Metagenomic approaches for examining the diversity of large DNA viruses in the biosphere

Roxanna Farzad

Thesis submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Biological Sciences

Frank O. Aylward, Chair
Bryan B. Hsu
Jeremy A. Draghi

July 12th, 2023

Blacksburg, Virginia

Keywords: Giant Viruses, *Nucleocytoviricota,* viral metagenomics, metagenome-assembled genomes (MAGs), large DNA viruses

Metagenomic approaches for examining the diversity of large DNA viruses in the biosphere

Roxanna Farzad

ABSTRACT

The discovery of large DNA viruses has challenged the traditional perception of viral complexity due to their enormous genome size and physical dimensions. Previously, viruses were considered small, filterable agents until the discovery of large DNA viruses. Among large DNA viruses, the phylum *Nucleocytoviricota* and its members, which are often called "giant viruses" have large genome sizes (up to 2.5 Mbp) and virion sizes (up to 1.5 um). Due to having large virion and genome sizes, these viruses were often excluded from viral surveys and remained understudied for years. Luckily, the advancement of metagenomic analysis has facilitated the study of large DNA viruses by analyzing them directly from their environment without cultivating them in the lab, which could be challenging for viruses. In the first chapter of the thesis, I investigated 11 metagenome-assembled genomes (MAGs) of giant viruses previously surveyed from Station ALOHA in the Pacific Ocean. St. ALOHA is located near Hawaii and represents oligotrophic gyres which the majority of the ocean is made of them. I focused on 11 MAGs of giant viruses to get insight into their phylogenetic characteristics, genomic repertoire, and global distribution patterns. Despite the fact that metagenomic analysis has facilitated the study of genetic materials of microbes and viruses on a huge scale, it is essential to benchmark the performance of metagenomic tools and understand the associated biases, particularly in viral metagenomics. In the second chapter, I evaluated the performance of metagenomic tools (contigs assembler and binning tool) in recovering viral genomes using annotated dataset. We used a metagenome simulator (CAMISIM) to generate simulated short reads with known composition to assess these processes. Moreover, I emphasized the importance of binning contigs for viral genomes to fully recover the genomes of viruses along with discussing how diversity metrics were differed for contigs, bins populations.

Metagenomic approaches for examining the diversity of large DNA viruses in the biosphere

Roxanna Farzad

GENERAL AUDIENCE ABSTRACT

Viruses are generally thought to be small biological agents with small genome (genetic material) sizes and tiny physical structures; for instance, the genome length of a Human Immunodeficiency Virus (HIV) is around 10 kilobase pair (a unit for measuring genetic material in an organism), and the virion size (physical dimension of a virus) can go up to 120 nm. The discovery of large DNA viruses has challenged the idea of considering viruses as small biological entities, as their genome sizes and physical dimensions can be up to 2.5 megabase pairs and 1500 nm, respectively. Famous members of large DNA viruses from the phylum *Nucleocytoviricota* are often known as “Giant Viruses" because they have enormous genome sizes and physical dimensions. Due to having large viral particles, these viruses may usually be excluded from viral surveys. For instance, in field studies, samples must be filtered through a fraction (e.g., 0.2 um) to eliminate bacterial and archaeal genomes and cellular debris, which also results in excluding larger viruses. Since these viruses remain understudied for several years because of biases associated with having large viral particles, there is a solid need to discover and investigate more about them. Growing and cultivating viruses in the laboratory may be challenging, as they need specific hosts to be dependent on to produce more viral progeny and some specific laboratory environments. Luckily, with the advancement of biotechnology, scientists could find ways to evade the need for cultivating viruses in the lab and study them with computational tools such as metagenomic analysis and bioinformatic tools.

Metagenomics analysis helps to study the genetic materials of microbial or viral populations directly from their habitat without growing them in a laboratory. In short, metagenomic analysis has multiple steps, including collecting and filtering samples, fragmenting DNA within the samples, generating short DNA sequences (short-read sequences) with NGS (Next Generation Sequencing) technology, assembling short-read sequences into large DNA fragments which can be contigs (contiguous DNA fragments) and metagenome-assembled genome (MAGs). With metagenomic analysis, we can recover the genome of multiple organisms, and we name the recovered genome as metagenome-assembled genome (MAGs) as it is generated through metagenomic

processes. The metagenomic analysis will allow us to study microbes and viruses in their environment and gain insight into their taxonomic details, genomic content, and how widespread they are.

In the first chapter, I studied 11 MAGs of giant viruses previously surveyed from St. ALOHA, Hawaii. St. ALOHA is a good field site for examining microbial processes and diversity and a good representative of oligotrophic waters (low in nutrients). I examined 11 MAGs of giant viruses to investigate their taxonomic characteristics to clarify which order they belong to within their phylum, their genomic content, and their global distribution pattern. Although studies have successfully recovered the genome of large DNA viruses from their habitats and then analyzed them, all these metagenomic processes need to be evaluated so the results will be valid to consider as the genome of our interested organisms. In the second chapter, I developed a workflow for viral metagenomic analysis to assess metagenomic tools' performance in recovering reliable viral genomes, particularly for large DNA viruses. Most of these benchmarking workflows are done for bacterial and archaeal genomes, and in this thesis, I used these metagenomic tools and applied them to recover large DNA viruses genomes. Also, I emphasized the importance of using binning tools to fully recover large DNA viruses genomes, as due to their large genome size, their genomes might remain fragmented into different contigs, which are longer sequences than reads but shorter than MAGs.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Frank Aylward, for his support and kindness throughout my academic journey. His patience, guidance, and teachings have been instrumental throughout these 2 years. I am truly grateful for his mentorship and all the essential skills and knowledge he taught me during my master's studies.

I would also appreciate my committee members, Dr. Draghi, and Dr. Hsu, for their valuable comments and helpful suggestions. Their expertise and insights have greatly impressed me to understand the concepts in my field better.

To my current and former lab members; Dr. Alaina Weinheimer, Dr. Carolina Martinez Gutierrez, Dr. Mohammad Moniruzzaman, Dr. Anh Ha, Dr. Uri Sheyn, Dr. Zach Barth, Sangita Karki, Paula Erazo, Abdeali Jivaji. I am thankful for their support, assistance, and the enjoyable moments we shared together. I would particularly like to thank Dr. Ahn Ha, who helped me a lot through the challenges of my research projects.

I would like to thank my relatives, Dr. Bahareh Nojabaei, and Dr. Hossein Foroutan, and my cousins, Yashar and Aida Foroutan, for their constant love, encouragement, and support throughout these years.

A special mention goes to my friends, particularly Justus Hargett, whose presence has made my journey in graduate school, especially as an international student, much more pleasant. Their friendship, understanding, and shared experiences have been a source of strength and comfort.

Lastly, I would like to thank my parents (Laleh Nojabaei, Babak Farzad, and my brother Danial Farzad) for their support, unconditional love, and unwavering faith in me. Despite the geographical distance between us, their belief in my abilities and strengths has been a constant motivation, and I am forever grateful for their presence in my life.

Attributions

Chapter 1: Diversity and Genomics of Giant Viruses in the North Pacific Subtropical Gyre

Farzad R, Ha AD, Aylward FO. Diversity and genomics of giant viruses in the North Pacific Subtropical Gyre. Front Microbiol. 2022;13: 1021923. doi:10.3389/fmicb.2022.1021923

This chapter has two additional authors:

Anh D. Ha[1] & Frank O. Aylward[1,2]
1 Department of Biological Sciences, Virginia Tech, Blacksburg, VA, 24061
2 Center for Emerging, Zoonotic, and Arthropod-Borne Infectious Disease, VirginiaTech, Blacksburg VA, 24061

Chapter 2: Benchmarking metagenomic approaches for recovering large DNA viruses

This chapter has two additional authors:

Anh D. Ha[1] & Frank O. Aylward[1]
1 Department of Biological Sciences, Virginia Tech, Blacksburg, VA, 24061

Table of Contents

List of Figures

Chapter 1

Chapter 2

List of Tables

Chapter 1

Chapter 2

Introduction

For many years in traditional microbiology studies, viruses were commonly perceived as small biological entities with no complex physical and genomic structures [1–4]. However, this perception was challenged upon discovering large DNA viruses, which unveiled that viruses can have large genomes and physical dimensions comparable to bacteria and archaea [5,6]. The largest groups of large DNA viruses are eukaryotic viruses from the phylum *Nucleocytoviricota,* often referred to as “giant viruses" [7–9], and a group of bacteriophages referred to as “jumbo bacteriophages" from the class *Caudoviricetes* which have redefined the historical definition of viral complexity. The most notable examples of the phylum *Nucleocytoviricota* are poxviruses, which include Variola major, the causative agent of smallpox as well as vaccinia virus, which was used to make the first vaccines [9–11]. Throughout the 20th century, the diversity of large DNA viruses in the *Nucleocytoviricota* continued to expand with the discovery of the African Swine Fever Virus (ASVF), large viruses that infect green algae of the genus *Chlorella* [12,13].

Although many members of the Nucleocytoviricota were discovered throughout the 20th Century, the discovery of *Acanthamoeba polyphaga* mimivirus in 2003 is often considered the first example of a truly “giant” virus due to its remarkable 750 nm virion and 1.2 Mbp genome [5,14,15]. This was a watershed moment in virology because it demonstrated that many viruses are large enough to be viewed in a light microscope and that some even have genomes larger than common free-living bacteria. After discovering mimivirus, extensive research and cultivation efforts were done with a particular focus on isolating a wide range of other giant viruses using diverse species of amoeba as their hosts [16–18]. These endeavors led to the discovery of numerous additional giant viruses, including some virion sizes reaching 1.5 um (*Pithovirus sibericum*) and genome sizes exceeding 2.5 Mbp (*Pandoravirus salinus*) [16–18].

Isolating novel viruses can present challenges due to their requirements for specific laboratory culture conditions and growth requirements of the host and the need for a specific phenotype (i.e., lysis) for viral detection. As a result, many viruses remain undiscovered using cultivation-based approaches alone. To overcome many of these obstacles, cultivation-independent approaches have been developed, which include several methods such as read mapping-based techniques, marker gene surveys, genome-resolved metagenomics, and single-cell genomics [2–4]. Metagenomic analysis has been considered one of the most successful methods to recover unknown viral genomes from environmental samples and reveal their community composition and genomic repertoires. Although many early metagenomic studies focused primarily on

bacterial and archaeal diversity, recent work has shown that these approaches are equally valid for examining viral diversity [19–24].

In the first chapter of the thesis, I examine metagenomic data generated from Station ALOHA in the Pacific Ocean to investigate the diversity of giant viruses (phylum *Nucleocytoviricota*) in this environment. This study aimed to gain insights into the phylogenetic characteristics, global and seasonal distribution patterns, and genomic repertoires of giant viruses in the open ocean. Station ALOHA, situated near Hawaii, was selected because it represents oligotrophic waters typical of those that cover 40% of the Earth's surface across the oceanic gyres of the planet. Through this analysis, we recovered eleven draft genomes of giant viruses, most of which represent novel lineages that had never been found before. Several of these viruses are globally distributed, particularly those belonging to *Imitervirales* and *Algavirales* orders. Moreover, in this chapter, I discuss the genomic repertoire of the eleven MAGs of giant viruses and show the presence of several genes that are commonly known to be found in cellular lineages (e.g., citrate synthase, aconitase, rhodopsin), which are presumably used by viruses to manipulate host during viral infection.

While metagenomics analysis has become indispensable for exploring the genetic material of microbes across diverse environments, from the deep sea to the human gut, it is essential to evaluate the performance of metagenomics tools and understand the biases of these approaches. Although several benchmarking studies exist to assess the efficacy of metagenomics, they are mostly focused on bacterial and archaeal genome recovery. Therefore, there is a strong need to develop a workflow to benchmark viral metagenomics, particularly for large DNA viruses. In the second chapter of the thesis, I evaluated the performance of metagenome-based viral genome recovery using a metagenome simulator that generates short-read metagenomic datasets of known composition. I assembled the short reads from these simulated metagenomes into larger contiguous fragments ("contigs") and then binned contigs together using commonly-used tools to generate draft viral genomes. I assess the performance of contig binning to recover large DNA viruses and provide a workflow for benchmarking purposes. This work highlights the importance of binning tools to fully recover viral genomes, particularly for large DNA viruses. These findings are particularly important given that most current viral metagenomic workflows assess only contigs and do not perform binning.

References

1. Scholthof K-BG, Shaw JG, Zaitlin M. Tobacco Mosaic Virus: One Hundred Years of Contributions to Virology. American Phytopathological Society; 1999. Available: https://books.google.com/books/about/Tobacco_Mosaic_Virus.html?hl=&id=573wAAAAMAAJ

2. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020;11: 1710. Available: https://www.nature.com/articles/s41467-020-15507-2

3. Schulz F, Abergel C, Woyke T. Giant virus biology and diversity in the era of genome-resolved metagenomics. Nat Rev Microbiol. 2022;20: 721–736. doi:10.1038/s41579-022-00754-5

4. Ha AD, Moniruzzaman M, Aylward FO. Assessing the biogeography of marine giant viruses in four oceanic transects. ISME Commun. 2023;3: 43. doi:10.1038/s43705-023-00252-6

5. Aylward FO, Moniruzzaman M. Viral Complexity. Biomolecules. 2022;12. doi:10.3390/biom12081061

6. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6: 315–319. doi:10.1038/nrmicro1858

7. Fischer MG. Giant viruses come of age. Curr Opin Microbiol. 2016;31: 50–57. doi:10.1016/j.mib.2016.03.001

8. Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 2006;117: 156–184. doi:10.1016/j.virusres.2006.01.009

9. Crawford DH. Deadly Companions: How Microbes Shaped Our History. Oxford University Press; 2018. Available: https://books.google.com/books/about/Deadly_Companions.html?hl=&id=KvJKDwAAQBAJ

10. Fenner F. Adventures with poxviruses of vertebrates. FEMS Microbiol Rev. 2000;24: 123–133. doi:10.1016/S0168-6445(00)00027-9

11. Crawford DH. Viruses: A Very Short Introduction. Oxford University Press; 2018. Available: https://books.google.com/books/about/Viruses.html?hl=&id=e2pNDwAAQBAJ

12. Goebel SJ, Johnson GP, Perkus ME, Davis SW, Winslow JP, Paoletti E. The complete DNA sequence of vaccinia virus. Virology. 1990;179: 247–66, 517–63.

doi:10.1016/0042-6822(90)90294-2

13. Van Etten JL, Meints RH. Giant viruses infecting algae. Annu Rev Microbiol. 1999;53: 447–494. doi:10.1146/annurev.micro.53.1.447

14. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, et al. A giant virus in amoebae. Science. 2003;299: 2033. doi:10.1126/science.1081867

15. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-megabase genome sequence of Mimivirus. Science. 2004;306: 1344–1350. doi:10.1126/science.1101485

16. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A. 2014;111: 4274–4279. doi:10.1073/pnas.1320670111

17. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, et al. Diversity and evolution of the emerging Pandoraviridae family. Nat Commun. 2018;9: 2285. doi:10.1038/s41467-018-04698-4

18. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A. 2011;108: 17486–17491. doi:10.1073/pnas.1110889108

19. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, et al. Hidden diversity of soil giant viruses. Nat Commun. 2018;9: 4881. doi:10.1038/s41467-018-07335-2

20. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. Proc Natl Acad Sci U S A. 2019;116: 20574–20583. doi:10.1073/pnas.1907517116

21. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, et al. Giant viruses with an expanded complement of translation system components. Science. 2017;356: 82–85. doi:10.1126/science.aal4657

22. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. MBio. 2019;10. doi:10.1128/mBio.02497-18

23. Andreani J, Verneau J, Raoult D, Levasseur A, La Scola B. Deciphering viral presences: two novel partial giant viruses detected in marine metagenome and in a mine drainage metagenome. Virol J. 2018;15: 66. doi:10.1186/s12985-018-0976-9

24. Wilson WH, Gilg IC, Moniruzzaman M, Field EK, Koren S, LeCleir GR, et al.

Genomic exploration of individual giant ocean viruses. ISME J. 2017;11: 1736–1745. doi:10.1038/ismej.2017.61

Chapter 1: Diversity and Genomics of Giant Viruses in the North Pacific Subtropical Gyre

Keywords: Giant Viruses, *Nucleocytoviricota*, *Mesomimiviridae*, *Prasinoviridae*, Station ALOHA, marine viruses

Abstract

Large double-stranded DNA viruses of the phylum *Nucleocytoviricota,* often referred to as “giant viruses”, are ubiquitous members of marine ecosystems that are important agents of mortality for eukaryotic plankton. Although giant viruses are known to be prevalent in marine systems, their activities in oligotrophic ocean waters remain unclear. Oligotrophic gyres constitute the majority of the ocean and assessing viral activities in these regions is therefore critical for understanding overall marine microbial processes. In this study, we generated 11 metagenome-assembled genomes (MAGs) of giant viruses from samples previously collected from Station ALOHA in the North Pacific Subtropical Gyre. Phylogenetic analyses revealed that they belong to the orders *Imitervirales* (n=6), *Algavirales* (n=4), and *Pimascovirales* (n=1). Genome sizes ranged from ~119-574 kbp, and several of the genomes encoded predicted TCA cycle components, cytoskeletal proteins, collagen, rhodopsins, and proteins potentially involved in other cellular processes. Comparison with other marine metagenomes revealed that several have broad distribution across ocean basins and represent abundant viral constituents of pelagic surface waters. Our work sheds light on the diversity of giant viruses present in oligotrophic ocean waters across the globe.

Introduction

Nucleo-cytoplasmic large DNA viruses (NCLDVs, phylum *Nucleocytoviricota*), also known as “giant viruses”, are a lineage of eukaryotic viruses that include many animal and protist pathogens. In addition to several well-known families that infect vertebrates (e.g. *Poxviridae*, *Asfaviridae*, and *Iridoviridae*), several families in this phylum infect a variety of algae and other protists (e.g. *Phycodnaviridae, Marseilleviridae,* and *Mimiviridae*) [1–5]. Appreciation of the environmental prevalence of viruses within the *Nucleocytoviricota* somewhat lagged behind other viral groups because the large capsid sizes of many members of this phylum often precluded their recovery in diversity surveys that focused on particles that could pass through a 0.2 μm filter. Nevertheless, pioneering studies focusing on marker genes provided early evidence that these large DNA viruses are widespread in the environment [6–8], and later metagenomic studies revealed an enormous diversity in this group, particularly in marine environments [9–14]. Recent estimates suggest that there are at least 32 different families of giant

viruses that reside in diverse ecosystems across the globe, and more will almost certainly be identified in the future [15].

Giant viruses have complex and chimeric genomes that are the product of widespread gene exchange with various cellular lineages [16–18]. Besides the core machinery involved in virion structure and DNA replication, giant viruses also commonly encode genes involved in translation, glycolysis, TCA cycle, cytoskeletal dynamics, light-harvesting, nutrient transport, and other pathways involved in nutrient homeostasis [9,19]. Rhodopsins are also common in a wide range of marine giant viruses [20–22]. Rhodopsins are light-driven ion pumps that can be involved in energy production and signal transduction [23,24]. Viral rhodopsins may permit viruses to modify host phototaxis during infection, which may promote their proliferation [25]. Proteins involved in cytoskeletal dynamics have also been found to be quite common in a variety of marine giant viruses; viral homologs to actin, myosin, and kinesin genes could potentially benefit viruses by manipulating the host's cytoskeleton by using host motor proteins to traffic virions or maintain the localization of viral machinery during infection [26–28]. These recent findings collectively suggest that giant viruses use a broad assortment of functional genes to manipulate host physiology and alter the intracellular environment to promote virion propagation.

Although giant viruses are globally distributed in a variety of habitats, they appear to be particularly diverse and abundant in the ocean [9,13,29,30]. The majority of the ocean is made up of oligotrophic oceanic gyres, and it is therefore of particular interest to examine viral dynamics in these systems. One field site that has been particularly useful for examining microbial diversity in oceanic gyres is Station ALOHA (A Long-term Oligotrophic Habitat Assessment), located at 22°45'N, 158°W, nearly 100 kilometers north of the Hawaiian island of Oahu [31]. Several recent studies have recently elucidated a rich diversity of viruses that are present at or near Station ALOHA [32–35]. In this study, we surveyed previously-sequenced metagenomes generated from Station ALOHA to characterize the diversity of giant viruses in this habitat [36]. Although metagenomes derived from <0.2 um size fractions are typically used to evaluate viral diversity in marine systems, recent studies have found that many giant viruses are often found in larger size fractions along with bacteria and archaea [9,13]. We also analyzed the encoded functions in the draft giant virus genomes that we recovered to gain insight into possible mechanisms they employ to manipulate their hosts during infection. Lastly, by examining publicly-available metagenomes from other marine environments we examined the distribution and biogeography of these viruses on a global scale. Because of Station ALOHA’s location in the North Pacific Subtropical Gyre, examination of giant viruses found here provides a window into those lineages that are likely broadly

distributed in oligotrophic ocean waters and may play important roles in marine ecological dynamics.

Methods

Metagenomes Used

We analyzed metagenomes that were generated in a previous study [36]. This dataset consists of 107 samples from depths ranging from 25m to 1000m that were collected over a 1.5 year sampling period at Station ALOHA on 11 cruises of the Hawaii Ocean Time-series (HOT). The methods used for sample collection and processing have been previously described [36]. Briefly, water was filtered onto 0.2 µm filters, and after DNA extraction, libraries were created with Illumina TruSeq LT Nano kit, and metagenomes were sequenced using Illumina MiSeq and NextSeq 500 systems.

Generation of Metagenome-Assembled Genomes (MAGs)

MAGs were generated using a workflow developed previously [37]. Briefly, metagenomes were assembled with MegaHit v. 1.2.9 (with parameters --min-contig-len 5000), and contigs were subsequently binned using MetaBat2 v. 2.12.1 (with parameters -s 100000, -m 10000, –minS 75, –maxEdges 75) [38]. All bins were analyzed with ViralRecall v. 2.0 to identify those that corresponded to giant viruses, and only those that contained 4 of 5 NCLDV marker genes were retained. The marker genes used for this were the A32-like ATPase (A32), B-family DNA polymerase (PolB), virus late transcription factor 3 (VLTF3), major capsid protein (MCP), and superfamily II helicase (SFII). This resulted in 11 NCLDV bins ranging in size from 119,690 to 574,081 bp (Table 1). Genome statistics were compiled with SeqKit v. 2.2.0 [39], and we predicted proteins using Prodigal v. 2.6.3 [40] with default parameters (Table 1).

Phylogenetic construction

In order to explore the phylogenetic placement of the reconstructed MAGs, we used the protein predictions of the 11 MAGs generated in this study together with proteins from all reference giant viruses that have previously been compiled in the Giant Virus Database [15]. Subsequently, we used the program ncldv_markersearch to generate a concatenated alignment of 7 marker genes, as previously described (https://github.com/faylward/ncldv_markersearch) [15]. Briefly, this tool use HMMER3 to identify 7 conserved marker genes (superfamily II helicase (SFII), virus-like transcription factor (VLTF3), B-family DNA polymerase (PolB), and A32-like ATPase (A32), a DNA-dependant RNA polymerase (RNAP) subunit, transcription elongation factor II-S (TFIIS), and a family II topoisomerase), and then uses Clustal Omega v1.2.4 to produce multi-sequence alignments, which are then concatenated. Proteins that are absent in a MAG are replaced with a series of X characters in the concatenated alignment. We then

generated a maximum-likelihood phylogenetic tree using IQ-TREE v. 1.6.12 with 1000 ultrafast bootstraps (parameters -m LG+F+I+G4 -bb 1000 -wbt -nt AUTO --runs 3) [41,42]. The tree was visualized in the interactive Tree of Life (iTOL) [43]; (Fig. 1).

Average nucleotide and amino acid identity

In order to assess divergence from reference genomes, we calculated one-way average amino acid identity (AAI) and average nucleotide identity (ANI) of the 11 MAGs against all genomes in the Giant Virus Database. Both comparisons were done with LAST v. 959, and results were parsed with a custom Python script. Results are shown in Data Set S1[44].

Sequence similarity search and protein annotation

To perform sequence homology searches, we used LAST v. 959 (parameters -m 5000, -f BlastTab, -P 32, -u 2, -Q 0)to compare all protein predictions against a protein database that included RefSeq 207 as well as all protein predictions in the Giant Virus Database [15,45]. We indicated the sequence similarities between 11 MAGs and multiple organisms (bacteria, eukaryotes, viruses, and others) together with those which have no hits (Fig. 2a). Moreover, we retained all the best hits to viruses and searched for homology between MAGs and different viral orders and families (Fig. 2b). Subsequently, plots to visualize the results were made with ggplot2 v. 3.3.6 in R software [46] (Fig. 2) and the final results of the sequence similarity search are accessible in Data Set S2. For protein functional prediction, we annotated all predicted proteins in each genome by searching them against the Pfam database v. 34 [47] using HMMER3 v. 3.3 (parameter “–cut_nc”) with all hits retained. These annotations are available in Data Set S3. Protein annotations were manually inspected to detect the presence of genes involved in central carbon metabolism, DNA processing, light harvesting, amino acid metabolism, cytoskeleton dynamics, and other functions of interest.

Read-mapping analysis

We examined the distribution of our 11 MAGs by mapping the reads from other marine metagenomes onto them using coverM 0.6.1 (parameters --min-read-percent-identity 0.95, minimum 20% covered fraction; available from (https://github.com/wwood/CoverM). We used a breadth cutoff of 20% (i.e. 20% covered fraction), consistent with a recent study that used read-mapping to determine the distribution of large bacteriophages [48]. We used the metagenomic datasets corresponding to the GA02, GA03, GP13, and GA10 cruises from the bioGEOTRACES dataset [49]. These metagenomes were sequenced from 480 samples collected during 2010-2011 from 91 stations. The location and the sampling date of each of the transects are as follows: GA02 (from North to South Atlantic, May 2010 - March 2011), GA03

(North Atlantic, October 2010 - December 2011), GA10 (South Atlantic, October 2010 - November 2010), GP13 (South Pacific Ocean, May 2011 - June 2011). In addition, we evaluated the distribution of 11 MAGs in St. ALOHA across different depths (0-1000m) and across a 1.5-year time series at that location to evaluate their depth distribution and seasonal abundance in this oligotrophic ecosystem. The Hawaii Ocean Time-series (HOT) metagenomes were previously generated [36] during approximate monthly sampling periods from August 2010 to December 2011 at St. ALOHA. Reads from these metagenomes were mapped onto the 11 MAGs using the same coverM parameters as described above. For visualization, we used R and R-based tools (v. 3.3.2; https://www.R-project.org) to draw world map plots and Ocean Data View software (v. 4.7.10; http://odv.awi.de) for contour and Hawaii Ocean Time-series (HOT) plots. Depth distribution was interpolated in Ocean Data View in the DIVA gridding mode. The raw read mapping statistics of 11 MAGs on a global scale and in Station ALOHA are available in Data Set S4 and Data Set S5, respectively.

Results and discussion

Phylogenetic analysis of the Giant Virus MAGs

Based on our multi-locus phylogenetic analysis, four of the MAGs can be placed within the order *Algavirales* (HOT_MAGs 4, 14, 20, and 30), six could be assigned to the *Imitervirales* (HOT_MAGs 3, 5, 60, 12, 13, and 10), and one could be placed in the *Pimascovirales* (HOT_MAG 22*)* (Fig. 1)**.** Among the MAGs that fall within the *Algavirales*, MAGs 14, 30, and 20 belong to the family *Prasinoviridae* [AG_01] and the same genus-level group (g177); (Table 1). Viruses in this family are known to infect prasinophytes of the genera *Bathycoccus*, *Micromonas*, and *Ostreococcus* [50]. Prasinophytes are picoeukaryotic algae that are broadly distributed in the ocean [51,52]. Prasinoviruses are correspondingly abundant in marine waters and play a crucial role in regulating the populations of their plankton hosts [53]. Cultivated representatives of these viruses have genome sizes ranging from 184 to 198 kbp and %GC contents from 37% to 45%. This is generally consistent with the prasinovirus MAGs that we recovered here, which range in size from ~120-172 kbp and in %GC content from 33.8-37.5% (Table 1). The somewhat smaller size of HOT_MAGs 14 and 20 is potentially an indication that these genomes are not complete, though it may also be a sign of genome reduction compared to their relatives. The last MAG that could be placed within the *Algavirales*, HOT_MAG 4, clustered within the family-level lineage AG_04 and the genus g175, both of which were recently demarcated [15]; (Fig. 1 & Table 1). The sole cultivated representative of this lineage is *Heterosigma akashiwo virus* (HaV), which infects the eponymous raphidophyte that is responsible for causing harmful algal blooms [54]. Although raphidophytes are commonly associated with algal blooms in

coastal waters, they have also been reported in oligotrophic gyres [55], and it is, therefore, possible that the host of this MAG lies within this group.

Of the six *Imitervirales* MAGs, three fall within the proposed *Mesomimiviridae* family (IM_01: HOT_MAGs 10, 12, 13); (Fig. 1 & Table 1). HOT_MAGs 10 and 13 are clustered within the same genus g342 while HOT_MAG12 falls within the genus g336; (Table 1). Members of the *Mesomimiviridae* are particularly widespread in global marine and freshwater environments, and phylogenomic analysis suggests that roughly half of all currently available MAGs can be placed in this group [15]. The size of these MAGs ranges from 380-430 kbp (Table 1), which is consistent with those of cultivated viruses in this family that infect haptophyte hosts of the genera *Chrysochromulina* and *Phaeocystis* [56]. *Phaeocystis globosa* virus 16T (PgV-16T) is one of the most well-studied members of this family; it has a 150 nm diameter icosahedral virion with a 470 kbp genome size, which is comparable to the related viruses that infect *Chrysochromulina ericina* virus *C. parva [56,57]*.

