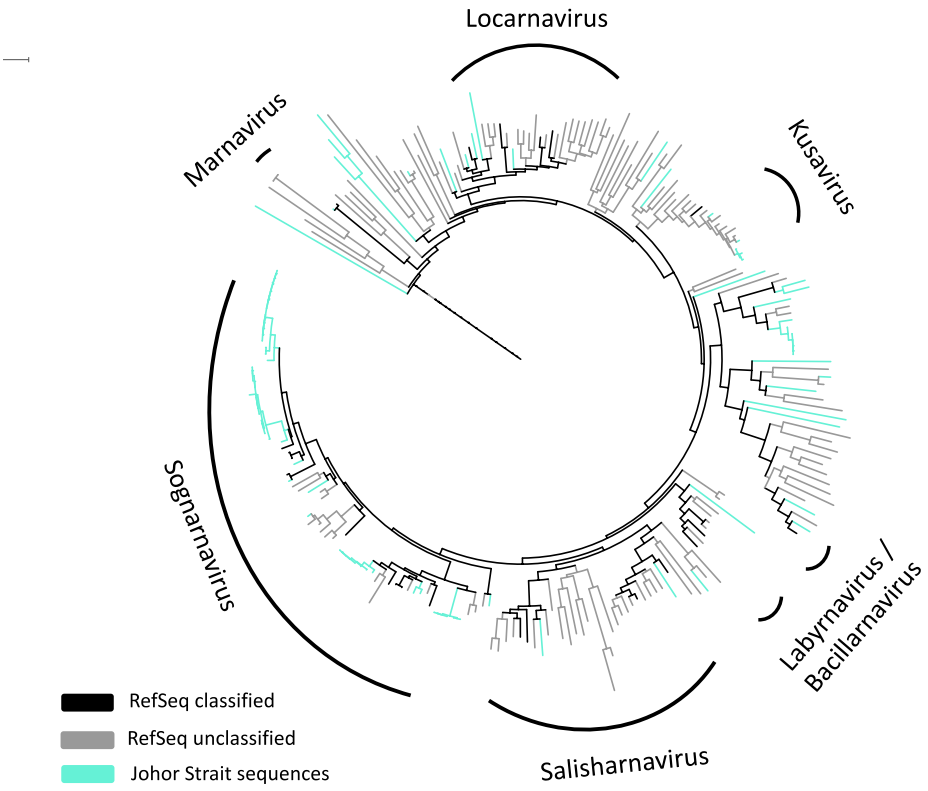


Figure A 4.5. Maximum likelihood phylogenetic tree of the family *Marnaviridae* (phylum *Pisuviricota* | order *Picornavirales*). Seven formally recognized by the ICTV genera are shown in black, unclassified marna-like sequenes from the RefSeq database in grey and the newly recovered Johor sequences highlighted in blue.

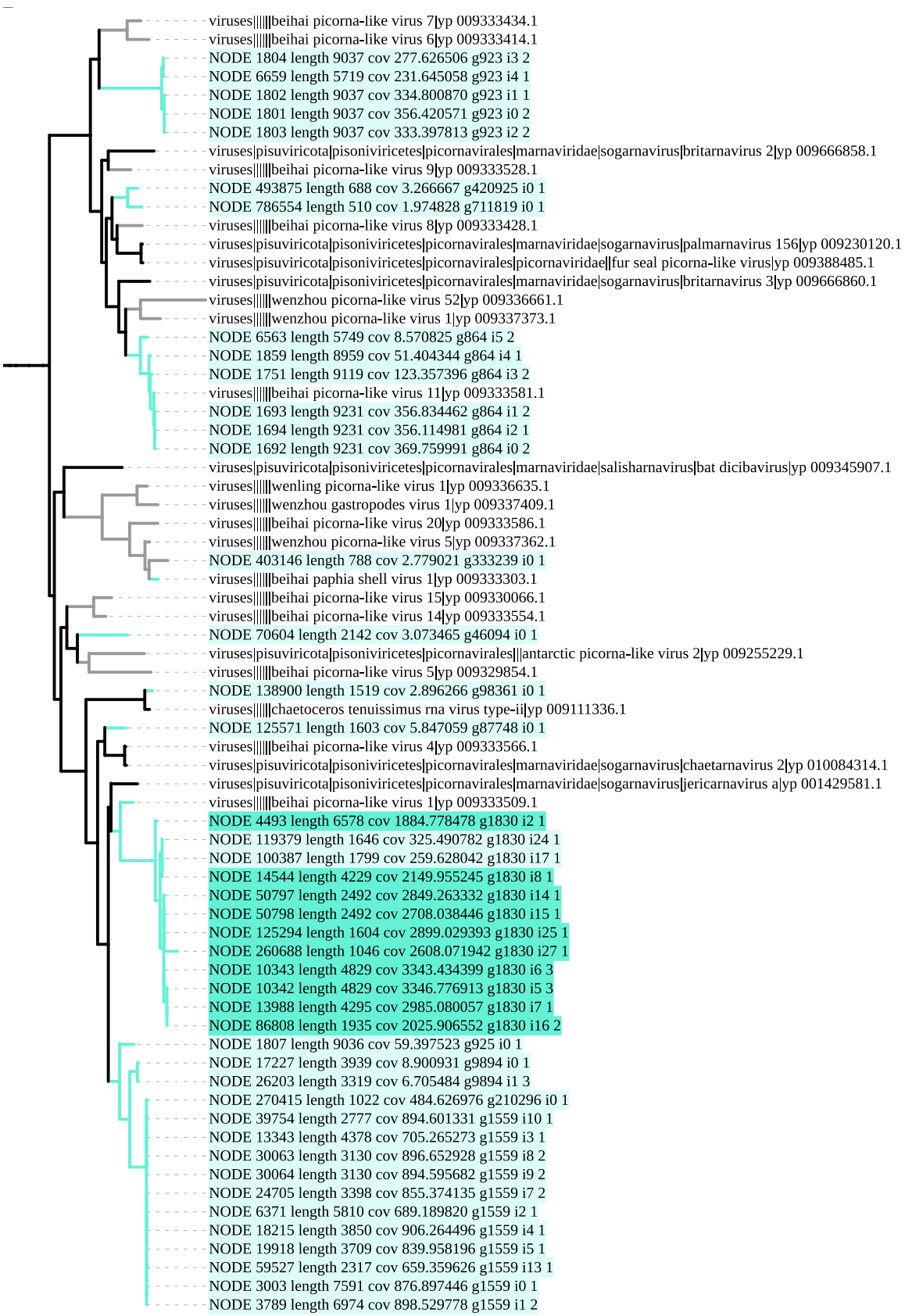


Figure A 4.6. *C*lose-up of the Marnaviridae phylogenetic tree. The dominant RNA viral groups in the Johor Strait metatranscriptome cluster within the genus Sogarnavirus (n=45). Sequences from the Johor Strait are highlighted in blue: sogarna-like contigs with low coverage in light blue, sogarna-like contigs with high coverage in bright blue.

CHAPTER 5 RNA AND GIANT DNA VIRUSES INFECTING EUKARYOTIC PHYTOPLANKTON IN EUTROPHIC COASTAL WATERS DISPLAY DISTINCT LIFE-CYCLE DYNAMICS

5.1. Abstract

The Strait of Johor is an enclosed equatorial coastal area heavily influenced by anthropogenic activities with recurrent phytoplankton blooms. During a 55-day time series, the Strait waters were consistently eutrophic, yet experienced extreme changes in phytoplankton biomass, reaching a near-bloom threshold on two occasions. The main environmental factors driving changes in eukaryotic phytoplankton community structure were concentrations of silicate, and phosphate and inorganic nitrogen, supporting the notion that diatoms as the dominant taxonomic group. RNA and giant DNA viruses infecting phytoplankton displayed distinct temporal dynamics, possibly reflecting differences in their life cycles. The majority of RNA viruses underwent extremely quick changes in transcriptional activity, indicative of acute "bloom and bust" infection dynamics. The transcriptional activity of most giant viruses displayed much lower variation throughout the sampling period, a trend that was consistent with the flow cytometric counts. With the exception of prasinoviruses, lytic infection of giant DNA viruses seems to be less acute, a phenomenon that I will refer to as "gleaner" infection dynamics. Three temporally separated "bloom and bust" cycles of marna-like virus populations were positively correlated with 3 most abundant diatoms during first chlorophyll-a maximum in nutrient-replete conditions. The second chlorophyll-a maximum, was accompanied with silica limitation, but did not exhibit a similar elevated marna-like viral expression, indicating that the diatom collapse was induced by silica

limitation rather than viral infection. Putative virus-host pairs identified may be an essential component of regulatory mechanisms that prevent frequent bloom development in the eutrophic Johor Strait coastal environment. This study underscores the potential of metatranscriptomic analysis to capture the full spectrum of viral life-cycles and their potential ecological consequences and emphasises the role of viruses in daily regulation of the ecosystem.

Key words: metatranscriptomics, marine RNA virus, giant viruses, virus-host dynamics, phytoplankton bloom

5.2. Introduction

While marine photosynthetic microbial eukaryotes (or phytoplankton) support half of the worldwide primary production and supply energy for marine food webs (Behrenfeld et al., 2006; Simon et al., 2009), the whole global phytoplankton population is consumed by grazers or eliminated by viruses every 2–6 days (Behrenfeld, 2014). Grazing transports carbon captured by the phytoplankton to higher trophic levels. Viral lysis of phytoplankton promotes the export of carbon to the deep ocean (a process known as viral shuttle) or its recycling in surface waters (a process known as viral shunt) (Kolundžija et al., 2022). The relative importance of the viral shunt and shuttle fluctuates geographically and temporally with phytoplankton composition and environmental factors, regulating the biological carbon pump (Kaneko et al., 2021). Therefore, studies of how phytoplankton diversity and dynamics are influenced by viruses are essential to understand oceanic carbon flows and their impact on global climate.

The full diversity of phytoplankton viruses and virus-phytoplankton interactions in the ecosystem cannot be adequately captured just by size-selected viromics, which normally concentrates either only on DNA or only on RNA viruses in the viral-size fraction.

Additionally, the large size of giant viruses excludes them from the viral-size fraction during the viral enrichment procedure (Kolundžija, Cheng, and Lauro 2022). In contrast to viromics, metranscriptomics enables simultaneous, unbiased detection of the mRNA of all viral genome types (Cobbin et al., 2021). Additionally, metatranscriptomics allows studying not only the immediate virus populations changes, but it can also identify subtler variations in viral stocks, given that the intracellular virus population changes more rapidly than the free virus population ("bank of viruses") (Aylward et al., 2017; Hevroni et al., 2020).

Marine viral metatranscriptomic studies almost always rely on weekly or monthly sampling, and given the short viral lifecycles, are unable to entirely capture the ecological dynamics of viruses and phytoplankton hosts during phytoplankton blooms (reviewed in Kolundžija et al., 2022). For example, two diatom bloom samples off the coast of the West Antarctic Peninsula collected a month apart revealed only that the transcriptional activity of diatom-infecting (+) ssRNA viruses increased during the bloom as compared to low-chlorophyll days (Alarcón-Schumacher et al., 2019). In another study, weekly sampling of *Aureococcus anophagefferens* brown tide bloom also revealed "bloom and bust" transcriptional dynamics of several giant viruses, including the virus infecting the dominant bloom species. At the final point of this bloom, the relative expression of (+) ssRNA picornaviruses dramatically increased, suggesting that other phytoplankton species present in the bloom succession are regulated by these viruses (Moniruzzaman et al., 2017). In an metatranscriptomic study that sampled every four hours, but only for a short period of 2.5 days, phytoplankton-infecting DNA and RNA viruses exhibited diel cycles of transcriptional activity (Kolody et al., 2019). The frequency of sampling thus clearly impacts the insight we gain into virus-host dynamics.

Acute viral infections of high abundance phytoplankton hosts ("opportunists") during the decline phase of the phytoplankton blooms in the above-mentioned studies are examples of bloom and bust dynamics; results which are fully consistent with the classic "Kill the Winner" model (Thingstad, 2000). On the contrary, it was suggested recently that low abundance phytoplankton hosts ("gleaners") and their viruses may not follow the same acute dynamics (Sandaa et al., 2022). The existence of "slow and steady" dynamics was observed during *Aureococcus anophagefferens* brown tide bloom, where some giant DNA viruses were continuously expressed at low levels (Moniruzzaman et al., 2017). A study of diel patterns revealed that RNA virus expression was negatively correlated with the abundance of their host while the expression of DNA viruses was positively correlated with their hosts (Kolody et al., 2019). A fine-scale, long-term monitoring study following the dynamics of hosts and viruses could lead to a better understanding of these complex processes.

Here, I present a 55-day high-resolution (sampling every two days) survey of diversity of phytoplankton-infecting RNA and giant DNA viruses and of the potential role of viral infection in structuring the phytoplankton community. This is a first monitoring campaign combining metatranscriptomics and metagenomics over an extended time and in non-bloom settings. With this approach I have gained insights into distinct temporal activity patterns, lifestyles and abundances of specific viral groups and even into virus-host interactions that keep fast-growing phytoplankton in check. Knowledge of the underlying mechanisms of phytoplankton community regulation and of factors that may contribute to or prevent bloom development is crucial for making ecologically and economically relevant decisions to prevent ecosystem deterioration as well as economic losses.

5.3. Materials and methods

5.3.1. Study site and physical, chemical and biological data collection

The study site, sampling frequency, environmental data and samples collected were described in details in Chapter 4. Briefly, the sampling was conducted at Sembawang for 2 months during the rainy season with bi-daily frequency (November -December 2020). Temperature, salinity and turbidity were measured on-site as described, and samples for biological and chemical analysis were collected and further processed in the lab. Microbial community samples were processed as described in Chapter 4. The chlorophyll-a concentration, concentrations of dissolved nutrients and concentrations of viruses ad prokaryotes were processed as described below.

5.3.2. Concentration of chlorophyll-a and dissolved inorganic nutrients

For chlorophyll-a concentration, unfiltered and 150μm-prefiltered seawater was filtered in duplicate through onto a sterile 25mm glass fibre filter (GF/F, Whatman), encased in aluminium foil and frozen at -80 °C until analysis. Chlorophyll-a was extracted from the GF/F filters by immersion in 90% acetone at 4 °C in the dark overnight. Samples were centrifuged at 500xG for 10 minutes. Chlorophyll-a fluorescence of the supernatant was measured with a Floromax 4 spectrophotometer (Horiba). For dissolved inorganic nutrient analysis, water samples were filtered through 0.22um Sterivex filter into a 15mL acid-washed collection tube in duplicate and stored at -20 °C until analysis. Silicate (Si), dissolved inorganic phosphate (PO_4), NO_x, and nitrite (NO_2) and ammonia (NH_4) concentrations were analysed on a SEAL AA3 High-resolution AutoAnalyser (Seal Analytical). Nitrate (NO_3) was expressed as $NO_x - NO_2$. Total dissolved inorganic nitrogen (DIN) was calculated as a sum of NO_x and ammonium concentrations.

5.3.2. Enumeration of viruses and prokaryotes with flow cytometry

Cellular fraction (<150μm) and viral fraction (<0.22μm) seawater samples for enumeration of viruses and prokaryotes were fixed with glutaraldehyde (EM-grade, Sigma-Aldrich) in final concentration 0.5% for 30 minutes in the dark, flash frozen in liquid nitrogen in the field and stored at -80 °C until analysis. Upon thawing, the samples were diluted 20x-50x with 0.02μm filtered TE buffer (10mM Tris, 1mM EDTA, pH 8). Viral fraction samples were stained at 80 °C for 10 minutes in the dark with Sybr Green I (Life Technologies) in 5 x 10^{-5} dilution of the original stain stock. Prokaryotic fraction samples were stained for 15 minutes in the dark at room temperature with Sybr Green I (Life Technologies) in 1 x 10^{-4} dilution of the original stock (Brussaard, 2004a; Brussaard & Payet, 2010). Viruses and prokaryotes were enumerated on a CytoFlex flow cytometer (Beckman Coulter), equipped with blue (488nm) and violet (405nm) laser, with trigger set on green fluorescence. The configuration of the filters was modified for the violet side scatter (VSSC) detection, enabling more sensitive detection of light scatter for smaller particles (Zhao et al., 2022). The samples were analysed with a slow flow rate (10μl/min) and abort rate was kept under 5%. Prokaryotic and phytoplankton viruses differ in flow cytometric signatures of fluorescence and side scatter and were split in 5 subpopulations (Vir1-Vir5) based on these characteristics (Figure A5.1, panels (A) and (B). Viral subpopulations Vir1-Vir2 had lower fluorescence and side scatter are mostly considered to be bacteriophages. Viral subpopulation Vir3 had higher fluorescence and contains both phytoplankton viruses and bacteriophages. Viral subpopulations Vir4 and Vir5 had higher side scatter and contain large phytoplankton viruses (Baudoux et al., 2006; Biggs et al., 2021; Brussaard & Payet, 2010; Jacquet et al., 2002). Blanks consisted of TE buffer or 0.02μm-filtered

seawater sample stained under same conditions and were subtracted from the samples to correct for the instrument's background noise (Figure A3.1, panel (C)).

5.3.4 Data analysis and visualization

To compare relative abundance of viruses between samples, metatranscriptome reads were mapped to the assembled contigs (for RNA viruses) or giant DNA viral genome database (for giant DNA viruses) with Bowtie2 (Langmead and Salzberg, 2012). Read mapping counts were normalized for sequencing depth and contig/genome length, and expressed as Reads Per Kilobase of Contig per Million reads. Beta diversity in. active viral community (metatranscriptome) was calculated with Bray-Curtis dissimilarity index and temporal patterns in beta diversity were visualized with non-metric multidimensional scaling (NMDS) ordination (monoMDS function, parameters (k=2, trymax=100)). Sampling period was split into 3 phases, early (low chlorophyll-a), middle (chlorophyll-a maximum 1), and late (chlorophyll-a maximum 2) which correspond to sampling days (10, 29) when the chlorophyll-concentrations reached the maxima ("almost-bloom events"). Permutational multivariate analysis of variance (PERMANOVA) and analysis of similarities (ANOSIM) were performed on Bray-Curtis dissimilarity matrices to test for temporal differences among the periods. Correlations between Bray-Curtis dissimilarity matrices and environmental variables were explored with canonical correlation analysis (CCA), envfit and Mantel tests.

5.4. Results

5.4.1. Fluctuations of the abiotic and biotic variables in the Johor Strait

Seawater temperature and salinity remained high and stable throughout the sampling period. The average temperature was 29.5°C, and the salinity averaged around 26, with two major dips at sampling day 3 and 55, decreasing to 16.7 and 21 (Table A.5.1). The

phytoplankton biomass, expressed as total chlorophyll-a concentration, fluctuated between 1.11µg/L up to 37.07µg/L, with an average concentration of 12.91 µg/L and a standard deviation of 9.67µg/L. Two total chlorophyll-a concentration maxima, on the sampling day 10 and 29, reached the "near-bloom conditions" defined as increase in total chlorophyll-a concentration >2 standard deviations from the average (Hii et al., 2021). Chlorophyll-a concentration of the filtered fraction (<150 µm) closely followed the total chlorophyll-a concentration (Pearson correlation coefficient, R=0.62) (Figure 5.1, Table A5.1).

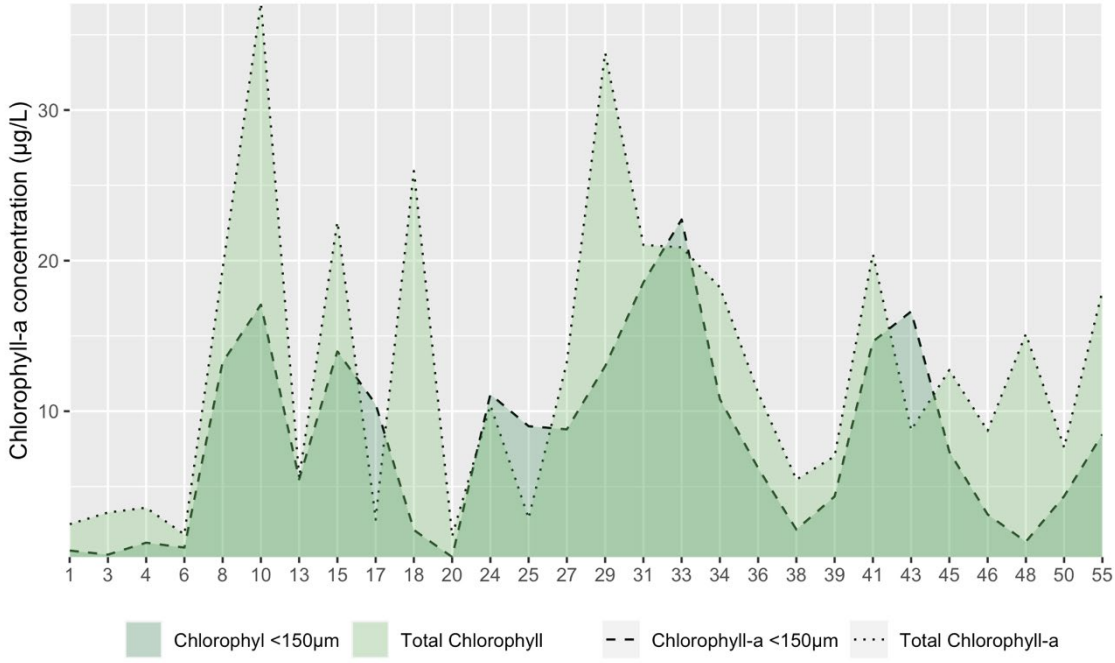


Figure 5.1. Fluctuations of the chlorophyll-a concentrations during the 55-day sampling period in the Johor Strait. The dotted lines represent the total chlorophyll-a concentrations and the dashed lines represent the chlorophyll-a concentration in 150µm-0.22µm fraction.

