

Seasonal Anti-cyclonic Vortex Event Reveals Connections between Environmental Features and Oligotrophic Oceanic Virio- and Bacterioplankton Communities

K. Eric Wommack^{1,*}, Ling-yi Wu², Gonalo J. Piedade³, Ana M. Martins⁴, Kay D. Bidle⁵ and Shawn Polson^{1*}

¹Center for Bioinformatics and Computational Biology, University of Delaware, 15 Innovation Way Newark, DE, 19711 United States, ²Department of Water Science and Policy, College of Agriculture and Natural Resources, University of Delaware, Newark, DE, 19716, United States, ³Department of Marine Microbiology, NIOZ Royal Netherlands Institute for Sea Research, Den Burg, Texel, 1790 AB, Netherlands, ⁴Department of Oceanography and Fisheries, University of the Azores, Ponta Delgada, Regiao Autonoma dos Acores, Portugal and ⁵Department of Marine and Coastal Sciences, Rutgers University, 71 Dudley Rd., New Brunswick, NJ, 08901, United States

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

Virioplankton are major players within oceanic ecosystems, impacting biogeochemical cycling of nutrients and energy. Virioplankton can only reproduce themselves through host infection, mostly bacterioplankton, and thus influence marine ecosystems through interactions with bacterioplankton. Observations of temporal and spatial variations in viral and bacterial communities are important for understanding the patterns of aquatic microbial diversity and how these changes may affect ocean biogeochemical processes. This study focuses on characterizing the patterns of marine viral and bacterial diversity in the pelagic and oligotrophic ocean. Seawater down to 75 meters was collected from a fixed station off Faial Island (station SCFP) during late spring and summer of 2016. Virioplankton and bacterioplankton communities were characterized using Ribonucleotide triphosphate reductase (RTPR) and 16S rDNA amplicon sequences, respectively. RTPR is one of the most abundant ribonucleotide reductase (RNR) genes, carried mostly by lytic viruses. We studied the dynamics of this group in conjunction with the co-occurring bacterial community at one location in order to (a) contextualize observations of virioplankton community composition and diversity in environmental changes and microbial abundances, (b) investigate beta and alpha diversity of virioplankton and bacterioplankton communities across the sampling days and sampling depths and compare their patterns, (c) investigate seasonal and spatial dynamics of distinct RTPR-carrying viruses and bacteria populations, (d) detect statistical associations between distinct viruses and bacteria populations to understand how viruses and bacteria community may influence each other and determine potential phage-host relationships. Statistics methods and bioinformatics tools that account

for the compositional characteristics of high-throughput sequencing (HTS) datasets are used to overcome challenges brought by biases that are inherent to HTS datasets derived from environmental samples. We found that seasonal anti-cyclonic vortex, which causes water stratification accompanying with temperature and other environmental changes, influenced bacterioplankton community diversity, and subsequently influenced changes in virioplankton community diversity. Among the multitude of discovered virioplankton and bacterioplankton populations, only a few populations appear to be most responsible for changes in their community diversity. Close correlations are found between viruses from the same phylogenetic clades and bacteria from the same order indicated that phylogenetically similar viruses or bacteria thrive in similar environments.

INTRODUCTION

Virioplankton are major players in ocean ecosystems, impacting biogeochemical cycles of nutrient elements and energy ((Suttle, 2005)). Viruses are the most abundant biological entities on Earth. The number of viruses has been estimated to be on the order of 10^{31} individuals, which is roughly about one billion times the number of stars in the observable universe (10^{22}) ((Hatfull et al., 2006)). Viruses cannot reproduce by themselves. They depend on their hosts machinery to replicate the viral genome and produce viral proteins to make new capsids ((Whittaker and Helenius, 1998); (Doms et al., 1993)). The hosts of virioplankton are mainly bacterioplankton. Thus, the behaviors of virioplankton populations influence the lives of bacterioplankton.

Virioplankton in lytic cycles can kill bacterioplankton through viral lysis. The mortality of bacterioplankton caused by viral lysis is one of the major top-down controls on

*To whom correspondence should be addressed. Tel: +01 302-831-4362; Fax: +44 000 0000000; Email: wommack@dbi.udel.edu; wommack@udel.edu

bacterioplankton productivity it is responsible for 10 to 40% of the mortality rate of the bacterial community ((Fuhrman, 1999)). Nutrients inside of the bacterioplankton cells are released into the environment after viral lysis and thus contribute to nutrient pools within marine ecosystems. Bacterioplankton lysis comprises the largest carbon flux process between biomass and dissolved organic matter (DOM) on the globe (150 gigatons of C/yr) ((Wilhelm and Suttle, 1999); (Suttle, 2005)). The conversion of biomass to DOM performed by viral lysis is five times more common than that performed by other biological mechanisms ((Wilhelm and Suttle, 1999); (Suttle, 2005)). The flux from biomass to DOM sustains the cycles of nutrients, such as carbon, nitrogen, phosphate, iron and etc., which are reused by living bacterioplankton and other organisms. The availability and limitation of different nutrients manipulates the bacterioplankton populations through bottom-up control ((Wilhelm and Suttle, 1999); (Suttle, 2005)). Besides shaping the composition and diversity of marine bacterial communities through top-down and bottom-up controls, viruses can influence bacterial behaviors (such as improving fitness in the face of environmental changes (((Jiang and Paul, 1998)); ((Schwartz and Lindell, 2017))) and altered metabolic pathways(((Ochman et al., 2000));(Rosenwasser et al., 2016))) through horizontal gene transfer and expression of auxiliary metabolic genes (AMGs) during viral infection.

In summary, viroplankton plays a key role in marine ecosystems mainly through the interactions with their host bacterioplankton. Different viroplankton populations can shape the ecology and evolution of host populations in unique ways. However, the behavior of individual viroplankton populations and the detailed interaction between specific viroplankton and bacterioplankton populations remain largely unknown. This study focuses on characterizing the patterns of marine microbial diversity in pelagic ocean to acquire a better understanding of marine ecosystems. Diversity of viroplankton communities are expected to dynamically respond to environmental changes and bacterioplankton communities.

The 16S rRNA gene has proven to be a robust marker gene to comprehensively characterize community structure and diversity of bacterioplankton from the environment. This is largely because the 16S rRNA gene is found in all bacterial species. However, the viroplankton lacks a universal marker gene because of the polyphyletic nature of viral evolution. There is no single viral gene existing in all of the viroplankton. However, a few genes are widespread across a number of viral populations and thus can be used as marker genes to characterize the diversity of a subset of viroplankton in the environment, such as DNA polymerase A gene (encoding DNA polymerase I) ((Schmidt et al., 2014)), major capsid protein gp20 gene ((Zhong et al., 2002)), major capsid protein gp23 gene ((Comeau and Krisch, 2008)) and Ribonucleotide Reductase (RNR) genes ((Sakowski et al., 2014)). DNA Polymerase I is an essential protein for DNA replication and repair in prokaryotes and is found in many tailed bacteriophages. Gp20 and gp23 are structural proteins of tailed Myoviruses. RNR is an enzyme that catalyzes the formation of deoxyribonucleotides from ribonucleotides, thus prime DNA synthesis ((Elledge et al., 1992)).

