Molecular ecological studies on the effect of viral infection on abundant marine prokaryotes

Kento Tominaga

2021

Table of contents

Chapter 1

General introduction

The marine environment covers 70% of the Earth's surface and thus, marine biogeochemical cycle such as carbon cycle plays a critical role in Earth's habitability. In the photic zone, phytoplankton (cyanobacteria and eukaryotic microalgae) are responsible for the majority of oceanic primary production and comparable to approximately 50% of earth's primary production (Field, 1998). Approximately half of the fixed carbon through the primary production is released from marine phytoplankton cells into the environment as dissolved organic matters (DOMs), then catabolized and re-mineralized by heterotrophic prokaryotes, which comprise approximately 10^{29} cells in the ocean (Cole *et al.*, 1988; Whitman *et al.*, 1998). The heterotrophic prokaryotes are preyed by eukaryotic unicellular phagotrophs and mixotrophs and thereby the fixed organic carbon is channeled back to the classic food chain comprising phytoplankton, zooplankton, and multicellular larger eukaryotes such as fish (Azam *et al.*, 1983; Buchan *et al.*, 2014; Worden *et al.*, 2015). This flux taking heterotrophic prokaryote and unicellular eukaryotes into account is called "microbial loop" and considered as an ecologically significant pathway of marine carbon flux (Azam *et al.*, 1983; Falkowski *et al.*, 2008)

The other key players in the marine biogeochemical cycling are marine viruses (Suttle, 2005, 2007; Zimmerman *et al.*, 2020). Viruses in the ocean are estimated to be approximately one or two orders of magnitude more abundant than prokaryotes (~ 10^{31} particles) (Bergh *et al.*, 1989; Proctor and Fuhrman, 1990). The marine viruses are estimated to contribute to 10~40% of prokaryotic cell mortality per day and give rise to the release of their cellular compounds and metabolites to the DOM pools, which can be

readily taken up by other microorganisms (Suttle, 1994, 2005, 2007; Breitbart *et al.*, 2018). The role of viruses in the biogeochemical cycle has been conceptualized as "viral shunt" through which up to 25% of oceanic fixed organic carbon is predicted to be recycled (Wilhelm and Suttle, 1999). Viral infection also leads to qualitative and quantitative changes of cellular organic matters of infected cells via metabolic reprograming for viral effective propagation (Jover *et al.*, 2014). Thus, viruses are responsible for the biogeochemical cycling in terms of not only quantity but also stoichiometry and composition (Jover *et al.*, 2014). Further, the marine viruses are responsible for the carbon removal via cell lysis which give rise to cell debris; such aggregated cell debris could sink and result in carbon removal from the surface layer (Guidi *et al.*, 2016; Zimmerman *et al.*, 2020). This mechanism is called the "viral shuttle" and drives the sequestration of carbon from the atmosphere to the ocean interior and seafloor sediments (Sullivan *et al.*, 2017; Zimmerman *et al.*, 2020).

In line with the biogeochemical importance of prokaryotes and their viruses introduced above, the host-specific viral infection is also believed to maintain the diversity of host prokaryotic community. Basically, it is thought that viruses infect their specific hosts (often restricted to strain within a species) in a frequency-dependent manner according to the increasing encounter rate between viruses and their hosts (Fuhrman and Suttle, 1993; Winter *et al.*, 2010). Especially, it is postulated that the high host cell density ($>10^4$ cells/ml) is required for effective viral propagation (Wiggins and Alexander, 1985). Thus, viruses have a role in modulating prokaryotic diversity through host-specific infection which prevents to their prokaryotic host populations (stain or genotype) to become dominant and which maintains diversity among genetically closely related prokaryotic populations through increased viral-induced mortality (Thingstad, 2000;

Rodriguez-Valera *et al.*, 2009). The proposed mechanism which is responsible for maintaining diversity among genetically closely related prokaryotic populations is called "constant diversity (CD) dynamics" (Rodriguez-Valera *et al.*, 2009). Actually, various mesocosm and field studies in aquatic ecosystems have supported the frequency-dependent viral predation manner (Schwalbach *et al.*, 2004; Bouvier and Del Giorgio, 2007; Yoshida *et al.*, 2008; Rodriguez-Brito *et al.*, 2010; Kuno *et al.*, 2012; Parsons *et al.*, 2012; Kimura *et al.*, 2013; Needham *et al.*, 2013; Cram *et al.*, 2016). If true, these influences of viral infection introduced above seem to be more important in abundant prokaryotic populations (Fuhrman and Suttle, 1993), however, the prediction has not yet been comprehensively evaluated.

Mathematical models of viral and host abundance have described that a prokaryotic species (or lineage) with a faster growth rate than others is susceptible to viral infection (Thingstad, 2000). The idea often explained the relationship of viral and its host abundance in the marine environment in connection with *r*-/*K*- selection theory, which describes the trade-off between population growth rate (*r*, intrinsic rate of natural increase) and its sustainable maximum population size (*K*, carrying capacity) (Andrews and Harris, 1986; Suttle, 2007). In the explanation, *r*-strategists (e.g. members of *Flavobacteriaceae*) having higher growth rate and higher metabolic activities undergo more frequent viral infection, eventually leading to their relatively low abundance and/or frequent fluctuation (Suttle, 2007). In contrast, *K*-strategists (e.g. SAR11) having lower growth rates and lower metabolic activities would be more resistant to viral infection and become dominant in a given ecosystem (Suttle, 2007). However, the discovery of SAR11 viruses as the most abundant marine viruses raised a question to the prediction (Zhao *et al.*, 2013; Zhang *et al.*, 2020). Thus, it is still unknown whether the viral infection

generally occurs in *K*-strategist populations and whether viral infection is prevalent in abundant prokaryotic populations according to their density.