The last three *Imitervirales* MAGs clustered within the recently demarcated family IM_09 (Fig. 1 & Table 1), which contains *Aureococcus anophagefferens* virus (AaV) and *Prymnesium kappa* virus (PkV). AaV is the smallest member of the *Imitervirales* isolated to date, with a genome size of 371 kbp and a virion diameter of 140 nm [58,59]. In contrast to this, PkV is rather large, with a 1.4 Mbp genome and a ~310 nm diameter virion [60,61], underscoring the large range of genome and virion sizes within this group. HOT_MAGs 3 and 5 fall within the same genus g279 while HOT_MAG 60 clustered within genus g274. The three MAGs that fall within this group have genome sizes ranging from 470 to 575 kbp (Table 1), which is larger than AaV but quite a bit smaller than PkV, suggesting that they represent intermediate-sized members of this lineage.

The only MAG that clustered within the *Pimascovirales* is HOT_MAG 22, which falls within the recently demarcated family-level lineage PM_01; (Fig. 1 & Table 1). Interestingly, HOT_MAG 22 falls outside of previously demarcated genus-level groups and had low AAI to any references (highest match of 40% AAI to GVMAG-M-3300023184-117, Data Set S1), suggesting that this MAG represents a new genus-level group. The order *Pimascovirales* includes several lineages that infect amoeba, such as Pithoviruses and Marseilleviruses, as well as other viruses that infect metazoan hosts ranging from insects to fish and frogs, such as the *Iridoviridae/Ascoviridae*. Pimascoviruses are not commonly viewed as widespread marine viruses, although, in the previous studies, other MAGs that fall within the PM_01 group have also been found in marine environments [15], and in some cases have been found to be transcriptionally active [26]. One study also reported several Pimascovirus

MAGs from deep sea sediments, including one with phylogenetic placement near the Marseilleviruses and a notably large genome (>700 kbp) [11]. The prevalence of these viruses in marine metagenomes suggests that they infect as-yet unknown protist hosts, and further research identifying their host range will be a necessary step towards clarifying the ecological impacts of marine pimascoviruses.

The phylogenetic placements of the 11 reconstructed MAGs are in agreement with the result from the homology search results of all proteins encoded by each MAG. Proteins in each MAG had best matches to viruses in the same order in which they were classified in the phylogeny (Fig. 2). The only possible exception is the *Pimascovirales* MAG (HOT_MAG 22) which had a larger proportion of proteins with no known homologs or hits to the *Imitervirales*. This may be due to the relatively poor representation of marine *Pimascovirales,* together with the large number of *Imitervirales* in our reference genome databases.

Environmental distribution of the giant virus MAGs

Previous studies have shown that the proposed *Mesomimiviridae* family is particularly prevalent in aquatic systems [15,57,62]. Consistent with this, we found that the three MAGs that fall within this clade were particularly widespread in the marine metagenomes that we surveyed here (HOT_MAGs 12, 13, and 10; Fig. 3). HOT_MAGs 12, 13, and 10 were found in 12, 8, and 2 bioGEOTRACES sample locations, respectively. HOT_MAG 3, which can be classified into the family-level group IM_09, was detected in 15 distinct locations and is the most broadly distributed of the MAGs that fall within the *Imitervirales* MAGs. HOT_MAGs 20 and 30, which fall within the *Prasinoviridae* family, were the most abundant MAGs in the bioGEOTRACES datasets (Fig. 3). HOT_MAGs 20 and 30 were also quite widespread, occurring in 23 and 30 distinct locations, respectively. The third prasinovirus (HOT_MAG 14), and the other MAG that places within the *Algavirales* (HOT_MAG4, family AG_04) were each found in only 3 sample sites within the Atlantic Ocean (Fig. 3). It is worth noting that HOT_MAGs 60 and 22 were not found in any reference bioGEOTRACES metagenomic dataset. HOT_MAG60 can be placed within the order *Imitervirales* (family IM_09) while HOT_MAG22 is the only member of the *Pimascovirales* we identified. This suggests that these viruses are either quite rare or typically found in low relative abundances and cannot be readily detected with metagenomic methods. Although these HOT_MAGs could only be detected at Station ALOHA, it seems unlikely that they are endemic to this region given that the prevailing conditions at this station are representative of widespread oligotrophic waters.

We examined the bioGEOTRACES GA02 transect in more detail because it traverses the north and south Atlantic and therefore allows for examination of trends in viral abundance across both latitude and depth (Fig. 4). Many of the MAGs that fall within the *Imitervirales* were mostly concentrated in oligotrophic surface waters (above 100m), consistent with their initial identification at Station ALOHA. These MAGs were identified across a wide range of latitudes; while HOT_MAG3 showed the highest prevalence near the equator, HOT_MAGs 10, 12, and 13 could be detected at latitudes near 40 degrees. Among the MAGs within the *Algavirales*, we detected only the three *Prasinoviridae* MAGs in the GA02 transect metagenomes. Two of these HOT_MAGs (20 and 30) showed high abundance both in equatorial waters as well as high northern latitudes (Fig. 4). In equatorial waters all prasinovirus MAGs were found in waters above 100m, but in the north Atlantic HOT_MAGs 20 and 30 could be found in waters near 200m, possibly due to sinking water masses in this area (i.e. overturning circulation).

Lastly, in order to evaluate temporal trends in viral abundance, we also examined the presence of the MAGs across a 1.5 year time-series at Station ALOHA from August 2010 to December 2011, which encompasses the same metagenomes used to generate these MAGs (Fig. 5 & Fig. 6). Here the *Algavirales* MAGs seem to be more prevalent than the *Imitervirales* MAGs during the sampling period from St. ALOHA. Among those MAGs that fall within the family *Mesomimiviridae*, all are present in surface waters (above 100m), and HOT_MAG12 was prevalent throughout almost the entire 1.5 year sampling period (Fig. 5). Within *Agavirales* MAGs, two members of the family *Prasinoviridae* (HOT_MAGs 20, 30) have almost the same pattern of their distribution and more likely to be present during August and September, while HOT_MAG14 is abundant in greater depths (100-500m) and prevalent in almost all seasons except for spring. HOT_MAG4, another *Agavirales* MAG, also has the same distribution trend as HOT_MAGs (20, 30), however, this MAG is highly abundant during November and December (Fig. 6). Overall, all of the MAGs within the *Algavirales* were prevalent in the 125m samples, which was just below the DCM for the majority of cases. HOT_MAG30, HOT_MAG20 and HOT_MAG4 were found most frequently at 125m, while HOT_MAG14 was also prevalent in several 200m samples; (Fig. 6). This transition below the DCM is consistent with the large-scale microbial community turnover that occurs in this region at Station ALOHA [36]. The only *Pimascovirales* MAG (HOT_MAG22), is distributed in shallow depths (0-100m) and is highly concentrated between December and January (Fig. 6).

The genomic repertoire encoded in the 11 MAGs

Giant viruses have complex genomes that encode various genes that are not commonly found in viral lineages, including components of central carbon metabolism and translation-associated proteins. As expected, we found giant viruses core genes

involved in DNA replication, transcription, and virion structure (Fig. 7). We did not find RNAP subunits in the prasinovirus MAGs, consistent with previous studies showing that this lineage lacks this enzyme [51]; (Fig. 7). The absence of RNAP subunits suggests that these viruses have a nuclear stage to their infection in which host RNAP is used for viral gene expression.

Asparagine synthase genes were identified among the *Imitervirales* MAGs, in particular among those members of the *Mesomimiviridae* (Fig. 7). This enzyme was previously identified in Phaeocystis globosa virus, though it remains unclear what role it may play during infection [62]. Collagen encoded genes were only found in *Imitervirales* MAGs belonging to the family IM_09, suggesting that these proteins may be a part of the structure of the virions of these viruses. Among the genes involved in DNA processing, chaperones from the heat shock protein families (HSP70, HSP90) which are presumably used for capsid protein folding [63], DNA mismatch repair (MutS), and histone acetyltransferase known to be functional for packaging DNA within the capsid [63,64], were mostly frequent among *Imitervilares* and *Pimascovirales* MAGs but were mostly absent from the *Algavirales* MAGs (Fig. 7).

Many of the MAGs encode multiple genes involved in central carbon metabolism, such as aldolase, 1-Deoxy-D-xylulose 5-phosphate synthase (DXP synthase), malate synthase, aconitase, and citrate synthase. Our findings indicate genes that belong to central carbon metabolism are mostly detected in *Mesomimiviridae* MAGs, notably those that are predicted to be involved in TCA (citrate synthase and aconitase); (Fig. 7). Interestingly, malate synthase and DXP synthase were found in *Pimascovirales* MAG and *Algavirales* MAG, respectively. Previous work has shown that enzymes involved in central carbon metabolism are quite common in many giant viruses [9], and our results here are consistent with those findings.

Previous studies have shown that rhodopsins and chlorophyll-binding proteins are quite common in a wide range of marine giant viruses [9,20–22]. Consistent with this, we found chlorophyll binding proteins in all three mesomimiviruses, and rhodopsin homologs in two (HOT_MAGs 10 and 13); (Fig. 7). This suggests that these viruses likely infect phototrophic or mixotrophic hosts and manipulate light harvesting machinery during infection. All mesomimivirus MAGs were detected in shallow waters at Station ALOHA, consistent with the prevalence of their hosts in well-lit surface waters. Interestingly, we detected three rhodopsin homologs in HOT_MAG10; it is unclear what role these three enzymes would play during infection, but the presence of three distinct homologs suggests that they are an important component of the infection strategy of many marine giant viruses.

Conclusion

Our study sheds light on the phylogenetic diversity, genomics, and distribution of giant viruses in oligotrophic marine waters. We present 11 MAGs of giant viruses that we reconstructed from metagenomes generated from Station ALOHA in the North Pacific Subtropical Gyre. These MAGs fall within five families in the orders *Imitervirales*, *Algavirales*, and *Pimascovirales*. Those MAGs that fall within the *Prasinoviridae* and *Mesomimiviridae* families are the most widespread and abundant, and several of these MAGs were detected in diverse bioGEOTRACES metagenomes that were collected in different ocean basins. Several of the MAGs were found consistently at Station ALOHA over a 1.5 year period, suggesting they are persistent community members in oligotrophic waters. The MAGs encoded a diverse range of functions, including genes involved in central carbon metabolism and light harvesting, suggesting that they use a variety of strategies to manipulate the physiology of their hosts during infection. Our work contributes to a growing body of research that suggests that large DNA viruses are abundant and widespread components of marine systems that play key roles in ecological dynamics and biogeochemical cycling.

Figure 1. Multi-locus phylogenetic tree of the 11 MAGs together with 1,381 reference genomes from the Giant Virus Database. The phylogenetic tree was constructed using 7 maker genes that are highly conserved in giant viruses (see Methods for details). According to the constructed phylogenetic tree, HOT_MAG22 belongs to the order *Pimascovirales* [PM_01]. HOT_MAG4 [AG_04], HOT_MAG14 [AG_01], HOT_MAG20 [AG_01] and HOT_MAG30 [AG_01] clustered within the order *Algavirales*. The rest of the MAGs, HOT_MAG3 [IM_09], HOT_MAG5 [IM_09], HOT_MAG60 [IM_09], HOT_MAG12 [IM_01], HOT_MAG13 [IM_01] and HOT_MAG10 [IM_01] belong the order *Imitervirales*.

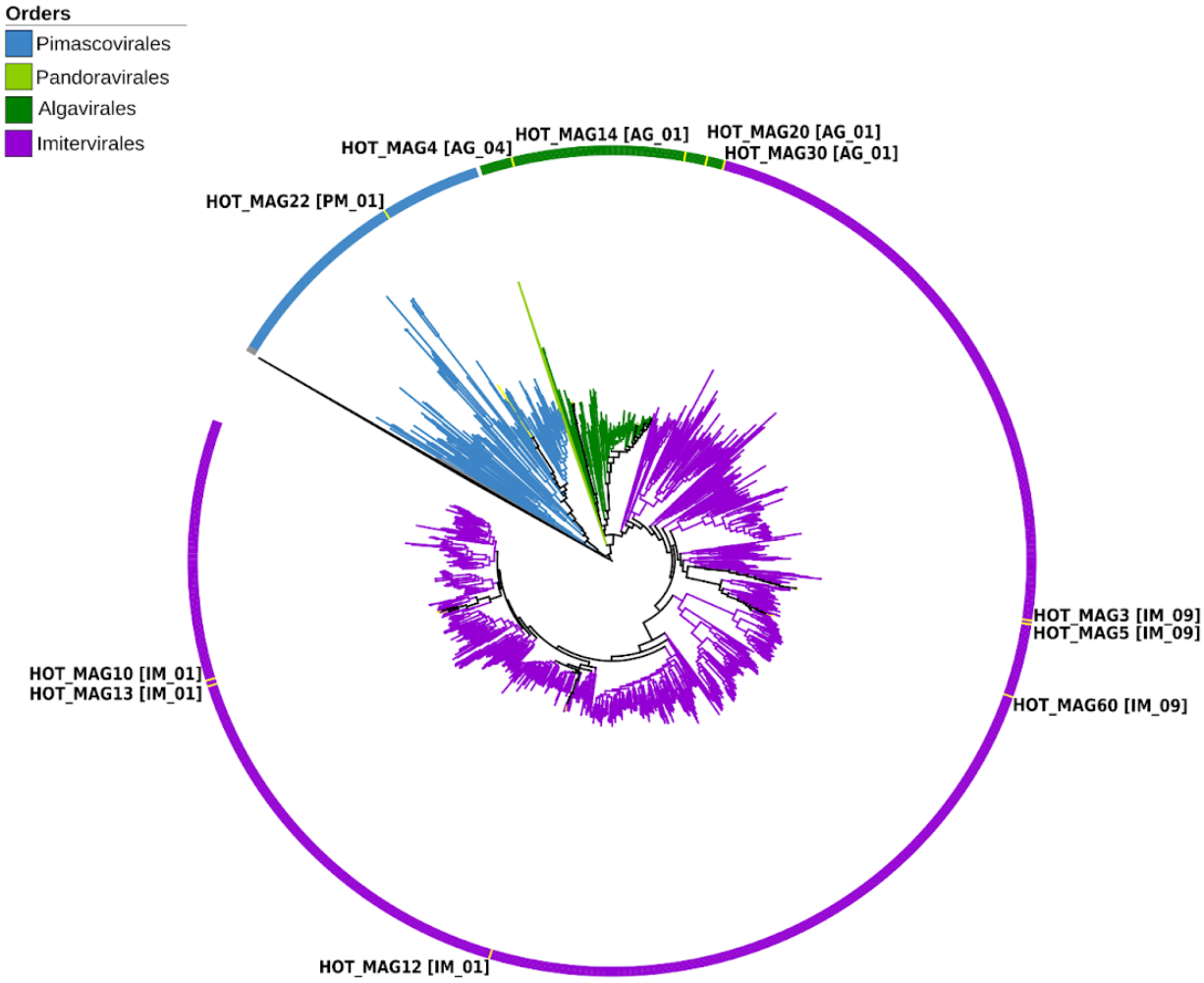


Figure 2. Best matches of predicted proteins identified in each HOT MAG. The results are based on a LAST-based homology search, with best hits retained (see Methods for details).