RNR genes are promising marker genes to characterize viral diversity ((Sakowski et al., 2014)) because they are functionally non-redundant, abundant and widely distributed among marine viruses. RNRs are the only known enzymes able to reduce ribonucleotides to deoxyribonucleotides, a key step for DNA synthesis ((Lundin et al., 2010); (Elledge et al., 1992)). **Thus, RNRs are vital to nucleotide biosynthesis, under rigorous evolutionary selection pressure, and among the most abundant annotated genes in marine virome libraries ((Dwivedi et al., 2013)). Importantly, RNR genes are present in all three families of tailed phages in the order Caudovirales and have been identified in viruses infecting hosts within all three domains of life.** Lytic viruses usually carry RNRs so that they can synthesize DNA and replicate fast by ensuring a steady supply of deoxyribonucleotides for DNA synthesis ((Sullivan et al., 2005)). Therefore, RNRs are representative of many lytic marine phages, which significantly influence nutrient cycles within the ocean ((Brussaard et al., 2008)).

In addition, RNRs are biologically informative and form three physiological classes according to reactivity with oxygen. Class I RNRs are oxygen-dependent. Class II RNRs are oxygen-independent and rely upon an adenosylcobalamin (vitamin B12) co-factor for ribonucleotide reduction. Class III RNRs are sensitive to oxygen. Class I and Class II RNRs, not Class III RNRs, are expected to be present in the surface water of pelagic ocean because of the abundance of oxygen ((Nordlund and Reichard, 2006)). Ribonucleotide triphosphate reductase (RTPR) reduce ribonucleotide triphosphate to deoxyribonucleotide triphosphate. Among all of the clades of known RNRs, RTPR RNRs were the most abundant RNR clade in studied viromes, with an estimated 34 to 54% of all RNR-carrying viruses carrying a RTPR RNR gene ((Sakowski et al., 2014)). Degenerated RTPR primers developed by Sakowski were used in the study to amplify RTPR genes ((Sakowski et al., 2014)).

Thanks to the advancement of sequencing techniques, second-and third- generation sequencing technologies can now provide massive amounts of sequencing data to characterize the diversity of viral and microbial communities. However, the traditional ways to analyze sequencing data, such as rarefaction, Bray-Curtis similarity matrix, principal coordinates analysis (PCoA) and correlation tests, are problematic ((McMurdie and Holmes, 2014); (Weiss et al., 2017)). These methods treat the sequencing data as ecological data where the counts of reads assigned to organisms are often normalized to a constant area or volume.

In a typical ecological study it is possible that different species co-exist, and the presence of one individual of one species does not influence the ability to detect the presence of another individual from another species. For example, an ecological survey intends to count the numbers of tigers and the numbers of ladybugs in a certain area. The numbers of total counts of tigers and ladybugs are not fixed, and the numbers of tigers are not correlated to the numbers of ladybugs during the survey ((Gloor et al., 2017)).

In contrast, datasets collected by high-throughput sequencing (HTS) are compositional because the numbers of sequences obtained from a sample have an arbitrary maximum imposed by the instrument ((Gloor et al., 2017)). Compositional data consist of vectors whose components are

the proportion or percentages of some whole. The peculiarity of compositional data is that their sum is constrained to be some constant, equal to 1 for proportions, 100 for percentages or possibly some other constant c for other situations such as parts per million (ppm) in trace element compositions ((Aitchison, 1982)).

As for sequencing, an Illumina MiSeq sequencer can generate 25 million reads per run at maximum. Pacbio RSII sequencer can generate about a million reads per SMRT-cell at maximum and Pacbio Sequel sequencer can generate up to seven million reads per unit instrument run. Since the total read counts observed in an HTS run is fixed, the read counts gained for a certain composition (bacterioplankton or virioplankton population in this case) from a sample are influenced by the read counts gained for other compositions from the same sample or from other samples.

Moreover, bias is inherent to HTS datasets derived from environmental samples as each step in the experimental workflow preferentially detects some taxa over others ((Brooks et al., 2015); (Sinha et al., 2017), (Brooks, 2016)). As a result, the measured relative abundances of each population within the sample can be systematically distorted from their true values. Bias can lead to qualitatively incorrect conclusions about differential relative abundance, compositional similarities, and correlations between different compositions from environmental samples. Both of the absolute and relative read counts are irrelevant and cannot be used to interpret microbial diversity ((Gloor et al., 2017)). Recent studies found that analyses based on proportions between compositions from different samples are less sensitive to bias than analyses based on relative abundance of each composition, which emphasizes the importance of using compositional data analysis tools to analyze HTS datasets derived from microbial/viral studies ((McLaren et al., 2019)).

Overall, HTS datasets derived from viral or microbial studies are compositional and should be treated as compositional data at all stages of analysis. Thus, we used statistics analysis methods designed for compositional data to analyze the sequencing dataset. To date, compositional data analysis paradigms have been applied to the study of human microbiomes ((Bian et al., 2017)) but have not been widely used in biogeography studies of marine viruses and microbes.

MATERIALS AND METHODS

Study Site and Sample Collection

Seawater was collected from the South of the Channel Faial-Pico (SCFP) Island, Azores, Portugal at 0 m, 5 m, 25 m, 50 m and 75 m depths on April 14th, May 12th, June 8th, June 22nd, July 11th, July 27th, August 5th and September 8th, 2016 (Julian Day: 103, 132, 159, 173, 192, 208, 217 and 251). SCFP is one of the islands in the Azores archipelago in Northeastern Atlantic, located at 38 °23'11.67"N and 28 °35'6.14"W, 8.6 nautical miles off Horta city (Faial island) and 6.4 nautical miles off So Mateus (Pico island).

Surface seawater was collected by submerging a 22 liter carboy into the water directly. Seawater at other depths was collected by sending a 2.5 liter Niskin bottle to each depth for three times and transfer all the water from the same depth

to a 5 liter carboy. All containers (including carboys and bottles) were washed using deionized water, followed by 0.1 M HCl rinse and deionized water rinse at the laboratory before sampling. **Containers were washed twice with seawater at the sampling site just before collecting seawater. More details here.**

In addition, a MIDAS CTD profiler (Valeport, Totnes, United Kingdom) was used to measure, **in triplicate**: the water temperature (°C), salinity (SPC), current direction (°), and current velocity (m/s). Temperature and salinity data from the CTD profiler were visualized using Ocean Data View (<https://odv.awi.de>). Temperature and salinity data between sampling days were extrapolated using the weighted average grid method ((Schlitzer, 2002)). Chlorophyll *a*, sea surface temperature, particulate organic carbon, particulate inorganic carbon and photosynthetically available radiance data in the year of 2016 were accessed from the NASA Ocean Color website (<http://oceancolor.gsfc.nasa.gov/>) at 1 km resolution. The Level 2 data was calculated using SeaDAS 7 for the nine km² area around the sampling station SCFP (Figure supplementary figure1). Data was extracted using the pixel extraction tool. The average values of each day and month were calculated in Excel. The Minimum-Maximum Normalization (equation 1) method was performed on the data to improve data visualization ((Larose and Larose, 2014)). x_i is the value of the i_{th} data from dataset x . $\min(x)$ is the minimum of dataset x and $\max(x)$ is the maximum of dataset x . z_i is the normalized value of data x_i .

$$z_i(\%) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \times 100 \quad (1)$$

Sample Preparation

Two mL of seawater was fixed with 0.5% glutaraldehyde and collected into 5 mL vials for flow cytometry analysis. Two hundred and fifty mL of seawater was fixed with 0.4% formalin and collected into a 1 L dark bottle for phytoplankton identification and enumeration. Seawater samples were transported to the laboratory as quickly as possible for further processing. Seawater for pigment analyses was filtered onto GF/F glassfiber filters (0.7 µm pore-size, 25 mm diameter), dried with absorbent paper, and stored at -80 °C. Samples for pigments and phytoplankton counts are still waiting further processing

Viral and microbial fractions were separated through peristaltic filtration. Each water sample was filtered sequentially through a 0.22 µm pore-size Sterivex filter (Milipore, SVGP01050) followed by a 0.02 µm pore-size Anotop filter (Whatman, WHA68092002) (Steward and Culley, 2010; Mueller, Culley and Steward, 2014a). Filtration through the 0.22 µm pore-size Sterivex filters was performed using a vacuum pump (Milipore, SD1P014M04). Filtration through the 0.02 µm pore-size Anotop filters was performed using a fabricated calking gun, which generated even and steady pressure for water filtration. After filtration, both filters (0.22 µm and 0.02 µm) were sealed and stored at -80°C until DNA extraction.