Measuring the abundance of viruses infecting each prokaryotic population in the environment is difficult because of the complex interactions among diverse prokaryotes and their viruses (Brum and Sullivan, 2015). So far, sequencing of the 16S rRNA gene of the environmental prokaryotic community has revealed over 35,000 species-level operational taxonomic units (OTUs, based on 97% sequence identity) in the ocean (Sunagawa *et al.*, 2015). Although most oceanic prokaryotic species fall into 13 major lineages corresponding to phyla (or class for proteobacteria) such as α-proteobacteria (e.g. SAR11 clade, SAR116 clade, and *Roseobacter* clade), γ-proteobacteria (e.g. SAR86 clade and SAR92 clade), Bacteroides (e.g. members of *Flavobacteriaceae*), and Cyanobacteria (e.g. *Synechococcus*) (Pommier *et al.*, 2006; Sunagawa *et al.*, 2015), metabolic capacity, physiologies, and ecology among these species are highly divergent and often distinct between interspecies-level populations (strains or genotypes) (Chafee *et al.*, 2018; Sieradzki *et al.*, 2019; Van Rossum *et al.*, 2020).

The vast majority of marine prokaryotes could not be cultivated using standard techniques (Rappé and Giovannoni, 2003) and approximately 50% of the class to genus-level taxonomic groups still remain uncultivated (Lloyd *et al.*, 2018). Thus, only a limited number of culture-based marine virus-host model systems have been studied except for picocyanobacteria-virus systems, which have characterized more than 100 viruses (Rappé and Giovannoni, 2003; Brum and Sullivan, 2015; Lloyd *et al.*, 2018). Accordingly, the diversity of marine viruses remains to be underrepresented by culture-based approaches (Brum and Sullivan, 2015). Furthermore, cultivation-independent studies for environmental viruses also had a limitation to evaluate its diversity because viruses lack

a universally conserved marker gene in contrast with 16S rRNA gene for prokaryotes (Edwards and Rohwer, 2005).

Recently, high-throughput sequencing offers metagenomic-based studies to investigate viral genomic information on a community-wide scale (Edwards and Rohwer, 2005). Nevertheless, the majority (63–93%) of sequences in marine viral metagenomes did not have closely related genomes in public databases, indicating the majority of uncultured viruses were distantly related to cultured ones (Hurwitz and Sullivan, 2013). Therefore, determining which virus infects which host populations was difficult on the basis of similarity with cultured viruses (so-called Who infects whom problem, Brum and Sullivan, 2015). For example, a previous study characterized 1,811 circular viral genomes from the marine viromes, however, 78.4% (1,420 genomes) of them were not assigned with their putative host (Nishimura, Watai, *et al.*, 2017). Therefore, the influence of viral infection on abundant marine prokaryotes in environments remains to be not well understood.

To this end, I aimed to examine whether the viral infection is prevalent in uncultured, abundant marine prokaryotic populations. For this purpose, first, I improved *in silico* host prediction methods for uncultured viruses to overcome the limitation to predict virus-host interactions in environments. As a model case, I focused on the viruses infecting marine Bacteroidetes, which is one of the most abundant heterotrophic prokaryotic phyla in marine ecosystems. I developed methods of *in silico* prediction of putative Bacteroidetes viruses from recently reported 1,811 marine viral genomes, by using microbial metagenome-assembled genomes (MAGs). This provided novel 81 genomes that were newly recognized as Bacteroidetes virus including those phylogenetically distant from the cultured marine viruses. Second, I applied the host

prediction method to diverse marine prokaryotic taxa. Then, I examined the temporal dynamics of the prokaryotes and viruses based on the host prediction at Osaka Bay for 2 years. To investigate whether the viral infection increased according to the putative host frequency, I analyzed the statistical association of the dynamics of viruses and their putative host by co-occurrence network analysis. Viral abundance increased with the increasing of putative host abundance between the predicted pairs, suggesting that frequency-dependent viral infection prevailed in the abundant populations of marine prokaryotes. Further, the faster temporal succession of the viral community than prokaryotes suggests that different viruses can infect a continuously dominated *K*-strategist host population at different time points.

Chapter 2

In silico prediction of virus-host interactions for marine Bacteroidetes with the use of metagenome-assembled genomes

Abstract

Bacteroidetes is one of the most abundant heterotrophic bacterial taxa in the ocean and play crucial roles in recycling phytoplankton-derived organic matter. Viruses of Bacteroidetes are also expected to have an important role in the regulation of host communities. However, knowledge on marine Bacteroidetes viruses is biased towards cultured viruses from a few species, mainly fish pathogens or Bacteroidetes not abundant in marine environments. In this study, I investigated the recently reported 1,811 marine viral genomes to identify putative Bacteroidetes viruses using various *in silico* host prediction techniques. Notably, I used microbial metagenome-assembled genomes (MAGs) to augment the marine Bacteroidetes reference genomic data. The examined viral genomes and MAGs were derived from simultaneously collected samples. Using nucleotide sequence similarity-based host prediction methods, I detected 31 putative Bacteroidetes viral genomes. The MAG-based method substantially enhanced the predictions (26 viruses) when compared with the method that is solely based on the reference genomes from NCBI RefSeq (7 viruses). Previously unrecognized genus-level groups of Bacteroidetes viruses were detected only by the MAG-based method. I also developed a host prediction method based on the proportion of Bacteroidetes homologs in viral genomes, which detected 321 putative Bacteroidetes virus genomes including 81 that were newly recognized as Bacteroidetes virus genomes. The majority of putative Bacteroidetes viruses were detected based on the proportion of Bacteroidetes homologs

in both RefSeq and MAGs; however, some were detected in only one of the two datasets. Putative Bacteroidetes virus lineages included not only relatives of known viruses but also those phylogenetically distant from the cultured viruses, such as marine Far-T4 like viruses known to be widespread in aquatic environments. The MAG and protein homology-based host prediction approaches enhanced the existing knowledge on the diversity of Bacteroidetes viruses and their potential interaction with their hosts in marine environments.