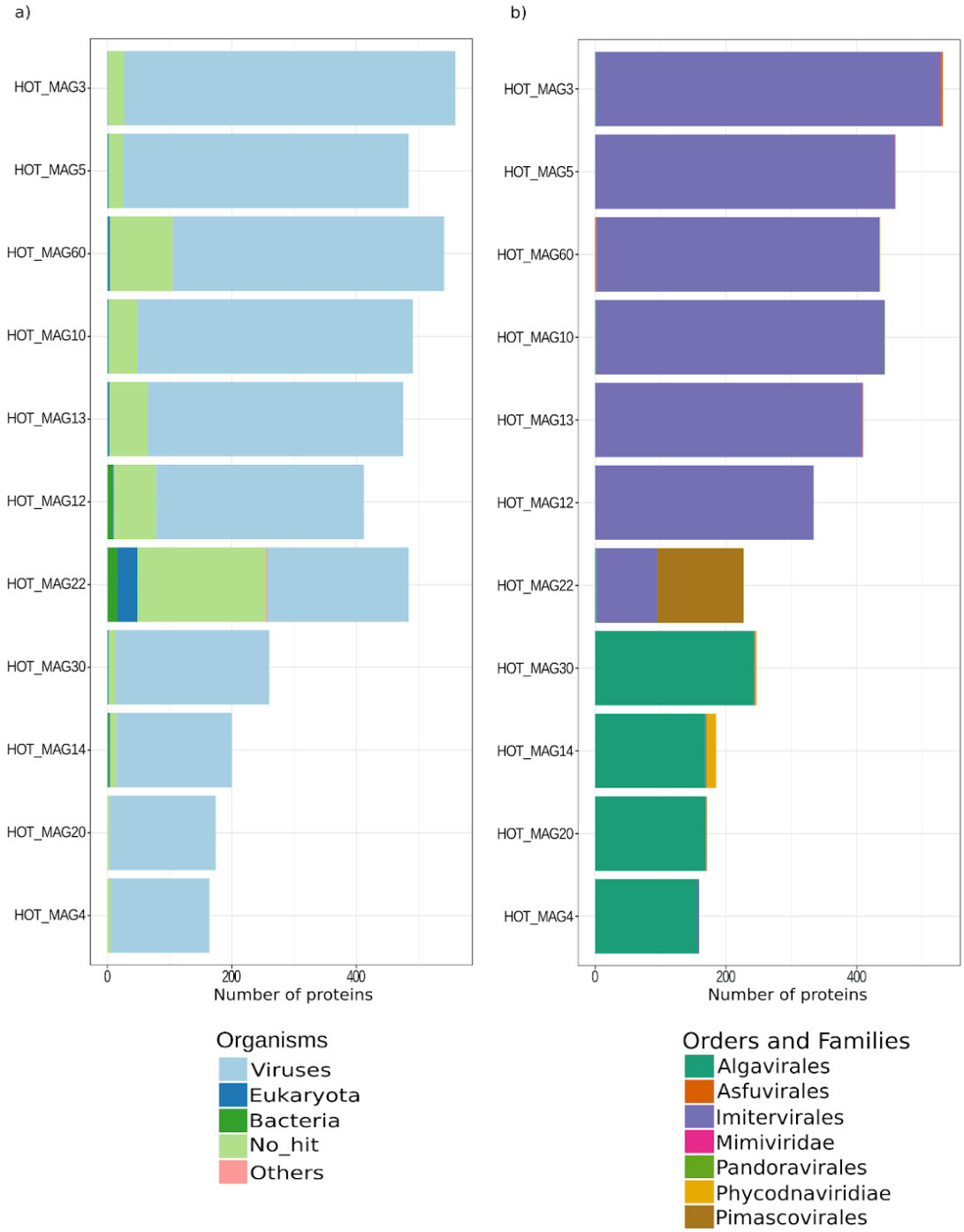


Figure 3. Environmental distribution of the HOT MAGs based on read-mapping of bioGEOTRACES metagenomes. HOT_MAGs (22,60) are not shown as they were not identified in bioGEOTRACES metagenomic dataset. HOT_MAG5 is not presented as it is closely related to HOT_MAG3 and its abundance is nearly the same as HOT_MAG3. Bubbles indicate the global distribution of the MAGs which was calculated based on RPKM. Moreover, the abundance of the MAGs in distinct sites with same latitude and longitude are reported in parenthesis; HOT_MAG3 (15), HOT_MAG12 (12), HOT_MAG13 (8), HOT_MAG10 (2), HOT_MAG30 (23), HOT_MAG20 (30), HOT_MAG14 (3), and HOT_MAG4 (3).

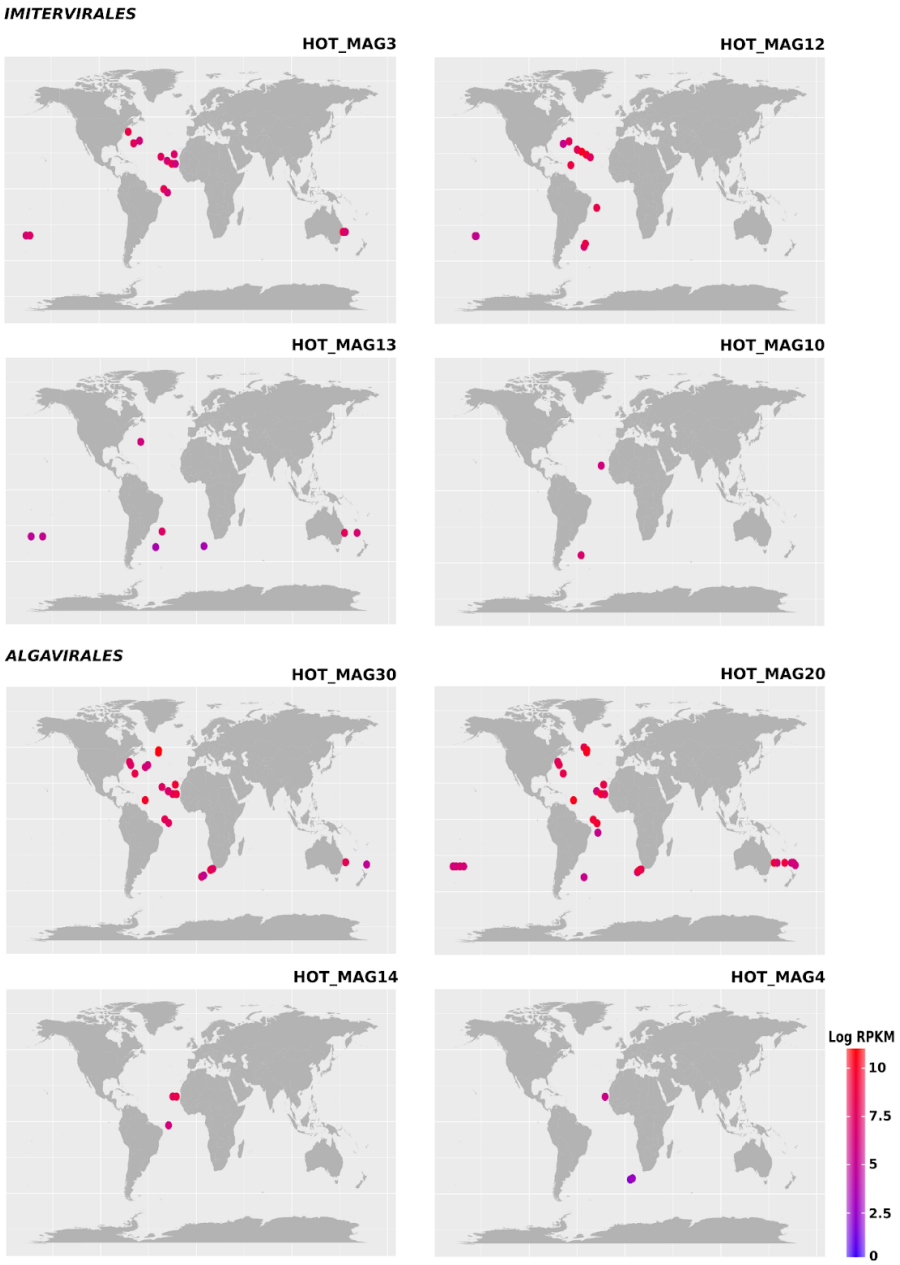


Figure 4. Environmental distribution of the HOT MAGs across a depth profile of the GA02 transect (North Atlantic to South Atlantic). The contour plots were drawn based on the results from mapping GA02 metagenome dataset onto the 11 MAGs (see methods for details). Total abundance of the MAGs were calculated based on RPKM. Y-axis and x-axis represent depths(m) and latitudes, respectively and the colorful bar refers to the log of total abundance (RPKM) for each of the MAGs (high abundance = red and low abundance = purple). HOT_MAGs (22, 4, 60) are not shown as they were not found in the GA02 metagenomic dataset. HOT_MAG5 is also not presented as it is closely related to HOT_MAG3 and its abundance is almost the same as HOT_MAG3.

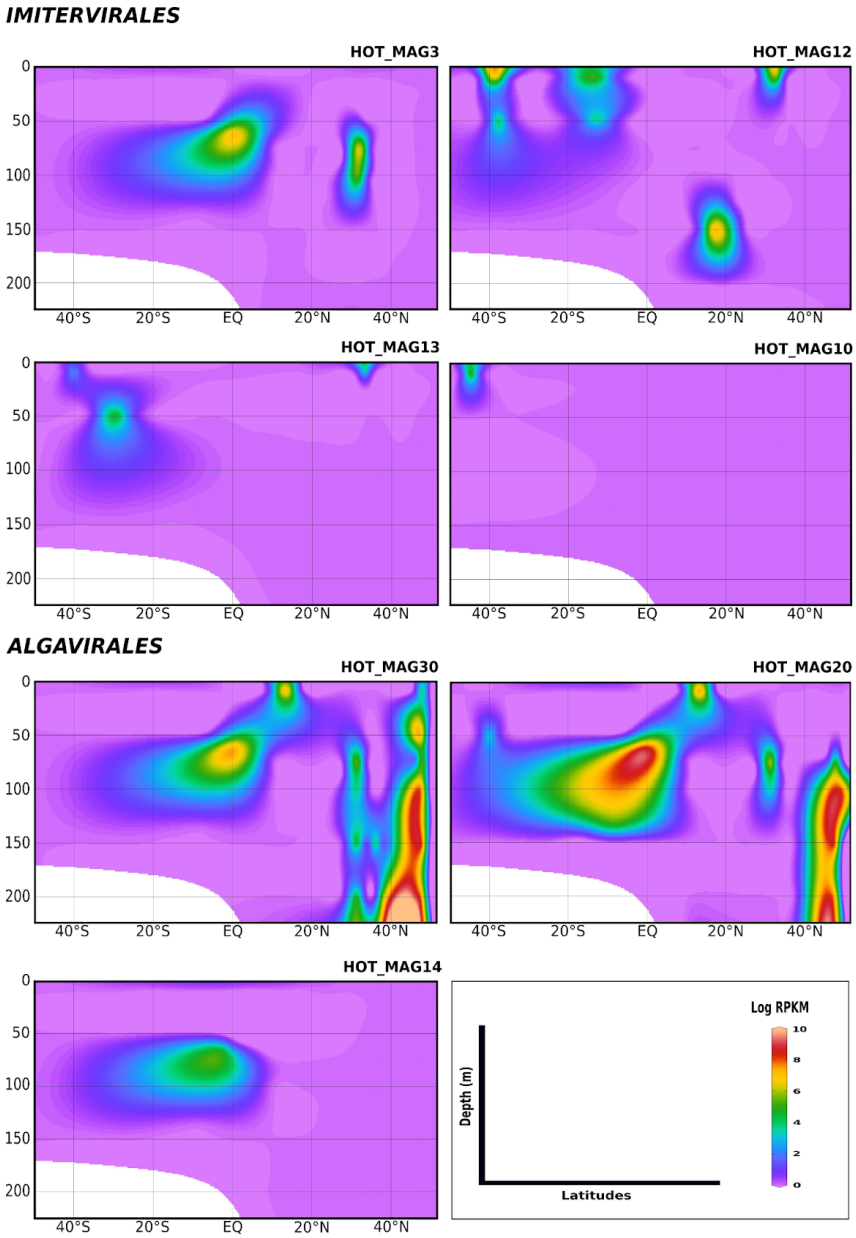


Figure 5. Distribution of 11 MAGs in the order *Imitervirales* in a 1.5-year sampling from St. ALOHA. Time series plots were drawn based on the results from mapping Hawaiian Ocean Time series metagenome dataset onto the 11 MAGs (see methods for details). Total abundance of the MAGs were calculated based on RPKM. Y-axis and x-axis represent depths(m) and latitudes, respectively and the colorful bar refers to the log of total abundance (RPKM) for each of the MAGs (high abundance = red and low abundance = purple). Only the distribution of the MAGs between 0-600m were shown in the plots.