The Johor Strait was continuously eutrophic, with particularly high concentrations of nitrogen species. The concentrations of the dissolved nutrients ranged between 0.2µM to 25.6 µm for silicate, 0.29 µM to 2.4 µM for phosphate, 12.43 µM to 31.67 µM for NO_x (nitrate+nitrite), 1.04 µM to 49.84 µM for ammonia. Total dissolved inorganic

nitrogen (DIN) fluctuated between 13.5µM and 69.37 µM (Figure 5.3; Table A.5.2.). Total DIN, silicate and phosphate were negatively correlated with phytoplankton biomass expressed as filtered chlorophyll-a concentration (R=-0.46, $p<0.05$; R=-0.55, $p<0.01$; R=-0.62, $p<0.001$, respectively). Salinity was negatively correlated with concentrations of silicate (R=-0.54, $p<0.001$), phosphate (R=-0.24, $p<0.01$), ammonia (R=-0.63, $p<0.05$) and total DIN (R=- 0.58, $p<0.05$). It also negatively affected bacterial and viral abundances, with most prominent effect on Vir3 population (R=-0.67, $p<0.001$) (Figure A5.4). No significant autocorrelations between environmental variables were found as described in Wijaya et al. (2022).

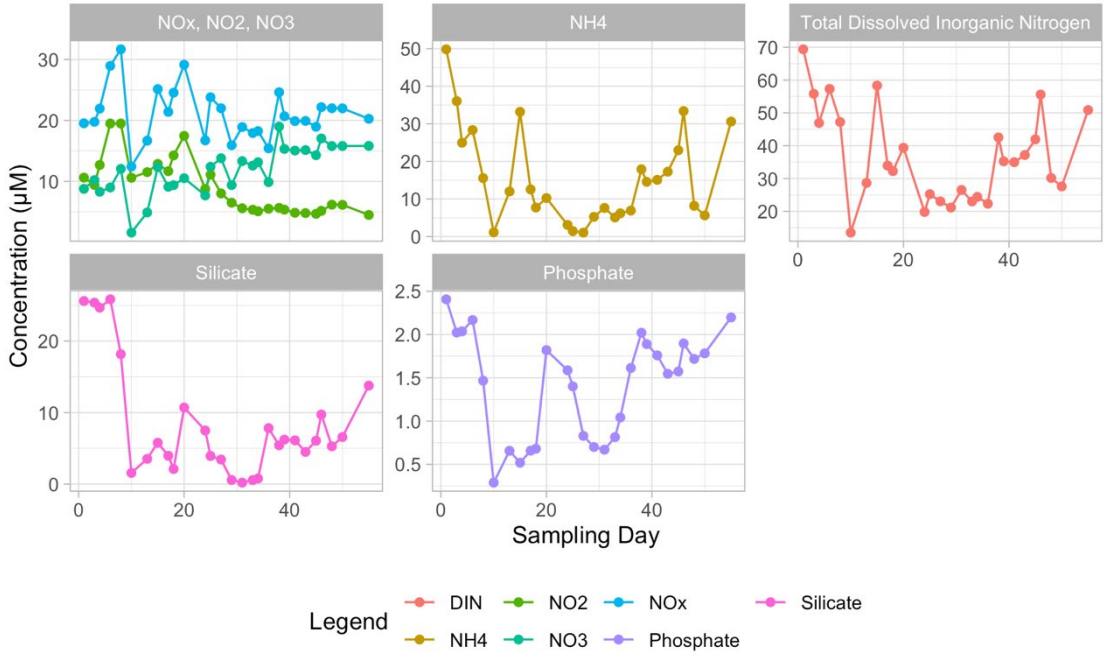


Figure 5.2. Fluctuations of the nutrient concentrations during the 55-day sampling period in the Johor Strait: Nitrite (NO_2)and Nitrate (NO_3), also summed together as NOx; ammonia (NH_4); total dissolved inorganic nitrogen (DIN); Silicate- Si $(OH)_4$;Phosphate (PO_4). All concentration were expressed as micromole per litre (µM).

Prokaryotic abundances ranged from 1.94 x10^6 prokaryotes/mL to 6.33 x 10^6 prokaryotes/mL, reaching the highest values on the sampling days 38 and 55, 6.31 x 10^6 prokaryotes/mL and 6.33 x 10^6 prokaryotes/mL, respectively. Total virus-like particles

(VLPs) abundance ranged from 2.07 x 10^7 VLPs/mL to 7.88 x 10^7 VLPs/mL, with the peak on the last sampling day reaching a maximum of 1.04 x 10^8 VLPs/mL (Figure A5.2). To quantify phytoplankton viral populations, which are significantly lower in number (roughly one order of magnitude) than prokaryotic viruses and, hence, overshadowed in total virus counts, I gated a total of 5 distinct viral subpopulations, named Vir1-Vir5, based on the difference in fluorescence intensity and side scatter properties (Figure A5.1). Viral subpopulations Vir1 and Vir2, which are considered to be prokaryotic viruses, had the biggest contribution to the total VLP counts and were tightly coupled/correlated with the total VLP abundances and prokaryotic abundances (R=0.59, p=0.00028 for V1-V2 population, R=0.62, p=0.00055 for total VLP) (Figure A5.2 and Figure A5.3). Viral subpopulation Vir3, containing both prokaryotic and phytoplankton viruses peaked at the beginning of the sampling period, reaching a maximum of 2.12x10^7 Vir3/mL and 1.98 x10^7 Vir3/mL on sampling days 3 and 4 and remained high until sampling day 10. Throughout the rest of the sampling period, the Vir3 subpopulation had continuously low abundance (mid 10^6 per mL), until the last day when it reached another maximum (Figure 5.3, panel (A). The putative phytoplankton virus subpopulation Vir 4 reached maxima (>10^6 Vir4/mL) on sampling days 1, 3 and 25. Similarly, the phytoplankton virus subpopulation Vir5 reached highest values (>10^6 V5/mL) on sampling days 25, 31 and 38 (Figure 5.3, panel (C)). Throughout the rest of the sampling period the two phytoplankton virus subpopulations fluctuated in the mid-range of 10^5 Vir4-Vir5/mL. The putative phytoplankton virus populations did not significantly correlate with chlorophyll-a concentrations.

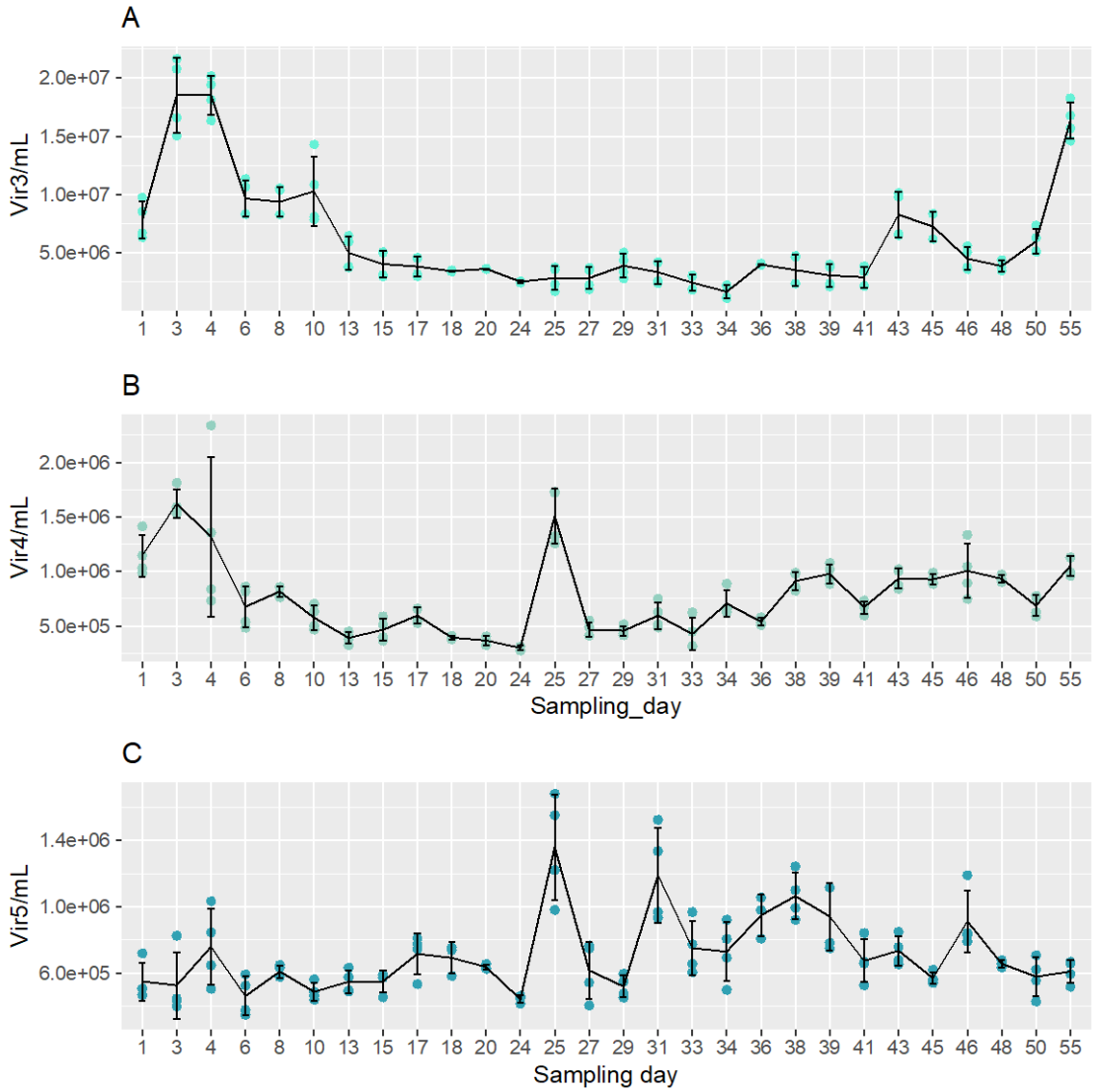


Figure 5.3. Temporal dynamics of distinct viral subpopulations of putative phytoplankton DNA viruses (*Phycodnaviridae*) distinguished with flow cytometry. A) Vir3 B) Vir4 C) Vir5. Blue dots represent single measurements, the line represent the average concentration.

5.4.2. Temporal dynamics of eukaryote-infecting RNA and giant DNA viruses in the Johor Strait

RNA viruses exhibited prominent temporal shifts in taxonomic composition (Figure 5.4). At the beginning of the sampling period, the relative expression of marna-like transcripts was low, also revealing the presence of amalga-like and partiti-like viruses from double-stranded order *Durnavirales (Pisuviricota)* and dicistro-like viruses from the order *Picornavirales*. During the first chlorophyll-a peak, the relative expression of marna-like viruses surged 10-fold and remained stable and high for 2 weeks. On

sampling days 24-25 relative expression dominated by few marna-like transcripts (Figure 5.4). These contigs are phylogenetically most similar to the *Sogarnaviruses* and are possibly isoforms of the same RNA viral genome (Chapter 4, Figure A4.6). On a higher resolution, the expression of marna-like viruses exhibited multiple "bloom and bust" peaks (marna-like_11 on day 10; marna-like cluster 1 on sampling days 13-20; marna -like cluster 2 on sampling days 25-29) (Figure A5.6). Tombus-like cluster 1 (sampling days 15-24), unclassified *Kitrinoviricota* viruses (sampling day 10) also followed the "bloom and bust" dynamics. After the second large chlorophyll-a peak on sampling day 29, marna-like viruses did not surge again. The relative expression of double-stranded picobirna-like sequences was higher during that period. Animal pathogens like hepe-like and noda-like viruses stayed low through the sampling period. Community structure of active RNA viruses was significantly different throughout the early sampling period (low chlorophyll-a concentration), middle (first chlorophyll-a maximum) and late (second chlorophyll-a maximum) (PERMANOVA, $R^2 = 0.33201$, p-value=0.001; ANOSIM ,R= 0.7282, p-value = 0.001).

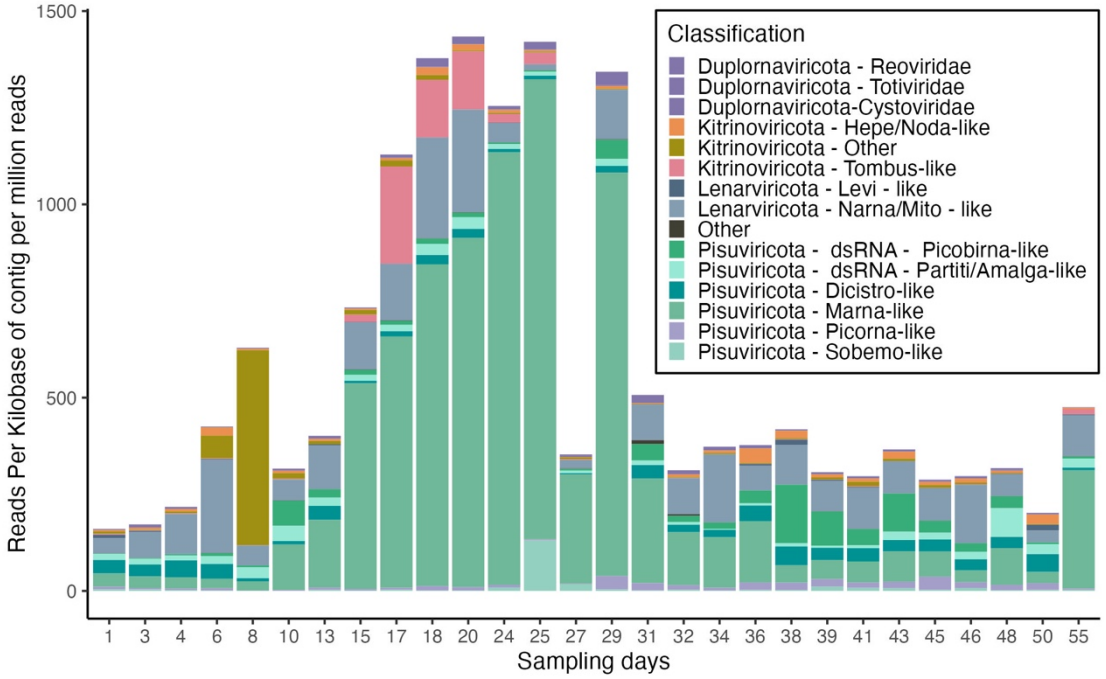


Figure 5.4. Normalized relative expression of major groups of RNA viruses during the 55-day time series in the Johor Strait expressed as Reads Per Kilobase of contig per Million reads. *Duplornaviricota* (Cysto-like, Toti-like, Reo like) are shown in purple; *Kitrinoviricota* (hepe-like/noda-like, tombus-like) in orange, brown; *Lenarviricota* (narna-like, mito-like) in grey; *Pisuvircota* (partiti/amalga-like, picobirna-like, dicistro-like, marna-like, sobemo-like and picorna like) in blue green palette.

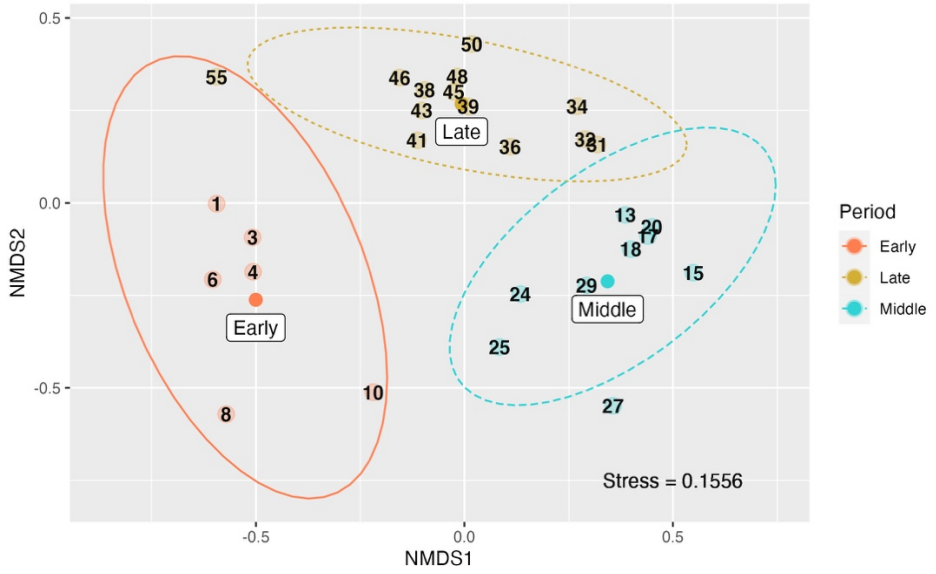


Figure 5.5. Non-metric multidimensional scaling (NMDS) ordination plot based on Bray-Curtis distance of RNA viral expression in 55-day time-series in the Johor Strait (k=2, trymax=100). Early sampling period (sampling days 1-10) are shown in red, middle sampling period (sampling days 13-29) in green and late sampling period (sampling days 31-55) in blue. Stress= 0.1556452. Centroids represent the multivariate means and the ellipses show 95% confidence intervals.

The giant DNA viruses showed highest expression in the beginning and in the end of the sampling period, with a mid-time-series peak at day 25. PERMANOVA and ANOSIM analysis confirmed that the diversity of giant DNA viruses shifted significantly during the time-series, but less of community variation could be explained by the groupings. (PERMANOVA, R^2 =0.22425 , p-value=0.002; ANOSIM ,R=. 0.2079, p-value = 0.005). Prasinoviruses and unclassified prasinoviruses had the highest contribution to the relative expression of giant DNA viruses at the beginning and the end of the sampling period (Figure 5.6). The expression peaks of unclassified prasinoviruses were tightly synchronized with the flow cytometry V3 virus particle subpopulation that contains smaller phytoplankton viruses (R=0.82, p=1.1e-07, Figure A5.3.). In flow cytometry tests with cultured phytoplankton viruses, prasinovirus Micromonas Pusilla Virus (MpV) had a flow cytometric signature within the Vir3 populations (Brussaard, 2004). The expression of the MGVL-67 family of giant viruses had a weak correlation with the concentrations of the flow cytometry Vir4 viral populations, containing large phytoplankton viruses (R=0.19, not significant) (Figure A5.3).

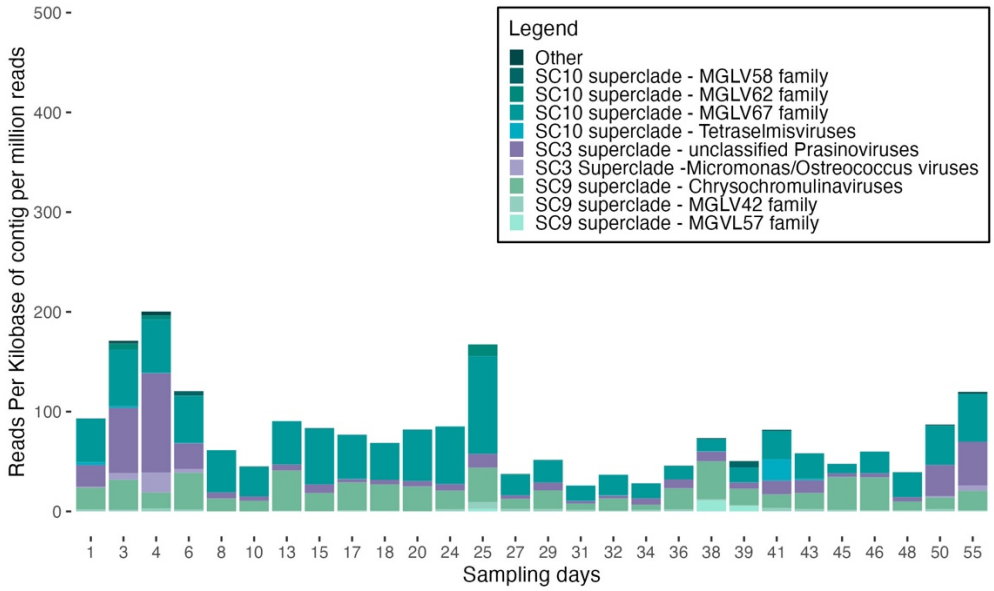


Figure 5.6. Normalized relative expression of detected giant DNA viruses during the 55-day time-series in the Johor Strait expressed as Reads Per Kilobase of contig per Million reads. SC10 superclade members (MGLV62 family, MGLV67 family, MLVG 58, Tetraselmisviruses) are shown in teal; SC3 superclade members (unclassified prasinoviruses, *Micromonas* and *Ostreococcus* viruses) in purple and SC9 superclade members (Chrysochromulinaviruses, MGLV42 family, MGLV57 family) are shown in blue-green palette.