Introduction

Marine heterotrophic prokaryotes are responsible for processing almost half of the organic matter that is fixed by marine phytoplankton, thus playing an important role in the global carbon cycle (Azam and Malfatti, 2007). Members of the phylum Bacteroidetes are the most abundant heterotrophic prokaryotes in the ocean along with those belonging to Proteobacteria (Glöckner *et al.*, 1999; Kirchman, 2002). Bacteroidetes inhabit various marine environments ranging from coastal water to open ocean habitats (Alonso et al., 2007; Pommier et al., 2006). They are especially abundant during and after the phytoplankton blooms and believed to have an important role in the decomposition and remineralization of the phytoplankton biomass (Teeling *et al.*, 2012). A previous study suggests that there are 1,200 species of marine planktonic Bacteroidetes and only about half of their global diversity has been described by cultivation (Alonso *et al.*, 2007). Despite being the abundant species during phytoplankton blooms, isolated marine Bacteroidetes strains are rarely observed in environment; therefore, most abundant lineages of marine Bacteroidetes remain poorly characterized (Unfried *et al.*, 2018).

Marine viruses are being increasingly recognized as important factors affecting the structure and function of the prokaryotic community through diverse virus-host interactions, which drive the global biochemical cycle in the ocean (Suttle, 2007; Yoshida *et al.*, 2019). Considering the importance of Bacteroidetes in the marine biochemical cycle, their viruses also likely have significant impact on the process. To date, 58 genomes have been reported for Bacteroidetes viruses isolated from aquatic environments (Puig and Girones, 1999; Borriss *et al.*, 2007; Cheng *et al.*, 2012; Kang, Jang, *et al.*, 2012; Kang, Kang, *et al.*, 2012; Holmfeldt *et al.*, 2013; Luhtanen *et al.*, 2014; Castillo *et al.*, 2014; Kang *et al.*, 2015, 2016; Laanto *et al.*, 2015; Castillo and Middelboe, 2016; Mihara *et al.*,

2016). However, their hosts cover only seven species of Bacteroidetes. Moreover, the host species of these viruses were biased toward coastal rare taxa (e.g., *Cellulophaga baltica*) or fish pathogen *Flavobacterium*. Therefore, our understanding on marine Bacteroidetes viruses brought by cultivation-based approaches are limited to less abundant taxa in the ocean..

Owing to the recent development of sequencing technology, viral metagenomes (viromes) have become a powerful tool to characterize the diversity of viruses as an alternative of the classical cultivation strategy (Brum and Sullivan, 2015). For example, Nishimura *et al.* (2017) recently constructed 1,600 complete environmental viral genomes (EVGs) from marine viromes. Among them, the authors identified 239 viral genomes which were classified into two groups, referred to as groups 1 and 2, likely infecting *Flavobacteriaceae*, a major group of marine Bacteroidetes(Nishimura, Watai, *et al.*, 2017). Although these groups include highly diverse viruses (representing 29 and 25 genus-level OTUs (gOTUs) based on genomic similarity), they showed a significant genomic similarity with the cultured siphoviruses infecting *Nonlabens* (group 1) or the podovirus phi38:1 infecting *Cellulophaga baltica* (group 2; one of the most globally abundant type of virus in the oceans), respectively (Roux *et al.*, 2016; Nishimura, Watai, *et al.*, 2017). Thus, our knowledge of the genome repertoire of marine Bacteroidetes viruses are still limited to the relatives of cultured Bacteroidetes viruses even after the application of viral metagenomics approaches.

Since viromes revealed enormous diversity of viruses with no isolated relatives, linking these viruses with their putative hosts by culture independent methods has become important to gain insights into the ecology of viruses. Recently, several *in silico* host prediction approaches using viral and microbial genomes have been developed

(Edwards *et al.*, 2016; Ahlgren *et al.*, 2017). These methods detect virus-host signals in viral and microbial genomes, which are shaped by virus-host co-evolutionary processes such as acquisition of CRISPR spacer sequences (Edwards *et al.*, 2016). However, genomic information of uncultured microorganisms is still limited (Rappé and Giovannoni, 2003; Locey and Lennon, 2016) and represents a major hurdle to expand our knowledge of virus-host interaction even though such in silico approaches.

Recently, metagenome assembled genomes (MAGs), which can aid us in overcoming this limitation, are receiving increasing attention. Development of metagenomic assembly, binning, and curation techniques have enabled us to construct nearly complete genomes of uncultured microorganisms from various environments (Anantharaman *et al.*, 2016; Bowers *et al.*, 2017; Parks *et al.*, 2017; Tully *et al.*, 2017, 2018; Delmont *et al.*, 2018; Stewart *et al.*, 2018; Almeida *et al.*, 2019; Nayfach *et al.*, 2019; Pasolli *et al.*, 2019). Recent studies have reported over 3000 microbial MAGs including over 500 putative Bacteroidetes MAGs (Tully *et al.*, 2017, 2018; Delmont *et al.*, 2018) from metagenomic samples obtained from the Tara Oceans expedition (Sunagawa *et al.*, 2015).