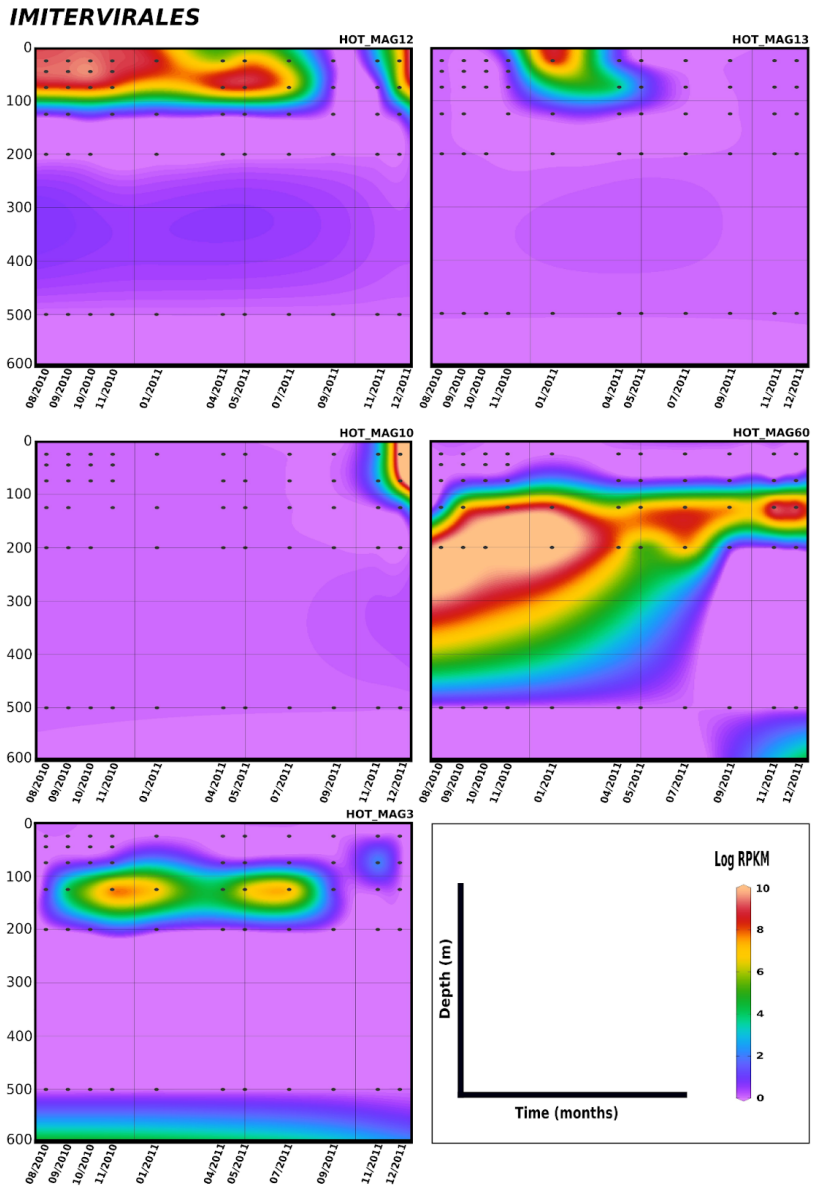


Figure 6. Distribution of 11 MAGs in the orders *Algavirales* and *Pimascovirales* in a 1.5-year sampling from St. ALOHA. Time series plots were drawn based on the results from mapping Hawaiian Ocean Time series metagenome dataset onto the 11 MAGs (see methods for details). Total abundance of the MAGs were calculated based on RPKM. Y-axis and x-axis represent depths(m) and latitudes, respectively and the colorful bar refers to the log of total abundance (RPKM) for each of the MAGs (high abundance = red and low abundance = purple). Only the distribution of the MAGs between 0-600m were shown in the plots.

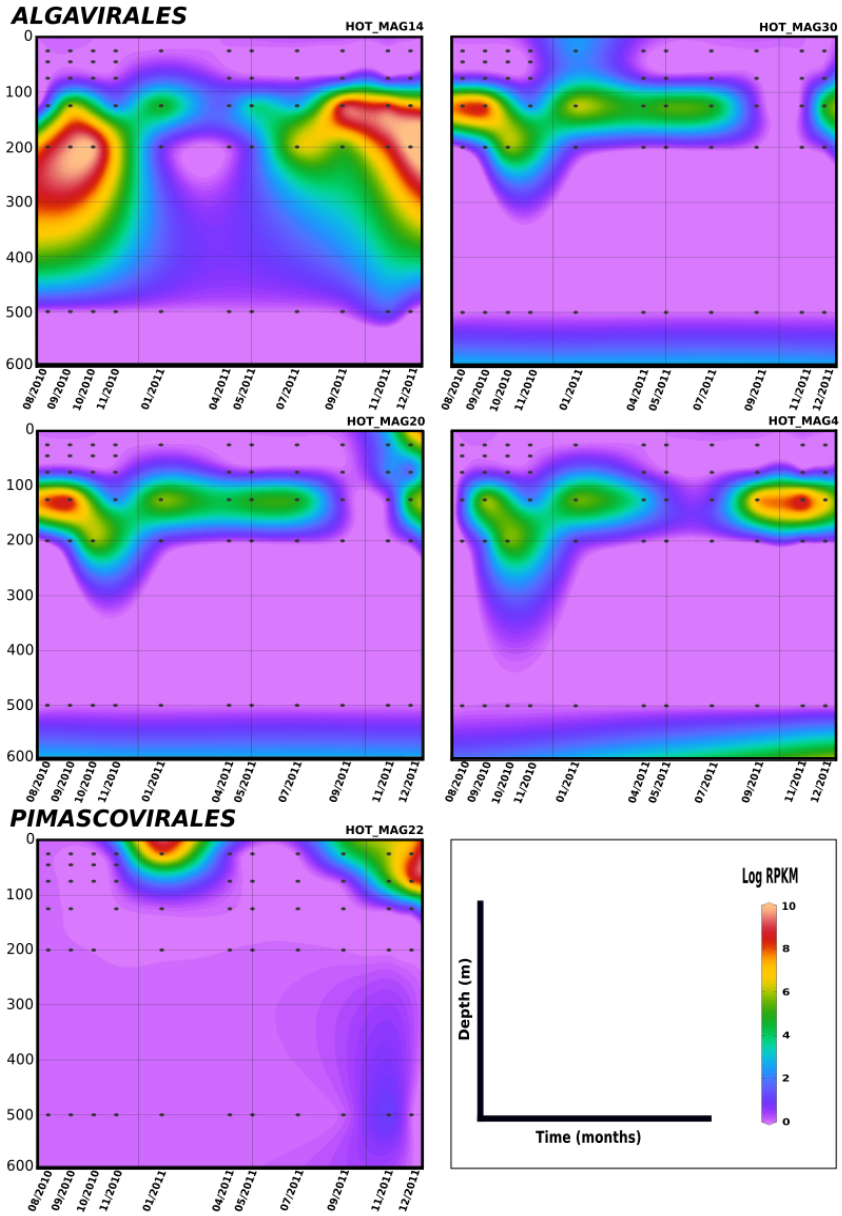


Figure 7. Distribution of selected genes within 11 MAGs of giant viruses.
The x axis indicates the 11 MAGs of giant viruses together with the viral families and orders that they belong to. The size of the bubbles indicate the total abundance of the genes within the genomes of 11 MAGs. The abbreviations of the encoded genes are as follows: MutS: [mismatch DNA repair], RNAP: [DNA-dependent RNA polymerase **L**arge and **S**mall subunits], SFII: [superfamily II helicase], TFIIB: [transcription factor IIB], A32: [A32 ATPase], VLTF: [viral late transcription factor 3], MCP: [major capsid protein], DXP synthase: [1-Deoxy-D-xylulose 5-phosphate synthase].

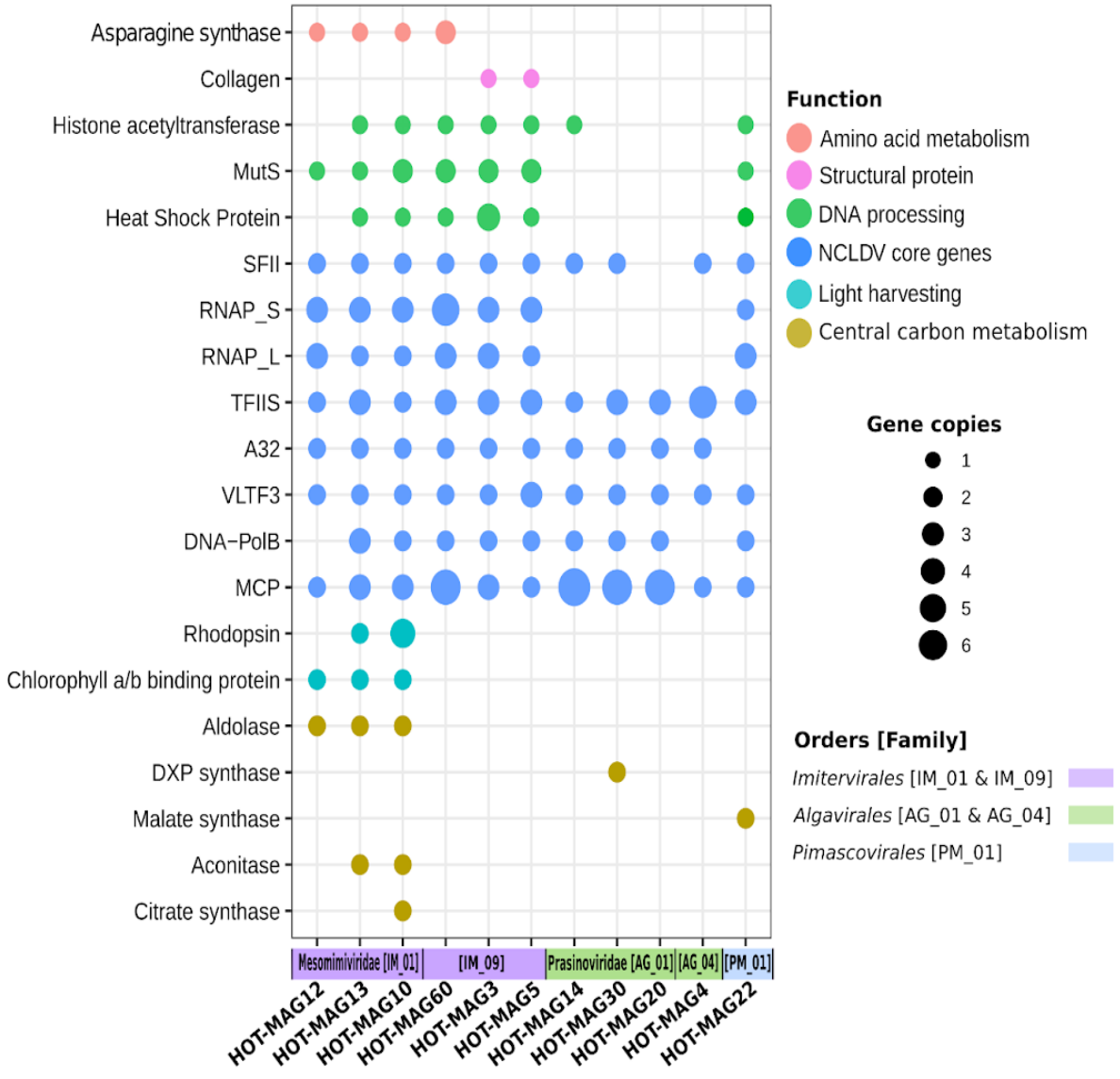


Table 1. General characteristics of 11 metagenome_assembeled genomes (MAGs) of giant viruses generated from St. ALOHA.

Genome	Genome Length	GC Content %	Num of protein coding genes	Order	Family	Genus
HOT_MAG4	122,808	38.47	164	*Algavirales*	AG_04	g175
HOT_MAG14	147,238	37.49	200	*Algavirales*	AG_01	g177
HOT_MAG30	171,883	34.21	260	*Algavirales*	AG_01	g177
HOT_MAG20	119,690	33.82	174	*Algavirales*	AG_01	g177
HOT_MAG12	386,441	33.29	412	*Imitervirales*	IM_01	g336
HOT_MAG13	433,885	32.35	475	*Imitervirales*	IM_01	g342
HOT_MAG10	426,436	29.37	491	*Imitervirales*	IM_01	g342
HOT_MAG3	574,081	31.33	559	*Imitervirales*	IM_09	g279
HOT_MAG5	477,804	30.13	484	*Imitervirales*	IM_09	g279
HOT_MAG60	489,708	24.69	541	*Imitervirales*	IM_09	g274
HOT_MAG22	471,006	34.86	484	*Pimascovirales*	PM_01	NA

References

1. Koonin EV, Yutin N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Adv Virus Res. 2019;103: 167–202. doi:10.1016/bs.aivir.2018.09.002

2. Fischer MG. Giant viruses come of age. Curr Opin Microbiol. 2016;31: 50–57. doi:10.1016/j.mib.2016.03.001

3. Karki S, Moniruzzaman M, Aylward FO. Comparative Genomics and Environmental Distribution of Large dsDNA viruses in the family *Asfarviridae*. doi:10.1101/2021.01.29.428683

4. Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M, Stough JMA. Standing on the Shoulders of Giant Viruses: Five Lessons Learned about Large Viruses Infecting Small Eukaryotes and the Opportunities They Create. PLoS Pathog. 2016;12: e1005752. doi:10.1371/journal.ppat.1005752

5. Aylward FO, Moniruzzaman M. Viral Complexity. Biomolecules. 2022;12. doi:10.3390/biom12081061

6. Chen F, Suttle CA, Short SM. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Appl Environ Microbiol. 1996;62: 2869–2874. doi:10.1128/aem.62.8.2869-2874.1996

7. Short SM, Suttle CA. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Appl Environ Microbiol. 2002;68: 1290–1296. doi:10.1128/AEM.68.3.1290-1296.2002

8. Short SM. The ecology of viruses that infect eukaryotic algae. Environ Microbiol. 2012;14: 2253–2271. doi:10.1111/j.1462-2920.2012.02706.x

9. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020;11: 1710. doi:10.1038/s41467-020-15507-2

10. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020. pp. 432–436. doi:10.1038/s41586-020-1957-x

11. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. MBio. 2019;10. doi:10.1128/mBio.02497-18

12. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, et al. Virophage control of antarctic algal host-virus dynamics. Proc Natl Acad Sci U S A. 2011;108: 6163–6168. doi:10.1073/pnas.1018221108

13. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, et al. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. Nat Ecol Evol. 2020;4: 1639–1649. doi:10.1038/s41559-020-01288-w

14. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J. 2013;7: 1678–1695. doi:10.1038/ismej.2013.59

15. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. A phylogenomic framework for charting the diversity and evolution of giant viruses. PLoS Biol. 2021;19: e3001430. doi:10.1371/journal.pbio.3001430

16. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. Nature. 2020;588: 141–145. doi:10.1038/s41586-020-2924-2

17. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci U S A. 2009;106: 21848–21853. doi:10.1073/pnas.0911354106

18. Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, et al. Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. J Virol. 2019;93. doi:10.1128/JVI.02130-18

19. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, et al. Giant viruses with an expanded complement of translation system components. Science. 2017;356: 82–85. doi:10.1126/science.aal4657

20. Yutin N, Koonin EV. Proteorhodopsin genes in giant viruses. Biol Direct. 2012;7: 34. doi:10.1186/1745-6150-7-34

21. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. Proc Natl Acad Sci U S A. 2019;116: 20574–20583. doi:10.1073/pnas.1907517116

22. Rozenberg A, Oppermann J, Wietek J, Fernandez Lahore RG, Sandaa R-A, Bratbak G, et al. Lateral Gene Transfer of Anion-Conducting Channelrhodopsins between Green Algae and Giant Viruses. Curr Biol. 2020;30: 4910–4920.e5. doi:10.1016/j.cub.2020.09.056

23. Ernst OP, Lodowski DT, Elstner M, Hegemann P, Brown LS, Kandori H. Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. Chem Rev. 2014;114: 126–163. doi:10.1021/cr4003769

24. Govorunova EG, Sineshchekov OA, Li H, Spudich JL. Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications. Annu Rev Biochem. 2017;86: 845–872. doi:10.1146/annurev-biochem-101910-144233

25. Gallot-Lavallée L, Archibald JM. Evolutionary Biology: Viral Rhodopsins Illuminate Algal Evolution. Current biology: CB. 2020. pp. R1469–R1471. doi:10.1016/j.cub.2020.10.080