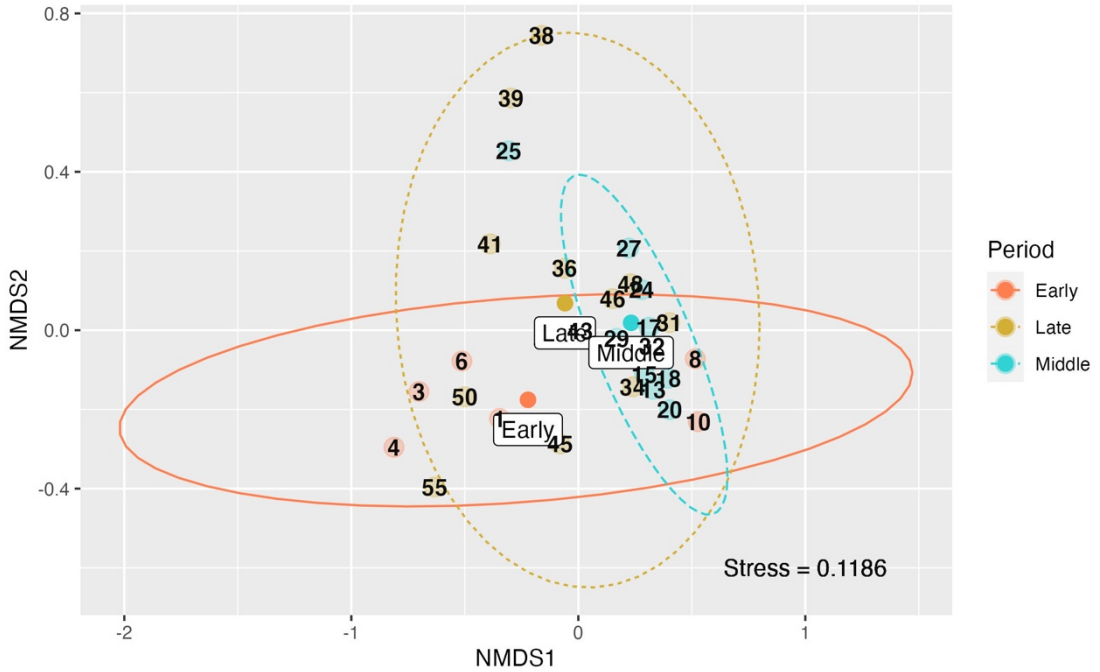


Figure 5.7. Non-metric multidimensional scaling (NMDS) ordination analysis based on Bray-Curtis distance of transcription of giant DNA viruses in 55-day time-series in the Johor Strait (k=2, trymax=100). Early sampling period (sampling days 1-10) are shown in red, middle sampling period (sampling days 13-29) in green and late sampling period (sampling days 31-55) in blue. Stress= 0.1186414. Centroids represent the multivariate means and the ellipses show 95% confidence intervals.

5.4.3. Environmental variables influencing the eukaryotic community composition in the Johor Strait

Microbial eukaryote community exhibited significant variation in diversity between sampling periods (PERMANOVA R2=0.2466 p=0.002, ANOSIM R=0.317, P=0.001). Envfit analysis of the whole eukaryotic community indicated that elevated nutrient concentrations (DIN, phosphate, silicate) and phytoplankton virus counts (flow cytometry populations Vir3-Vir4) were structuring the early period community. Physical parameters (rain, temperature, salinity, turbidity) did not show significance in envfit analysis (Figure 5.8)

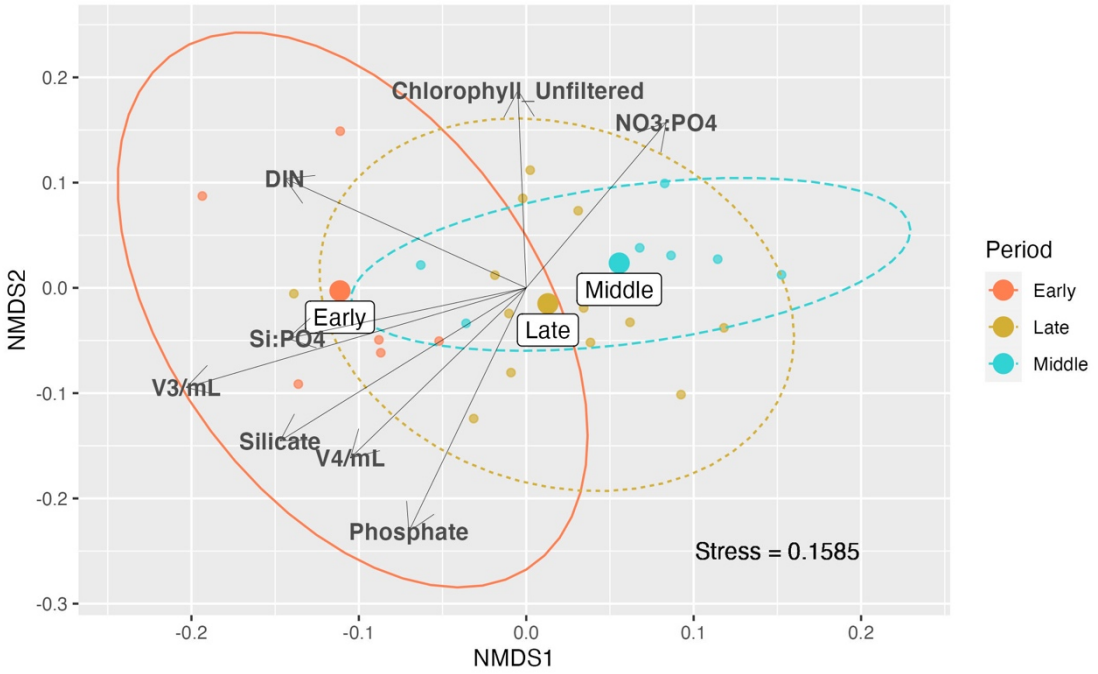


Figure 5.8. Non-metric multidimensional scaling (NMDS) ordination analysis based on Bray-Curtis distance of phytoplankton community in 55-day time-series in the Johor Strait (k=2, trymax=100). Red denotes early sampling period (days 1-10), green denotes middle sampling period (days 13-29) and blue denotes late sampling period (days 31-55). 95% confidence intervals are shown in the ellipses and multivariate means by the centroids. Envfit vectors (shown as arrows) fitted to the ordination plot indicate environmental parameters that were significantly correlated ($p < 0.05$) with the ordination.

The canonical correlation analysis (CCA) between most abundant phytoplankton taxa (n=10) and environmental variables as predictor variables indicated strong correlations between the datasets. The first two canonical roots of CCA explained 78.2% of total inertia (Monte Carlo permutation test, n=999, F=5.40, p=0.001). Nutrients played a significant role in the distribution of phytoplankton taxa. Silicate was the primary

environmental variable affecting the top phytoplankton taxa in Johor Strait, followed by phosphate and NOx. The first canonical root (CCA1) was strongly negatively correlated to silicate (-0.755, p=0.001) and phosphate (-0.521394, p=0.002). The second canonical root had significant positive correlation with NOx(0.50, p=0.05) and NO2(0.79, p=0.001) and almost significant negative correlation with NO_3 (-0.38, p=0.09). *Thalassiosira* and *Skeletonema* were positively associated with NO_2, and *Chaetoceros* with NO_3. All of these three diatom taxa were negatively associated with silicate. High phosphate concentration had most influence on the dinoflagellate *Symbodinium*, ciliate *Stylonychia* and diatom *Chaetoceros*.

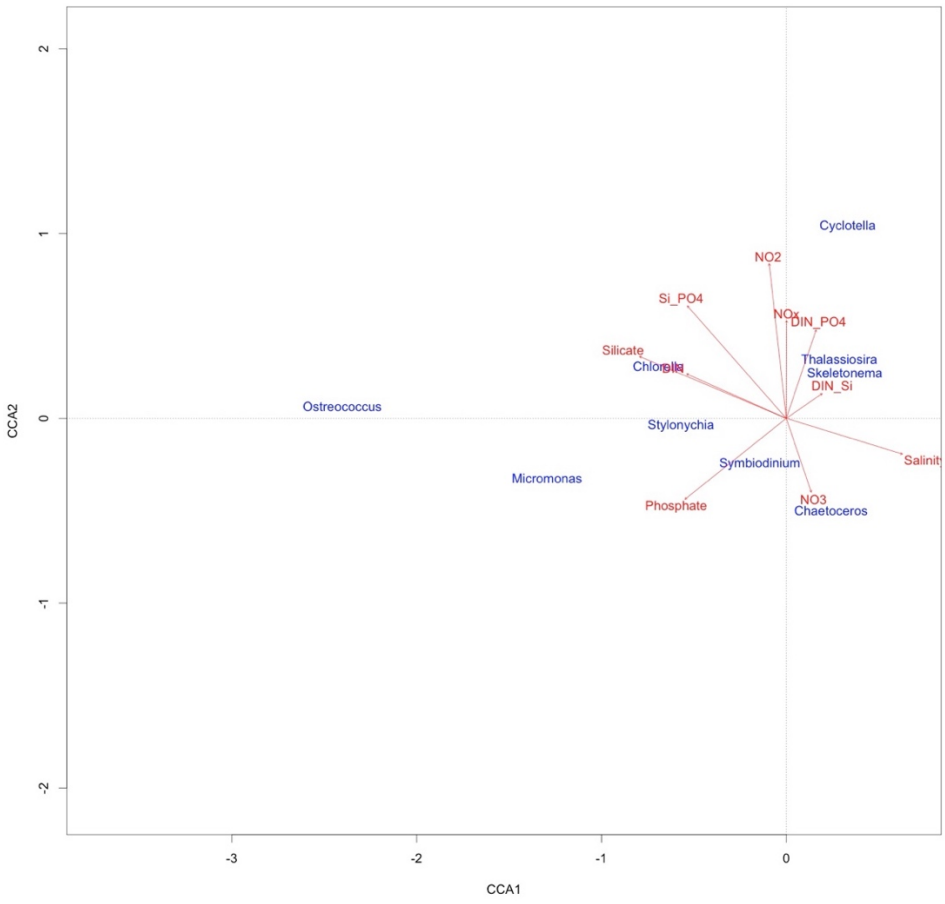


Figure 5.9. Canonical correlation analysis (CCA) of phytoplankton community (top 10 most abundant taxa) and environmental variables in the Johor Strait.

To explore the relationship between viral communities and phytoplankton community structure, canonical correlation analysis was performed using giant DNA virus and RNA virus dissimilarity matrices (top 20 taxa) as predictor variables (Figure A5.5). The canonical correlation analyses revealed significant contribution of giant DNA viruses to the phytoplankton community variation, with first two canonical axes accounting for 54.78% and 17.67% of observed variance. Likewise, the first two canonical axes accounted for 32.89% and 15.52% of the observed variance between phytoplankton and RNA viruses. Correlation between specific virus - host pairs revealed are discussed in details in the next paragraph. Mantel tests performed on the full dissimilarity matrices identified the strongest positive correlations between eukaryotes and RNA viruses (Mantel statistic R: 0.4231, p=1e-04), followed by eukaryotes and giant DNA viruses (Mantel statistic R: 0.372, p=7e-04) and the lowest correlation between eukaryotes and environmental variables (Mantel statistic R: 0.1642, p=0.0904).

5.4.4. Putative virus - eukaryotic host pairs in the Johor Strait

To explore the potential roles of viruses in regulating the eukaryotic community, I co-correlated relative abundances of top most abundant phytoplankton taxa present in the metagenomes and relative expression of top 20 most abundant giant DNA and RNA viruses. RNA viruses and diatoms accounted for most positive correlations, indicative of co-expression patterns (Figure 5.10.(A)). Diatom *Thalassiosira* was positively correlated with viral contig Marna-like_11 (R=0.29) and unclassified virus from phylum *Kitrinoviricota* (R=0.33), but the correlation was not statistically significant. Diatom *Chaetoceros* was positively correlated with viral contig Marna-like_11 (R=0.55, P <0.001). Diatom *Skeletonema* was weakly positively correlated two marna-like cluster (Marna-like cluster 1, R=0.13; Marna-like cluster 2, R=0.31), but the correlation was not statistically significant. Diatom *Fragilariopsis* was positively correlated with contig

Marna-like_11 (R=0.50,p<0.01) and an unclassified virus from the phylum *Kitrinoviricota* (R=0.51, p<0.01). Of non-diatom taxa, prasinophyte *Ostreococcus* and ciliate *Stylonychia* were correlated with Narna-like_2 viral contig (both displaying a correlation coefficient of R=0.44, p<0.01). Dinoflagellate *Symbodinium* was positively correlated Picobirna-like_1 viral contig (R=0.53, p<0.01). In general, the giant DNA viruses correlated positively with non-diatom phytoplankton (Figure 5.10. (B)). Two prasinophytes, *Ostreoccocus* and *Micromonas* were tightly correlated with expression of prasinoviruses (R=0.96, p=0.001) and unclassified MGLV67 family_1 virus (R=0.68, p<0.001 and R=0.53, p<0.01, respectively).

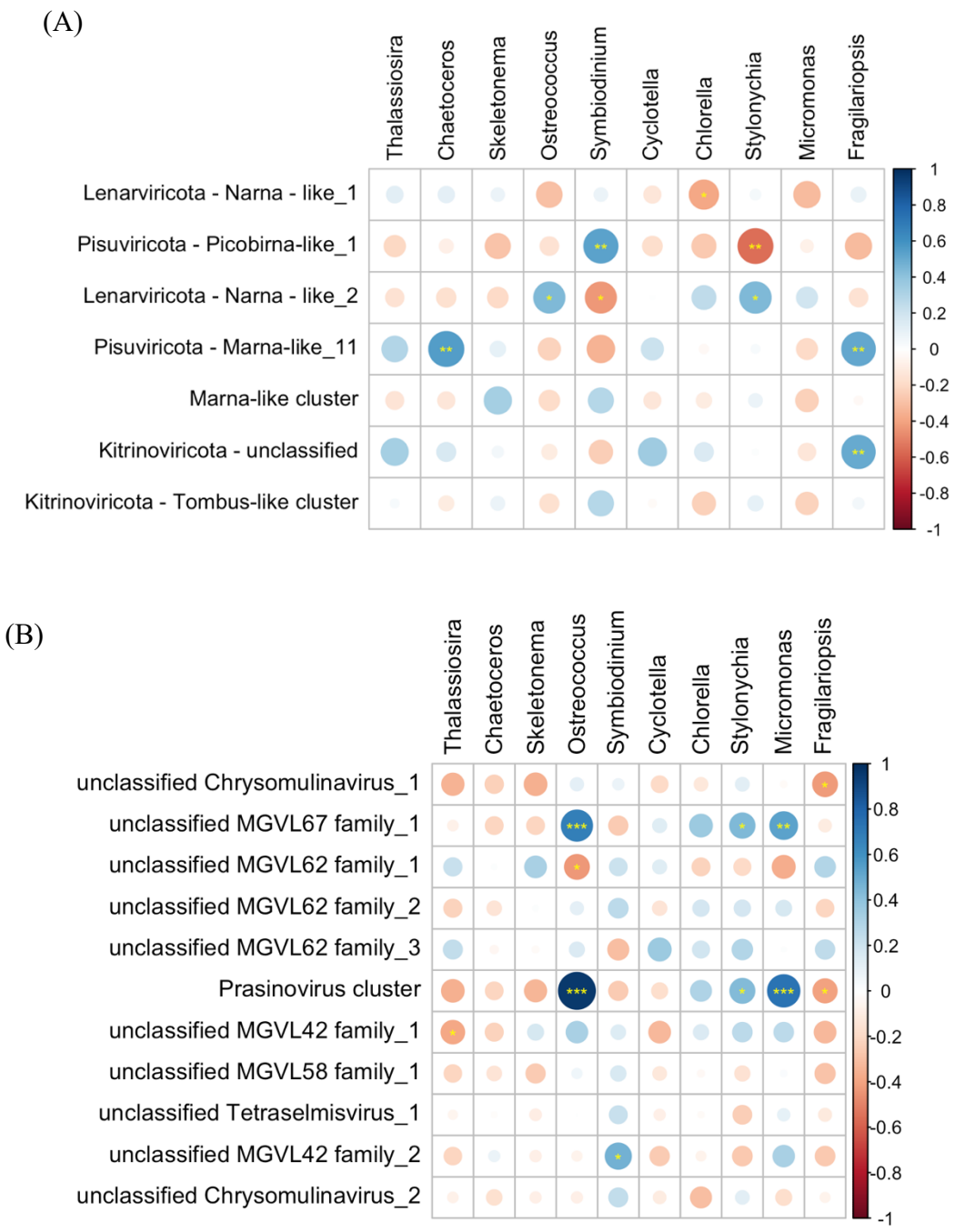


Figure 5.10. (A) Corrplot showing significant correlations between the top 10 phytoplankton taxa and top 20 RNA viruses. RNA viruses with similar abundance patterns were collapsed into clusters. (B) Corrplot showing significant correlations between top 10 phytoplankton taxa and top 20 giant DNA viruses. Unclassified prasinoviruses (n=10) with similar abundance patterns were collapsed into prasinovirus clusters. Stars indicate significance levels (*** p=0.001, ** p=0.01, * p=0.05). Significant positive correlation was considered to be an indication of a potential co-expression of virus-host pair.

5.5. Discussion

5.5.1. Fluctuations of the abiotic and biotic variables in the Johor Strait

The high nutrient concentrations during my time-series reflect the eutrophication of the Johor Strait, consistent with the data from other published studies (Chénard et al., 2019; Mohd-Din et al., 2022). The area east of Singapore-Johor causeway is considered one of the hotspots for phytoplankton blooms (Trottet et al., 2021; Wijaya et al., 2022), with blooms reported in October 2018, and from January to September 2019 (Hii et al., 2021). Chlorophyll concentration averages around 21.5 μg/L for the whole area of the Johor Strait (Gin et al., 2000). Extremely high chlorophyll concentrations (up to 160μg/L) were recorded during phytoplankton blooms (Chénard et al., 2019; Gin et al., 2000; Hii et al., 2021; Lim et al., 2014; Mohd-Din et al., 2020). The fluctuations of chlorophyll-a concentration recorded in my study are lower in comparison, possibly due sampling nearer to the coastline. The viral and prokaryotic abundance recorded in this study are within ranges for marine ecosystems, as well the abundances of putative DNA phytoplankton viruses (Biggs et al., 2021; Mojica et al., 2016; Wigington et al., 2015). Viral subpopulations with higher side scatter are (Vir4-Vir5) have been reported closely correlated to the chlorophyll-a concentration (Suttle, 2007), but I detected none that were statistically significant. Since both chlorophyll-a and flow cytometry counts are bulk measurements, a higher correlation may be revealed when looking at the specific virus-host pairs in the metatranscriptomic data. There may be a correlation between the abundances of lytic RNA viruses that infect phytoplankton and reductions in chlorophyll-a. However, the current flow cytometers are not able to accurately quantify abundances of RNA viruses due to their low fluorescence (small genomes) and low side scatter (small particle size) (Brussaard, 2004a; Tomaru and Nagasaki, 2007).

5.5.2. Temporal dynamics of eukaryote-infecting RNA and giant DNA viruses in the Johor Strait

Metatranscriptomics is able to detect immediate changes in the viral activity, but it cannot distinguish between transcripts and genomes of (+) ssRNA viruses (Chapter 2, Kolundžija et al., 2022). For this reason, it is possible that I occasionally detected low abundance (+) ssRNA viruses adsorbed to the cellular organisms or organic matter. In contrast to RNA viruses, the presence of RNA transcript always implies an active ongoing infection for giant DNA viruses.

A recently proposed theory suggested that lytic infection of phytoplankton may exist on a spectrum and mimicking their host growth strategy (Sandaa et al., 2022). Phytoplankton with high growth rates ("opportunists") host "aggressive" viruses with acute infections that release thousands of infectious viral particles during the lysis of the host cells, also known as "bloom and bust" dynamics. Slow growing, low abundance phytoplankton ("gleaners") host less acute viruses that do not kill the cell immediately. The continuous and low release of viral particles is an evolutionary adaptation to scarcity of hosts (Sandaa et al., 2022). My high-resolution temporal study revealed clear temporal differences in expression patterns of giant DNA viruses and RNA viruses infecting phytoplankton and provided evidence supporting this theory.

Multiple RNA viruses engaged in "bloom and bust" dynamics, disappearing after strong, but ephemeral increase in expression. The prominent 10-fold increase in the abundance of marna-like transcripts is likely tied to their host, which are considered to be diatoms (Lang et al., 2021). A similar "bloom" of unclassified RNA viruses dominated the metatranscriptome in the aftermath of a brown tide bloom (Moniruzzaman et al., 2017). As the magnitude of burst size affects the viral abundances and the likelihood of new viral infections, I speculate that large burst sizes of RNA

viruses could mediate the rapid turnover characteristic for “bloom and bust” dynamics. (Edwards et al., 2021).

Giant DNA viruses contain many cellular metabolic genes and can fine-tune host physiology during infection to better control viral production (Moniruzzaman et al., 2020; Schulz et al., 2020). This could explain the continuous expression of some giant DNA viral transcripts inside cells throughout the sampling period and accompanying stable populations of extracellular phytoplankton viruses (Vir3, Vir4 and Vir5) detected with flow cytometry. A short 2-day metatranscriptome study in the California Current also noted continuous expression of the giant DNA viral core genes (Ha et al., 2021). In the weekly metatranscriptome study of brown tide bloom, more giant DNA viruses displayed continuous expression than the “bloom and bust” dynamics (Moniruzzaman et al., 2017). I have observed a “bloom and bust” of prasinoviruses at the beginning and the end of my sampling campaign. Prasinoviruses have the smallest genomes of all giant DNA viruses (Weynberg et al., 2017) which could explain why they are more prone to follow “bloom and bust” dynamics.

5.5.3. Abiotic and biotic variables influencing the eukaryotic community composition

A spatial study along the Johor Strait identified salinity, silicate and rainfall as the main environmental parameters structuring phytoplankton community during North-East monsoon (Hii et al., 2021), and salinity-induced changes in phytoplankton structure were observed in temporal study by Chenard et al. (2019). In my study, nutrient concentrations had the strongest influence on the phytoplankton community structure. Silicate had the strongest negative effect on *Chaetoceros*, and lesser effect on *Skeletonema* and *Thalassiosira*. Silicate limitation (N:Si ratio higher than 4:1) coincided with the diatom-dominated second chlorophyll-a maximum and may have contributed

to the diatom biomass decline, supporting the negative relationship between silicate and diatoms. High concentrations of phosphate are known to shift the community structure towards an increased proportion of dinoflagellates (Er et al., 2018; Lau et al., 2017). Elevated phosphate concentration contributed to the bloom of dinoflagellate *Scrippsiella* in 2018 in the Johor Strait (Mohd-Din et al., 2020), which may explain positive relationship of dinoflagellate *Symbodinium* with the phosphate concentration observed in my analysis. Strong correlation between diatoms *Thalassiosira*, *Skeletonema* and *Chaetoceros* with inorganic nitrogen sources may be due to their known appetite for nitrate, preferring inorganic nitrogen sources over organic sources (Gilbert and Burford, 2017; Gilpin et al., 2004; Mohd-Din et al., 2022; Tilstone et al., 2000).