In this study, I performed a computational host prediction analysis for a thousand of EVGs, using the Bacteroidetes MAGs as potential host genomes, to overcome the bottleneck of viral host prediction and expand our knowledge of the diversity of Bacteroidetes viruses. The MAG based prediction approach is expected detect lineage-specific interactions between EVGs and their hosts, which will be compared with the previous family level host prediction of *Flavobacteriaceae* EVG group 1 and 2. Considering the locality of marine virus-host interaction (Brum *et al.*, 2015; Yoshida *et al.*, 2018), these microbial MAGs likely represent ideal host candidates for the EVGs,

because most of the MAGs and EVGs were obtained from simultaneously sampled metagenomes of the *Tara* Oceans expedition (Brum *et al.*, 2015; Sunagawa *et al.*, 2015). A recent study successfully detected viruses-host interactions by such an approach in samples from a freshwater lake (Okazaki *et al.*, 2019). I also applied a protein homology-based method after carefully examining prediction parameters for prediction of Bacteroidetes viruses, which enabled a more sensitive signal detection than previously proposed nucleotide similarity-based in silico methods.

Materials and Methods

Collection of viral and Bacteroidetes genomes

I used the previously assembled 1,811 environmental viral genomes (EVGs; all being circularly assembled genomes) derived from marine viromes (Nishimura, Watai, *et al.*, 2017). Genus-level genomifc operational taxonomic units (gOTUs) were assigned to these EVGs as previously described (Nishimura, Watai, *et al.*, 2017). I also collected 58 isolated Bacteroidetes viral genomes and 100 randomly selected isolated prokaryotic viral genomes infecting non-Bacteroidetes prokaryotes (e.g., Proteobacteria) as reference viral genomic data from NCBI RefSeq (as of April 2019).

Bacteroidetes genomes that were publicly available prior to April 2019 were collected from NCBI RefSeq (total 3,695 genomes representing 2,148 species) and used as references for the host prediction analysis. I also collected 3,882 MAGs from the *Tara* Oceans metagenomic datasets (here after referred to as TARA-MAGs), which include 518 MAGs assigned to the phylum Bacteroidetes in the original studies (here after referred to as Bacteroidetes-MAGs) (Tully *et al.*, 2017, 2018; Delmont *et al.*, 2018). To remove the contamination of virus-like contigs from TARA-MAGs, 11,537 contigs predicted as viral-like sequence (category 1, 2, and 3) by VirSorter (Roux, Francois

Enault, *et al.*, 2015) were discarded from 1,732 MAGs. Taxonomy of the Bacteroidetes-MAGs predicted as hosts of EVGs were further confirmed based on the conserved maker genes in bacterial genomes by GTDB-Tk with classify mode (Chaumeil *et al.*, 2019).

Host prediction by nucleotide similarity-based methods

I used four computational host prediction strategies that are frequently used to identify potential virus-host interactions. All of these methods utilize nucleotide sequence similarity for prediction, and details of these methods are reviewed elsewhere (Edwards *et al.*, 2016). (i) CRISPR spacers match: CRISPR spacer sequences from Bacteroidetes genomes were predicted by CRISPR Recognition Tool (Bland *et al.*, 2007). Sixty-nine thousand one hundred and seventy-two and 2,004 spacer sequences were extracted from the Bacteroidetes genomes in NCBI RefSeq and MAGs, respectively. Detected spacer sequences were queried against EVGs using the BLASTn-short function with these parameters: at least 95% identity over the whole spacer length and only 1–2 SNPs at the 5′end of the sequence was allowed. (ii) tRNA match (Paez-Espino *et al.*, 2016) : tRNAs were recovered from bacterial genomes and EVGs by ARAGORN with '-t' option (Laslett and Canback, 2004). tRNAs (192,217, 13,018, and 6,322) were recovered from the Bacteroidetes genomes in NCBI Refseq, MAGs, and EVGs, respectively. The recovered tRNAs were compared by BLASTn (Camacho *et al.*, 2009) and only a perfect match (100% length and 100% sequence identity) was considered indicative of putative Bacteroidetes-virus pairs. (iii) Nucleotide sequence homology of Bacteroidetes genomes and EVGs: EVGs were queried against Bacteroidetes genomes using BLASTn (Camacho *et al.*, 2009). Only the best hits above 70% identity across alignment with length ≥1000 bp were indicative of Bacteroidetes-virus pairs. (iv) Oligo nucleotide frequency (ONF) distance: Oligo nucleotide frequency and distance between MAGs and EVGs were

calculated by VirHostMatcher with a dissimilarity score <0.13 as an indication of Bacteroidetes-virus pairs (Ahlgren *et al.*, 2017).

I performed taxonomic validation for each contig in Bacteroidetes-MAG showing similarity with EVGs in the above methods (CRISPR, tRNA, and nucleotide sequences homology) by the following procedures as previously described with slight modification (Coutinho *et al.*, 2017). Open reading frames (ORFs) of each contig were predicted by MetaGeneMark with -p 0 option (Zhu *et al.*, 2010) and queried against RefSeq database (as of May 2018) by BLASTp (E-value <1e-10, identity >30%, and bit score >50). The sum of the bit score of the all best hits from each contig was calculated, and if >80% of the total bit score was consistently assigned to Bacteroidetes, the contig of the MAG was considered to be derived from Bacteroidetes genomes; otherwise it was considered as a contaminant contig from other taxa (i.e. not Bacteroidetes). If a contig was regarded as contaminant contigs, the EVG showing similarity with the contig were removed from candidates of Bacteroidetes virus. Similarly, to remove viral contamination-like contigs in RefSeq Bacteroidetes genomes, the contigs predicted as viruses by VirSorter (Roux, Francois Enault, *et al.*, 2015) were discarded.