26. Ha AD, Moniruzzaman M, Aylward FO. High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System. mSystems. 2021;6: e0029321. doi:10.1128/mSystems.00293-21

27. Kijima S, Delmont TO, Miyazaki U, Gaia M, Endo H, Ogata H. Discovery of Viral Myosin Genes With Complex Evolutionary History Within Plankton. Front Microbiol. 2021;12:

683294. doi:10.3389/fmicb.2021.683294

28. Da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO, Forterre P. Giant Viruses Encode Actin-Related Proteins. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac022

29. Ghedin E, Claverie J-M. Mimivirus relatives in the Sargasso sea. Virol J. 2005;2: 62. doi:10.1186/1743-422X-2-62

30. Monier A, Claverie J-M, Ogata H. Taxonomic distribution of large DNA viruses in the sea. Genome Biol. 2008;9: R106. doi:10.1186/gb-2008-9-7-r106

31. Karl DM, Church MJ. Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol. 2014;12: 699–713. doi:10.1038/nrmicro3333

32. Luo E, Aylward FO, Mende DR, DeLong EF. Bacteriophage Distributions and Temporal Variability in the Ocean's Interior. mBio. 2017. doi:10.1128/mbio.01903-17

33. Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. ISME J. 2020;14: 1304–1315. doi:10.1038/s41396-020-0604-8

34. Luo E, Leu AO, Eppley JM, Karl DM, DeLong EF. Diversity and origins of bacterial and archaeal viruses on sinking particles reaching the abyssal ocean. ISME J. 2022;16: 1627–1635. doi:10.1038/s41396-022-01202-1

35. Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proc Natl Acad Sci U S A. 2017;114: 11446–11451. doi:10.1073/pnas.1714821114

36. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2: 1367–1373. doi:10.1038/s41564-017-0008-3

37. Aylward FO, Moniruzzaman M. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. Viruses. 2021. p. 150. doi:10.3390/v13020150

38. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7: e7359. doi:10.7717/peerj.7359

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

40. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

41. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534. doi:10.1093/molbev/msaa015

42. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the

Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35: 518–522. doi:10.1093/molbev/msx281

43. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259. doi:10.1093/nar/gkz239

44. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21: 487–493. doi:10.1101/gr.113985.110

45. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45. doi:10.1093/nar/gkv1189

46. Wickham H. Ggplot2. Wiley Interdiscip Rev Comput Stat. 2011;3: 180–185. doi:10.1002/wics.147

47. Nguyen CD, Gardiner KJ, Cios KJ. Protein annotation from protein interaction networks and Gene Ontology. J Biomed Inform. 2011;44: 824–829. doi:10.1016/j.jbi.2011.04.010

48. Weinheimer AR, Aylward FO. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. ISME J. 2022;16: 1657–1667. doi:10.1038/s41396-022-01214-x

49. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, et al. Marine microbial metagenomes sampled across space and time. Sci Data. 2018;5: 180176. doi:10.1038/sdata.2018.176

50. Weynberg KD, Allen MJ, Wilson WH. Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review. Viruses. 2017;9. doi:10.3390/v9030043

51. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J Virol. 2010;84: 12555–12563. doi:10.1128/JVI.01123-10

52. Lopes dos Santos A, Gourvil P, Tragin M, Noël M-H, Decelle J, Romac S, et al. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. ISME J. 2016;11: 512–528. doi:10.1038/ismej.2016.120

53. Bellec L, Grimsley N, Derelle E, Moreau H, Desdevises Y. Abundance, spatial distribution and genetic diversity of Ostreococcus tauri viruses in two different environments. Environ Microbiol Rep. 2010;2: 313–321. doi:10.1111/j.1758-2229.2010.00138.x

54. Nagasaki K, Shirai Y, Tomaru Y, Nishida K, Pietrokovski S. Algal viruses with distinct intraspecies host specificities include identical intein elements. Appl Environ Microbiol. 2005;71: 3599–3607. doi:10.1128/AEM.71.7.3599-3607.2005

55. Leles SG, Mitra A, Flynn KJ, Tillmann U, Stoecker D, Jeong HJ, et al. Sampling bias misrepresents the biogeographical significance of constitutive mixotrophs across global oceans. Global Ecology and Biogeography. 2019. pp. 418–428. doi:10.1111/geb.12853

56. Gallot-Lavallée L, Blanc G, Claverie J-M. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate

Evolutionary Relationship with the Established Mimiviridae Family. Journal of Virology. 2017. doi:10.1128/jvi.00230-17

57. Stough JMA, Yutin N, Chaban YV, Moniruzzaman M, Gann ER, Pound HL, et al. Genome and Environmental Activity of a Chrysochromulina parva Virus and Its Virophages. Front Microbiol. 2019;10: 703. doi:10.3389/fmicb.2019.00703

58. Gann ER, Jackson Gainer P, Reynolds TB, Wilhelm SW. Influence of light on the infection of Aureococcus anophagefferens CCMP 1984 by a “giant virus.” PLoS One. 2020;15: e0226758. doi:10.1371/journal.pone.0226758

59. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, et al. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology. 2014;466-467: 60–70. doi:10.1016/j.virol.2014.06.031

60. Johannessen TV, Bratbak G, Larsen A, Ogata H, Egge ES, Edvardsen B, et al. Characterisation of three novel giant viruses reveals huge diversity among viruses infecting Prymnesiales (Haptophyta). Virology. 2015;476: 180–188. doi:10.1016/j.virol.2014.12.014

61. Blanc-Mathieu R, Dahle H, Hofgaard A, Brandt D, Ban H, Kalinowski J, et al. A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. J Virol. 2021. doi:10.1128/JVI.02446-20

62. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, et al. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc Natl Acad Sci U S A. 2013;110: 10800–10805. doi:10.1073/pnas.1303251110

63. Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, et al. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. Genome Research. 2010. pp. 664–674. doi:10.1101/gr.102582.109

64. Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology. 2010;53: 284–292. doi:10.1159/000312913

Chapter 2: Benchmarking metagenomic approaches for recovering large DNA viruses

Abstract

Metagenomics has transformed our understanding of the microbial world and led to the discovery of many previously unknown microbial lineages. Viral metagenomics in particular, has allowed for the identification of viruses that have remained elusive due to the challenges of laboratory cultivation. To date most viral metagenomic studies focus on the analysis of individual sequenced contigs that are produced through short-read assembly programs. However, it has recently become clear that many viral genomes, in particular those belonging to large DNA viruses, are highly fragmented in metagenome assemblies. Due to this complication, it is often necessary to bin multiple contigs together to recover draft viral genomes rather than relying on single-contig analyses alone. This study highlights the importance of binning approaches in viral metagenomics, particularly for recovering large DNA virus genomes, and emphasizes on the need for developing benchmarking workflows for viral metagenomics by investigating the correlation between assemblies completeness and viral genome size and their coverage depth. Moreover, biodiversity metrics (Shannon diversity, evenness, and richness) were evaluated for contigs, bins and genomes populations of each sample, and biodiversity metrics of contigs and bins populations were compared against the true Shannon diversity, evenness and richness.

Keywords: Large DNA viruses, Viral metagenomics, CAMISIM, metaBAT2, Shannon diversity, evenness, richness

Introduction

Over the past years, studies have underscored the importance of large DNA viruses in the ocean and their role as major regulators in marine environments. It has indicated that large DNA viruses are involved in many crucial processes, such as biogeochemical cycling, nutrient cycling, termination of algal blooms, host evolution, and gene exchange [1,2]. Despite their crucial role in aquatic ecosystems, little is known about the characteristics of these viruses due to the few cultivated microbial hosts and the difficulties associated with cultivating them in the lab [3]. Metagenomic analysis has addressed this issue by allowing researchers to examine microbial diversity and genomics without needing cultivation methods [4,5].

Metagenomics is a powerful tool for evaluating microbial diversity in the environment. A major advance in the last ~15 years has been the ability to reconstruct draft genomes of

microbes from environmental samples through direct short-read shotgun sequencing. In a typical metagenomic analysis, samples are collected from an environment (e.g., soil, water, human gut), and then DNA is extracted. Extracted DNA is fragmented and sequenced with NGS technology to create a library consisting of short-read sequences. Lastly, short-read assemblers combine reads and form contigs. Subsequently, contigs with similar base composition and coverage will be clustered to generate bins.

Given the widespread distribution of large DNA viruses, such as jumbo phages and giant viruses, limited methods have been developed for isolating and examining the diversity of these viruses [2,6–8]. Traditionally, viruses were considered small biological agents [9], and the discovery of large DNA viruses has challenged this conventional view of viral complexity. Due to potential biases and difficulties associated with cultivating large DNA viruses in laboratory settings [3,10,11], bioinformatic and metagenomic analysis can facilitate understanding these viruses [12].

With metagenomics analysis, several studies have been able to recover genomes of large DNA viruses from their habitats, and they often assemble up to a contig or scaffold level. This approach might be efficient for smaller viral genomes, while the genomes of large DNA viruses might remain fragmented in contigs and not fully recovered. Genome recovery is often accomplished at a binning level for bacterial and archeal genomes, where contigs or scaffolds are assembled into bins or metagenome-assembled genomes (MAGs) [13]. Previous studies have proven that a similar approach, using a binning tool, can obtain complete genomes of large DNA viruses [6,8,14,15].

While binning tools have proven successful in recovering viral genomes, standardized methods for benchmarking viral metagenomics and evaluating the performance of binning tools are still lacking. Existing benchmarking methods primarily rely on algorithms to detect mis assemblies and calculate genome completeness within bins.These methods have predominantly focused on bacterial and archaeal genomes, therefore, there is a strong need to develop efficient software or workflows to benchmark viral metagenomics and assess binning processes. There are complications associated with binning tools and not only viral genomes, but also bacterial and archeal genomes might be negatively affected in terms of being correctly recovered during metagenomics analysis. For instance, in the binning process bins may contain genome fragments from other organisms, which is known as mis-binning [16]. It is worth noting that the risk of mis-binning is amplified when the length of contigs or scaffolds is small (e.g., < 5kp), as these are often considered unreliable fragments or noise in metagenomic analysis [16]. Moreover, in the field of metagenomics for bacterial, archeal and viral genomes, one of the main challenges is that microbial genomes are often low complex and the presence of repeated sequences can highly affect the efficiency of the

assemblers. In this study, we particularly focus on viral genomes and viral metagenomics. We aim to employ contigs of sufficient length to minimize mis-binning and evaluate the quality of our bins based on their completeness within their reference genome and their contamination rate.

Research objectives

In this study, we aim to develop a workflow for benchmarking viral metagenomic analysis, particularly for large DNA viruses (>200 kp genome length). Our goal is to evaluate the performance of the current binning tools in recovering complete genomes and develop best practices for this research. For benchmarking purposes, there are several computational and laboratory-based approaches. For example, a previous study proposed a workflow that added spiked particles of a large DNA virus (Fadolivirus) into a wastewater sample and assessed the performance of multiple binning tools [17]. As a result, They could recover an almost complete length of the Fadolivirus genome. Despite their success in regenerating a full-length viral genome, their analysis encountered some challenges. Depending on different binning tools, some contigs with lengths from 15 kb to 351 kb were wrongly assigned to the Fadovirus reference genome, which highlights the limitations and challenges regarding the performance of binning tools and metagenomic analysis as a whole [16,17]. In general, laboratory-based benchmarking workflows can be time-consuming and challenging to evaluate a broad scale of metagenomic samples [18–21].

To avoid the need for laboratory experiments, we aim to use metagenome simulators to generate realistic mock community datasets from microbial genomes with known compositions [22,23]. Employing an annotated benchmark dataset will allow us to understand the metagenome content better and benchmark the underlying dataset in a relatively short time. Also, having an annotated benchmark dataset allows us to know the exact viral composition in the samples and thus the performance of the assembly and binning tools can be precisely evaluated.

We examined the performance of metaBAT2 [24] as a binning tool for this analysis. Based on previous findings [8,14], metaBAT2 is effective at binning viral contigs. This tool bins contigs together if they have similar nucleotide composition (assessed through calculation of tetranucleotide frequencies (TNFs)) and sequence coverage (a proxy for relative abundance). Although metaBAT2 was developed for recovering bacterial and archaeal genomes, binning based on nucleotide composition and sequence coverage would be expected to be equally effective for viral genomes.

Methods

Used dataset

We used a publicly available database called INPHARED (INfrastructure for a PHAge REference Database), which contains a large set of complete bacteriophage genomes [25]. It is worth noting that we did not use giant viruses genomes as they have more complex genomes and in fact, INPHARED dataset contains the most up to date bacteriophage sequences (viral genomes) which makes it a beneficial dataset to be used for bioinformatic analysis and in viral metagenomics. Briefly, INPHARED contains 14,244 genomes with genome sizes ranging from 3.1 to 624.4 kp [25]. In addition, 2.2% of the INPHARED database consists of jumbo bacteriophages, which are known to have genomes larger than 200 kbp [26].

Metagenome simulation

CAMISIM (Critical Assessment of Metagenome Interpretation) used the INPHARED dataset as an input to generate 1GB of simulated paired-end reads for each sample. A total of 7 samples were produced with different sets of log sigma values; table 1 [23,27]. The relative abundance of reads in each sample was derived from a log-normal distribution where log mu is set to 1 and standard deviation (log sigma) was different for each sample. By utilizing CAMISIM, we can provide simulated reads from viral communities with known composition, which enable us to benchmark bioinformatic methods in recovering viral genomes.

Read assembly

The *de novo* assembly of short reads is an essential component of any viral metagenomic analysis [28–34]. We used metaSPAdes [35], a variant of SPAdes assembly known to be efficient for metagenomics, with default parameters to assemble the simulated reads to create contigs. MetaSPAdes platform uses de Bruijn graphs to solve challenges regarding microdiversity or the presence of multiple strains and unequal sequence coverage [35–37].

Binning contigs

For binning contigs, metaBAT2 [24] requires a depth of coverage between generated contigs and the reference genomes; therefore, we used coverM contig v. 0.6.1, with metabat option, and mapped back the contigs onto the reference genome and retained those with more than 95% coverage. Subsequently, metaBAT2 [24] (parameters -s 10000 -m 5000 --minS 75 --maxEdges 75 --unbinned) was used to cluster similar contigs to form bins based on TNF (tetranucleotide frequency) and sequence coverage which are consistent among viral genomes. MetaBAT2 collected bins and contigs larger than 10000 bp and 5000bp, respectively.