Microbial and Viral Abundances

Place Holder for Goncalo

Viral DNA Extraction and Tag-Encoded RTPR Amplicon Sequencing

For the viral RTPR gene analysis, DNA was extracted from 500 µL of the treated viral concentrate using the MasterPure complete DNA/RNA purification kit (Epicentre, Madison, WI, United States) per the manufacturers instructions with adaptations and quantified with a HS DNA Qubit fluorescent concentration assay, described in detail in the Supplementary1. For each sample, DNA was used in the initial amplification PCR with 35 cycles, followed by 22 cycles of limited PCR using distinctive barcodes designed in Wommack Lab. The final libraries were then quantitated by KAPA SYBR FAST qPCR kit and sequenced on the PacBio RSII (Pacific Biosciences, Menlo Park, CA, United States).

Bacterial DNA Extraction and Tag-Encoded 16S Amplicon Sequencing

For the bacterial 16S rDNA gene analysis, DNA was extracted from each of the 0.22 µm filters utilizing an adapted phenol/chloroform method ((Dez et al., 2001)) described in detail in the Supplementary1. The V3-V4 hypervariable region of the 16S rRNA gene was PCR-amplified and sequenced on the Illumina MiSeq (Illumina, San Diego, CA, United States) utilizing a dual-indexing strategy for multiplexed sequencing developed at the Institute for Genome Sciences for twice (Fadrosh et al., 2014), described in detail in the Supplementary2. Obtained bacterial 16S rDNA gene amplicon sequences from the two libraries were analyzed individually for core diversity, described in detail in Supplementary2. The results showed that the two datasets are similar to each other. Thus, obtained bacterial 16S rDNA gene amplicon sequences from the two libraries were combined together for downstream analysis.

Virioplankton RTPR Gene Data Analysis

RTPR gene reads from virioplankton were initially screened for low quality bases and short or long read lengths at PacBio Data Repository of Delaware Biotechnology Institute (<http://pacifica.dbi.udel.edu>). Circular consensus sequences were generated. Reads with less than three full passes, less than 98% of minimum predicted accuracy and with a length shorter than 250 bp or larger than 5000 bp were excluded. Generated circular consensus sequences were demultiplexed using `split_libraries.py` command from QIIME1. Barcode-demultiplexed reads were screened for forward and reverse degenerate primer sequences using Daniel Naskos `demultiplex.pl` script (https://github.com/dnasko/biobin/tree/master/pcr_products/adaptersearching).

Reads containing both forward and reverse primer sequences with at least 85% of similarity on their first 60 base pairs were retained. Barcode and primer sequences were trimmed from the reads. Barcode-primer-demultiplexed reads were translated into amino acid sequences and the frame shifts were corrected using Daniel Naskos `frameshift_polisher` script (https://github.com/dnasko/frameshift_polisher). RTPR amino acid sequences from rrndb ((Lundin et al., 2009))

and Tara Ocean ((Pesant et al., 2015)) databases were used as references. Translated RTPR amino acid sequences were screened for key catalytic sites using Protein active site validator (PASV) designed by Ryan Moore (<https://github.com/mooreryan/pasv>). An RTPR amino acid sequence with all of the active key residues indicates a functional RTPR. RTPR Amino acid sequences with key residues C439, E441 and C46 (*E. coli* numbering) were aligned using the FFT-NS-1 algorithm in MAFFT alignment ((Katoh et al., 2002)) in Geneious 10.0.9 (<https://www.geneious.com>) and trimmed to the region of interest (H23 to C783 in *E. coli*). Quality-controlled RTPR amino acid sequences were clustered at 95% *de novo* into operational taxonomic units (OTUs) using `cluster_fast` command in USEARCH ((Edgar, 2010)). OTUs occurring no more than twice among all samples were removed.

Beta diversity was determined through principal component analysis (PCA) plots of Aitchison distances between samples calculated using the R package `compositions` ((van den Boogaart et al., 2018)) and tested for significance with `beta-group-significance` command ((Anderson, 2001)) in QIIME2. It should be noted that zero count OTUs were replaced by an imputed value using the count zero multiplicative method from the `zCompositions` R package ((Palarea-Albaladejo and Martín-Fernández, 2015)). The centered log ratio (clr) transform was applied to the zero replaced data set, and the data were subsequently used as input for a singular value decomposition (SVD). This approach returns data where the samples are separated by the variance in the OTUs rather than by differences in abundant OTUs ((Gloor and Reid, 2016); (Gloor et al., 2016)).

The dendrogram of virioplankton samples was made based on Aitchison distances between samples using Hierarchical Cluster Analysis in the R package `stats` ((Team, 2013)) and visualized using Iroki ((Moore et al., 2018)). Correlations between virioplankton diversity and associated metadata, such as environmental features, microbial abundances and viral abundance, was tested using Mantel test ((Mantel, 1967); (Pearson, 1895); (Spearman, 2010);) in QIIME2. Associated metadata sets that have significant correlations with virioplankton diversity (p-value < 0.05) were plotted next to the dendrogram of virioplankton samples using Iroki ((Moore et al., 2018)).

Alpha diversity was calculated using the R packages: `DivNet` ((Willis and Martin, 2018)), `breakaway` ((Willis and Bunge, 2015)), `phyloseq` ((McMurdie and Holmes, 2013)), `tidyverse` ((Wickham, 2017)) and `ggplot2` ((Wickham, 2016)) on raw viral RTPR OTU table and tested for significance using `betta` function in R package `breakaway` ((Willis et al., 2017)). Estimated Shannon indices and Estimated Simpson indices were converted to Hill numbers for better visualisation ((Kang et al., 2016)).

Correlation analysis was performed on samples from the same depth. Three pseudo-replicates were generated for each sample. Correlation analyses were done using a symmetric modification of the metric ((Quinn et al., 2017)), which measures the variance in the ratios between OTUs. OTUs with a low ratio variance are said to be compositionally associated. The metric ((Erb and Notredame, 2016)) was calculated as an expected value across clr-transformed Monte Carlo Dirichlet replicates ((Quinn et al., 2017); (Fernandes et al., 2013))

generated by the `aldex.clr` function, and values of 0.65 were taken as indicating association between pairs of OTUs.

Differential abundance analysis were performed on samples from the same depth on from the same day.