Calculation of the proportion of Bacteroidetes homologs in viral genomes

ORF for the viral genomes was predicted by MetaGeneMark with -p 0 option (Zhu *et al.*, 2010). Homology search was conducted using BLASTp against the RefSeq database (as of May 2019, bit score >50). Similarly, BLASTp search was conducted against the ORFs of TARA-MAGs predicted by MetaGeneMark with -p 0 option (Zhu *et al.*, 2010). Taxonomic validation to the matched contigs of the MAGs was performed as described in the previous section. Among the most closely matched cellular homologs of a viral genome, proportion of the Bacteroidetes homologs was calculated. To check the

possible origin of the Bacteroidetes homologs, putative provirus regions in the Bacteroidetes genomes were checked by VirSorter (category 4, 5, and6) (Roux, Francois Enault, *et al.*, 2015). If the Bacteroidetes homologs were encoded within the provirus region, the Bacteroidetes homologs were regarded as provirus origin.

Proteomic tree calculation

The viral proteomic tree (Rohwer and Edwards, 2002) was calculated between 4,240 viral genomes in a previous study (Nishimura, Watai, *et al.*, 2017) or constructed based on their genome similarity scores derived from all-against-all tBLASTx computation as previously described (Bhunchoth *et al.*, 2016; Nishimura, Watai, *et al.*, 2017; Nishimura, Yoshida, *et al.*, 2017). Parts of the proteomic tree were visualized from ViPTree webserver (Nishimura, Yoshida, *et al.*, 2017) and an interactive visualization server of viral genomes developed in a previous study (Nishimura et al., 2017a; https://www.genome.jp/tools/mg_viewer2/.)

Gene prediction and annotation

Gene prediction and functional annotation of the EVGs were obtained from a previous study (Nishimura, Watai, *et al.*, 2017). Additionally, to explore the auxiliary metabolic genes (AMGs), ORFs were queried against the Pfam domain database v.31 (Finn et al., 2016) with hmmsearch (threshold 10^{-5} for E-value) (Eddy, 2011) and annotated by eggNOG-mapper (Huerta-Cepas *et al.*, 2017) using eggNOG 5.0 database (Huerta-Cepas *et al.*, 2019). Protein motifs found in the AMGs were defined according to the previous studies (Roux *et al.*, 2016; Luo *et al.*, 2017)

Phylogenetic trees of Gp23 of Far-T4 like viruses

Far-T4 reference genomic fragments assembled from freshwater viromes were obtained from Metavir web server under project "FarT4 / Far-T4 Lake Pavin" (Roux, François Enault, *et al.*, 2015). Other reference sequences were obtained from the NCBI RefSeq database of complete viral genomes. Multiple sequences were aligned using the MAFFT program (version 7.245) (Katoh *et al.*, 2002), with the FFT-NS-2 mode and a maximum of 1,000 iterations (--retree 2, --maxiterate 1000). Conserved positions in the alignments were selected with the trimAl program (version 1.3) (Capella-Gutierrez *et al.*, 2009). Approximately maximum likelihood trees were constructed by FastTree (Price *et al.*, 2010) and visualized by iTOL (Letunic and Bork, 2019).

Virome read mapping

Forty-three *Tara* Oceans viromes were downloaded from the European Nucleotide Archive (www.ebi.ac.uk/ena/) under accession numbers reported in the original study (Brum *et al.*, 2015) and quality control was performed as previously described (Nishimura, Watai, *et al.*, 2017). The quality controlled sequences were mapped against the 1,811 EVGs using Bowtie2 with a parameter "--score-min L,0,-0.3" (Langmead and Salzberg, 2012). Fragments per kilobase per mapped million reads (FPKM) values were calculated by in-house ruby scripts (Nishimura, Watai, *et al.*, 2017).

Results

Detection of Bacteroidetes viruses by nucleotide similarity-based methods

To identify novel Bacteroidetes-virus pairs, I first conducted host prediction analyses on the 1,811 EVGs based on CRISPR spacer sequences, tRNA genes, sequence similarity (BLASTn) and ONF distance, by using 3,695 Bacteroidetes genomes in NCBI RefSeq and 518 Bacteroidetes-MAGs (**Table 2-1**). In total, I detected 57 signals of virus-

host interactions between EVGs and Bacteroidetes-MAGs or Bacteroidetes genomes in RefSeq. An EVG (TARA_ERS490053_N000309) was predicted as Bacteroidetes virus with both datasets. After removal of redundancy, 35 EVGs including 18 previously described as members of *Flavobacteriaceae* viruses were predicted as putative Bacteroidetes viruses. Of these, OBV_N00073 and OBV_N00010 were previously predicted as viruses infecting SAR 11 and Marine group II archaea, respectively. I discarded these two EVGs as false positives from further analysis, taking into consideration the limitation of computational host prediction accuracy (Edwards *et al.*, 2016) and the previous detailed analysis (Nishimura, Watai, *et al.*, 2017). The remaining 33 Bacteroidetes EVGs were classified into 18 genus-level groups (gOTUs) based on the viral genome similarity (**Table 2-2**)

Table 2-1. The number of EVGs assigned to Bacteroidetes viruses according to nucleotide based-methods (i.e., CRISPR, tRNA, BLASTn, and oligonucleotide frequency) using Bacteroidetes genomes.