Biodiversity metrics calculation (Shannon/alpha diversity, richness, and evenness)

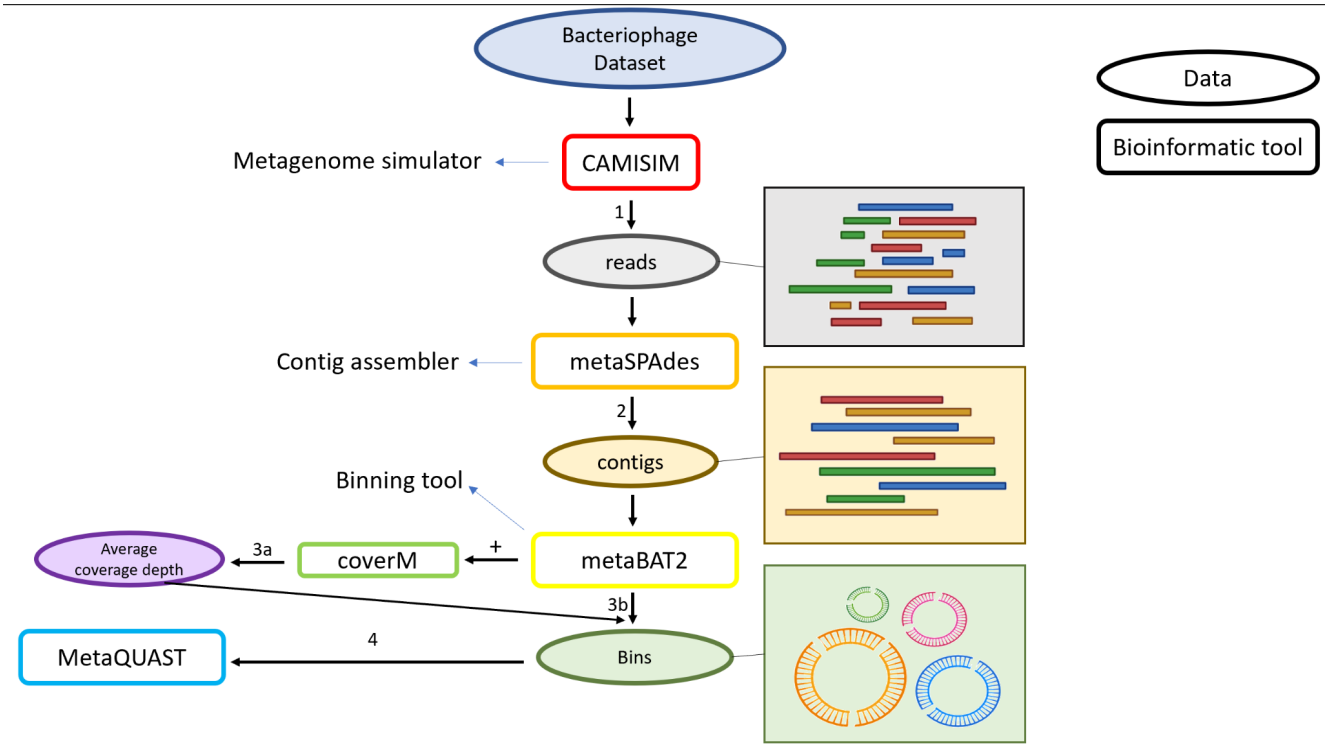
To assess the quality of the bins and the efficiency of the binning tool across our simulated samples with diverse standard deviations (log sigma), it is crucial to compute biodiversity metrics (Shannon/alpha diversity, richness, and evenness) for both contigs and bins in each sample. We mapped contigs and bins of each sample on their original reads with coverm (coverm genome/contig -m covered_fraction rpkm --min-covered-fraction 20) and retained those that are longer than 5kp (contigs) and 10kp (bins) and have more than 20% coverage. Subsequently, we calculated biodiversity metrics (Shannon/alpha diversity, evenness, and richness) in R studio with the “vegan” package [38] and plotted them against each other in base R; table 2, figure 1.

Additionally, CAMISIM provides the relative abundance of each genome in a given sample. Therefore, we calculated the biodiversity metrics using CAMISIM-generated datasets to get the true values of each of these metrics and then compared them with biodiversity metrics of bins and contigs; table 3, figure 1.

Quality assessment of bins

To have more insight into the assemblies' lengths and be able to compare them to their reference genome, MetaQUAST is a tool that we used to check the quality of our bins. MetaQUAST is an updated version of QUAST, and it provides general statistics of our bins and it examines the assignment of each assemblies to its reference genome. MetaQUAST output files contain a comprehensive assembly report, including general statistics of aligned assemblies, a comparison of all assemblies, and metrics information. In order to detect good quality bins, we calculated the completeness and contamination of each bin in a given sample.

Workflow



Results and discussion

Binning is essential to get an accurate view of viral diversity

In accordance with previous study, the retrieval of the genetic material of large DNA viruses (e.g., giant viruses and jumbo bacteriophages) can be challenging due to their large genome size and it necessitates the implementation of binning approaches to fully or almost fully recover their genomes [13]. Without a binning step, their genomes will often remain fragmented into different contigs.

For benchmarking purposes, we used metagenome simulated short reads generated by CAMISIM and compiled the simulated reads in 7 individual samples. MetaSPAdes and metaBAT2 were used as the contig assembler and the binning tool, respectively. Results have shown that all of the samples contain bins with multiple contigs, ranging from ~60% (sample 1) to 9% (sample 7); table 1. These results point out the crucial role of binning steps to recover the genomes of large DNA viruses in metagenomics studies.

Bin quality assessment

Calculating Shannon diversity (alpha diversity) for bins and contigs

Metagenomics is a valuable tool for assessing the diversity of viruses in a particular environment, but evaluating the diversity of a community in an unbiased way has always been challenging [39]. We calculated the relative abundance of contigs and bins in each sample by mapping them against their original sets of reads, and those with more than 20% coverage were retained for further analysis. The 20% coverage cutoff might exclude rare viruses and makes it difficult to detect them if they have coverage of less than 20%.

Based on the results, overall, Shannon diversity indexes were higher for contigs than bins; table 2. This observation aligns with the fact that the Shannon diversity index is typically more inflated for contigs than bins, as contigs are more dispersed and diverse in each sample; table 2. Furthermore, since each sample generated by CAMISIM has drawn from a log-normal distribution and has diverse standard deviations, samples with higher standard deviations tend to be more spread. Consequently, as the standard deviation increases for each sample, Shannon diversity decreases, indicating that contigs/bins in samples with higher log sigma (standard deviation) are less diverse; table 1.

Lastly, we compared the Shannon diversity of contigs/bins with the CAMISIM Shannon diversity (true Shannon diversity); tables 2 & 3, figure 1. The results revealed that Shannon diversity of contigs compared to true Shannon diversity in their respective sample, remained similar or higher; tables 1 & 2, figure 1. For the bins population in each sample, nearly half of the samples exhibit the same Shannon diversity index compared to the true Shannon diversity index, while the remainder had lower Shannon diversity index.

Calculating evenness and richness for bins and contigs and their relation with bin recovery rate

We measured other aspects of biodiversity (evenness and richness) and discussed their relations with the bins' recovery rate (see materials and methods). Both evenness and richness positively correlate with the recovery rate of bins with more than 50% completeness and less than 10% contamination for most of the samples; tables 1&2. Sample 1, with the highest evenness, richness and Shannon diversity for bins and contigs, has an 18.5% recovery rate; tables 1&2. Although metaBAT2 recovered many bins, even in the best case, the recovery rate is low (~20%). It is important to note that metagenomics primarily captures the most abundant viruses, and numerous members of the rare viruses cannot be effectively recovered.

In addition, we compared the richness and evenness of bins and contigs with true richness and true evenness (information derived from CAMISIM); figure 1. Contigs and bins evenness are similar to each other indicating that bins and contigs have almost similar distribution patterns ; figure 1, table 2 & 3. Moreover, the evenness of contigs and bins are higher than true evenness (CAMISIM_evenness), indicating that contigs and bins populations only contain highly abundant viruses while CAMISIM generated dataset consists of both high and rare abundant viruses which result in less even population. For richness, overall CAMISIM richness is higher than richness for contigs and bins, except for sample 1, where the number of contigs is more than the number of genes; figure 1, table 2 & 3.

Bin features (completeness and contamination)

Evaluating the quality of the bins is essential as contaminated bins can lead us to wrong ecological and evolutionary insights. In this study, we aimed to recover bins with more than 50% completeness and less than 10% contamination. Completeness, in this context, refers to whether an assembly (bin) is fully covered by its reference genome and is measured with total alignment length (between the reference genome and bin) divided by the whole reference genome size. The other term, contamination, explains a condition when multiple contigs belonging to different organisms are wrongly assembled together. A pure bin will be defined as a bin that consists of similar contigs and fully assigned and covered by its reference genome. We calculated contamination by dividing the total alignment length by the total bin size subtracted by 100.

$$Completeness\% = \frac{total\ aligned\ length}{reference\ genome\ size} \times 100$$

$$Contamination\% = 100 - \left(\frac{total\ aligned\ length}{total\ bin\ size} \times 100\right)$$

Furthermore, we examined the effects of genome size and genome coverage on completeness and contamination as a whole for all the samples and the results are presented with plots made by the ggplot2 package in R software [40]. Completeness exhibits a negative correlation with genome size, as the genome size increases, the completeness rate decreases, making it more challenging to recover larger genomes. While higher genome coverage between the reference genome positively affect the completeness; figure 2a. In addition, it seems contamination might increase more with larger genomes, whereas it decreases with high genome coverage; figure 2b.

Conclusion and future directions

Binning is critical for short-read metagenomes

Viral metagenomics has greatly expanded the study of unknown viral communities in their natural habitats. Shotgun metagenomic sequencing has provided valuable insights

into microbial composition. For instance, the short-read metagenomic assembly has facilitated the discovery of large DNA viruses and shed light on their ecological and biogeochemical roles, taxonomic diversity, and genomic repertoire. Although, over the past years, short-read metagenomic approaches have been helpful in discovering many of unknown DNA viruses by recovering their genomes, this method still has its limitations and needs to be evaluated effectively.

One approach, particularly suitable for large DNA viruses, is using binning tools to create bins that recover the full length of large DNA viruses' genomes (>200kbp). We employed metaBAT2 to recover high-quality bins (> 50% completeness & < 10% contamination). However, only the most abundant viruses with higher richness and diversity were more likely to be recovered and identified (the highest recovery rate was less than 20%). Long-read metagenomes may help resolve complete genomes and obviate the need for binning [41].

Short-read metagenomic assemblies may increase the rate of mis assemblies in viral metagenomics, where only the most abundant viruses are likely to be recovered. To overcome these challenges, the usage of long-read metagenomic reads has been proposed. Long-read metagenomes may eliminate the need for binning and have successfully regenerated viral genomes [41]. This method has also successfully identified regions of viral genomes (e.g. termini, DTR, rare genomic island regions) that have traditionally been difficult to detect [42,43]. However, long-read metagenomic approaches have their limitations in terms of acquiring high concentrations of viral loads to initiate the process [44] and having high error rates [45]. Considering the limitations of short and long read assemblies, previous study indicated that using short reads for correcting errors of long reads will result in more reliable long reads which are able to capture highly abundant viruses that were previously not detected due to the short-read metagenomics biases [46].

Although metagenomics has made significant progress, we are still in its early development stages. The application of long-read metagenomic sequencing holds promise as it may eliminate the necessity for binning to recover large DNA viruses’ genomes. Combination of long and short read sequences in metagenomics might increase the recovery rate of a broad range of viruses. Continued research and efforts are necessary to fully understand the capabilities and limitations of long-read metagenomics and viral metagenomics in general, to produce valid results.

Figure 1. Comparison of biodiversity metrics (Shannon diversity index, evenness, and richness) of three individual populations (contigs, bins, original genomes (CAMISIM dataset)) for all the seven simulated samples. a) Shannon diversity index of contigs and bins are similar to the true Shannon diversity index (CAMISIM_Shannon diversity). Contigs Shannon diversity is inflated compared to bins Shannon diversity indicating a more diverse population of contigs. b) Bins and contigs populations seem to have the same relative abundance, and are much higher than the true evenness (CAMISIM_evenness), suggesting that CAMISIM contains both high and low abundant species while contigs and bins populations only represent the highly abundant ones. c) True_richness (CAMISIM_richness) is higher than almost all the contigs and bins populations. Contigs richness seems to be always higher than bins richness.

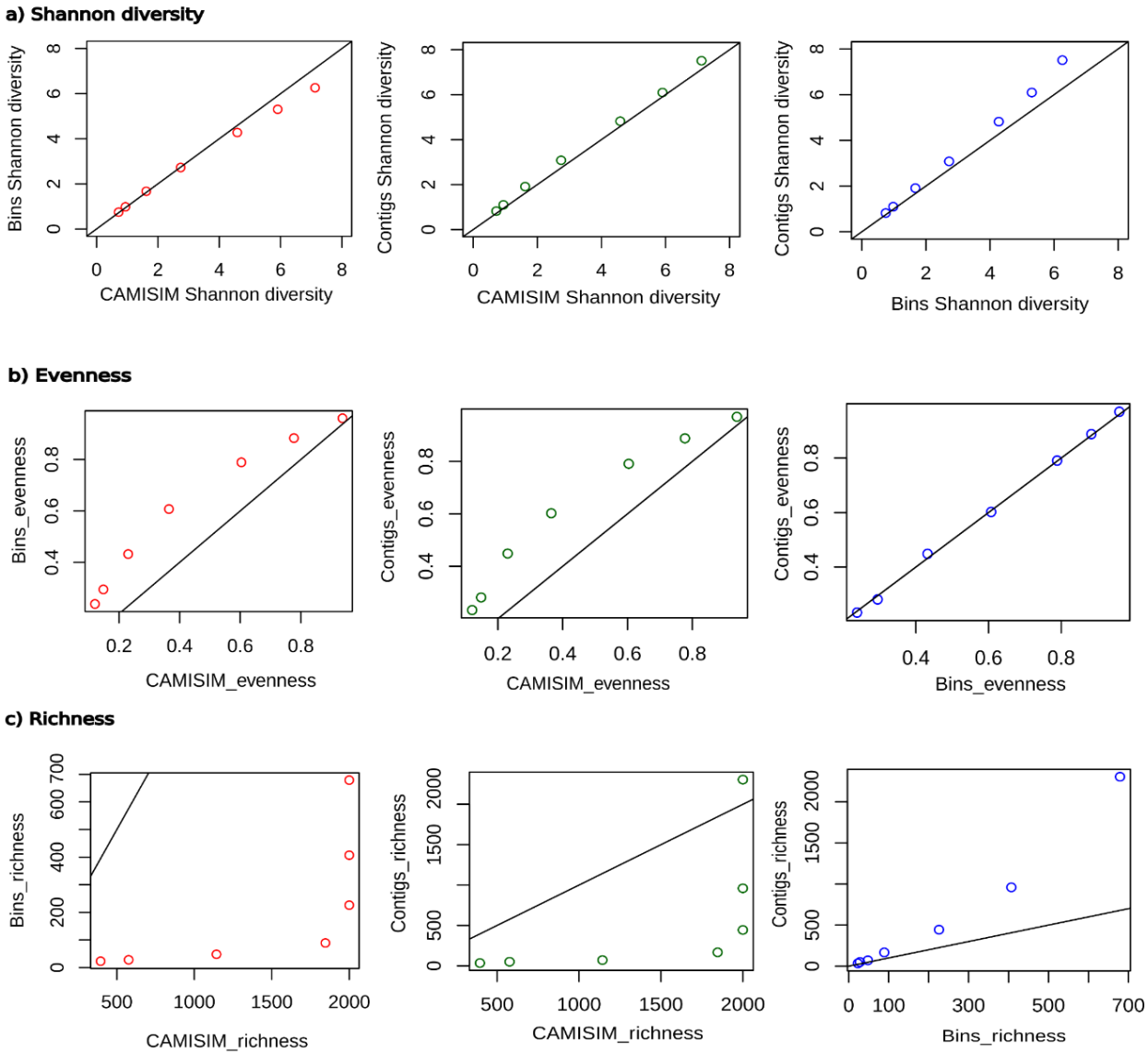


Figure 2. Correlations between the bin completeness and bin contamination versus genome size (kb) and genome coverage (log 10). a) completeness can be negatively affected by the larger genome size, while higher genome coverage can positively affect the completeness. b) contamination seems to be affected more by the larger genome size, whereas, higher genome coverage can reduce the contamination.