Mantel correlation tests suggest that both giant DNA viruses and RNA viruses are actively shaping the eukaryotic phytoplankton community composition. Stronger Mantel correlations between RNA viruses and eukaryotic phytoplankton suggest that RNA virus may be more important in regulating quick, swift responses to phytoplankton biomass increase and preventing the shift from diverse community to a one-species dominance. This could be explained with RNA viruses display acute "bloom and bust" dynamics with high burst sizes which is less common in giant DNA viruses (Sandaa et al., 2022). This process could play a fundamental role in coastal habitats with significant levels of eutrophication where the top-down controls might be crucial for maintaining ecosystem diversity and function (Wijaya et al., 2022).

5.4.4. Putative virus - eukaryotic host pairs in the Johor Strait

To explore the role of viral infection in regulation of the phytoplankton community structure, I investigated possible interactions between most abundant phytoplankton taxa and viruses. Prasinophytes *Ostreoccocus.* and *Micromonas* correlated very tightly

(R=0.96, p<0.0001) with expression of unclassified prasinoviruses in my dataset. Similar temporal synchronization in expression of giant DNA viruses and their host was observed in a diel metatranscriptome study in California Current system (Kolody et al., 2019). Besides the clear and strong correlation of prasinophytes and prasinoviruses, other groups of giant viruses did not show strong correlation patterns with the most abundant phytoplankton taxa, which was also supported the CCA analysis.

Three most abundant diatoms - *Thalassiosira*, *Skeletonema, and Chaetoceros* displayed clear grouping with particular viral groups in the CCA analysis, indicating possible virus-host pairs. The positive correlation between diatoms and their putative viruses, marnaviruses were weaker than expected. This could be due to their "bloom and bust" life style, often described as "Kill the Winner" dynamics (Thingstad, 2000). It should be noted that transcriptional temporal decoupling or anti-correlation, where virus abundance lags after host abundance, has been observed for RNA viruses and their host (Kolody et al., 2019). To clarify this, my dataset necessitates further investigation using time-delayed (lag) correlations.

Two narna-like viruses and unclassified *Kitrinoviricota* virus were strongly synchronized with *Ostreococcus* and diatom *Fragilariopsis,* respectively. Similar viral groups were detected within metatranscriptomes of cultured phytoplankton, suggesting these potential virus-host pairs might have a persistent, non-lytic infection style (Charon et al., 2020, 2021; Chiba et al., 2020). These viruses completely lack the extracellular stage, are transmitted vertically, and can have beneficial effect on their host (Roossinck, 2019). A high diversity of viruses with this type of lifestyle was observed in the Johor Strait metatranscriptome suggesting they might be very common in marine environment (Chapter 4).

To reveal more virus-host pairs with this dynamics, the virus-host prediction analysis should not be limited to only abundant taxa like in my analysis, as these interactions can be overshadowed by the acute infections. All host predictions based on bulk metagenomic and metatranscriptomic approaches must be treated with caution until similar coupling is proven experimentally using isolated virus - host system or by transcriptomic sequencing of the host culture. Nevertheless, this is an important step in expanding the classic "Kill the Winner" view and shift from the perspective that phytoplankton viruses are preferentially lytic. It provides evidence for the widespread presence of less acute lytic infections as displayed by the giant DNA viruses and truly persistent infections of multiple (+) ssRNA and dsRNA viruses.

5.6. Conclusions

This high-resolution, long-term monitoring study offers valuable insight into abiotic and biotic factors affecting the eukaryotic phytoplankton dynamics in coastal eutrophic environment. Both RNA and giant DNA viruses varied significantly, albeit with distinct temporal patterns and intensity and exhibited strong associations with the phytoplankton community structure. This is not an isolated process, rather one of my co-regulating mechanism, together with grazing and nutrient depletion to maintain the diversity of the marine ecosystem. This chapter provides the first evidence of the role of eukaryotic viruses in bloom prevention in a tropical marine ecosystem and may provide the foundation for studies aimed at bloom control measures and serve as an important reference for other eutrophic coastal systems.

5.7. Acknowledgements

I would like to thank Woo Yi Hui for the sampling assistance and Oon Yee Woo for the nutrient analysis. I would like to acknowledge Prof Corina Brussaard for her advice on flow cytometry optimization. This study was funded by the Intra-CREATE Seed Collaboration Grant (Award NRF2018-ITS004-0001) and Competitive Research Programme (Award CRP21-2018-0005) awarded by the National Research Foundation (NRF) of Singapore. Sandra Kolundžija is supported by a Singapore International Graduate Award (SINGA) of the Agency for Science, Technology & Research (A*STAR).

APPENDIX 5

Table A 5.1. Precipitation, chlorophyll-a concentration, temperature and salinity during the 55-day sampling campaign in the Johor Strait.

Date	Sampling Day	Rain (mm)	Chlorophyll-a filtered (µg/L)	Chlorophyll-a unfiltered (µg/L)	Temperature (°C)	Salinity (PSU)
4/11/20	1	24	0.74062272	2.48823214	28.8	22.7
6/11/20	3	16.4	0.47137132	3.26334547	30.2	16.7
7/11/20	4	0	1.26329823	3.58463979	29.5	25.3
9/11/20	6	62.2	0.93596469	1.80438233	29.9	26.5
11/11/20	8	6.8	13.2498746	19.4228424	30.5	25.5
13/11/20	10	0	17.0750631	37.0735334	30.8	25
16/11/20	13	52.6	5.47427532	5.54780737	30.6	25.3
18/11/20	15	11.2	13.9681359	22.5663296	30.3	22.8
20/11/20	17	5.2	10.4602668	2.78698041	28.8	23.7
21/11/20	18	7.8	2.10777981	25.9518168	30.4	25.7
23/11/20	20	85.6	0.31252958	1.69022398	29.9	24.2
25/11/20	22	1.2	0.65177261	1.1105091	29.4	26.5
27/11/20	24	1.8	11.1156054	10.3909497	29	26.7
28/11/20	25	1.8	8.99992748	2.86991568	28.6	26.7
30/11/20	27	2	8.79198791	13.2375871	31.6	27.2
2/12/20	29	1.6	12.9810512	33.79588	29.4	27.5
4/12/20	31	31.8	18.5414337	21.0365813	29.2	27.2
5/12/20	33	36.8	22.7301233	20.8767077	29.3	27.5
7/12/20	34	0.2	10.7943731	18.2151087	29.3	27.7
9/12/20	36	6.6	6.2175353	11.2173717	29.9	28.4
11/12/20	38	40	2.13455672	5.44807135	29.1	25.9
12/12/20	39	19.8	4.33464974	7.00097249	29.4	26.8
14/12/20	41	27.4	14.6169887	20.4247373	30.3	26.4
16/12/20	43	21.6	16.6209664	8.73213786	29	26.7
18/12/20	45	24.2	7.29259563	12.7320944	27.5	25.5
19/12/20	46	14.4	3.13924272	8.68513239	28.5	26.8
21/12/20	48	2.6	1.33007663	15.1193177	30	26.3
23/12/20	50	0.2	4.34811293	7.57783608	28.9	26.6
26/12/20	51	11.2	2.17471708	2.80994808	29.8	25.6
28/12/20	55	60.4	8.46217175	17.9491719	29.5	21

Table A 5.2. Nutrient concentrations and nutrient ratios during the 55-day sampling campaign in the Johor Strait.

Date	Sampling Day	NO_x (μM)	NO_2 (μM)	NO_3 (μM)	NH_4 (μM)	DIN (μM)	Phosphate (μM)	Silicate (μM)	NO_3:PO_4	NO_3:Si	Si:PO_4
4/11/20	1	19.528	10.61	8.768	49.84	69.368	2.407	25.6	3.64270877	0.3425	10.635646
6/11/20	3	19.766	9.412	10.215	36.05	55.816	2.023	25.38	5.04943154	0.40248227	12.5457242
7/11/20	4	21.942	12.69	8.284	24.974	46.916	2.037	24.683	4.06676485	0.33561561	12.1173294
9/11/20	6	28.959	19.49	8.996	28.34	57.299	2.168	25.844	4.14944649	0.34808853	11.9206642
11/11/20	8	31.675	19.49	12.047	15.549	47.224	1.467	18.175	8.21199727	0.66283356	12.3892297
13/11/20	10	12.424	10.588	1.588	1.11	13.534	0.289	1.562	5.49480969	1.01664533	5.40484429
16/11/20	13	16.675	11.51	4.896	11.985	28.66	0.656	3.527	7.46341463	1.38814857	5.37652439
18/11/20	15	25.145	12.86	12.328	33.19	58.335	0.519	5.777	23.7533719	2.13397957	11.1310212
20/11/20	17	21.4	11.67	9.103	12.534	33.934	0.658	3.966	13.8343465	2.29525971	6.02735562
21/11/20	18	24.559	14.26	9.377	7.705	32.264	0.681	2.106	13.7694567	4.45251662	3.09251101
23/11/20	20	29.149	17.46	10.518	10.238	39.387	1.82	10.709	5.77912088	0.98216454	5.88406593
25/11/20	22	29.997	16.9	11.361	6.946	36.943	2.018	13.311	5.62983152	0.85350462	6.59613479
27/11/20	24	16.723	8.752	7.691	3.086	19.809	1.587	7.492	4.84625079	1.02656167	4.72085696
28/11/20	25	23.803	11.085	12.39	1.444	25.247	1.401	3.93	8.84368308	3.15267176	2.80513919
30/11/20	27	21.997	8.011	13.799	1.042	23.039	0.829	3.431	16.6453559	4.02185952	4.13872135
2/12/20	29	15.923	6.497	9.387	5.237	21.16	0.701	0.571	13.3908702	16.4395797	0.81455064
4/12/20	31	18.914	5.579	13.318	7.606	26.52	0.67	0.204	19.8776119	65.2843137	0.30447761
5/12/20	33	17.937	5.336	12.582	5.07	23.007	0.815	0.555	15.4380368	22.6702703	0.6809816
7/12/20	34	18.243	5.088	13.141	6.175	24.418	1.042	0.786	12.6113244	16.7188295	0.75431862
9/12/20	36	15.421	5.503	9.869	6.911	22.332	1.614	7.821	6.11462206	1.2618591	4.84572491

Date	Sampling Day	NO_x (μM)	NO_2 (μM)	NO_3 (μM)	NH_4 (μM)	DIN (μM)	Phosphate (μM)	Silicate (μM)	NO_3:PO_4	NO_3:Si	Si:PO_4
11/12/20	38	24.622	5.641	18.999	17.901	42.523	2.021	5.432	9.40079169	3.49760678	2.68777833
12/12/20	39	20.684	5.358	15.32	14.558	35.242	1.89	6.221	8.10582011	2.46262659	3.29153439
14/12/20	41	19.896	4.847	15.055	15.082	37.169	1.76	6.121	8.55397727	2.45956543	3.47784091
16/12/20	43	19.921	4.794	15.136	17.248	41.932	1.546	4.503	9.79042691	3.36131468	2.91267788
18/12/20	45	18.962	4.671	14.295	22.97	55.58	1.573	6.065	9.08773045	2.3569662	3.85568977
19/12/20	46	22.18	5.158	17.042	33.4	30.175	1.897	9.728	8.98365841	1.75185033	5.128097
21/12/20	48	22	6.168	15.797	8.175	27.6	1.718	5.285	9.19499418	2.98902554	3.07625146
23/12/20	50	21.984	6.145	15.804	5.616	39.613	1.783	6.583	8.86371284	2.40072915	3.69209198
26/12/20	51	19.949	4.795	15.166	19.664	50.871	2.4	13.332	6.31916667	1.13756376	5.555
28/12/20	55	20.271	4.486	15.819	30.6	69.368	2.198	13.773	7.19699727	1.14855151	6.26615105

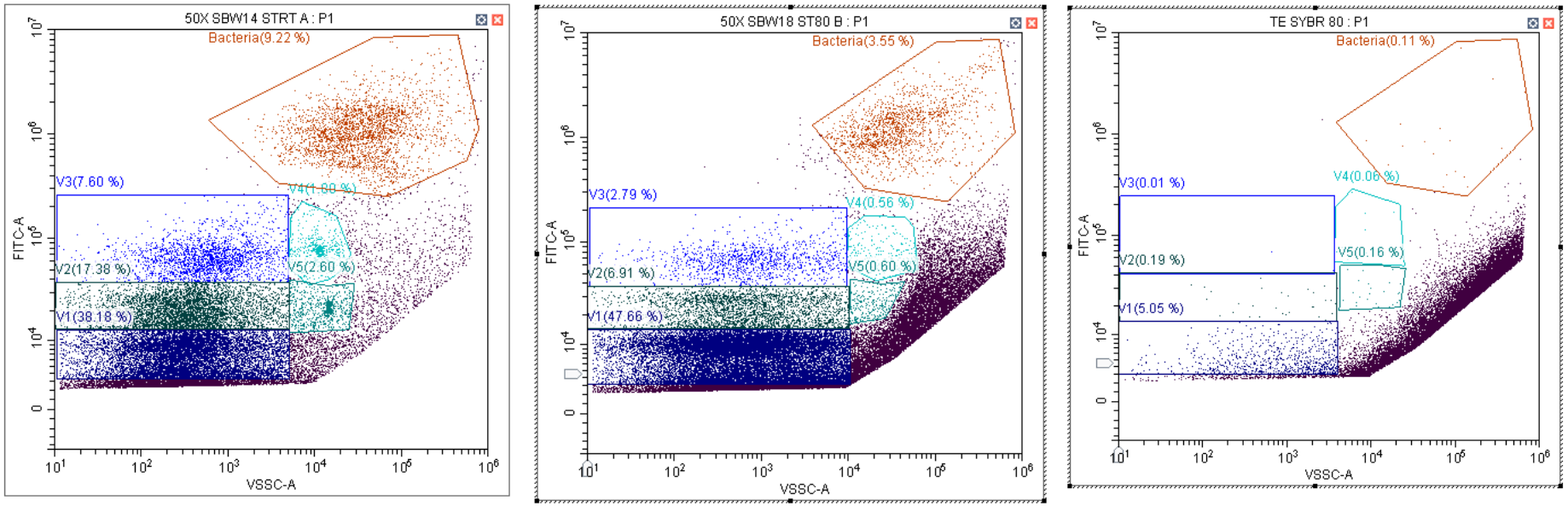


Figure A 5.1. Flow cytograms indicating prokaryotes (orange) and viruses (shades of blue). Viral subpopulations Vir1-Vir2 represent prokaryotic viruses, viral subpopulation Vir3 presents a mix of prokaryotic and phytoplankton viruses. Viral subpopulation Vir4-Vir5 are phytoplankton viruses. (A) Sample with high abundance of Vir4-Vir5 subpopulations (B) Sample with low abundance of Vir4-Vir5 subpopulations. (C) blank run. Y-axis is green fluorescence; X-axis is violet side scatter.

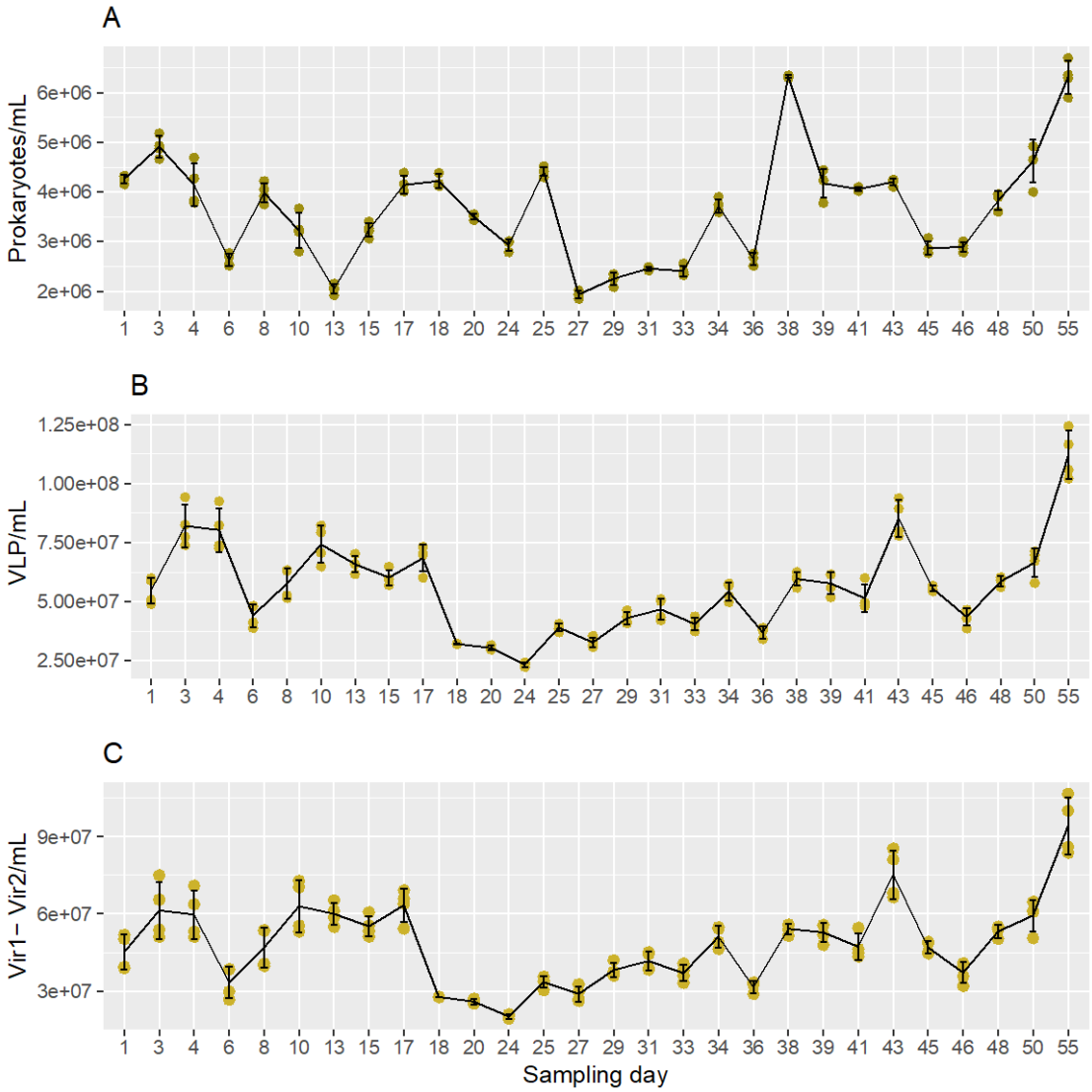


Figure A5.2. Temporal dynamics of prokaryotes (A), total virus-like particle (VLP) and Vir1-Vir2 virus subpopulation (B) abundances during the 55-day sampling campaign in the Johor Strait determined with flow cytometry. Virus-like particles (VLPs) represent total viral particles count and viral populations Vir1-Vir2 are considered to be prokaryotic viruses. Yellow dots represent single measurements and the black line represents the average.

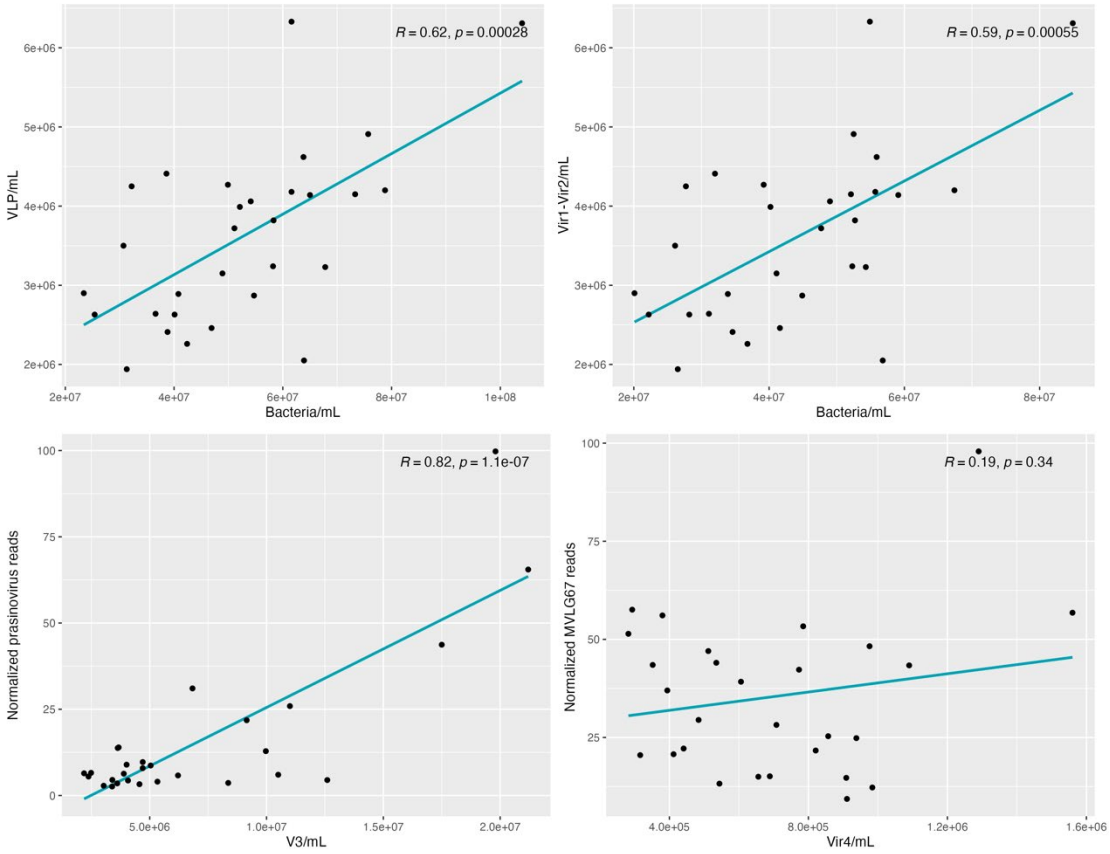


Figure A 5.3. Linear correlations between viral populations captured by FCM, chlorophyll-a concentration and relative expression of giant DNA viruses. VLP represent total virus-like particles, Vir1-Vir2 represent the bacteria infecting viruses, and Vir3 and Vir4 phytoplankton virus populations in flow cytometry. Normalized relative expression of giant DNA viral population (prasinoviruses, viruses from MVLG67 family) represents reads per kilobase of contig per million reads.