Bacterioplankton 16S rDNA Gene Data Analysis

16S rDNA gene reads were initially screened for low quality bases and short read lengths (Fadrosh et al., 2014), paired reads were merged using PANDAseq (Masella et al., 2012), de-multiplexed, trimmed of artificial barcodes and primers, and assessed for chimeras using UCHIME in de novo mode, as implemented in Quantitative Insights Into Microbial Ecology (QIIME; version 1.9.1-20150604) (Caporaso et al., 2010). Quality-controlled reads were clustered at 97% de novo into operational taxonomic units (OTUs) with the SILVA 16S database (Quast et al., 2012) in QIIME (Caporaso et al., 2010). All sequences taxonomically assigned to chloroplasts were removed from further downstream analysis. When appropriate, data was normalized using metagenomeSeqs cumulative sum scaling to account for uneven sampling depth (Paulson et al., 2013b). To visualize the relative abundance of bacterial phyla, stacked bar charts were generated using `ggplot2` (Wickham, 2009). In addition, bacterial taxa were summarized at the genera level in QIIME (level = 6) and those with a maximum relative abundance greater than 1% in at least one sample were used to build a heatmap via R (ver. 3.3.2) and `vegan` `heatmap` (Ploner, 2012). Significance tests were conducted using a Tukeys HSD Test between filter size fractions and among sampling dates. Additionally, Pearson correlation coefficients were calculated to identify associations between the water characteristics and the relative abundance of the bacterial phyla/genera. Alpha diversity was calculated using the R packages: Bioconductor (Huber et al., 2015), `metagenomeSeq` (Paulson et al., 2013a), `vegan` (Oksanen et al., 2007), `phyloseq` (McMurdie and Holmes, 2013), `fossil` (Vavrek, 2011), `biomformat` (McMurdie and Paulson, 2016), and `ggplot2` (Wickham, 2009) on data rarefied to an even sampling depth (55,307 sequences) and tested for significance using a Tukeys HSD Test. Beta diversity was determined through principle coordinates analysis (PCoA) plots of BrayCurtis, Weighted and Unweighted UniFrac distances calculated using the R package `phyloseq` and tested for significance with ANOSIM (9,999 permutations) (Beals, 1984; Clarke, 1993; Lozupone et al., 2011; McMurdie and Holmes, 2013).

RESULTS

Seawater Characteristics

Temperature and salinity of surface seawater were homogeneous prior to June, especially in the upper 40 m of the water column, indicating seawater was well-mixed in the beginning of sampling period (Figure 1). Temperature started to increase in early June, continuing throughout the summer. Coincidentally, salinity increased slightly with a distinct increase centered at 75 m depth, especially on Jul11. The stratification patterns of temperature and salinity suggested the presence of a downwelling anticyclonic vortex in the region. Downwelling anticyclonic vortex is likely to prevent nutrient from deeper and colder seawater from

coming up to the sea surface, causing nutrient depletion in the region. Salinity tended to be homogeneous again on the surface seawater in August, potentially indicating an upwelling vortex from the deep ocean to the surface water. The upwelling vortex is likely to bring nutrient from deeper and colder seawater to the sea surface and supply nutrient for the food web. Unfortunately, temperature and salinity profile data is missing for the last sampling day due to equipment failure. The stratification intensity was assumed to mitigate according to data from similar studies in nearby areas (**need to confirm**).

Bacterial 16S rDNA Gene and Viral RTPR Gene Sequencing Effort

In total, 35 bacterioplankton samples were PCR-amplified for the 16S rDNA gene and sequenced. After sequence quality filtering, 1,779 ASVs were identified from a total of 0.9 million sequences across all samples, with an average number of ca. 26,251 sequences per sample and an average of ca. 281 ASVs.

29 viroplankton samples were PCR-amplified for the 16S rDNA gene and sequenced. Amplification was failed in six of the viroplankton samples. After sequence quality filtering, 220 OTUs (95% identity) were identified from a total of 30 thousand sequences across all samples, with an average number of ca. 1,034 sequences per sample and an average of ca. 87 OTUs.

Bacterial and Viroplankton Community Composition and Temporal Variations

The most abundant bacterioplankton phyla were Proteobacteria, Bacteroidetes and Cyanobacteria in all samples as order using CLR transformed abundance (Supplementary Table S1), however, the most abundant bacterioplankton phyla were Proteobacteria, Cyanobacteria and Bacteroidetes in all samples as order using relative abundance (Supplementary Table S2).

Bacterioplankton and Viroplankton Beta Diversity

Beta diversity comparisons using PCA plots of Aitchison distances computed for all samples revealed obvious clustering before (excluding) and after (including) sampling day Jun22 along PC1 and PC2, which accounted for nearly 38% and 37% of the variation of bacterioplankton and viroplankton, respectively. For both bacterioplankton and viroplankton, samples collected before Jun22 clustered together on the left part of PC1 while samples collected after Jun22 clustered together on the middle-bottom part of PC2. The clustering was significant for all the bacterioplankton samples taken from May12 and some of the bacterioplankton samples taken from Jun8 with both p and q value < 0.05 (Supplementary Table S3). The clustering is not significant for any viroplankton samples (Supplementary Table S3).

Obvious clustering was also revealed across different sampling depths for bacterioplankton (Supplementary Figure S1A and Table S4). Samples clustered from 0 m depth distributed along the diagonal of the PCA plot. Sample clustering shifted to the left and bottom of the PCA plot gradually as the depth went deeper. All samples collected

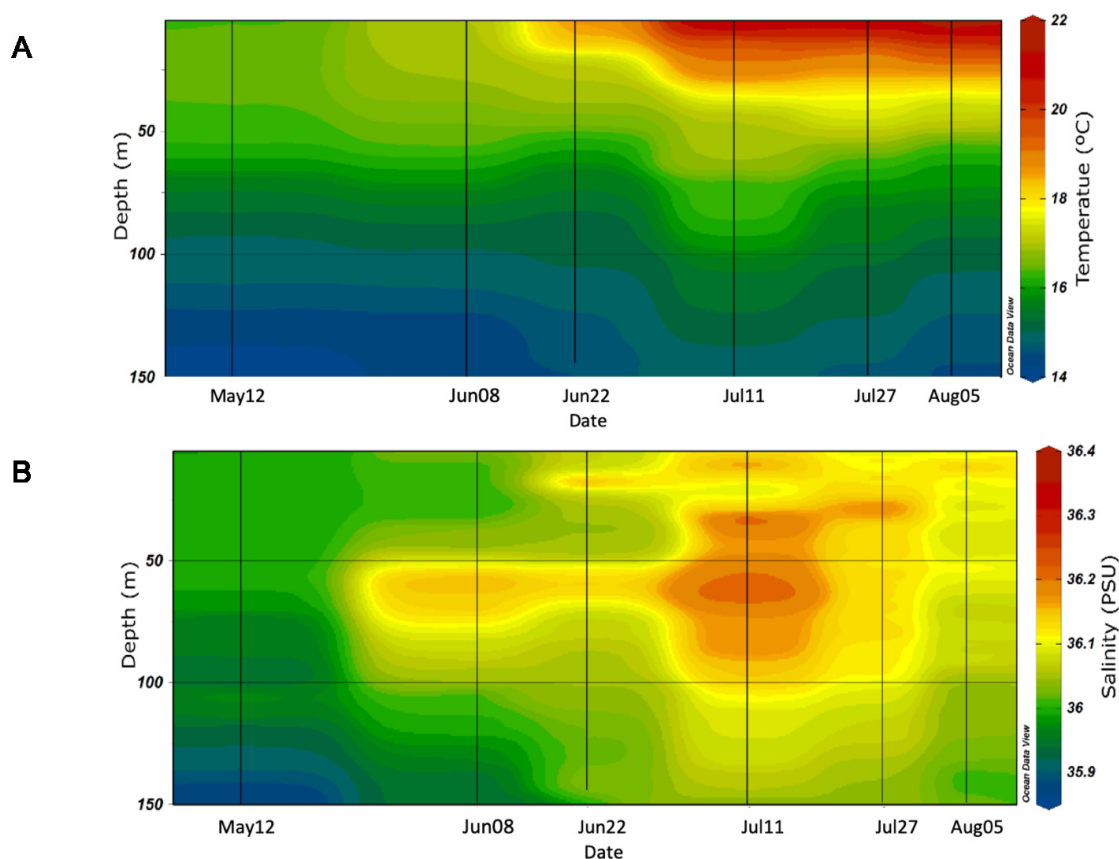


Figure 1. CTD-derived vertical profile contours of (A) temperature , (B) salinity at station SCFP from May12 to Aug05 in 2016.

from 75 m depth clustered together on the left of PC1 and on the bottom of PC2 except for sample May12_75m. No obvious clustering was seen across different sampling depths for virioplankton (Supplementary Figure S1B and Table S4).