	CRISPR	tRNA	BLASTn (> 1 kb)	Oligo nucleotide frequency	Total
3,695 Refseq Bacteroidetes genomes	3	0	16	0	19
518 TARA Bacteroidetes MAGs	1	14	18	5	38

The nucleotide similarity-based approaches for the EVGs and Bacteroidetes genomes in RefSeq revealed 20 signals of virus-host interactions (between 6 EVGs and 18 Bacteroidetes genomes in RefSeq; **Table 2-2**). All the 6 EVGs were classified as the members of the *Flavobacteriaceae* EVG group 1, previously identified by their genomic similarity to cultured Bacteroidetes viruses (**Table 2-2**). Putative host Bacteroidetes of

these EVGs were members of *Flavobacteriaceae* isolated from marine environments such as sea water (Nedashkovskaya *et al.*, 2005; Yu *et al.*, 2014; Dai *et al.*, 2015; Xing *et al.*, 2015), marine sediment (Miyazaki *et al.*, 2010; Lee *et al.*, 2014), sponges (Esteves *et al.*, 2013; Morrissey *et al.*, 2015), and coral reef (Keller-Costa *et al.*, 2016) samples. Our results not only support the previous host prediction studies based on genomic similarity with cultivated Bacteroidetes viruses and genomic context (Nishimura, Watai, *et al.*, 2017), but also offer additional clues for lineage specific interaction between *Flavobacteriaceae* EVGs and Bacteroidetes. For example, two EVGs classified into a genus-level genomic OTU (G490 in the previous study, Nishimura, Watai, *et al.*, 2017) were paired with *Aquimarina* species which is associated with marine sponge or coral reef (**Table 2-2**).

The nucleotide similarity-based approaches for the EVGs and Bacteroidetes-MAGs revealed 37 signals between 26 EVGs and 13 MAGs (**Table 2-2**). Although Bacteroidetes-MAG data were seven-folds smaller in size than the genomic data from RefSeq, Bacteroidetes-MAGs have twice as many significant signals with EVGs. Among the 26 putative Bacteroidetes EVGs, two and 11 EVGs were members of the *Flavobacteriaceae* EVG group 1 and group 2, respectively (Nishimura, Watai, *et al.*, 2017). Also, TARA_ERS490388_N000065 showed nearly genus-level similarity with *Cellulophaga* viruses classified into Cba41likevirus (Holmfeldt *et al.*, 2013). In addition to these previously described Bacteroidetes EVGs, I detected 12 new candidates of Bacteroidetes EVGs classified into five genus-level groups from MAG-based prediction (**Table 2-2**). I performed taxonomic classification of the putative host MAGs by genome-based phylogeny (Parks *et al.*, 2018). I could not classify some of these putative host MAGs because of the low completeness. However, the classification of high

completeness MAGs suggests that most of the putative host MAGs are members of marine uncultured Bacteroidetes lineages, from which no viruses have been previously described (**Table 2-2**). For example, three MAGs were classified into candidates genus SHAN690 mostly composed of marine MAGs (Parks *et al.*, 2017) and one MAG was classified into another candidates genus MS024-2A mostly composed of marine single cell genomes (Woyke *et al.*, 2009).

Detection of Bacteroidetes viruses by protein homology-based approach

The nucleotide similarity-based approaches enabled us to detect a large number of Bacteroidetes viruses when combined with the TARA-MAG data than when it was solely based on cultured strain genomes. However, most members of the previously described 239 *Flavobacteriaceae* EVGs were still not detected by the nucleotide similarity-based methods (Nishimura, Watai, *et al.*, 2017). This was due to the fact that the nucleotide similarity-based prediction methods rely on rare and/or strain specific evolutionary events such as acquisition of CRISPR spacer or horizontal gene transfer (Edwards *et al.*, 2016). Further, nucleotide sequence-based comparison can detect only recent evolutionary events because nucleotide sequences can change more rapidly than protein sequences because of redundancy in the genetic code (Edwards *et al.*, 2016). I therefore developed a more sensitive method to detect Bacteroidetes viruses based on protein-homology. Bacterial homologues (i.e. the match with the lowest E-value) of viral encoded proteins are frequently found in Bacterial genomes in the same phylum as the host of the viruses (Mahmoudabadi and Phillips, 2018). Actually, 10% to 92% of proteins encoded in the genomes of the *Flavobacteriaceae* EVG groups 1 and 2 were most similar to Bacteroidetes genes (Nishimura, Watai, *et al.*, 2017). However, the proportion of Bacteroidetes homologs was not tested in other EVGs and the prediction method was not

standardized in the previous study. I hypothesized that the Bacteroidetes viruses have more Bacteroidetes homologs than other prokaryotic viruses, and thereby the proportion of Bacteroidetes homologs in viral genomes may be a useful genetic signal of Bacteroidetes viruses.

Firstly, we examined the proportion of proteins that best hit to Bacteroidetes proteins (defined as the most similar protein detected by BLASTp; E-value <1e-10, identity >30%, and bit score >50) for cultured Bacteroidetes viruses (**Figures 2-1A, 2-1B**, **2-2**). As expected, most of the cultured Bacteroidetes viruses have many homologs of Bacteroidetes in RefSeq (average 35.8%) or TARA-MAGs (average 11.6%) in their genomes (**Figures 2-1**, **2-2**). Among the possible homologs-sharing mechanisms between bacteria and viruses, we examined the contribution of provirus and AMGs to the shared homologs. Provirus-like regions in Bacteroidetes genomes appeared to mainly contribute (**Figure 2-4**, average: 55.5%, maximum: 96%) to these homologs. This trend was observed not only in the lysogenic viruses or viruses having putative integrase homologs but also in the lytic Bacteroidetes viruses (**Figures 2-2A, 2-2B**). In contrast, AMGs rarely contributed (**Figure 2-4**, average: 3%, maximum: 11%) to the detection of Bacteroidetes homologs (**Figure 2-4**). The viruses infecting other prokaryotes (i.e., non-Bacteroidetes viruses) rarely showed Bacteroidetes homologs (**Figures 2-1A**, **2-1B** , **2-3A, 2-3B**, at most 7.9% and 4.2% to Bacteroidetes in RefSeq and TARA-MAGs, respectively). According to the comparison of the result between Bacteroidetes viruses and non-Bacteroidetes viruses, we chose the following criteria for the prediction of putative Bacteroidetes EVGs. We considered EVGs that satisfy all the following three criteria as Bacteroidetes EVGs: (i) At least 7.9% or 4.2% of viral genes should be homologs of Bacteroidetes genes in RefSeq or TARA-MAGs, respectively,