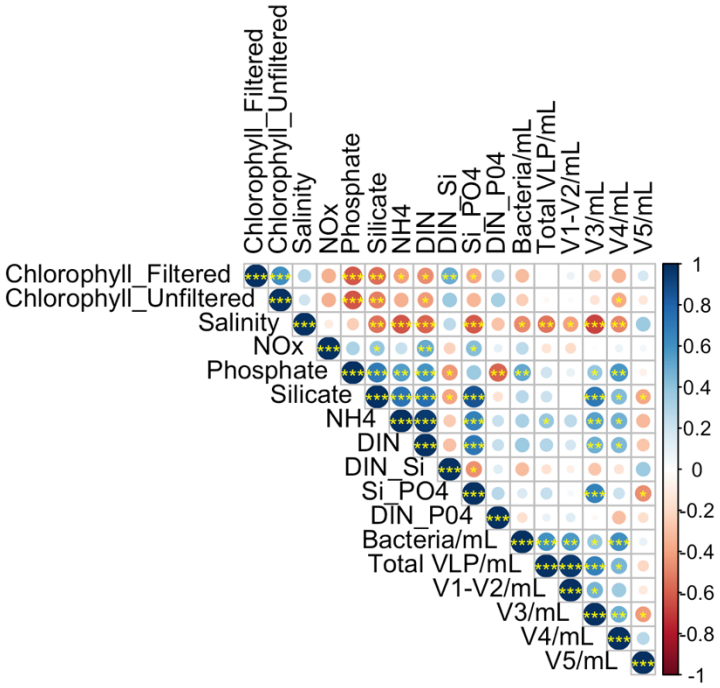


Figure A 5.4. Correlation plot showing significant correlations between environmental variables at my sampling location in the Johor Strait. Stars indicate significance levels (*** p=0.001, ** p=0.01, * p=0.05).

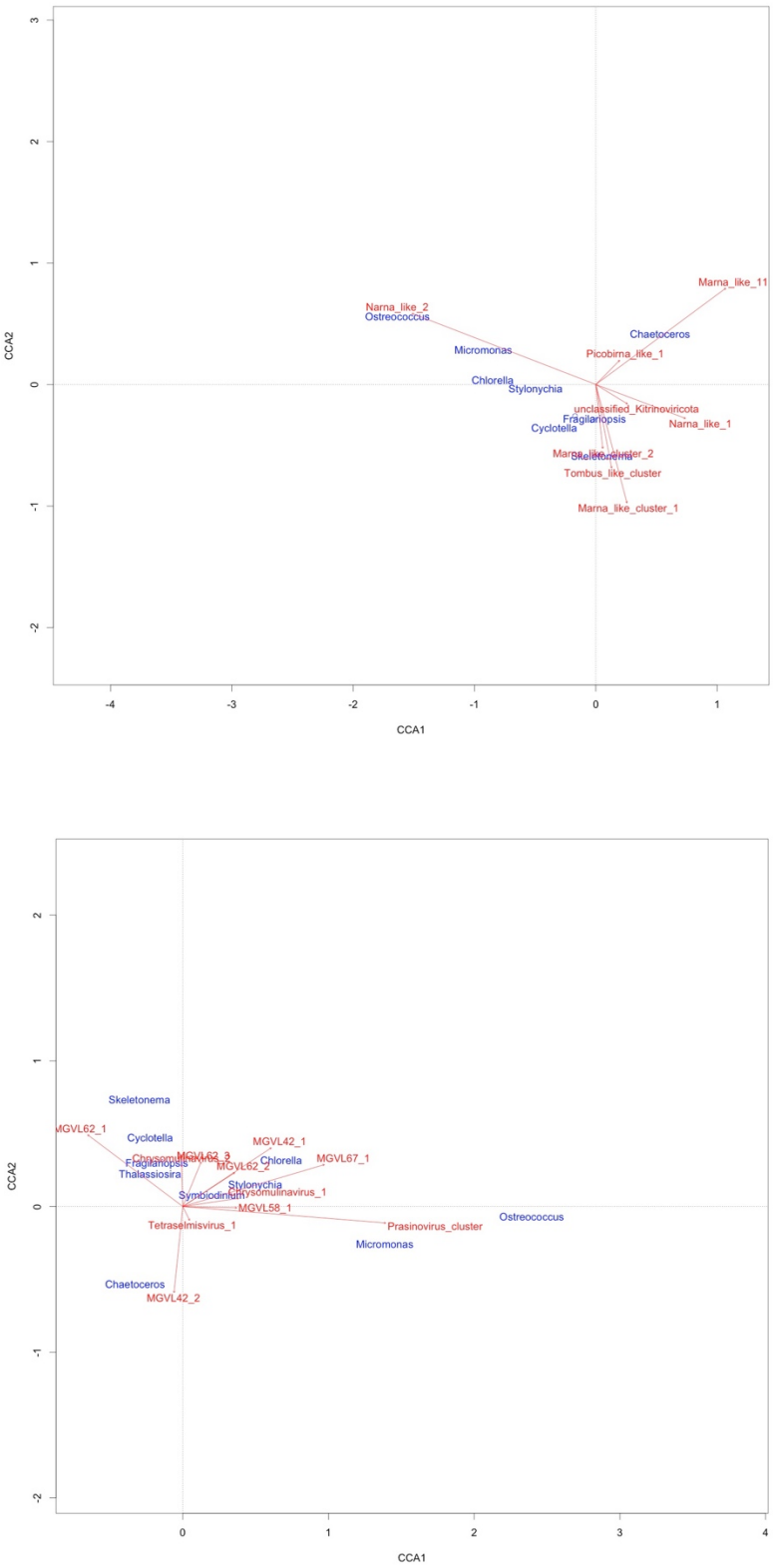


Figure A 5.5. Canonical correlation analysis (CCA) of phytoplankton community (top 10 most abundant taxa) and phytoplankton viruses in the Johor Strait: (A) RNA viruses (B) giant DNA viruses

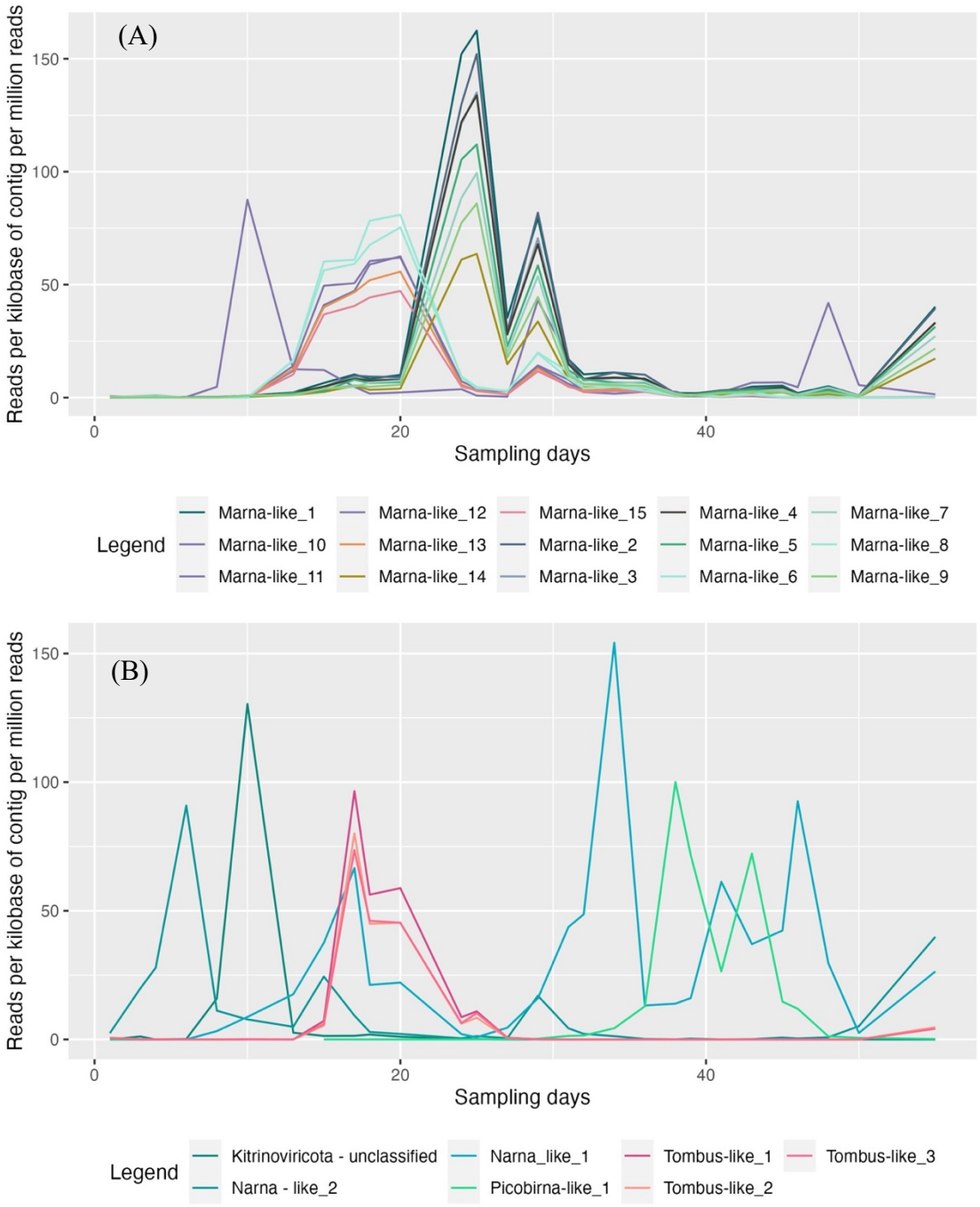


Figure A5.6. Temporal dynamics of top 20 most abundant RNA viral transcripts in the Johor Strait metatranscriptome. (A) Three separate populations of marna-like viruses were present on different sampling days: (1) Marna-like_11 contig peaked on sampling day 10 (2) Marna-like cluster 1 peaked on sampling days 13-20 (3) Marna -like cluster 2 peaked on sampling days 25-29 (B) The temporal dynamics of remaining top 20 non-marna-like RNA viral transcripts.

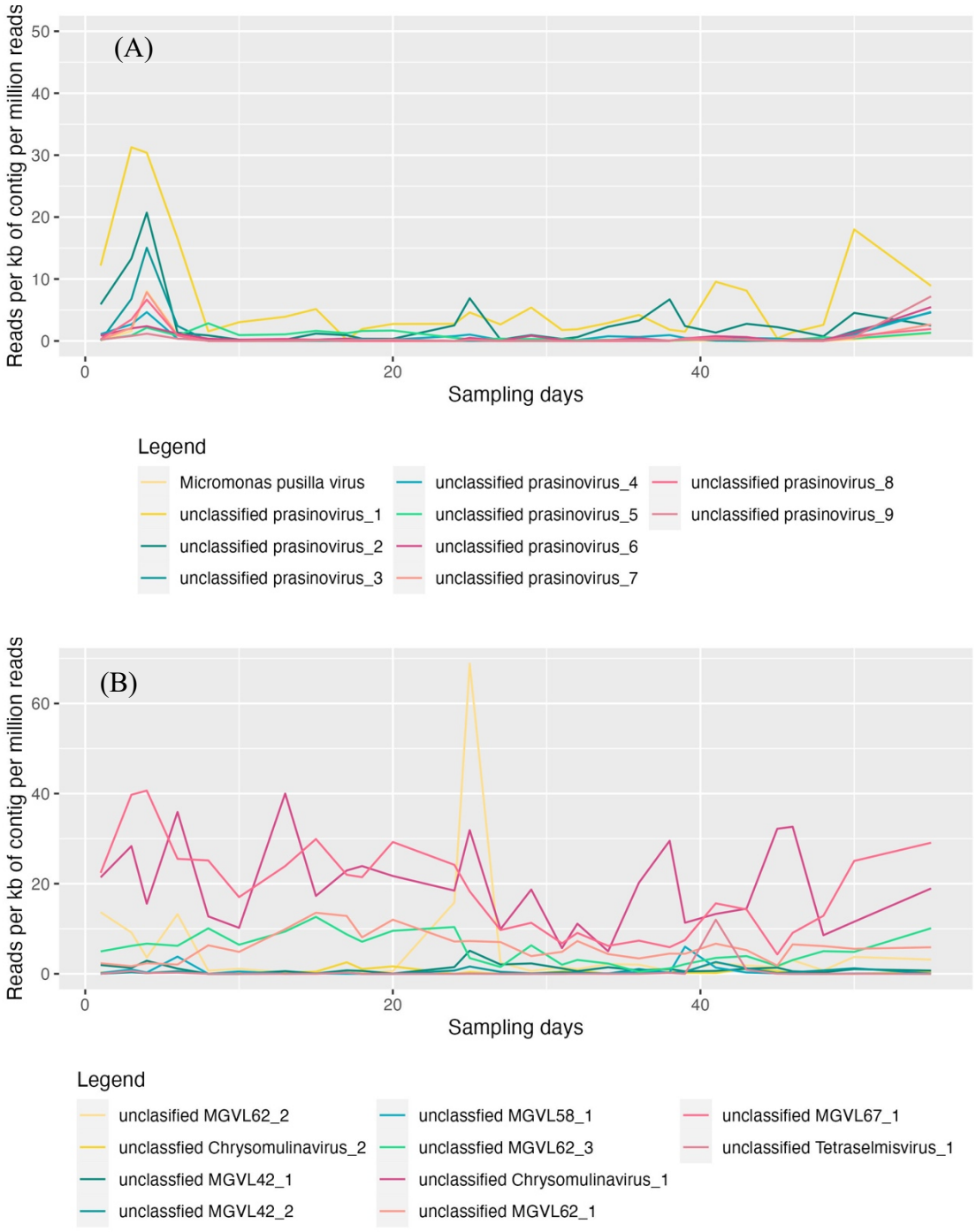


Figure A 5.7. Temporal dynamics of top 20 most abundant giant DNA viral transcripts in the Johor Strait metatranscriptome(A) prasinoviruses (B) other giant DNA viruses,

CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS

6.1. Synthesis and significance of this work

Viruses are highly dynamic, diverse, and abundant in marine ecosystems. Virus-mediated cell lysis affects the host community structure and redirects carbon from higher trophic levels, either to fuel the microbial loop in surface waters ("virus shunt") or sequestration to deep sea storage ("viral shuttle") (Weitz et al., 2015; Wilhelm et al., 2013; Zimmerman et al., 2020). Diversity-focused surveys of marina RNA viromes have been performed at a global scale (Dominguez-Huerta et al., 2022; Wolf et al., 2020; Zayed et al., 2022; Zhang, 2022), whereas ecological studies of virus-host dynamics have been significantly lagging behind and performed at very low temporal resolution (reviewed in Chapter 2, Kolundžija et al., 2022). The ecological study presented in this thesis, with its high resolution and long-term sampling, is one of the most detailed studies of virus-host dynamics performed up to date.

My research first had to optimize the sample preparation for robust detection of RNA viruses in the marine environments. The improved viral-fraction "sample-to-sequence" workflow presented in Chapter 3 can recover > 1μg of RNA, avoiding random amplification procedures and its biases. Physical and *in-silico* rRNA depletion reduces host background and increases number of recovered viral reads. Wide phylogenetic range of RNA viruses captured suggest that this protocol broadens the access to novel and/or divergent RNA viral genomes in environmental samples. Detailed analysis of selected genomes revealed a high divergence of structural proteins, implying the existence of a broad range of hosts for these new viruses. Additionally, by obtaining 50 near-complete RNA viral genomes from one environmental sample, the developed viral-size fraction protocol proves to be robust in recovering RNA viral genomes and can be

leveraged in future studies. Optimization of protocols for metatranscriptomics of cellular-size fraction included tests of RNA recovery with different extraction approaches and careful considerations of other factors (e.g. amount of biomass filtered) to achieve good quality and quantity of RNA, essential for metatranscriptomic studies. The optimized metatranscriptomic lab protocol was used in the ecological study performed in the Johor Strait.

In chapter 4, I was able to capture a high diversity of phytoplankton-infecting RNA and giant DNA viruses with a metatranscriptomic approach. Analysis of the viral sequences revealed completely new clades of viruses, with deep phylogenetic branches, suggesting high evolutionary divergence from known viruses. Efforts like these will advance our knowledge on evolutionary origin, phylogenetic range, and ecological roles of phytoplankton viruses.

Chapter 5 further consolidated the knowledge of viral-host dynamics in Singapore's coastal waters and highlighted the potential role of viral infection in co-regulation of the biomass increase of phytoplankton community during the Northeast monsoon season. The phytoplankton community in the Johor Strait was dominated by *Thalassiosira, Chaetoceros* and *Skeletonema,* chain-forming diatoms with high growth rates that can form high biomass blooms. However, despite high levels of eutrophication, no bloom was observed during the sampling period. Pronounced spikes in diatom biomass were kept in check either by marna-like virus expression during the period of nutrient availability or by silica limitation. This provides a baseline knowledge for managing phytoplankton blooms as well as monitoring of aquaculture pathogens in the Johor Strait. Final finding coming from this body of work is the empirical evidence for the coexistence of three different infection strategies of viruses in the marine environment. RNA viruses have pronounced "bloom and bust" lifestyle, consisting of strong changes

in community composition and abundance of RNA viruses, rapidly responding to changes in host abundance. Giant DNA viruses, whose hosts are not abundant, have less lytic lifestyle and exhibit slower and lesser community changes, described here as a "gleaner" lifestyle. Another type of infection seemingly common in marine ecosystem are low key infections with persistent, vertically transmitted, RNA viruses that do not have a lytic stage in their lifecycle. By broadening our knowledge of viral life strategies, we can begin to fully appreciate the complex and dynamic relationship between viruses and their hosts in the marine environment.

6.2. Future work

To expand the work presented in this thesis and get a deeper understanding of processes that influence phytoplankton dynamics in coastal ecosystems, further data analysis of this comprehensive Johor Strait dataset should be performed. This environmental time-series dataset offers four exciting possibilities for complementing the research presented in this thesis.

1) Compare the active taxa (metatranscriptome) to the environmental reservoir (metagenome of the cellular fraction) within the giant virus community of Johor Strait. The active taxa that were explored in this thesis constituted only heterotrophic protist-infecting and phytoplankton-infecting giant viruses from the orders *Algavirales* and *Imitervirales*. A tremendous diversity of giant viruses exists outside of the above-mentioned orders (Aylward et al., 2021; Matsuyama et al., 2020; Schulz et al., 2020) and could be captured metagenomes of the cellular fraction. Giant DNA viruses that could potentially be detected in the Johor Strait metagenomes include iridoviruses, herpesviruses and asfarviruses that infect fish, bivalves, and sea snails (abalone) and have been associated with aquaculture outbreaks with high mortality rates (Kibenge, 2019; Matsuyama et al., 2020; Wei et al., 2019). This is especially important given the

high proximity of the aquaculture farms in the Johor Strait. As is always the case with environmental datasets, identification of novel giant viruses is highly possible as well.

2) Analyze the activity of both bacteria and bacteriophages in the cellular metatranscriptomes of the Johor Strait dataset, which could help to gain deeper understanding of the complex processes in this ecosystem. Based on analysis of 16S metabarcoding study that was performed concurrently with my study, the family *Saprospiraceae*, a bacterial taxon involved with the breakdown of organic matter was a key taxon in the microbial community (Wijaya et al., 2022), pointing out at high activity of the microbial loop in this ecosystem, thriving on released organic matter after viral lysis of the phytoplankton.

3) Leverage the potential of a combined long-read Nanopore sequencing and deep short-read Illumina sequencing to recover eukaryotic and giant virus metagenome-assembled genomes (MAGs). Marine microbial eukaryotes are burdensome to isolate, often completely unculturable and are severely underrepresented in culture collections. Hence, the exploration of environmental dataset, such as this, could contribute to the discovery of novel lineages of both eukaryotes and giant viruses. Despite the larger size and more complex genome architecture of eukaryotes, more than 700 eukaryotic MAGs were successfully reconstructed from the Tara Ocean dataset, with the smallest MAGs size around 10Mb (Delmont et al., 2022; Ha et al., 2021). Given the average read length of 4kb in my long read sequencing runs (as opposed to 150bp Illumina read length), inclusion of long reads should substantially improve MAG recovery as demonstrated for prokaryotes and viruses (Overholt et al., 2020).

4) Reconstruct full-length viral RNA genomes from the virome dataset. The recovery of viral RNA genomes from other ecosystems has led to the description of new genera like the recently described *Locarnavirus*, *Salisharnavirus* and *Sogarnavirus* (family

Marnaviridae), which are completely based on environmental sequences (Lang et al., 2021). Full genome annotation of all full-length genomes recovered from the Raffles Marina viral-size fraction and the examination of the Johor Strait time-series metatranscriptome assemblies for full-length genomes is yet be performed. The recovery of multiple full-length genomes by this approach could markedly expand both the known diversity of viruses and their known host ranges and result in creation of new genera, families and possibly higher taxonomic groups.

On a broader scale, future work in this field should aim to incorporate multiple sequencing approaches, focus on exploring metatranscriptomic activity of phytoplankton viruses on shorter time scales. Studies of RNA viruses in the viral-size fraction of marine samples are comparatively rare and more technically challenging than metatranscriptome studies. However, it is essential to conduct such studies to thoroughly investigate the RNA viral diversity and biogeography in the marine ecosystem. Biodiversity is fundamental for maintaining healthy and resilient marine ecosystems. Eukaryotic viruses may play important roles in controlling the dominance of a single, bloom-forming phytoplankton species, thereby preventing the loss of biodiversity and its negative impacts. In an era of climate change, eutrophication and frequent harmful phytoplankton blooms, studies like this are timely and critical for understanding the resilience of marine ecosystems in response to the ever-increasing anthropogenic pressures.