Relationships between Water Characteristics and Bacterioplankton or Virioplankton Beta Diversity

Mantel test between sample composition similarity matrix and all metadata matrices showed that a few environmental features are correlated with sample compositions. Sample dendrogram based on Aitchison distance between sample compositions were plotted against associated environmental features (Figure 3).

Temperature was the best indicator of sample composition similarity of both bacterioplankton (Spearman: $r = 0.5$, $p = 0.001$; Pearson: $r = 0.53$, $p = 0.001$) and virioplankton (Spearman: $r = 0.32$, $p = 0.001$; Pearson: $r = 0.34$, $p = 0.001$) communities among all the tested parameters (Supplementary Table S5). Besides, sampling day and sampling depth were associated with sample compositions of both bacterioplankton and virioplakon communities. Salinity was associated with bacterioplankton composition similarity. No association was found between composition similarity and any microbial or viral abundance similarities.

Bacterioplankton and Virioplankton Alpha Diversity

Alpha diversity was computed for estimated Shannon diversity and tested for significance over time and over depth using DivNet. DivNet is a bioinformatics tool to estimate and compare ecosystem diversity across one of multiple environmental covariates and ignores the influence of other environmental covariates. DivNet focuses on differences between ecosystems rather than differences between samples. Thus, the hypothesis testing claims the the influence of certain environmental features to the diversities of communities.

An Inverse Relationship was observed between estimated Shannon diversity of bacterioplankton and virioplankton (Figure 4). Bacterioplankton Shannon diversity experienced a local minimum on Jun22 (Julian Day 192), coinciding with the start of seawater stratification and nutrient depletion. At the same time, virioplankton Shannon diversity experienced a local maximum. It is likely because that the dominant bacteria on Jun22 (Julian Day 192) had broad RTPR-carrying phage ranges. A slight increase of bacterioplankton Shannon diversity was observed after Jun22, likely causing by regulation effects on phage infection. Phage are thought to play a crucial role in the maintainance and stabilization of diversity in natural bacterial communities. The increase of bacterioplankton Shannon diversity decreased the virioplankton Shannon diversity in return as dominant bacteria

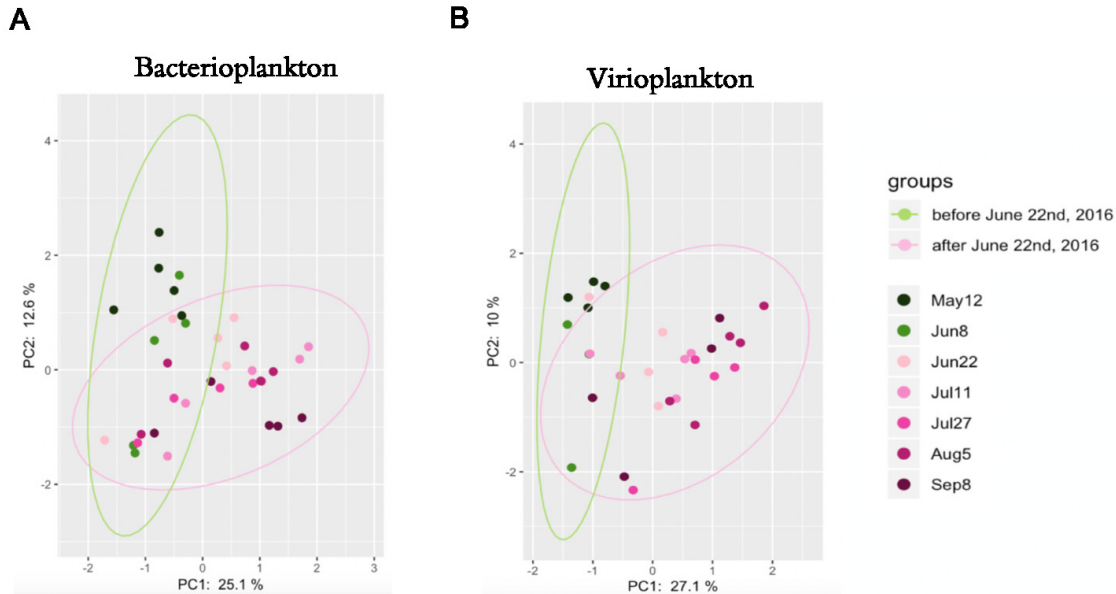


Figure 2. (A) bacterioplankton and (B) viroplankton responded similarly to sampling days.

having broad RTPR-carrying phage ranges were replaced with bacteria having narrower RTPR-carrying phage ranges.

Correlations between Individual Bacterioplankton and Viroplankton Populations

we used an expected value of metric to identify clusters of coassociated features (ASVs for bacterioplankton and OTUs for viroplankton) in the data set. This strength of association approach identifies features where both the direction and magnitude of variance are similar in the multidimensional data set. In a multivariate compositional sense, the metric is measuring both the slope of the correlation of the centered log ratio (clr)-transformed values and the correlation itself. A slope of 1 and a correlation of 1 are preferred. The metric coupled with a Bayesian estimation of feature relative abundance has the advantage of being agnostic to the level of sparsity in the data, unlike other recent approaches that depend on this sparsity. ASVs from bacterioplankton had stronger association to each other than OTUs from viroplankton. Each panel in (A) in **Figure 5** show the relative abundance of bacterioplankton ASVs assigned to one of 9 classes that had an expected value of 0.85, plotted against sampling day as a continuous variable. Each panel in (B) in **Figure 5** show the relative abundance of viroplankton OTUs assigned to one of 7 clades that had an expected value of 0.65, plotted against sampling day as a continuous variable.

This approach identified twelve distinct bacterioplankton clusters and four distinct viroplankton clusters, and many clusters contained ASVs classified into the same classes or OTUs classified into the same clades. And the ASVs related to the same classes or OTUs related to the same clades within the same clusters have approximately equal ratios between each other. It is likely that the most strongly associated groups include predominantly members of the same phylogenetic groups because different members of the same phylogenetic

groups have similar growth requirements, limitations, and interactions. On contrast, associated ASVs from different phylogenetic groups mostly had unequal ratios between each other along sampling days.

DISCUSSION

Environmental Changes Accompanying Temperature Changes Influence both Bacterioplankton and Viroplankton communities

Place holder for oceanographic knowledge

Seawater Stratification Influenced Bacterioplankton Diversity and Subsequently Influenced Changes in Viroplankton Diversity

Seawater was well-mixed in the beginning of the sampling period (prior to June) as shown in **Figure 1**. Water stratification started in early June, suggesting a down-welling anticyclonic vortex (**Figure 1**). The downwelling vortex prevented nutrient from deeper and colder water from welling up to the sea surface, causing nutrient depletion. Water stratification continued throughout the summer and mitigated in early August, suggesting that seawater could well up and bring nutrient from bottom water to the surface to supply the nutrient. Beta and alpha diversity of bacterioplankton and viroplankton communities changed with the changes of seawater characteristics.