Table 2-2. Predicted virus-host pairs between EVGs and Bacteroidetes genomes.

EVGs	gOTU	Prediction Method	Bacteroidetes genome ID	Matched Bacteroidetes Genome Contig	Bacteroidetes genome source	Host Taxonomy	Taxonomical classification of MAGs by GTDBtk
OBV_N00073	5	CRISPR	GCF_003984825	NZ_RYDM01000010	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Flavobacterium:Flavobacterium sp. RSP46	-
TARA_ERS492160_N000662	468	CRISPR	GCF_001683825+E4:G6	NZ_LXTR01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Flavobacterium:Flavoceae bacterium CP2B	-
TARA_ERS488499_N000464	468	CRISPR	GCF_002954665	NZ_MSCM01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Polaribacter:Polaribacter glomeratus ATCC 43844	-
TARA_ERS490320_N000023	490	blastn	GCF_900624725	NZ_UYXD01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina sp. Aq349	-
TARA_ERS490320_N000023	490	blastn	GCF_900299485	NZ_OMKB01000021	RefSeq	Bacteroidetes:Flavobacteriia:Flavoles:Flavoceae:Aquimarina:Aquimarina sp. Aq349	-

Table 2-2. Continued

TARA_ERS490320_N000023	490	blastn	GCF_900089995	NZ_FLRG01000005	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina megaterium EL33	-
TARA_ERS490285_N000146	490	blastn	GCF_900624725	NZ_UYXD01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina sp. Aq349	-
TARA_ERS490285_N000146	490	blastn	GCF_900299485	NZ_OMKB01000021	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina sp. Aq349	-
TARA_ERS490285_N000146	490	blastn	GCF_900089995	NZ_FLRG01000005	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina megaterium EL33	-
TARA_ERS490142_N000102	491	blastn	GCF_004364165	NZ_SOAY01000010	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Maribacter:Maribacter spongiicola DSM 25233	-

Table 2-2. Continued

TARA_ERS490142_N000102	491	blastn	GCF_000430665	NZ_KE387191	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina muelleri DSM 19832	-
TARA_ERS490142_N000102	491	blastn	GCF_003143755	NZ_QFRI01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Algibacter:Algibacter sp. ZY111	-
TARA_ERS490142_N000102	491	blastn	GCF_000799465	NZ_JUGU01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Psychroserpens:Psychroserpens sp. Hel_I_66	-
TARA_ERS490053_N000309	492	blastn	GCF_000766795	NZ_JPDS01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Polaribacter:Polaribacter sp. Hel1_85	-
TARA_ERS490053_N000309	492	blastn	GCF_001642835	NZ_LXEI01000001	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Tamlana:Tamlana agarivorans JW-26	-

Table 2-2. Continued

TARA_ERS490053_N000309	492	blastn	GCF_004310335	NZ_SIRS01000003	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Hyunsoonleella:Hyunsoonleella pacifica SW033	-
TARA_ERS490053_N000309	492	blastn	GCF_000520975	NZ_JACB01000005	RefSeq	Bacteroidetes:Flavobacteriia:Flavobacteriales:Flavobacteriaceae:Aquimarina:Aquimarina megaterium XH134	-
TARA_ERS490053_N000309	492	blastn	GCF_000520995	NZ_JACA01000049	RefSeq	Bacteroidetes; Flavobacteriia; Flavobacteriales;Flavobacteriaceae; Aquimarina.Aquimarina macrocephali JAMB N27	-
TARA_ERS490494_N000064	185	VHM	TARA_ANW_MAG_00076	-	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS488589_N000003	398	VHM	TARA_ASE_MAG_00029	-	Delmont e*t al.* 2018	Bacteroidetes;Sphingobacteriia;Sphingoles;Saprospiraceae;na;na	-
OBV_N00010	455	blastn	TARA_ANW_MAG_00082	TARA_ANW_MAG_00082_000000000427	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS490346_N000577	493	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000001219	Delmont et al. 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-

Table 2-2. Continued

TARA_ERS488673_N000208	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al*. 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS488836_N000045	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al*. 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS489943_N000203	504	blastn	TARA_ASE_MAG_00029	TARA_ASE_MAG_00029_000000000388	Delmont *et al.* 2018	Bacteroidetes;Sphingobacteriia;Sphingoles;Saprospiraceae;na;na	-
TARA_ERS490053_N000098	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS490120_N000192	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al*. 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS490346_N000191	504	blastn	TARA_ANW_MAG_00082	TARA_ANW_MAG_00082_000000000065	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS490953_N000029	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al*. 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS492198_N000066	504	blastn	TARA_ANW_MAG_00076	TARA_ANW_MAG_00076_000000000577	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	-
TARA_ERS488589_N001952	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-
TARA_ERS488589_N001952	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-