BIBLIOGRAPHY

Adriaenssens, E. M., Farkas, K., Harrison, C., Jones, D. L., Allison, H. E., & McCarthy, A. J. (2018). Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *MSystems*, *3*(3), 1–39. https://doi.org/10.1128/mSystems.00025-18

Alarcón-Schumacher, T., Guajardo-Leiva, S., Antón, J., & Llewellyn, C. A. (2019). Elucidating Viral Communities During a Phytoplankton Bloom on the West Antarctic Peninsula. *Frontiers in Microbiology*, *10*(May), 1–14. https://doi.org/10.3389/fmicb.2019.01014

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 1–16. https://doi.org/10.1186/s13059-020-1935-5

Argov, T., Azulay, G., Pasechnek, A., Stadnyuk, O., Ran-Sapir, S., Borovok, I., Sigal, N., & Herskovits, A. A. (2017). Temperate bacteriophages as regulators of host behavior. *Current Opinion in Microbiology*, *38*, 81–87. https://doi.org/10.1016/j.mib.2017.05.002

Arsenieff, L., Simon, N., Rigaut-Jalabert, F., Le Gall, F., Chaffron, S., Corre, E., Com, E., Bigeard, E., & Baudoux, A.-C. (2019). First Viruses Infecting the Marine Diatom Guinardia delicatula. *Frontiers in Microbiology*, *9*(January). https://doi.org/10.3389/fmicb.2018.03235

Attoui, H., Jaafar, F. M., Belhouchet, M., de Micco, P., de Lamballerie, X., & Brussaard, C. P. D. (2006). Micromonas pusilla reovirus: A new member of the family Reoviridae assigned to a novel proposed genus (Mimoreovirus). *Journal of General Virology*, *87*(5), 1375–1383. https://doi.org/10.1099/vir.0.81584-0

Aylward, F. O., Boeuf, D., Mende, D. R., Wood-Charlson, E. M., Vislova, A., Eppley,

J. M., Romano, A. E., & DeLong, E. F. (2017). Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11446–11451. https://doi.org/10.1073/pnas.1714821114

Aylward, F. O., & Moniruzzaman, M. (2021). Viralrecall—a flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses*, *13*(2), 15–17. https://doi.org/10.3390/v13020150

Aylward, F. O., Moniruzzaman, M., Ha, A. D., & Koonin, E. V. (2021). A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biology*, *19*(10 October), 1–18. https://doi.org/10.1371/JOURNAL.PBIO.3001430

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, *35*(3), 235–241. https://doi.org/10.1128/mmbr.35.3.235-241.1971

Baudoux, A. C., Noordeloos, A. A. M., Veldhuis, M. J. W., & Brussaard, C. P. D. (2006). Virally induced mortality of Phaeocystis globosa during two spring blooms in temperate coastal waters. *Aquatic Microbial Ecology*, *44*(3), 207–217. https://doi.org/10.3354/ame044207

Beaulaurier, J., Luo, E., Eppley, J. M., Uyl, P. Den, Dai, X., Burger, A., Turner, D. J., Pendelton, M., Juul, S., Harrington, E., & DeLong, E. F. (2020). Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Research*, *30*(3), 437–446. https://doi.org/10.1101/gr.251686.119

Behrenfeld, M. J. (2014). Climate-mediated dance of the plankton. *Nature Climate Change*, *4*(10), 880–887. https://doi.org/10.1038/nclimate2349

Behrenfeld, M. J., O'Malley, R. T., Siegel, D. A., McClain, C. R., Sarmiento, J. L.,

Feldman, G. C., Milligan, A. J., Falkowski, P. G., Letelier, R. M., & Boss, E. S. (2006). Climate-driven trends in contemporary ocean productivity. *Nature*, *444*(7120), 752–755. https://doi.org/10.1038/nature05317

Belshaw, R., Pybus, O. G., & Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research*, *17*(10), 1496–1504. https://doi.org/10.1101/gr.6305707

Beman, J. M., Arrigo, K. R., & Matson, P. A. (2005). Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean. *Nature*, *434*(7030), 211–214. https://doi.org/10.1038/nature03370

Biggs, T. E. G., Huisman, J., & Brussaard, C. P. D. (2021). Viral lysis modifies seasonal phytoplankton dynamics and carbon flow in the Southern Ocean. *ISME Journal*, *15*(12), 3615–3622. https://doi.org/10.1038/s41396-021-01033-6

Blindheim, S., Nylund, A., Watanabe, K., Plarre, H., Erstad, B., & Nylund, S. (2015). A new aquareovirus causing high mortality in farmed Atlantic halibut fry in Norway. *Archives of Virology*, *160*(1), 91–102. https://doi.org/10.1007/s00705-014-2235-8

Blouin, A. G., Ross, H. A., Hobson-Peters, J., O'Brien, C. A., Warren, B., & MacDiarmid, R. (2016). A new virus discovered by immunocapture of double-stranded RNA, a rapid method for virus enrichment in metagenomic studies. *Molecular Ecology Resources*, *16*(5), 1255–1263. https://doi.org/10.1111/1755-0998.12525

Brum, J. R., Sullivan, M. B., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., & Chaffron, S. (2015). Ocean Viral Communities. *Science*, *348*(6237), 1261498-1–11. https://doi.org/10.1126/science.1261498

Brussaard, C. P. D. (2004a). Optimization of Procedures for Counting Viruses by

Flow Cytometry Optimization of Procedures for Counting Viruses by Flow Cytometry. *Applied and Environmental Microbiology*, *70*(3), 1506–1513. https://doi.org/10.1128/AEM.70.3.1506

Brussaard, C. P. D. (2004b). Viral control of phytoplankton populations--a review. *The Journal of Eukaryotic Microbiology*, *51*(2), 125–138. https://doi.org/10.1111/j.1550-7408.2004.tb00537.x

Brussaard, C. P. D., Marie, D., & Bratbak, G. (2000). Flow cytometric detection of viruses. *Journal of Virological Methods*, *85*(1–2), 175–182. http://www.ncbi.nlm.nih.gov/pubmed/10716350

Brussaard, C., & Payet, J. (2010). Quantification of aquatic viruses by flow cytometry. *Manual of Aquatic Viral Ecology*, *2004*, 102–109. https://doi.org/10.4319/mave.2010.978-0-9845591-0-7.102

Bujarski, J., Gallitelli, D., Fernando, García-Arenal Vicente, P., Palukaitis, P., Reddy, M. K., & Wang, A. (2019). Virus Taxonomy The ICTV Report on Virus Classification and Taxon Nomenclature Asfarviridae Chapter Asfarviridae Virion Morphology Physicochemical and physical properties. *Journal of General Virology*, *100*, 1206–1207. https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001332

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology and Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), 1–13. https://doi.org/10.1093/gigascience/giz100

Callanan, J., Stockdale, S. R., Shkoporov, A., Draper, L. A., Ross, R. P., & Hill, C.

(2020). Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances*, *6*(6). https://doi.org/10.1126/sciadv.aay5981

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M. A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02342-1

Chappell, J. D., & Dermody, T. S. (2015). Biology of Viruses and Viral Diseases. In *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* (pp. 1681-1693.e4). https://doi.org/10.1016/B978-1-4557-4801-3.00134-X

Charon, J., Marcelino, V. R., Wetherbee, R., Verbruggen, H., & Holmes, E. (2020a). *Meta-transcriptomic detection of diverse and divergent RNA viruses in green and chlorarachniophyte algae*. 1–54. https://doi.org/10.1101/2020.06.08.141184

Charon, J., Marcelino, V. R., Wetherbee, R., Verbruggen, H., & Holmes, E. C. (2020b). Metatranscriptomic identification of diverse and divergent RNA viruses in green and chlorarachniophyte algae cultures. *Viruses*, *12*(10), 1–27. https://doi.org/10.3390/v12101180

Charon, J., Murray, S., & Holmes, E. C. (2021),Revealing RNA virus diversity and evolution in unicellular algae transcriptomes. *Virus Evolution*, *7*(2), 1–18. https://doi.org/10.1093/ve/veab070

Chen, J., Toh, X., Ong, J., Wang, Y., Teo, X. H., Lee, B., Wong, P. S., Khor, D., Chong, S. M., Chee, D., Wee, A., Wang, Y., Ng, M. K., Tan, B. H., & Huangfu, T. (2019). Detection and characterization of a novel marine birnavirus isolated from Asian seabass in Singapore. *Virology Journal*, *16*(1), 1–10.

https://doi.org/10.1186/s12985-019-1174-0

Chénard, C., Wijaya, W., Vaulot, D., Lopes dos Santos, A., Martin, P., Kaur, A., & Lauro, F. M. (2019). Temporal and spatial dynamics of Bacteria, Archaea and protists in equatorial coastal waters. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-52648-x

Cheng, D.-Q., Kolundžija, S., & Lauro, F. M. (2022). Global phylogenetic analysis of the RNA-dependent RNA polymerase with OrViT (OrthornaVirae Tree). *Frontiers in Virology*, *2*(September), 1–9. https://doi.org/10.3389/fviro.2022.981177

Chiba, Y., Tomaru, Y., Shimabukuro, H., Kimura, K., Hirai, M., Takaki, Y., Hagiwara, D., Nunoura, T., & Urayama, S. I. (2020). Viral rna genomes identified from marine macroalgae and a diatom. *Microbes and Environments*, *35*(3), 1–8. https://doi.org/10.1264/jsme2.ME20016

Claverie, J. M., & Abergel, C. (2018). Mimiviridae: An expanding family of highly diverse large dsDNA viruses infecting a wide phylogenetic range of aquatic eukaryotes. *Viruses*, *10*(9), 8–15. https://doi.org/10.3390/v10090506

Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., & Mahar, J. E. (2021). Current challenges to virus discovery by meta-transcriptomics. *Current Opinion in Virology*, *51*, 48–55. https://doi.org/10.1016/j.coviro.2021.09.007

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Ranst, M. Van, Heylen, E., & Matthijnssens, J. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. *Scientific Reports*, *5*(October), 1–14. https://doi.org/10.1038/srep16532

Corinaldesi, C., Tangherlini, M., & Dell'anno, A. (2017). From virus isolation to

metagenome generation for investigating viral diversity in deep-sea sediments. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-08783-4

Coy, S. R., Gann, E. R., Pound, H. L., Short, S. M., & Wilhelm, S. W. (2018). Viruses of eukaryotic algae: Diversity, methods for detection, and future directions. *Viruses*, *10*(9). https://doi.org/10.3390/v10090487

Cruaud, P., Vigneron, A., Fradette, M. S., Charette, S. J., Rodriguez, M. J., Dorea, C. C., & Culley, A. I. (2017). Open the SterivexTM casing: An easy and effective way to improve DNA extraction yields. *Limnology and Oceanography: Methods*, *15*(12), 1015–1020. https://doi.org/10.1002/lom3.10221

Culley, A. (2018). New insight into the RNA aquatic virosphere via viromics. *Virus Research*, *244*, 84–89. https://doi.org/10.1016/j.virusres.2017.11.008

Culley, A. I., Lang, A. S., & Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science*, *312*(5781), 1795–1798. https://doi.org/10.1126/science.1127404

Culley, A. I., Lang, A. S., & Suttle, C. A. (2007). The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virology Journal*, *4*, 69. https://doi.org/10.1186/1743-422X-4-69

Culley A. I., Mueller, J. A., Belcaid, M., Wood-Charlson, E. M., Poisson, G., & Steward, G. F. (2014). The Characterization of RNA Viruses in Tropical Seawater Using Targeted PCR and Metagenomics. *MBio*, *5*(3), 1–11. https://doi.org/10.1128/mbio.01210-14

Danovaro, R., Corinaldesi, C., Rastelli, E., & Dell'Anno, A. (2015). Towards a better quantitative assessment of the relevance of deep-sea viruses, Bacteria and Archaea in the functioning of the ocean seafloor. *Aquatic Microbial Ecology*, *75*(1), 81–90. https://doi.org/10.3354/ame01747

Danovaro, Roberto, Corinaldesi, C., Dell'Anno, A., Fuhrman, J. A., Middelburg, J. J., Noble, R. T., & Suttle, C. A. (2011). Marine viruses and global climate change. *FEMS Microbiology Reviews*, *35*(6), 993–1034. https://doi.org/10.1111/j.1574-6976.2010.00258.x

Danovaro, Roberto, Dell'anno, A., Corinaldesi, C., Rastelli, E., Cavicchioli, R., Krupovic, M., Noble, R. T., Nunoura, T., & Prangishvili, D. (2016). Virus-mediated archaeal hecatomb in the deep seafloor. *Science Advances*, *2*(10), 1–9. https://doi.org/10.1126/sciadv.1600492

Davenport, E. J., Neudeck, M. J., Matson, P. G., Bullerjahn, G. S., Davis, T. W., Wilhelm, S. W., Denney, M. K., Krausfeldt, L. E., Stough, J. M. A., Meyer, K. A., Dick, G. J., Johengen, T. H., Lindquist, E., Tringe, S. G., & McKay, R. M. L. (2019). Metatranscriptomic Analyses of Diel Metabolic Functions During a Microcystis Bloom in Western Lake Erie (United States). *Frontiers in Microbiology*, *10*(September), 1–16. https://doi.org/10.3389/fmicb.2019.02081

De Corte, D., Martínez, J. M., Cretoiu, M. S., Takaki, Y., Nunoura, T., Sintes, E., Herndl, G. J., & Yokokawa, T. (2019). Viral communities in the global deep ocean conveyor belt assessed by targeted viromics. *Frontiers in Microbiology*, *10*(AUG), 1–14. https://doi.org/10.3389/fmicb.2019.01801

Delmas, B., Attoui, H., Ghosh, S., Malik, Y. S., Mundt, E., & Vakharia, V. N. (2019a). ICTV virus taxonomy profile: Birnaviridae. *Journal of General Virology*, *100*(1), 5–6. https://doi.org/10.1099/jgv.0.001185

Delmas, B., Attoui, H., Ghosh, S., Malik, Y. S., Mundt, E., & Vakharia, V. N. (2019b). Ictv virus taxonomy profile: Picobirnaviridae. *Journal of General Virology*, *100*(2), 133–134. https://doi.org/10.1099/jgv.0.001186

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra,

A., Eren, A. M., Kourlaiev, A., D'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., de Vargas, C., … Speich, S. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, *2*(5), 100123. https://doi.org/10.1016/j.xgen.2022.100123

Depledge, D. P., Mohr, I., & Wilson, A. C. (2019). Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *Journal of Virology*, *93*(1). https://doi.org/10.1128/jvi.01342-18

Derelle, E., Yau, S., Moreau, H., & Grimsley, N. H. (2018). Prasinovirus Attack of Ostreococcus Is Furtive by Day but Savage by Night. *Journal of Virology*, *92*(4). https://doi.org/10.1128/JVI.01703-17

Dolja, V. V., & Koonin, E. V. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Research*, *244*(August 2017), 36–52. https://doi.org/10.1016/j.virusres.2017.10.020

Dominguez-Huerta, G., Zayed, A. A., Wainaina, J. M., Guo, J., Tian, F., Pratama, A. A., Bolduc, B., Mohssen, M., Zablocki, O., Pelletier, E., Delage, E., Alberti, A., Aury, J. M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., Bowler, C., Eveillard, D., … Sullivan, M. B. (2022). Diversity and ecological footprint of Global Ocean RNA viruses. *Science*, *376*(6598), 1202–1208. https://doi.org/10.1126/science.abn6358

Dong, X., Hu, T., Liu, Q., Li, C., Sun, Y., Wang, Y., Shi, W., Zhao, Q., & Huang, J. (2020). A Novel Hepe-Like Virus from Farmed Giant Freshwater Prawn Macrobrachium rosenbergii. *Viruses*, *12*(3), 1–7. https://doi.org/10.3390/v12030323

Duhaime, M. B., & Sullivan, M. B. (2012). Ocean viruses: Rigorously evaluating the

metagenomic sample-to-sequence pipeline. *Virology*, *434*(2), 181–186. https://doi.org/10.1016/j.virol.2012.09.036

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

Edwards, K. F., Steward, G. F., & Schvarcz, C. R. (2021). Making sense of virus size and the tradeoffs shaping viral fitness. *Ecology Letters*, *24*(2), 363–373. https://doi.org/10.1111/ele.13630

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., de Vargas, C., Sullivan, M. B., Bowler, C., Wincker, P., Karp-Boss, L., Sunagawa, S., & Ogata, H. (2020). Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology and Evolution*. https://doi.org/10.1038/s41559-020-01288-w

Er, H. H., Lee, L. K., Lim, Z. F., Teng, S. T., Leaw, C. P., & Lim, P. T. (2018). Responses of phytoplankton community to eutrophication in Semerak Lagoon (Malaysia). *Environmental Science and Pollution Research*, *25*(23), 22944–22962. https://doi.org/10.1007/s11356-018-2389-0

Eynde, B. Van, Christiaens, O., Delbare, D., Shi, C., Vanhulle, E., Yinda, C. K., Matthijnssens, J., & Smagghe, G. (2020). Exploration of the virome of the European brown shrimp (Crangon crangon). *Journal of General Virology*, *101*(6), 651–666. https://doi.org/10.1099/jgv.0.001412

Eyngor, M., Zamostiano, R., Tsofack, J. E. K., Berkowitz, A., Bercovier, H., Tinman, S., Lev, M., Hurvitz, A., Galeotti, M., Bacharach, E., & Eldar, A. (2014).

Identification of a novel RNA virus lethal to tilapia. *Journal of Clinical Microbiology*, *52*(12), 4137–4146. https://doi.org/10.1128/JCM.00827-14

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, *320*(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, *281*(5374), 237–240. https://doi.org/10.1126/science.281.5374.237

Frith, M. C. (2011). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*, *39*(4). https://doi.org/10.1093/nar/gkq1212

Fuhrman, J. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(0), 541–548.

Gaia, M., Benamar, S., Boughalmi, M., Pagnier, I., Croce, O., Colson, P., Raoult, D., & La Scola, B. (2014). Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS ONE*, *9*(4), 1–8. https://doi.org/10.1371/journal.pone.0094923

Gann, E. R., Kang, Y., Dyhrman, S. T., Gobler, C. J., & Wilhelm, S. W. (2021). Metatranscriptome Library Preparation Influences Analyses of Viral Community Activity During a Brown Tide Bloom. *Frontiers in Microbiology*, *12*(May). https://doi.org/10.3389/fmicb.2021.664189

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., … Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, *15*(3), 201–206. https://doi.org/10.1038/nmeth.4577

Gastrich, M. D., Leigh-Bell, J. A., Gobler, C. J., Anderson, O. R., Wilhelm, S. W., & Bryan, M. (2004). Viruses as potential regulators of regional brown tide blooms caused by the alga, Aureococcus anophagefferens. *Estuaries*, *27*(1), 112–119. https://doi.org/10.1007/BF02803565

Geoghegan, J. L., & Holmes, E. C. (2017). Predicting virus emergence amid evolutionary noise. *Open Biology*, *7*(10). https://doi.org/10.1098/rsob.170189

Gerashchenko, M. V., & Gladyshev, V. N. (2017). Ribonuclease selection for ribosome profiling. *Nucleic Acids Research*, *45*(2), e6. https://doi.org/10.1093/nar/gkw822

Ghabrial, S. A., & Suzuki, N. (2009). Viruses of plant pathogenic fungi. *Annual Review of Phytopathology*, *47*, 353–384. https://doi.org/10.1146/annurev-phyto-080508-081932

Gilbert, P. M. M., & Burford, M. A. A. (2017). Globally Changing Nutrient Loads and Harmful Algal Blooms. *Oceanography*, *30*(1), 58–69. https://doi.org/10.5670/oceanog.2017.110%0Ahttps://tos.org/oceanography/assets/docs/30-1_glibert.pdf

Gilpin, L. C., Davidson, K., & Roberts, E. (2004). The influence of changes in nitrogen: Silicon ratios on diatom growth dynamics. *Journal of Sea Research*, *51*(1), 21–35. https://doi.org/10.1016/j.seares.2003.05.005

Gin, K. Y. H., Holmes, M. J., Zhang, S., & Lin, X. (2006). Phytoplankton structure in the tropical port waters of Singapore. *The Environment in Asia Pacific Harbours*, 347–375. https://doi.org/10.1007/1-4020-3655-8_21

Gin, K. Y. H., Lin, X., & Zhang, S. (2000). Dynamics and size structure of phytoplankton in the coastal waters of Singapore. *Journal of Plankton Research*, *22*(8), 1465–1484. https://doi.org/10.1093/plankt/22.8.1465

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., … Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, *5*(5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Gorbalenya, A. E., & Lauber, C. (2022). Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. *Current Opinion in Virology*, *52*, 48–56. https://doi.org/10.1016/j.coviro.2021.11.003

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Roux, S. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, *177*(5), 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040

Greninger, A. L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Research*, *244*(July 2017), 218–229. https://doi.org/10.1016/j.virusres.2017.10.014

Gustavsen, J. A., Winget, D. M., Tian, X., & Suttle, C. A. (2014). High temporal and spatial diversity in marine RNA viruses implies that they have an important role in mortality and structuring plankton communities. *Frontiers in Microbiology*, *5*(DEC), 1–13. https://doi.org/10.3389/fmicb.2014.00703

Ha, A. D., Moniruzzaman, M., & Aylward, F. O. (2021). High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a

Coastal Marine System. *MSystems*, *6*(4). https://doi.org/10.1128/msystems.00293-21

Hassim, M. E. E., & Timbal, B. (2019). Observed rainfall trends over singapore and the Maritime Continent from the perspective of regional-scale weather regimes. *Journal of Applied Meteorology and Climatology*, *58*(2), 365–384. https://doi.org/10.1175/JAMC-D-18-0136.1

Herbert W. Virgin, M. D. (2014). The virome in mammalian physiology and disease. *Cell*, *157*(1), 142–150. https://doi.org/10.1016/j.cell.2014.02.032.The

Hevroni, G., Flores-Uribe, J., Béjà, O., & Philosof, A. (2020). Seasonal and diel patterns of abundance and activity of viruses in the Red Sea. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29738–29747. https://doi.org/10.1073/pnas.2010783117

Hewson, I., Bistolas, K. S. I., Button, J. B., & Jackson, E. W. (2018). Occurrence and seasonal dynamics of RNA viral genotypes in three contrasting temperate lakes. *PLoS ONE*, *13*(3), 1–19. https://doi.org/10.1371/journal.pone.0194419

Hii, K. S., Mohd-Din, M., Luo, Z., Tan, S. N., Lim, Z. F., Lee, L. K., Leong, S. C. Y., Teng, S. T., Gu, H., Cao, X., Lim, P. T., & Leaw, C. P. (2021). Diverse harmful microalgal community assemblages in the Johor Strait and the environmental effects on its community dynamics. *Harmful Algae*, *107*(July), 102077. https://doi.org/10.1016/j.hal.2021.102077

Hillary, L. S., Adriaenssens, E. M., Jones, D. L., & McDonald, J. E. (2022). RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Communications*, *2*(1), 1–10. https://doi.org/10.1038/s43705-022-00110-x

Hillman, B. I., & Cai, G. (2013). The Family Narnaviridae. Simplest of RNA Viruses.