Bacterioplankton samples taken before Jun22 are more similar to each other along PC1 but dissimilar to each other along PC2 (A in **Figure 2**). Bacterioplankton samples taken after Jun22 are more similar to each other along PC2 but dissimilar to each other along PC1. Viroplankton samples showed similar patterns with bacterioplankton samples (B in **Figure 2**). As seawater stratification started in early June, the similarity and dissimilarity patterns between samples of

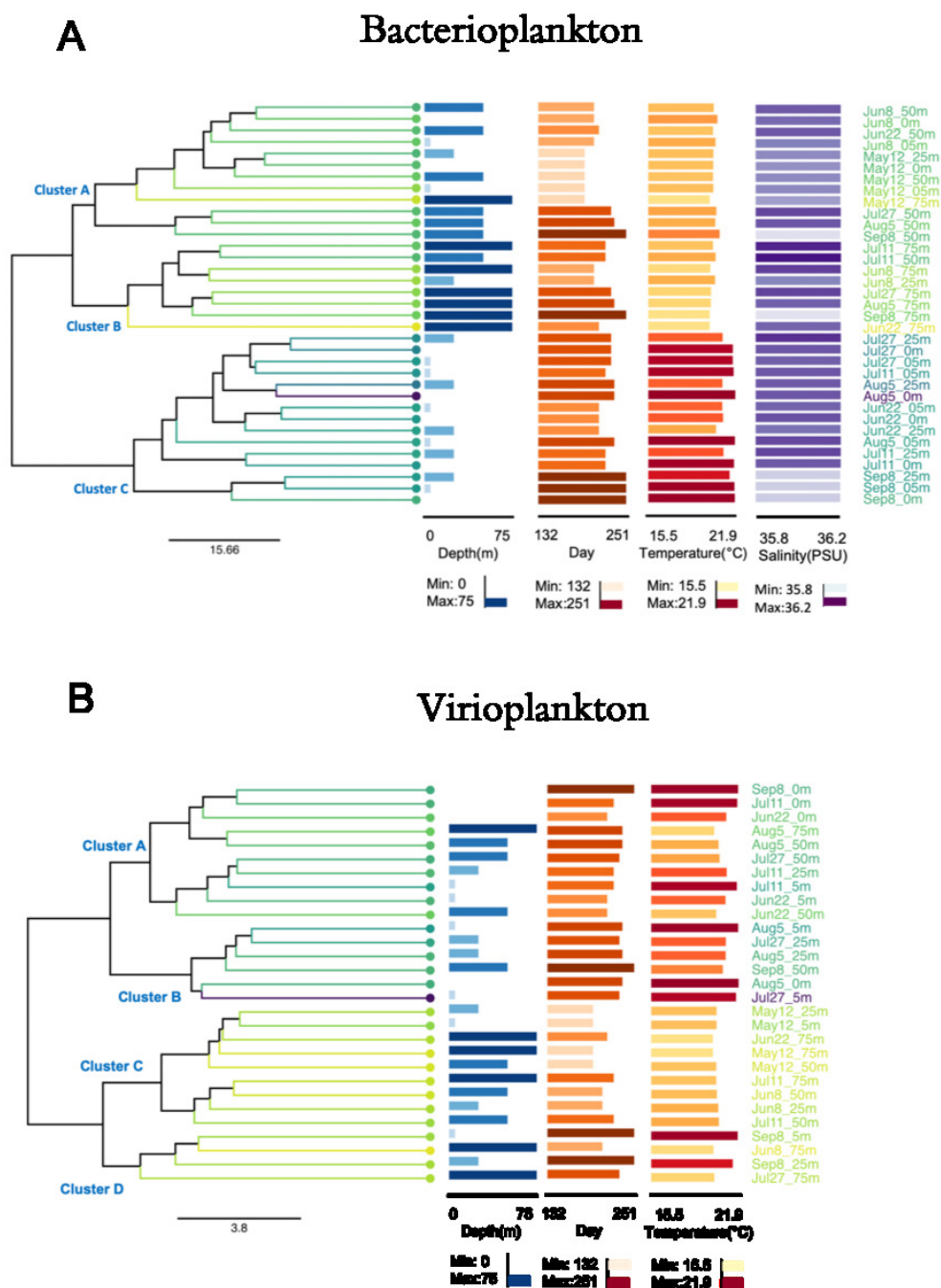


Figure 3. (A) bacterioplankton and (B) virioplankton composition similarity dendrogram with associated environmental features. Average-linkage hierarchical clustering based on sample RTPR compositions. Branch color is based on a scaled, 1-dimensional projection of sampling depths, sampling days, temperature and salinity onto the cube-helix color gradient. Samples that are more similar to each other in color represent those that are more similar to each other with respect to the environmental parameters in the ordination. The first bar series (blues) represents sampling depths (m), the second bar series (oranges) represent sampling days, the third bar series (red/orange) represents temperature (°C) and the fourth bar series (purples) represents salinity (PSU). Shorter bars and lighter colors indicate lower values, and longer bars with darker colors indicate higher values.

bacterioplankton and virioplankton communities were likely caused by the different seawater characteristics before and after the seawater stratification. Bacterioplankton samples taken from 75 m depth are very similar to each other except for sample May12_75m (A in Supplementary Figure S1).

The distinction might be caused by salinity or other covariate factors as temperature is homogenous across all sampling days at 75 m depth (A in Figure 1) while salinity is low on May12 but similarly higher in other sampling days at 75 m depth (B in Figure 1). Overall, bacterioplankton samples

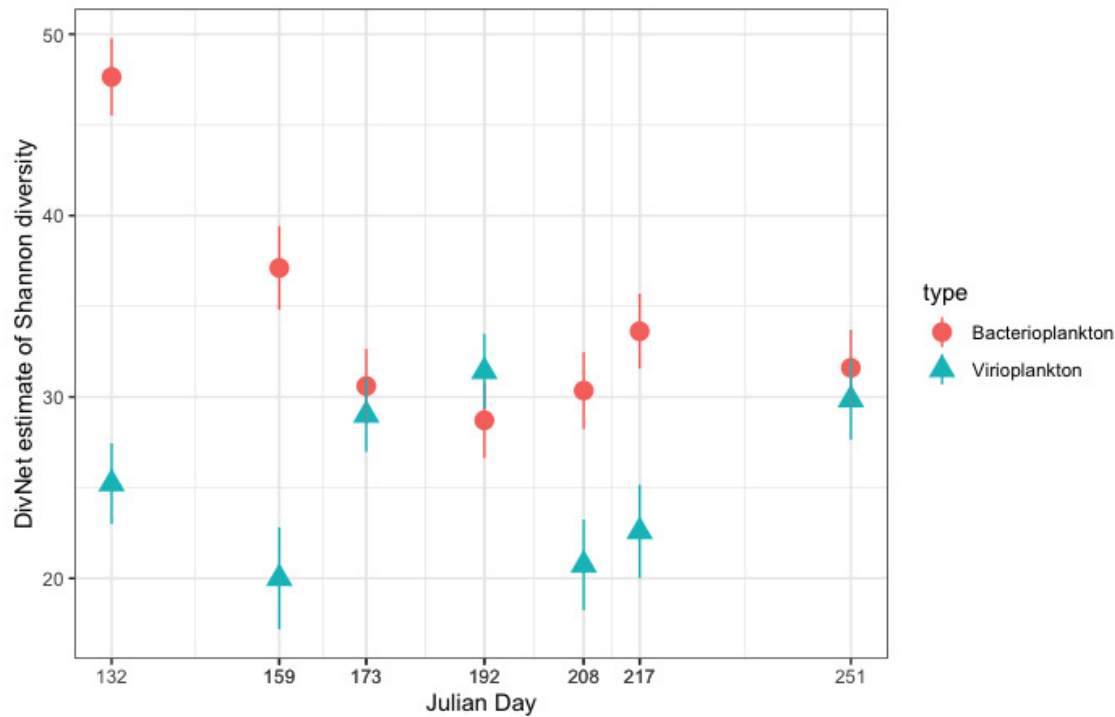


Figure 4. An Inverse Relationship was observed between estimated Shannon diversity of bacterioplankton (red) and viroplankton (cyan).