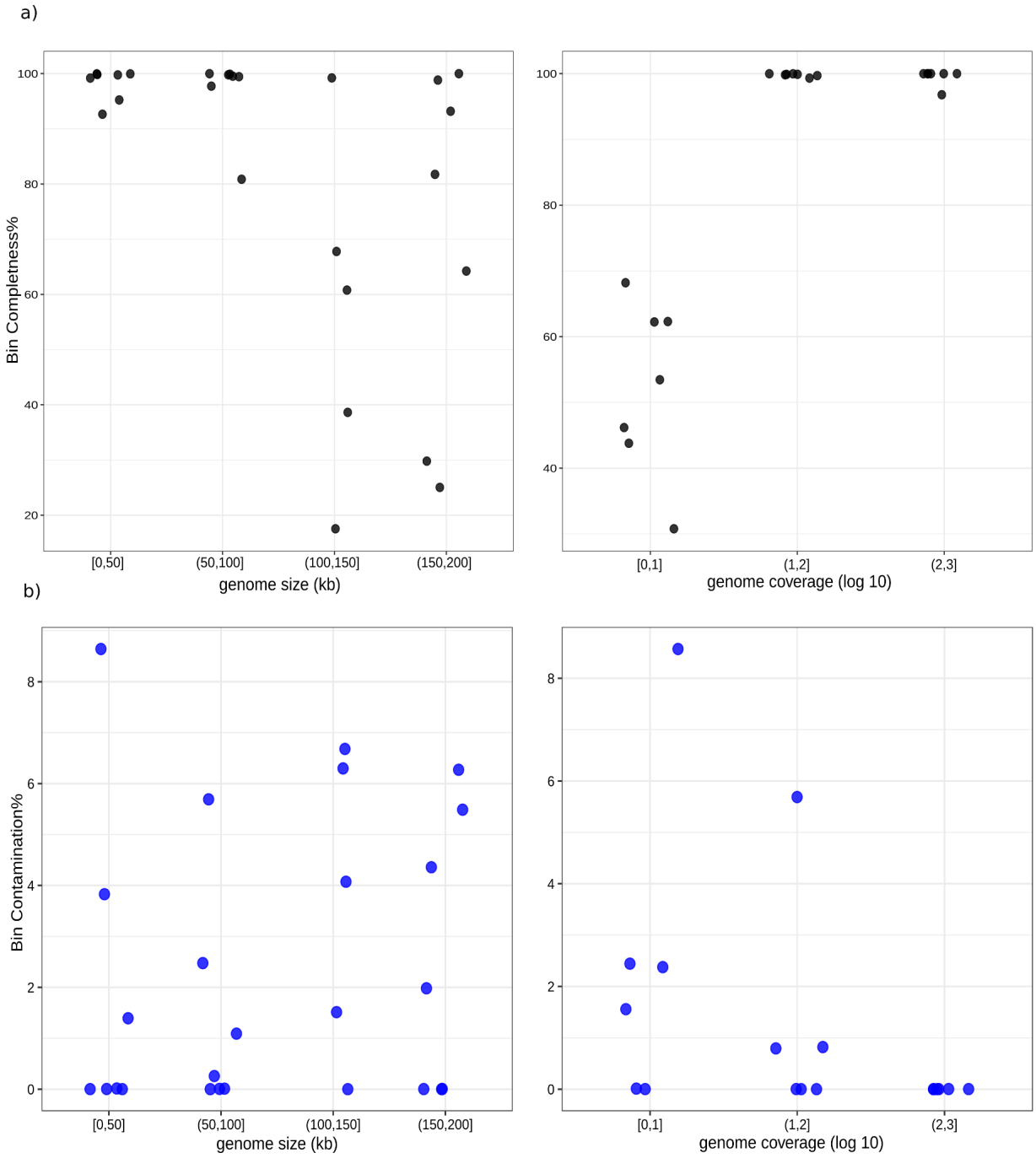


Table 1. General Information of all the seven simulated samples generated by CAMISIM. The abbreviations of the columns name are as follows: num of genes/contigs/bins: number of genes, contigs and bins in each sample, Avg num of contig(s) in a bin: average number of contigs in an individual bin for each sample, Bins (contig(s)>=1)% : the percentage of bins with multiple contigs in a given sample, num bins (>50% completeness & <10% contamination): number of bins which are more than 50% complete and less than 10% contaminate, recovery rate%: the percentage of bins that binning tool (metaBAT2) could recover for each sample, log_sigma: standard deviation that was used by CAMISIM to draw each sample from a log normal distribution, Shannon diversity: α diversity index for each sample.

Samples	num of genes	num of contigs	num of bins	Avg num of contig(s) in a bin	Bins (contig(s)>=1)%	num bins (>50% completeness & <10% contamination)	recovery rate%	log_sigma	Shannon diversity
S1	2000	2024	679	3	57.29%	370	18.5%	1	6.260296
S2	2000	809	407	2	37.10%	279	13.95%	2	5.305778
S3	2000	366	226	2	27.87%	174	8.7%	3	4.275578
S4	1846	143	89	2	29.21%	72	3.90%	5	2.725157
S5	1143	62	48	1	14.58%	33	2.88%	7	1.672453
S6	576	43	28	2	9.30%	23	3.99%	9	0.9816084
S7	395	31	23	2	21.7%	13	3.29%	10	0.7462127

Table 2. Biodiversity metrics (Shannon diversity, evenness, and richness) of contigs and bins in each sample. These results represent the differences of biodiversity metrics of contigs and bins population and their impacts on the recovered viral genomes if they were assembled up to a contig level versus the bin level.

Samples	Bins Shannon diversity	Contigs Shannon diversity	Bins richness	Contigs richness	Bins evenness	Contigs evenness
S1	6.260296	7.510723	679	2306	0.9600766	0.9699679
S2	5.305778	6.096791	407	960	0.8829994	0.8878477
S3	4.275578	4.820323	226	444	0.7887742	0.7907581
S4	2.725157	3.082129	89	167	0.6071235	0.6022144
S5	1.672453	1.911781	48	71	0.4320244	0.4484927
S6	0.9816084	1.099826	28	50	0.2945823	0.28114
S7	0.7462127	0.8211464	23	34	0.2379889	0.2328594

Table 3. Biodiversity metrics (Shannon diversity, evenness, and richness) of CAMISIM datasets for each sample. Biodiversity metrics of CAMISIM generated datasets represent the true value of Shannon diversity, evenness and richness as CAMISIM provided information of the genomes in each sample.

Samples	CAMISIM Shannon diversity	CAMISIM richness	CAMISIM evenness
S1	7.124629	2000	0.9373399
S2	5.906463	2000	0.7770739
S3	4.588208	2000	0.6036399
S4	2.742779	1846	0.3646936
S5	1.620362	1143	0.2301189
S6	0.9421016	576	0.1482199
S7	0.7200346	395	0.1204296

References

1. Jacquet S, Miki T, Noble R, Peduzzi P, Wilhelm S. Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. Advances in Oceanography and Limnology. 2010;1: 97–141. doi:10.1080/19475721003743843

2. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020;11: 1710. Available: https://www.nature.com/articles/s41467-020-15507-2

3. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348: 1261498. doi:10.1126/science.1261498

4. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15: 161–168. doi:10.1038/nrmicro.2016.177

5. Zhang Y-Z, Shi M, Holmes EC. Using Metagenomics to Characterize an Expanding Virosphere. Cell. 2018;172: 1168–1172. doi:10.1016/j.cell.2018.02.043

6. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020;578: 432–436. doi:10.1038/s41586-020-1957-x

7. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, et al. Clades of huge phages from across Earth's ecosystems. Nature. 2020;578: 425–431. doi:10.1038/s41586-020-2007-4

8. Weinheimer AR, Aylward FO. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. ISME J. 2022;16: 1657–1667. doi:10.1038/s41396-022-01214-x

9. Malone LM, Warring SL, Jackson SA, Warnecke C, Gardner PP, Gumy LF, et al. A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. Nature Microbiology. 2019;5: 48–55. doi:10.1038/s41564-019-0612-5

10. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537: 689–693. doi:10.1038/nature19366

11. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell. 2019;177: 1109–1123.e14. doi:10.1016/j.cell.2019.03.040

12. Aylward FO, Moniruzzaman M. Viral Complexity. Biomolecules. 2022;12. doi:10.3390/biom12081061

13. García-López R, Vázquez-Castellanos JF, Moya A. Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. Front Bioeng Biotechnol. 2015;3: 141. doi:10.3389/fbioe.2015.00141

14. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020;11: 1710. doi:10.1038/s41467-020-15507-2

15. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. MBio. 2019;10. doi:10.1128/mBio.02497-18

16. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res. 2020;30: 315–333. doi:10.1101/gr.258640.119

17. Schulz F, Andreani J, Francis R, Boudjemaa H, Bou Khalil JY, Lee J, et al. Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. mSystems. 2020;5. doi:10.1128/mSystems.00048-20

18. Brand MW, Wannemuehler MJ, Phillips GJ, Proctor A, Overstreet A-M, Jergens AE, et al. The Altered Schaedler Flora: Continued Applications of a Defined Murine Microbial Community. ILAR Journal. 2015. pp. 169–178. doi:10.1093/ilar/ilv012

19. Wagner RD. Effects of microbiota on GI health: gnotobiotic research. Adv Exp Med Biol. 2008;635: 41–56. doi:10.1007/978-0-387-09550-9_4

20. Lavin R, DiBenedetto N, Yeliseyev V, Delaney M, Bry L. Gnotobiotic and Conventional Mouse Systems to Support Microbiota Based Studies. Curr Protoc Immunol. 2018;121: e48. doi:10.1002/cpim.48

21. Kremer JM, Sohrabi R, Paasch BC, Rhodes D, Thireault C, Schulze-Lefert P, et al. Peat-based gnotobiotic plant growth systems for Arabidopsis microbiome research. Nat Protoc. 2021;16: 2450–2470. doi:10.1038/s41596-021-00504-6

22. Van Camp P-J, Porollo A. SEQ2MGS: an effective tool for generating realistic artificial metagenomes from the existing sequencing data. NAR Genom Bioinform. 2022;4: lqac050. doi:10.1093/nargab/lqac050

23. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome. 2019;7: 17. doi:10.1186/s40168-019-0633-6

24. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7: e7359. doi:10.7717/peerj.7359

25. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection

of Cultured Phage Genomes. Phage (New Rochelle). 2021;2: 214–223. doi:10.1089/phage.2021.0007

26. Hendrix RW. Jumbo bacteriophages. Curr Top Microbiol Immunol. 2009;328: 229–240. doi:10.1007/978-3-540-68618-7_7

27. Sczyrba A, Hofmann P, Belmann P, Koslicki D. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nature. 2017. Available: https://www.nature.com/articles/nmeth.4458

28. Altan E, Dib JC, Gulloso AR, Escribano Juandigua D, Deng X, Bruhn R, et al. Effect of Geographic Isolation on the Nasal Virome of Indigenous Children. J Virol. 2019;93. doi:10.1128/JVI.00681-19

29. Altan E, Kubiski SV, Burchell J, Bicknese E, Deng X, Delwart E. The first reptilian circovirus identified infects gut and liver tissues of black-headed pythons. Vet Res. 2019;50: 35. doi:10.1186/s13567-019-0653-z

30. Altan E, Kubiski SV, Boros Á, Reuter G, Sadeghi M, Deng X, et al. A Highly Divergent Picornavirus Infecting the Gut Epithelia of Zebrafish (Danio rerio) in Research Institutions Worldwide. Zebrafish. 2019;16: 291–299. doi:10.1089/zeb.2018.1710

31. Brito F, Cordey S, Delwart E, Deng X, Tirefort D, Lemoine-Chaduc C, et al. Metagenomics analysis of the virome of 300 concentrates from a Swiss platelet bank. Vox Sang. 2018. doi:10.1111/vox.12695

32. Ng TFF, Chen L-F, Zhou Y, Shapiro B, Stiller M, Heintzman PD, et al. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc Natl Acad Sci U S A. 2014;111: 16842–16847. doi:10.1073/pnas.1410429111

33. Sadeghi M, Altan E, Deng X, Barker CM, Fang Y, Coffey LL, et al. Virome of > 12 thousand Culex mosquitoes from throughout California. Virology. 2018;523: 74–88. doi:10.1016/j.virol.2018.07.029

34. Phan TG, da Costa AC, Zhang W, Pothier P, Ambert-Balay K, Deng X, et al. A new gyrovirus in human feces. Virus Genes. 2015;51: 132–135. doi:10.1007/s11262-015-1210-0

35. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27: 824–834. doi:10.1101/gr.213959.116

36. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

37. Gerner SM, Graf AB, Rattei T. Tamock: simulation of habitat-specific benchmark data in metagenomics. BMC Bioinformatics. 2021;22: 227. doi:10.1186/s12859-021-04154-z

38. Oksanen J. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Gavin L Simpson, Peter Solymos, M Henry H Stevens ….

39. Roswell M, Dushoff J, Winfree R. A conceptual guide to measuring species diversity. Oikos. 2021;130: 321–338. doi:10.1111/oik.07202

40. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016. Available: https://books.google.com/books/about/ggplot2.html?hl=&id=RTMFswEACAAJ

41. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, Burger A, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res. 2020;30: 437–446. doi:10.1101/gr.251686.119

42. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, et al. Genotypic diversity within a natural coastal bacterioplankton population. Science. 2005;307: 1311–1313. doi:10.1126/science.1106028

43. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ. 2019;7: e6800. doi:10.7717/peerj.6800

44. Jain, M., Koren, S., Miga, K. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36, 338–345 (2018). https://doi.org/10.1038/nbt.4060

45. Weirather JL, de Cesare M, Wang Y *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; peer review: 2 approved]. *F1000Research* 2017, **6**:100 (https://doi.org/10.12688/f1000research.10571.2)

46. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B. 2019. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* 7:e6800 https://doi.org/10.7717/peerj.6800