In *Advances in Virus Research* (1st ed., Vol. 86). Copyright © 2013, Elsevier Inc. All Rights Reserved. https://doi.org/10.1016/B978-0-12-394315-6.00006-4

Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmento, H., Villar, E., Lima-Mendez, G., Faust, K., Sunagawa, S., Claverie, J. M., Moreau, H., Desdevises, Y., Bork, P., Raes, J., De Vargas, C., Karsenti, E., … Ogata, H. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME Journal*, *7*(9), 1678–1695. https://doi.org/10.1038/ismej.2013.59

Hjelmsø, M. H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., Elsässer, D., Aarestrup, F. M., Löfström, C., Bofill-Mas, S., Abril, J. F., Girones, R., & Schultz, A. C. (2017). Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PLoS ONE*, *12*(1), 1–17. https://doi.org/10.1371/journal.pone.0170199

Holmes, E. C. (2003). Error thresholds and the constraints to RNA virus evolution. *Trends in Microbiology*, *11*(12), 543–546. https://doi.org/10.1016/j.tim.2003.10.006

Hurwitz, B. L., Deng, L., Poulos, B. T., & Sullivan, M. B. (2013). Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology*, *15*(5), 1428–1440. https://doi.org/10.1111/j.1462-2920.2012.02836.x

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-

2105-11-119

Iker, B. C., Bright, K. R., Pepper, I. L., Gerba, C. P., & Kitajima, M. (2013). Evaluation of commercial kits for the extraction and purification of viral nucleic acids from environmental and fecal samples. *Journal of Virological Methods*, *191*(1), 24–30. https://doi.org/10.1016/j.jviromet.2013.03.011

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

Jacquet, S., Heldal, M., Iglesias-Rodriguez, D., Larsen, A., Wilson, W., & Bratbak, G. (2002). Flow cytometric analysis of an Emiliana huxleyi bloom terminated by viral infection. *Aquatic Microbial Ecology*, *27*, 111–124.

Jain, M., Abu-Shumays, R., Olsen, H. E., & Akeson, M. (2022). Advances in nanopore direct RNA sequencing. *Nature Methods*, *19*(10), 1160–1164. https://doi.org/10.1038/s41592-022-01633-w

Johannessen, T. V., Larsen, A., Bratbak, G., Pagarete, A., Edvardsen, B., Egge, E. D., & Sandaa, R. A. (2017). Seasonal dynamics of haptophytes and dsDNA algal viruses suggest complex virus-host relationship. *Viruses*, *9*(4). https://doi.org/10.3390/v9040084

John, S. G., Mendez, C. B., Deng, L., Poulos, B., Kauffman, A. K. M., Kern, S., Brum, J., Polz, M. F., Boyle, E. A, & Sullivan, M. B. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports*, *3*(2), 195–202. https://doi.org/10.1111/j.1758-2229.2010.00208.x

Kaletta, J., Pickl, C., Griebler, C., Klingl, A., Kurmayer, R., & Deng, L. (2020). A rigorous assessment and comparison of enumeration methods for environmental

viruses. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-75490-y

Kaneko, H., Blanc-Mathieu, R., Endo, H., Chaffron, S., Delmont, T. O., Gaia, M., Henry, N., Hernández-Velázquez, R., Nguyen, C. H., Mamitsuka, H., Forterre, P., Jaillon, O., de Vargas, C., Sullivan, M. B., Suttle, C. A., Guidi, L., & Ogata, H. (2021). Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *IScience*, *24*(1). https://doi.org/10.1016/j.isci.2020.102002

Karlsson, O. E., Belák, S., & Granberg, F. (2013). The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecurity and Bioterrorism : Biodefense Strategy, Practice, and Science*, *11 Suppl 1*, S227-34. https://doi.org/10.1089/bsp.2013.0008

Kibenge, F. S. (2019). Emerging viruses in aquaculture. *Current Opinion in Virology*, *34*, 97–103. https://doi.org/10.1016/j.coviro.2018.12.008

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Knowles, B., Silveira, C. B., Bailey, B. a., Barott, K., Cantu, V. a., Cobián-Güemes, a. G., Coutinho, F. H., Dinsdale, E. a., Felts, B., Furby, K. a., George, E. E., Green, K. T., Gregoracci, G. B., Haas, a. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., … Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*. https://doi.org/10.1038/nature17193

Kohl, C., Brinkmann, A., Dabrowski, P. W., Radonić, A., Nitsche, A., & Kurth, A. (2015). Protocol for metagenomic virus detection in clinical specimens.

Emerging Infectious Diseases, *21*(1), 48–57. https://doi.org/10.3201/eid2101.140766

Kok, J. W. K., & Leong, S. C. Y. (2019). Nutrient conditions and the occurrence of a Karenia mikimotoi (Kareniaceae) bloom within East Johor Straits, Singapore. *Regional Studies in Marine Science*, *27*, 100514. https://doi.org/10.1016/j.rsma.2019.100514

Kolody, B. C., McCrow, J. P., Allen, L. Z., Aylward, F. O., Fontanez, K. M., Moustafa, A., Moniruzzaman, M., Chavez, F. P., Scholin, C. A., Allen, E. E., Worden, A. Z., Delong, E. F., & Allen, A. E. (2019). Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *ISME Journal*, *13*(11), 2817–2833. https://doi.org/10.1038/s41396-019-0472-2

Kolundžija, S., Cheng, D.-Q., & Lauro, F. M. (2022). RNA Viruses in Aquatic Ecosystems through the Lens of Ecological Genomics and Transcriptomics. *Viruses*, *14*(4). https://doi.org/10.3390/v14040702

Koonin, E. V., & Yutin, N. (2019). Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. In *Advances in Virus Research* (1st ed., Vol. 103). Elsevier Inc. https://doi.org/10.1016/bs.aivir.2018.09.002

Koonin, E. V, Dolja, V. V, Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2), e00061-19.

Koonin, E. V, Krupovic, M., & Agol, V. I. (2021). The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution?

Microbiology and Molecular Biology Reviews : MMBR, *85*(3), e0005321. https://doi.org/10.1128/MMBR.00053-21

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

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295. https://doi.org/10.1038/nmeth.1311

Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, *239*, 136–142. https://doi.org/10.1016/j.virusres.2017.02.002

Krishnamurthy, S. R., & Wang, D. (2018). Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology*, *516*(November 2017), 108–114. https://doi.org/10.1016/j.virol.2018.01.006

Krupovic, M., Dolja, V. V., & Koonin, E. V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology*, *17*(7), 449–458. https://doi.org/10.1038/s41579-019-0205-6

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., & Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, *455*(7209), 100–104. https://doi.org/10.1038/nature07218

Laber, C. P., Hunter, J. E., Carvalho, F., Collins, J. R., Hunter, E. J., Schieler, B. M., Boss, E., More, K., Frada, M., Thamatrakoln, K., Brown, C. M., Haramaty, L., Ossolinski, J., Fredricks, H., Nissimov, J. I., Vandzura, R., Sheyn, U., Lehahn, Y., Chant, R. J., … Bidle, K. D. (2018). Coccolithovirus facilitation of carbon

export in the North Atlantic. *Nature Microbiology*, *3*(5), 537–547. https://doi.org/10.1038/s41564-018-0128-4

Labonté, J. M., & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME Journal*, *7*(11), 2169–2177. https://doi.org/10.1038/ismej.2013.110

Lang, A. S., Rise, M. L., Culley, A. I., & Steward, G. F. (2009). RNA viruses in the sea. *FEMS Microbiology Reviews*, *33*(2), 295–323. https://doi.org/10.1111/j.1574-6976.2008.00132.x

Lang, A. S., Vlok, M., Culley, A. I., Suttle, C. A., Takao, Y., Tomaru, Y., Siddell, S. G., Lefkowitz, E., Simmonds, P., Zerbini, F. M., Smith, D. B., Orton, R., & Sabanadzovic, S. (2021). ICTV virus taxonomy profile: Marnaviridae 2021. *Journal of General Virology*, *102*(8), 1–2. https://doi.org/10.1099/JGV.0.001633

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

Lau, W. L. S., Law, I. K., Liow, G. R., Hii, K. S., Usup, G., Lim, P. T., & Leaw, C. P. (2017). Life-history stages of natural bloom populations and the bloom dynamics of a tropical Asian ribotype of Alexandrium minutum. *Harmful Algae*, *70*, 52–63. https://doi.org/10.1016/j.hal.2017.10.006

Lauro, F. M., Senstius, S. J., Cullen, J., Neches, R., Jensen, R. M., Brown, M. V., Darling, A. E., Givskov, M., McDougald, D., Hoeke, R., Ostrowski, M., Philip, G. K., Paulsen, I. T., & Grzymski, J. J. (2014). The Common Oceanographer: Crowdsourcing the Collection of Oceanographic Data. *PLoS Biology*, *12*(9). https://doi.org/10.1371/journal.pbio.1001947

Le Gall, O., Christian, P., Fauquet, C. M., King, A. M. Q., Knowles, N. J., Nakashima, N., Stanway, G., & Gorbalenya, A. E. (2008). Picornavirales, a proposed order of

positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. *Archives of Virology*, *153*(4), 715–727. https://doi.org/10.1007/s00705-008-0041-x

Letunic, I., & Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), 256–259. https://doi.org/10.1093/nar/gkz239

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, L., Deng, X., Mee, E. T., Collot-Teixeira, S., Anderson, R., Schepelmann, S., Minor, P. D., & Delwart, E. (2015). Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *Journal of Virological Methods*, *213*, 139–146. https://doi.org/10.1016/j.jviromet.2014.12.002

Li, Z., Pan, D., Wei, G., Pi, W., Zhang, C., Wang, J. H., Peng, Y., Zhang, L., Wang, Y., Hubert, C. R. J., & Dong, X. (2021). Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME Journal*, *15*(8), 2366–2378. https://doi.org/10.1038/s41396-021-00932-y

Liang, G., & Bushman, F. D. (2021). The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology*, *19*(8), 514–527. https://doi.org/10.1038/s41579-021-00536-5

Lightner, D. V. (2011). Virus diseases of farmed shrimp in the Western Hemisphere (the Americas): a review. *Journal of Invertebrate Pathology*, *106*(1), 110–130. https://doi.org/10.1016/j.jip.2010.09.012

Lim, H. C., Leaw, C. P., Tan, T. H., Kon, N. F., Yek, L. H., Hii, K. S., Teng, S. T.,

Razali, R. M., Usup, G., Iwataki, M., & Lim, P. T. (2014). A bloom of Karlodinium australe (Gymnodiniales, Dinophyceae) associated with mass mortality of cage-cultured fishes in West Johor Strait, Malaysia. *Harmful Algae*, *40*, 51–62. https://doi.org/10.1016/j.hal.2014.10.005

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., & Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, *438*(7064), 86–89. https://doi.org/10.1038/nature04111

Liu, D., & Graber, J. H. (2006). Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-77

Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R., & DeLong, E. F. (2020). Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME Journal*, *14*(5), 1304–1315. https://doi.org/10.1038/s41396-020-0604-8

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., De Vargas, C., Bittner, L., Zingone, A., & Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(11), E1516–E1525. https://doi.org/10.1073/pnas.1509523113

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

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S. W., & Wommack, K. E. (2014). Caught in the middle with multiple displacement

amplification: The myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome*, *2*(1), 1–8. https://doi.org/10.1186/2049-2618-2-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*. https://doi.org/10.14806/ej.17.1.200

Matsuyama, T., Takano, T., Nishiki, I., Fujiwara, A., Kiryu, I., Inada, M., Sakai, T., Terashima, S., Matsuura, Y., Isowa, K., & Nakayasu, C. (2020). A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-61492-3

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11257

Miranda, J. A., Culley, A. I., Schvarcz, C. R., & Steward, G. F. (2016). RNA viruses as major contributors to Antarctic virioplankton. *Environmental Microbiology*, *18*(11), 3714–3727. https://doi.org/10.1111/1462-2920.13291

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mohd-Din, M., Hii, K. S., Abdul-Wahab, M. F., Mohamad, S. E., Gu, H., Leaw, C. P., & Lim, P. T. (2022). Spatial-temporal variability of microphytoplankton assemblages including harmful microalgae in a tropical semi-enclosed strait (Johor Strait, Malaysia). *Marine Environmental Research*, *175*(February), 105589. https://doi.org/10.1016/j.marenvres.2022.105589

Mohd-Din, M., Mohd, &, Abdul-Wahab, F., Mohamad, S. E., Jamaluddin, H., Shahir, S., Ibrahim, Z., Kieng, &, Hii, S., Suh, &, Tan, N., Chui, &, Leaw, P., Gu, H., Po, &, & Lim, T. (2020). Prolonged high biomass diatom blooms induced formation of hypoxic-anoxic zones in the inner part of Johor Strait. *Environmental Science and Pollution Research*, *27*, 42948–42959. https://doi.org/10.1007/s11356-020-10184-6

Mojica, K. D. A., Huisman, J., Wilhelm, S. W., & Brussaard, C. P. D. (2016). Latitudinal variation in virus-induced mortality of phytoplankton across the North Atlantic Ocean. *ISME Journal*, *10*(2), 500–513. https://doi.org/10.1038/ismej.2015.130

Monalisha, K., R. Bharathi, R., & Gayatri, T. (2020). Nidoviruses in Aquatic Organisms - Paradigm of a Nascent Concern. *International Journal of Aquaculture and Fishery Sciences*, *6*, 082–088. https://doi.org/10.17352/2455-8400.000061

Moniruzzaman, M., Gann, E. R., LeCleir, G. R., Kang, Y., Gobler, C. J., & Wilhelm, S. W. (2016). Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, Aureococcus anophagefferens. *FEMS Microbiology Ecology*, *92*(5), 1–10. https://doi.org/10.1093/femsec/fiw058

Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R., & Aylward, F. O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-15507-2

Moniruzzaman, M., Wurch, L. L., Alexander, H., Dyhrman, S. T., Gobler, C. J., & Wilhelm, S. W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nature Communications*, *8*(May),

1–10. https://doi.org/10.1038/ncomms16054

Moya, A., Holmes, E. C., & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology*, *2*(4), 279–288. https://doi.org/10.1038/nrmicro863

Mueller, J. A., Culley, A. I., & Steward, G. F. (2014). Variables influencing extraction of nucleic acids from microbial plankton (viruses, bacteria, and protists) collected on nanoporous aluminum oxide filters. *Applied and Environmental Microbiology*, *80*(13), 3930–3942. https://doi.org/10.1128/AEM.00245-14

Munke, A., Kimura, K., Tomaru, Y., & Okamoto, K. (2020). Capsid Structure of a Marine Algal Virus of the Order Picornavirales . *Journal of Virology*, *94*(9), 1–20. https://doi.org/10.1128/jvi.01855-19

Mushegian, A. R. (2020). Are there 1031virus particles on earth, or more, or fewer? *Journal of Bacteriology*, *202*(9), 18–22. https://doi.org/10.1128/JB.00052-20

Mutsafi, Y., Fridmann-Sirkis, Y., Milrot, E., Hevroni, L., & Minsky, A. (2014). Infection cycles of large DNA viruses: Emerging themes and underlying questions. *Virology*, *466–467*, 3–14. https://doi.org/10.1016/j.virol.2014.05.037

Nagasaki, K., & Bratbak, G. (2010). Isolation of viruses infecting photosynthetic and nonphotosynthetic protists. *Manual of Aquatic Viral Ecology*, 92–101. https://doi.org/10.4319/mave.2010.978-0-9845591-0-7.92

Nasir, A., Forterre, P., Kim, K. M., & Caetano-Anollés, G. (2014). The distribution and impact of viral lineages in domains of life. *Frontiers in Microbiology*, *5*(APR), 1–5. https://doi.org/10.3389/fmicb.2014.00194

Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., Proal, A. D., Fischbach, M. A., Bhatt, A. S., Hugenholtz, P., & Kyrpides, N. C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut

microbiome. *Nature Microbiology*, *6*(7), 960–970. https://doi.org/10.1038/s41564-021-00928-6

Needham, D. M., Poirier, C., Hehenberger, E., Jiménez, V., Swalwell, J. E., Santoro, A. E., & Worden, A. Z. (2019). Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1786). https://doi.org/10.1098/rstb.2019.0086

Needham, D. M., Yoshizawa, S., Hosaka, T., Poirier, C., Choi, C. J., Hehenberger, E., Irwin, N. A. T., Wilken, S., Yung, C. M., Bachy, C., Kurihara, R., Nakajima, Y., Kojima, K., Kimura-Someya, T., Leonard, G., Malmstrom, R. R., Mende, D. R., Olson, D. K., Sudo, Y., … Worden, A. Z. (2019). A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(41), 20574–20583. https://doi.org/10.1073/pnas.1907517116

Neri, U., Wolf, Y. I., Roux, S., Camargo, A. P., Lee, B. D., Kazlauskas, D., Chen, I. M., Ivanova, N., Allen, L. Z., Paez-Espino, D., Bryant, D. A., Bhaya, D., Consortium, R. V. D., Krupovic, M., Dolja, V. V., Kyrpides, N. C., Koonin, E., & Gophna, U. (2022). A Five-Fold Expansion of the Global RNA Virome Reveals Multiple New Clades of RNA Bacteriophages. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4047248

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x

Overholt, W. A., Hölzer, M., Geesink, P., Diezel, C., Marz, M., & Küsel, K. (2020).

Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environmental Microbiology*, *22*(9), 4000–4013. https://doi.org/10.1111/1462-2920.15186

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425–430. https://doi.org/10.1038/nature19094

Palani, S., Balasubramanian, R., & Tkalich, P. (2012). A 3-D Model on the Possible Role of Atmospheric Deposition in Tropical Coastal Eutrophication. *Marine Science*, 11–21.