taken from 75 m depth are similar to each other because seawater was mostly stable at 75 m depth compared to other sampling depths. Bacterioplankton samples taken from 0 m and 5 m depth are highly dissimilar to each other, likely caused by the high disturbance on the sea surface along sampling days. Virioplankton samples taken from different depths did not show clear patterns (B in Supplementary Figure S1), likely caused by two reasons: Firstly, viroplankton living in the sampling depths carried functionally similar RTPR as the environments are mostly stable in the pelagic oceans. Secondly, RTPR, as a marker gene, does not provide as high resolution as the one provided by 16S rDNA provided because RTPR is a functional gene and not carried by all viruses while 16S rDNA is a conserved phylogenetic gene carried by all bacteria.

Bacterioplankton Shannon diversity decreased from May12 (Julian Day 132) to Jul11 (Julian Day 192) as seawater stratification started and intensified. It is likely because that down-welling anticyclonic vortex caused nutrient depletion at the surface seawater, subsequently caused less taxonomic groups of bacteria surviving in the environment. The existing dominant bacteria are likely to have broader phage ranges, as virioplankton Shannon diversity increased from Jun8 (Julian Day 159) to Jul11 (Julian Day 192). As a result, bacterioplankton Shannon diversity experienced a local minimum while virioplankton Shannon diversity experienced a local maximum on Jul11 (Julian Day 192). Later on, viral regulation effects increased bacterioplankton Shannon diversity. Virioplankton Shannon diversity decreased in return at the same time. Overall, bacterioplankton and virioplankton Shannon diversity have a reverse relationship.

Phylogenetically Similar Bacteria or Functionally Similar Viruses Exist in Similar Environments

CoDa Analysis Methods Produced More Reproducible Results Than Commonly Used Methods

CONCLUSION

[illegible]

ACKNOWLEDGEMENTS

[illegible]

Conflict of interest statement. None declared.

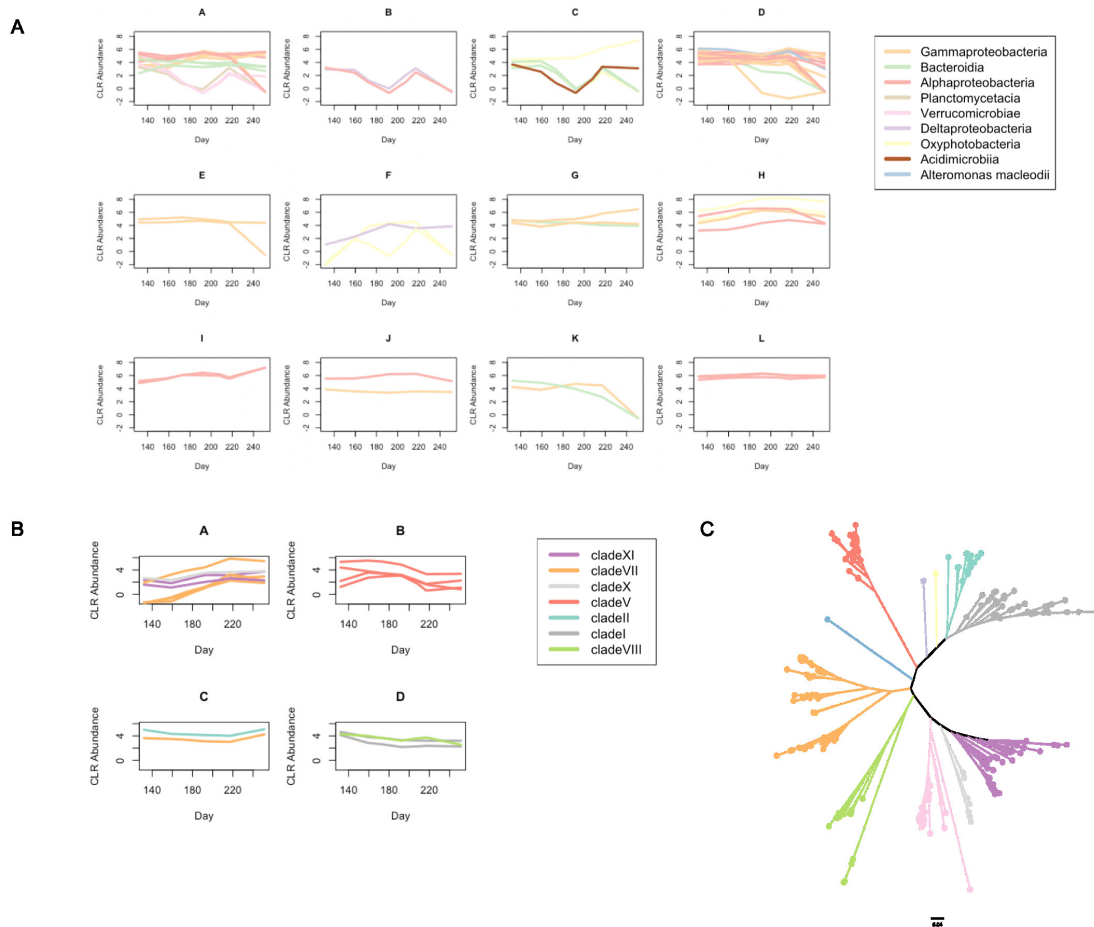


Figure 5. Proportionality test within (A) bacterioplankton and (B) viroplankton communities along sampling days at 50 m depth. Each panel represents clusters of concordant ASVs or OTUs with an expected value cutoff of 0.85 (bacterioplankton) or 0.65 (viroplankton). Each line in a cluster plot is the Loess line of best fit for the clr relative abundance (CLR Abundance on the y axis) of an individual ASV or OTU across sampling days (x axis). A 0 value indicates that the relative abundance of an ASV or OTU is equal to the mean log2 relative abundance of all ASVs or OTUs, while a positive or negative value indicates relative abundances greater or less than the mean log2 relative abundance, respectively. ASV lines of best fit are colored according to the class that the bacterioplankton ASV is classified into according to the key in (A). OTU lines of best fit are colored according to the clades that the viroplankton OTU is classified into according to the key in (B). Viroplankton clades are defined manually in the RTPR tree (C). Note that many of the clusters contain bacterioplankton ASVs related by the same class (A, C, D, E, G, H, I and L in (A)) and have approximately equal ratios between each other. And many of the clusters contain viroplankton OTUs related by the same clade (A, B and D in (B)) and have approximately equal ratios between each other as well. Associated bacterioplankton ASVs (A, C, D and F in (A)) related by different classes mostly had unequal ratios between each other along sampling days.