Patel, A., Noble, R. T., Steele, J. A., Schwalbach, M. S., Hewson, I., & Fuhrman, J. A. (2007). Virus and prokaeyote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols*, *7*. https://doi.org/10.1039/nprot.2007.006

Payne, S. (2017). Introduction to Animal Viruses. In *Viruses* (pp. 1–11). https://doi.org/10.1016/B978-0-12-803109-4.00001-5

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.-M., & Abergel, C. (2013). Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science*, *341*(6143), 281–286. https://doi.org/10.1126/science.1239181

Polinski, M. P., Vendramin, N., Cuenca, A., & Garver, K. A. (2020). Piscine orthoreovirus: Biology and distribution in farmed and wild fish. *Journal of Fish Diseases*, *43*(11), 1331–1352. https://doi.org/10.1111/jfd.13228

Pound, H. L., Gann, E. R., Tang, X., Krausfeldt, L. E., Huff, M., Staton, M. E., Talmy,

D., & Wilhelm, S. W. (2020). The “Neglected Viruses” of Taihu: Abundant Transcripts for Viruses Infecting Eukaryotes and Their Potential Role in Phytoplankton Succession. *Frontiers in Microbiology*, *11*(April). https://doi.org/10.3389/fmicb.2020.00338

Prakash, A., Jeffryes, M., Bateman, A., & Finn, R. D. (2017). The HMMER Web Server for Protein Sequence Similarity Search. *Current Protocols in Bioinformatics*, *60*(1), 3.15.1-3.15.23. https://doi.org/10.1002/cpbi.40

Prangishvili, D., Bamford, D. H., Forterre, P., Iranzo, J., Koonin, E. V., & Krupovic, M. (2017). The enigmatic archaeal virosphere. *Nature Reviews Microbiology*, *15*(12), 724–739. https://doi.org/10.1038/nrmicro.2017.125

Reche, I., D’Orta, G., Mladenov, N., Winget, D. M., & Suttle, C. A. (2018). Deposition rates of viruses and bacteria above the atmospheric boundary layer. *ISME Journal*, *12*(4), 1154–1162. https://doi.org/10.1038/s41396-017-0042-4

Rodríguez-Lázaro, D., Cook, N., Ruggeri, F. M., Sellwood, J., Nasser, A., Nascimento, M. S. J., D’Agostino, M., Santos, R., Saiz, J. C., Rzezutka, A., Bosch, A., Gironés, R., Carducci, A., Muscillo, M., Kovač, K., Diez-Valcarce, M., Vantarakis, A., von Bonsdorff, C. H., de Roda Husman, A. M., … van der Poel, W. H. M. (2012). Virus hazards from food, water and other contaminated environments. *FEMS Microbiology Reviews*, *36*(4), 786–814. https://doi.org/10.1111/j.1574-6976.2011.00306.x

Rohwer, F., & Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature*, *459*(7244), 207–212. https://doi.org/10.1038/nature08060

Roossinck, M. J. (2019). Evolutionary and ecological links between plant and fungal viruses. *New Phytologist*, *221*(1), 86–92. https://doi.org/10.1111/nph.15364

Roossinck, M. J., Martin, D. P., & Roumagnac, P. (2015). Plant virus metagenomics:

Advances in virus discovery. *Phytopathology*, *105*(6), 716–727. https://doi.org/10.1094/PHYTO-12-14-0356-RVW

Rosani, U., Shapiro, M., Venier, P., & Allam, B. (2019). A needle in a haystack: Tracing bivalve-associated viruses in high-throughput transcriptomic data. *Viruses*, *11*(3), 1–21. https://doi.org/10.3390/v11030205

Rosario, K., Duffy, S., & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. *Archives of Virology*, *157*(10), 1851–1871. https://doi.org/10.1007/s00705-012-1391-y

Rosseel, T., Ozhelvaci, O., Freimanis, G., & Van Borm, S. (2015). Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *Journal of Virological Methods*, *222*, 72–80. https://doi.org/10.1016/j.jviromet.2015.05.010

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum information about an uncultivated virus genome (MIUVIG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., … Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, *537*(7622), 689–693. https://doi.org/10.1038/nature19366

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar,

A., Matheus Carnevali, P. B., Cheng, J. F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C., & Eloe-Fadrosh, E. A. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology*, *4*(11), 1895–1906. https://doi.org/10.1038/s41564-019-0510-x

Roux, S., Krupovic, M., Debroas, D., Forterre, P., & Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology*, *3*(12), 130160. https://doi.org/10.1098/rsob.130160

Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F. (2012). Evolution and diversity of the microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE*, *7*(7), 1–12. https://doi.org/10.1371/journal.pone.0040418

Rozenberg, A., Oppermann, J., Wietek, J., Fernandez Lahore, R. G., Sandaa, R. A., Bratbak, G., Hegemann, P., & Béjà, O. (2020). Lateral Gene Transfer of Anion-Conducting Channelrhodopsins between Green Algae and Giant Viruses. *Current Biology*, *30*(24), 4910-4920.e5. https://doi.org/10.1016/j.cub.2020.09.056

Sadeghi, M., Tomaru, Y., & Ahola, T. (2021). RNA Viruses in Aquatic Unicellular Eukaryotes. *Viruses*, *13*(3). https://doi.org/10.3390/v13030362

Sadiq, S., Chen, Y. M., Zhang, Y. Z., & Holmes, E. C. (2022). Resolving deep evolutionary relationships within the RNA virus phylum Lenarviricota. *Virus Evolution*, *8*(1). https://doi.org/10.1093/ve/veac055

Hameed, A. S., Ninawe, A. S., Nakai, T., Chi, S. C., & Johnson, K. L. (2019). ICTV virus taxonomy profile: Nodaviridae. *Journal of General Virology*, *100*(1), 3–4. https://doi.org/10.1099/jgv.0.001170

Sandaa, R. A. (2008). Burden or benefit? Virus-host interactions in the marine environment. *Research in Microbiology*, *159*(5), 374–381. https://doi.org/10.1016/j.resmic.2008.04.013

Sandaa, R. A., Saltvedt, M. R., Dahle, H., Wang, H., Våge, S., Blanc-Mathieu, R., Steen, I. H., Grimsley, N., Edvardsen, B., Ogata, H., & Lawrence, J. (2022). Adaptive evolution of viruses infecting marine microalgae (haptophytes), from acute infections to stable coexistence. *Biological Reviews*, *97*(1), 179–194. https://doi.org/10.1111/brv.12795

Sandlund, L., Mor, S. K., Singh, V. K., Padhi, S. K., Phelps, N. B. D., Nylund, S., & Mikalsen, A. B. (2021). Comparative molecular characterization of novel and known piscine toti-like viruses. *Viruses*, *13*(6), 1–20. https://doi.org/10.3390/v13061063

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D. A., Denef, V. J., McMahon, K. D., Konstantinidis, K. T., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Woyke, T. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature*, *578*(7795), 432–436. https://doi.org/10.1038/s41586-020-1957-x

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, E. C., & Zhang, Y.-Z. (2016). Redefining the invertebrate RNA virosphere. *Nature*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*(October 2017), 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Short, S. M., Staniewski, M. A., Chaban, Y. V., Long, A. M., & Wang, D. (2020).

Diversity of viruses infecting eukaryotic algae. *Current Issues in Molecular Biology*, *39*, 29–62. https://doi.org/10.21775/cimb.039.029

Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B., & Fuhrman, J. A. (2019). Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09106-z

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(539). https://doi.org/10.1038/msb.2011.75

Simon, N., Cras, A. L., Foulon, E., & Lemée, R. (2009). Diversity and evolution of marine phytoplankton. *Comptes Rendus - Biologies*, *332*(2–3), 159–170. https://doi.org/10.1016/j.crvi.2008.09.009

Sivasankar, P., John, R., George, R., & Mansoor, M. (2017). *Rna Viral Diseases of Finfish: a Review*. *299*. http://www.journalcra.com/sites/default/files/26896.pdf

Sorensen, J. W., Zinke, L. A., ter Horst, A. M., Santos-Medellín, C., Schroeder, A., & Emerson, J. B. (2021). DNase Treatment Improves Viral Enrichment in Agricultural Soil Viromes. *MSystems*, *6*(5). https://doi.org/10.1128/msystems.00614-21

Spatharis, S., Tsirtsis, G., Danielidis, D. B., Chi, T. Do, & Mouillot, D. (2007). Effects of pulsed nutrient inputs on phytoplankton assemblage structure and blooms in an enclosed coastal area. *Estuarine, Coastal and Shelf Science*, *73*(3–4), 807–815. https://doi.org/10.1016/j.ecss.2007.03.016

Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F., & Firestone, M. K. (2019). Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(51), 25900–25908. https://doi.org/10.1073/pnas.1908291116

Steward, G. F., Culley, A. I., Mueller, J. a, Wood-Charlson, E. M., Belcaid, M., & Poisson, G. (2013). Are we missing half of the viruses in the ocean? *The ISME Journal*, *7*(3), 672–679. https://doi.org/10.1038/ismej.2012.121

Stough, J. M. A., Tang, X., Krausfeldt, L. E., Steffen, M. M., Gao, G., Boyer, G. L., & Wilhelm, S. W. (2017). Molecular prediction of lytic vs lysogenic states for Microcystis phage: Metatranscriptomic evidence of lysogeny during large bloom events. *PLoS ONE*, *12*(9), 1–17. https://doi.org/10.1371/journal.pone.0184146

Sunarto, A., & Naim, S. (2016). *Chapter 30 - Totiviruses of Crustaceans* (F. S. B. Kibenge & M. G. B. T.-A. V. Godoy (eds.); pp. 425–439). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-801573-5.00030-9

Suttle, C A, Chan, A. M., & Cottrell, M. T. (1991). Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Applied and Environmental Microbiology*, *57*(3), 721–726. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=182786&tool=pmcentrez&rendertype=abstract

Suttle, Curtis A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361. https://doi.org/10.1038/nature04160

Suttle, Curtis A. (2007). Marine viruses--major players in the global ecosystem. *Nature Reviews. Microbiology*, *5*(10), 801–812. https://doi.org/10.1038/nrmicro1750

Suzuki, N., Ghabrial, S. A., Kim, K. H., Pearson, M., Marzano, S. Y. L., Yaegashi, H., Xie, J., Guo, L., Kondo, H., Koloniuk, I., & Hillman, B. I. (2018). ICTV virus taxonomy profile: Hypoviridae. *Journal of General Virology*, *99*(5), 615–616. https://doi.org/10.1099/jgv.0.001055

Takao, Y., Nagasaki, K., Mise, K., Okuno, T., & Honda, D. (2005). Isolation and characterization of a novel single-stranded RNA virus infectious to a marine fungoid protist, Schizochytrium sp. (Thraustochytriaceae, Labyrinthulea). *Applied and Environmental Microbiology*, *71*(8), 4516–4522. https://doi.org/10.1128/AEM.71.8.4516-4522.2005

Tan, K. S., Acerbi, E., & Lauro, F. M. (2016). Marine habitats and biodiversity of Singapore's coastal waters: A review. *Regional Studies in Marine Science*, *8*, 340–352. https://doi.org/10.1016/j.rsma.2016.01.008

Tarutani, K., Nagasaki, K., & Yamaguchi, M. (2000). Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton: Heterosigma akashiwo. *Applied and Environmental Microbiology*, *66*(11), 4916–4920. https://doi.org/10.1128/AEM.66.11.4916-4920.2000

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, *45*(6), 1320–1328. https://doi.org/10.4319/lo.2000.45.6.1320

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J. A., & Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39). https://doi.org/10.1073/pnas.1102164108

Thurber, R. V., Payet, J. P., Thurber, A. R., & Correa, A. M. S. (2017). Virus-host interactions and their roles in coral reef health and disease. *Nature Reviews Microbiology*, *15*(4), 205–216. https://doi.org/10.1038/nrmicro.2016.176

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature Protocols*, *4*(4), 470–483. https://doi.org/10.1038/nprot.2009.10

Tilstone, G. H., Miguez, B. M., Figueiras, F. G., & Fermin, E. G. (2000). Diatom dynamics in a coastal ecosystem affected by upwelling: Coupling between species succession, circulation and biogeochemical processes. *Marine Ecology Progress Series*, *205*, 23–41. https://doi.org/10.3354/meps205023

Tomaru, Y, & Nagasaki, K. (2007). Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *Journal of Oceanography*, *63*, 215–221. http://link.springer.com/article/10.1007/s10872-007-0023-8

Touchon, M., Moura de Sousa, J. A., & Rocha, E. P. (2017). Embracing the enemy: The diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Current Opinion in Microbiology*, *38*, 66–73. https://doi.org/10.1016/j.mib.2017.04.010

Trottet, A., George, C., Drillet, G., & Lauro, F. M. (2021). Aquaculture in coastal urbanized areas: A comparative review of the challenges posed by Harmful Algal Blooms. *Critical Reviews in Environmental Science and Technology*, *0*(0), 1–42. https://doi.org/10.1080/10643389.2021.1897372

Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Progress in Oceanography*, *130*, 205–248. https://doi.org/10.1016/j.pocean.2014.08.005

Urayama, S. I., Takaki, Y., & Nunoura, T. (2016). FLDS: A comprehensive DSRNA sequencing method for intracellular RNA virus surveillance. *Microbes and Environments*, *31*(1), 33–40. https://doi.org/10.1264/jsme2.ME15171

Urayama, S., Takaki, Y., Nishi, S., Yoshida-Takashima, Y., Deguchi, S., Takai, K., & Nunoura, T. (2018). Unveiling the RNA virosphere associated with marine microorganisms. *Molecular Ecology Resources*, *18*(6), 1444–1455. https://doi.org/10.1111/1755-0998.12936

Vainio, E. J., Chiba, S., Ghabrial, S. A., Maiss, E., Roossinck, M., Sabanadzovic, S., Suzuki, N., Xie, J., & Nibert, M. (2018). ICTV virus taxonomy profile: Partitiviridae. *Journal of General Virology*, *99*(1), 17–18. https://doi.org/10.1099/jgv.0.000985

Valverde, R. A., Khalifa, M. E., Okada, R., Fukuhara, T., & Sabanadzovic, S. (2019). ICTV virus taxonomy profile: Endornaviridae. *Journal of General Virology*, *100*(8), 1024–1025. https://doi.org/10.1099/jgv.0.001277

Vincent, F., Sheyn, U., Porat, Z., Schatz, D., & Vardi, A. (2021). Visualizing active viral infection reveals diverse cell fates in synchronized algal bloom demise. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(11). https://doi.org/10.1073/PNAS.2021586118

Vlok, M., Lang, A. S., & Suttle, C. A. (2019). Marine RNA Virus Quasispecies Are Distributed throughout the Oceans. *MSphere*, *4*(2), 1–18. https://doi.org/10.1128/mspheredirect.00157-19

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., …

Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, *167*(11), 2429–2440. https://doi.org/10.1007/s00705-022-05516-5

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A. C., Allen, M. J., Sullivan, M. B., & Temperton, B. (2019). Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ*, *2019*(4), 1–28. https://doi.org/10.7717/peerj.6800

Weber, F., Wagner, V., Rasmussen, S. B., Hartmann, R., & Paludan, S. R. (2006). Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. *Journal of Virology*, *80*(10), 5059–5064. https://doi.org/10.1128/jvi.80.10.5059-5064.2006

Wei, J., Huang, Y., Zhu, W., Li, C., Huang, X., & Qin, Q. (2019). Isolation and identification of Singapore grouper iridovirus Hainan strain (SGIV-HN) in China. *Archives of Virology*, *164*(7), 1869–1872. https://doi.org/10.1007/s00705-019-04268-z

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, *28*(2), 127–181. https://doi.org/10.1016/j.femsre.2003.08.001

Weitz, J. S., Stock, C. A., Wilhelm, S. W., Bourouiba, L., Coleman, M. L., Buchan, A., Follows, M. J., Fuhrman, J. A., Jover, L. F., Lennon, J. T., Middelboe, M., Sonderegger, D. L., Suttle, C. A., Taylor, B. P., Frede Thingstad, T., Wilson, W. H., & Eric Wommack, K. (2015). A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME Journal*, *9*(6), 1352–1364. https://doi.org/10.1038/ismej.2014.220

Weitz, J. S., & Wilhelm, S. W. (2012). Ocean viruses and their effects on microbial

communities and biogeochemical cycles. *F1000 Biology Reports*, *4*(1), 2–9. https://doi.org/10.3410/B4-17

Weynberg, K. D., Allen, M. J., & Wilson, W. H. (2017). Marine prasinoviruses and their tiny plankton hosts: A review. *Viruses*, *9*(3), 1–20. https://doi.org/10.3390/v9030043

Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W., & Bae, J.-W. (2012). Metagenomic Characterization of Airborne Viral DNA Diversity in the Near-Surface Atmosphere. *Journal of Virology*, *86*(15), 8221–8231. https://doi.org/10.1128/JVI.00293-12

Wigington, C. H., Sonderegger, D., Brussaard, C. P. D., Buchan, A., Jan, F., Fuhrman, J., Lennon, J. T., Middelboe, M., Suttle, C. a, Stock, C., Wilson, W. H., Wommack, K. E., Wilhelm, S. W., & Weitz, J. S. (2015). *Re-examining the relationship between virus and microbial cell abundances in the global oceans*. *January*, 1–21. https://doi.org/10.1038/nmicrobiol.2015.24

Wijaya, W., Suhaimi, Z., Chua, E. X. C., Sunil, R. S., Kolundžija, S., Bin Rohaizat, M. A., Binti Azmi, N., Hazrin-Chong, N. H., & Lauro, F. M. (2022). *Frequent pulse disturbances influence resistance and resilience in tropical marine microbial communities*. *65*.

Wilcox, A. H., Delwart, E., & Díaz-Muñoz, S. L. (2019). Next-generation sequencing of dsRNA is greatly improved by treatment with the inexpensive denaturing reagent DMSO. *Microbial Genomics*, *5*(11), 0–7. https://doi.org/10.1099/mgen.0.000315

Wilhelm, S. W., Suttle, C. a, & Wilhelm, W. (2013). *Nutrient Cycles the aquatic food webs*. *49*(10), 781–788.

Williamson, K. E., Fuhrmann, J. J., Wommack, K. E., & Radosevich, M. (2017).

Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annual Review of Virology*, *4*, 201–219. https://doi.org/10.1146/annurev-virology-101416-041639

Wnuk, E., Waśko, A., Walkiewicz, A., Bartmiński, P., Bejger, R., Mielnik, L., & Bieganowski, A. (2020). The effects of humic substances on DNA isolation from soils. *PeerJ*, *8*, 1–15. https://doi.org/10.7717/peerj.9378

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., & Koonin, E. V. (2018). Origins and evolution of the global RNA virome. *MBio*, *9*(6). https://doi.org/10.1128/mBio.02329-18

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V. V., & Koonin, E. V. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology*, *5*(10), 1262–1270. https://doi.org/10.1038/s41564-020-0755-4

Wommack, K., Sime-Ngando, T., Winget, D. M., Jamindar, S., & Helton, R. R. (2010). Filtration-based methods for the collection of viral concentrates from large water samples. *Ma*, 110–117. https://doi.org/10.4319/mave.2010.978-0-9845591-0-7.110

Xia, J., Kameyama, S., Prodinger, F., Yoshida, T., Cho, K.-H., Jung, J., Kang, S.-H., Yang, E.-J., Ogata, H., & Endo, H. (2021). Tight association between microbial eukaryote and Imitervirales communities in the Pacific Arctic Ocean. *BioRxiv*, 2021.09.02.458798. https://doi.org/10.1101/2021.09.02.458798

Yamada, Y., Guillemette, R., Baudoux, A. C., Patel, N., & Azam, F. (2020). Viral attachment to biotic and abiotic surfaces in seawater. *Applied and Environmental Microbiology*, *86*(3), 1–18. https://doi.org/10.1128/AEM.01687-19

Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V, Thomas, T., Raftery, M. J.,

Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Gibson, J. a, & Cavicchioli, R. (2011). Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6163–6168. https://doi.org/10.1073/pnas.1018221108

Yilmaz, S., Allgaier, M., & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods*, *7*(12), 943–944. https://doi.org/10.1038/nmeth1210-943

Yoshida, M., Mochizuki, T., Urayama, S. I., Yoshida-Takashima, Y., Nishi, S., Hirai, M., Nomaki, H., Takaki, Y., Nunoura, T., & Takai, K. (2018). Quantitative viral community DNA analysis reveals the dominance of single-stranded DNA viruses in offshore upper bathyal sediment from Tohoku, Japan. *Frontiers in Microbiology*, *9*(FEB), 1–10. https://doi.org/10.3389/fmicb.2018.00075

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T., & Takai, K. (2013). Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0057271

Yuan, Y., & Gao, M. (2017). Jumbo bacteriophages: An overview. *Frontiers in Microbiology*, *8*(MAR), 1–9. https://doi.org/10.3389/fmicb.2017.00403

Zayed, A. A., Wainaina, J. M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A. A., Bolduc, B., Zablocki, O., Cronin, D., Solden, L., Delage, E., Alberti, A., Aury, J.-M., Carradec, Q., Da Silva, C., Labadie, K., Poulain, J., … Sullivan, M. B. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*, *in press*(April), 156–162.

Zeigler Allen, L., McCrow, J. P., Ininbergs, K., Dupont, C. L., Badger, J. H., Hoffman, J. M., Ekman, M., Allen, A. E., Bergman, B., & Venter, J. C. (2017). The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA

Viruses. *MSystems*, *2*(1), e00125-16. https://doi.org/10.1128/mSystems.00125-16

Zell, R., Delwart, E., Gorbalenya, A. E., Hovi, T., King, A. M. Q., Knowles, N. J., Lindberg, A. M., Pallansch, M. A., Palmenberg, A. C., Reuter, G., Simmonds, P., Skern, T., Stanway, G., & Yamashita, T. (2017). ICTV virus taxonomy profile: Picornaviridae. *Journal of General Virology*, *98*(10), 2421–2422. https://doi.org/10.1099/jgv.0.000911

Zhang, Q., Xu, T., Wan, X., Liu, S., Wang, X., Li, X., Dong, X., Yang, B., & Huang, J. (2017). Prevalence and distribution of covert mortality nodavirus (CMNV) in cultured crustacean. *Virus Research*, *233*, 113–119. https://doi.org/10.1016/j.virusres.2017.03.013

Zhang, X. (2022). Environmental viromes reveal global virosphere of deep-sea RNA viruses. Research Square. https://doi.org/10.21203/rs.3.rs-1294962/v1

Zhang, Y. Z., Chen, Y. M., Wang, W., Qin, X. C., & Holmes, E. C. (2019). Expanding the RNA Virosphere by Unbiased Metagenomics. *Annual Review of Virology*, *6*, 119–139. https://doi.org/10.1146/annurev-virology-092818-015851

Zhang, Y. Z., Shi, M., & Holmes, E. C. (2018). Using Metagenomics to Characterize an Expanding Virosphere. *Cell*, *172*(6), 1168–1172. https://doi.org/10.1016/j.cell.2018.02.043

Zhao, J., Jing, H., Wang, Z., Wang, L., Jian, H., Zhang, R., Xiao, X., Chen, F., Jiao, N., & Zhang, Y. (2022). Novel Viral Communities Potentially Assisting in Carbon, Nitrogen, and Sulfur Metabolism in the Upper Slope Sediments of Mariana Trench. *MSystems*, *7*(1). https://doi.org/10.1128/msystems.01358-21

Zhao, Yanlin, Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., Ellisman, M., Deerinck, T., Sullivan, M. B., & Giovannoni, S. J. (2013). Abundant SAR11 viruses in the ocean. *Nature*, *494*(7437), 357–360.

https://doi.org/10.1038/nature11921

Zhao, Yuan, Zhao, Y., Zheng, S., Zhao, L., Zhang, W., Xiao, T., & Grégori, G. (2022). Enhanced resolution of marine viruses with violet side scatter. *Cytometry Part A*. https://doi.org/10.1002/cyto.a.24674

Zheng, X., Liu, W., Dai, X., Zhu, Y., Wang, J., Zhu, Y., Zheng, H., Huang, Y., Dong, Z., Du, W., Zhao, F., & Huang, L. (2021). Extraordinary diversity of viruses in deep-sea sediments as revealed by metagenomics without prior virion separation. *Environmental Microbiology*, *23*(2), 728–743. https://doi.org/10.1111/1462-2920.15154

Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology*, *52*(1), 79–92. https://doi.org/10.1016/j.femsec.2004.10.006

Zimmerman, A. E., Howard-Varona, C., Needham, D. M., John, S. G., Worden, A. Z., Sullivan, M. B., Waldbauer, J. R., & Coleman, M. L. (2020). Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nature Reviews Microbiology*, *18*(1), 21–34. https://doi.org/10.1038/s41579-019-0270-x