REFERENCES

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.
- Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.
- Bian, G., Gloor, G. B., Gong, A., Jia, C., Zhang, W., Hu, J., Zhang, H., Zhang, Y., Zhou, Z., Zhang, J., Burton, J. P., Reid, G., Xiao, Y., Zeng, Q., Yang, K., and Li, J. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere*, 2(5).
- Brooks, J. P. (2016). Challenges for case-control studies with microbiome data. *Annals of Epidemiology*, 26(5):336 – 341.e1. The Microbiome and Epidemiology.
- Brooks, J. P., Edwards, D. J., Harwich, Michael D. J., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Consortium, V. M., Strauss, Jerome F. r., Jefferson, K. K., and Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16s rna studies. *BMC microbiology*, 15:66–66.
- Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K.,

- Paul, J. H., Schroeder, D. C., Suttle, C. A., Vaqué, D., and Wommack, K. E. (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *The Isme Journal*, 2:575 EP –.
- Comeau, A. M. and Krisch, H. M. (2008). The Capsid of the T4 Phage Superfamily: The Evolution, Diversity, and Structure of Some of the Most Prevalent Proteins in the Biosphere. *Molecular Biology and Evolution*, 25(7):1321–1332.
- Doms, R. W., Lamb, R. A., Rose, J. K., and Helenius, A. (1993). Folding and assembly of viral membrane proteins. *Virology*, 193(2):545 – 562.
- Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A., and Breitbart, M. (2013). A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC evolutionary biology*, 13:33–33.
- Dez, B., Pedrós-Alió, C., and Massana, R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rna gene cloning and sequencing. *Applied and Environmental Microbiology*, 67(7):2932.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.
- Elledge, S. J., Zhou, Z., and Allen, J. B. (1992). Ribonucleotide reductase:

- regulation, regulation, regulation. *Trends in Biochemical Sciences*, 17(3):119–123.
- Erb, I. and Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135(1):21–36.
- Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLOS ONE*, 8(7):e67019–.
- Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, 399(6736):541–548.
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcúe, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8:2224.
- Gloor, G. B. and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8):692–703.
- Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcúe, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329.
- Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E., Gonda, R. M., Houtz, J. M., Hryckowian, A. J., Kelchner, V. A., Namburi, S., Pajcini, K. V., Popovich, M. G., Schleicher, D. T., Simanek, B. Z., Smith, A. L., Zdanowicz, G. M., Kumar, V., Peebles, C. L., Jacobs, Jr., W. R., Lawrence, J. G., and Hendrix, R. W. (2006). Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLOS Genetics*, 2(6):1–13.
- Jiang, S. C. and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Applied and environmental microbiology*, 64(8):2780–2787.
- Kang, S., Rodrigues, J. L. M., Ng, J. P., and Gentry, T. J. (2016). Hill number as a bacterial diversity measure framework with high-throughput sequence data. *Scientific Reports*, 6(1):38263.
- Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.
- Larose, D. T. and Larose, C. D. (2014). *An Introduction to Data Mining*, chapter 1, pages 1–15. John Wiley & Sons, Ltd.
- Lundin, D., Gribaldo, S., Torrents, E., Sjöberg, B.-M., and Poole, A. M. (2010). Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. *BMC Evolutionary Biology*, 10(1):383.
- Lundin, D., Torrents, E., Poole, A. M., and Sjöberg, B.-M. (2009). Rnrd, a curated database of the universal enzyme family ribonucleotide reductase, reveals a high level of misannotation in sequences deposited to genbank. *BMC genomics*, 10:589–589.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.
- McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *bioRxiv*.
- McMurdie, P. J. and Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, 8(4):e61217–.
- McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, 10(4):1–12.
- Moore, R. M., Harrison, A. O., McAllister, S. M., and Wommack, K. E. (2018). Iroki: automatic customization and visualization of phylogenetic trees. *bioRxiv*.
- Nordlund, P. and Reichard, P. (2006). Ribonucleotide reductases. *Annual Review of Biochemistry*, 75(1):681–706.
- Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.
- Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2015). zcompositions — r package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96.
- Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.
- Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., Searson, S., Acinas, S. G., Bork, P., Boss, E., Bowler, C., De Vargas, C., Follows, M., Grimsley, N., Hingamp, P., Jaillon, O., Karp-Boss, L., Krzic, U., Ogata, H., Raes, J., Reynaud, E. G., Sardet, C., Sieracki, M., Stemmann, L., Sullivan, M. B., Sunagawa, S., Velayoudon, D., Weissenbach, J., Wincker, P., and Coordinators, T. O. C. (2015). Open science resources for the discovery and analysis of tara oceans data. *Scientific Data*, 2(1):150023.
- Quinn, T. P., Richardson, M. F., Lovell, D., and Crowley, T. M. (2017). propr: An r-package for identifying proportionally abundant features using compositional data analysis. *Scientific Reports*, 7(1):16252.
- Rosenwasser, S., Ziv, C., Creveld, S. G. v., and Vardi, A. (2016). Virocell metabolism: Metabolic innovations during host-virus interactions in the ocean. *Trends in Microbiology*, 24(10):821–832.
- Sakowski, E. G., Munsell, E. V., Hyatt, M., Kress, W., Williamson, S. J., Nasko, D. J., Polson, S. W., and Wommack, K. E. (2014). Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proceedings of the National Academy of Sciences*, 111(44):15786–15791.
- Schlitz, R. (2002). Interactive analysis and visualization of geoscience data with ocean data view. *Computers & Geosciences*, 28(10):1211–1218.
- Shareware and freeware in the Geosciences II. A special issue in honour of John Butler.
- Schmidt, H. F., Sakowski, E. G., Williamson, S. J., Polson, S. W., and Wommack, K. E. (2014). Shotgun metagenomics indicates novel family a dna polymerases predominate within marine viroplankton. *The ISME journal*, 8(1):103–114.
- Schwartz, D. A. and Lindell, D. (2017). Genetic hurdles limit the arms race between prochlorococcus and the t7-like podoviruses infecting them. *The ISME Journal*, 11(8):1836–1851.
- Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Consortium, T. M. Q. C. P., Abnet, C. C., Knight, R., White, O., and Huttenhower, C. (2017). Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium. *Nature Biotechnology*, 35:1077 EP –.
- Spearman, C. (2010). The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150.
- Sullivan, M. B., Coleman, M. L., Weigle, P., Rohwer, F., and Chisholm, S. W. (2005). Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS biology*, 3(5):e144–e144.
- Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356–361.
- Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- van den Boogaart, K. G., Tolosana-Delgado, R., and Bren, M. (2018). compositions: Compositional data analysis.
- Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittner, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.
- Whittaker, G. R. and Helenius, A. (1998). Nuclear import and export of viruses and virus genomes. *Virology*, 246(1):1–23.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wickham, H. (2017). Easily install and load the 'tidyverse'.
- Wilhelm, S. W. and Suttle, C. A. (1999). Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, 49(10):781–788.
- Willis, A. and Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, 71(4):1042–1049.
- Willis, A., Bunge, J., and Whitman, T. (2017). Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(5):963–977.
- Willis, A. D. and Martin, B. D. (2018). Divnet: Estimating diversity in networked communities. *bioRxiv*.
- Zhong, Y., Chen, F., Wilhelm, S. W., Poorvin, L., and Hodson, R. E. (2002). Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Applied and environmental microbiology*, 68(4):1576–1584.