Table 2-2. Continued

TARA_ERS488701_N001850	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-
TARA_ERS488701_N001850	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-
TARA_ERS488757_N001037	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-
TARA_ERS488757_N001037	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS488813_N001860	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS488813_N001860	794	tRNA	TMED46	TMED46_22	Tully *et al*. 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS488836_N001537	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS488836_N001537	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS489084_N002225	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS489084_N002225	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-
TARA_ERS489148_N002107	794	tRNA	TMED217	TMED217_3	Tully *et al.* 2017	Bacteroidetes/ChlorobiGroup;Novel Class A;na;na;na	-

Table 2-2. Continued

TARA_ERS489148_N002107	794	tRNA	TMED46	TMED46_22	Tully *et al.* 2017	Bacteroidetes/Chlorobi Group;Novel Class A;na;na;na	-
TARA_ERS489943_N000229	185	VHM	TARA_PSE_MAG_00127	-	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;1G12;SHAN690;
TARA_ERS490494_N000064	185	VHM	TARA_PSE_MAG_00127	-	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;1G12;SHAN690;
TARA_ERS488589_N000003	398	VHM	TARA_PSE_MAG_00127	-	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;1G12;SHAN690;
TARA_ERS490346_N000037	405	blastn	TARA_ASE_MAG_00025	TARA_ASE_MAG_00025_000000001206	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;Cryomorphaceae;;
TARA_ERS490452_N000394	471	blastn	TARA_PSE_MAG_00145	TARA_PSE_MAG_00145_000000000397	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;Flavobacteriaceae;MS024-2A;
TARA_ERS490053_N000309	492	blastn	TOBG_SP-3040	TOBG_SP-3040_11	Tully *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;novel Genus_F;null	Bacteroidota;Bacteroidia;o__Flavobacteriales;Flavobacteriaceae;GCA-2719315;GCA_002719315.1

Table 2-2. Continued

TARA_ERS490053_N000309	492	blastn	TOBG_EAC-674	TOBG_EAC-674_26	Tully *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;novelGenus_G;null	Bacteroidota;Bacteroidia;o__Flavobacteriales;Flavobacteriaceae;MS024-2A;GCA_002695445.1
TARA_ERS490053_N000309	492	blastn	TOBG_CPC-288	TOBG_CPC-288_2	Tully *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;novelGenus_H;null	Bacteroidota;Bacteroidia;o__Flavobacteriales;Flavobacteriaceae;MS024-2A;GCA_002705385.1
OBV_N00024	506	blastn	TARA_PSE_MAG_00127	TARA_PSE_MAG_00127_000000000260	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;1G12;SHAN690;
TARA_ERS488589_N000065	506	blastn	TMED12	TMED12_37	Tully *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;na;na	Bacteroidota;Bacteroidia;o__Flavobacteriales;Flavobacteriaceae;Muricauda;GCA_002167435.1
TARA_ERS488757_N000013	515	blastn	TARA_ASE_MAG_00025	TARA_ASE_MAG_00025_000000001175	Delmont *et al.* 2018	Bacteroidetes;Flavobacteriia;Flavobacteriales;na;na;na	;Bacteroidota;Bacteroidia;o__Flavobacteriales;Cryomorphaceae;;

(ii) The Bacteroidetes homologs should account for at least 18.8% or 38.9% of cellular homologs in RefSeq or TARA-MAGs, respectively, and (iii) At least 5 or 3 viral genes should be Bacteroidetes homologs in RefSeq or TARA-MAGs, respectively. Each threshold corresponds to the maximum value observed for non-Bacteroidetes viruses.

By applying these criteria to 1,811 EVGs, I identified 311 EVGs as putative Bacteroidetes viruses (**Figures 2-1C**, **2-1D**, **2-1E**, **2-1F**). All of the 239 EVGs that were previously described as members of *Flavobacteriaceae* group 1 and 2 (Nishimura, Watai, *et al.*, 2017) were included in these putative Bacteroidetes EVGs. Seventy-two EVGs were newly predicted as Bacteroidetes viruses.

Classification of Bacteroidetes EVGs and their genomic features

There are 21 overlaps between the Bacteroidetes EVGs predicted based on nucleotide similarity-based methods and protein homology-based method (**Figure 2-4**). Ten EVGs were only predicted by nucleotide similarity-based methods using MAGs and 290 EVGs were only predicted by protein homology-based method (**Figure 2-5**). In total, I identified 321 EVGs as putative Bacteroidetes EVGs including 81 EVGs which were not predicted as their host in previous studies. The 321 EVGs were classified into 29 gOTUs based on their genomic similarity (Nishimura et al., 2017a, **Figure 2-6**). In the following sections, I describe the genomic features of 81 EVGs, which are the newly identified putative Bacteroidetes viruses.

Novel Sub-Clade of *Flavobacteriaceae* EVG Group 1

Twenty-four EVGs of two gOTUs (G493 and G494) were located near the branches of the previously described *Flavobacteriaceae* EVGs group 1 in the viral proteomic tree (**Figure 2-6**). These EVGs were 27.5 kb to 50.5 kb with an average G+C content of 32.6% (**Table 2-3**). Putative viral structural protein genes (major capsid, prohead protease,

terminase, and portal) and putative DNA replication genes were well conserved within the viral group. Genome synteny of the tail like structure such as putative endosialidase tail spikes were also conserved but exhibited low sequence homology within the group (**Figure 2-7A**). They also shared portal gene homologs conserved in the members of the group 1 (**Figure 2-7A**). Therefore, I concluded that the twenty-four EVGs are new members of the subclade of *Flavobacteriaceae* EVGs group 1.

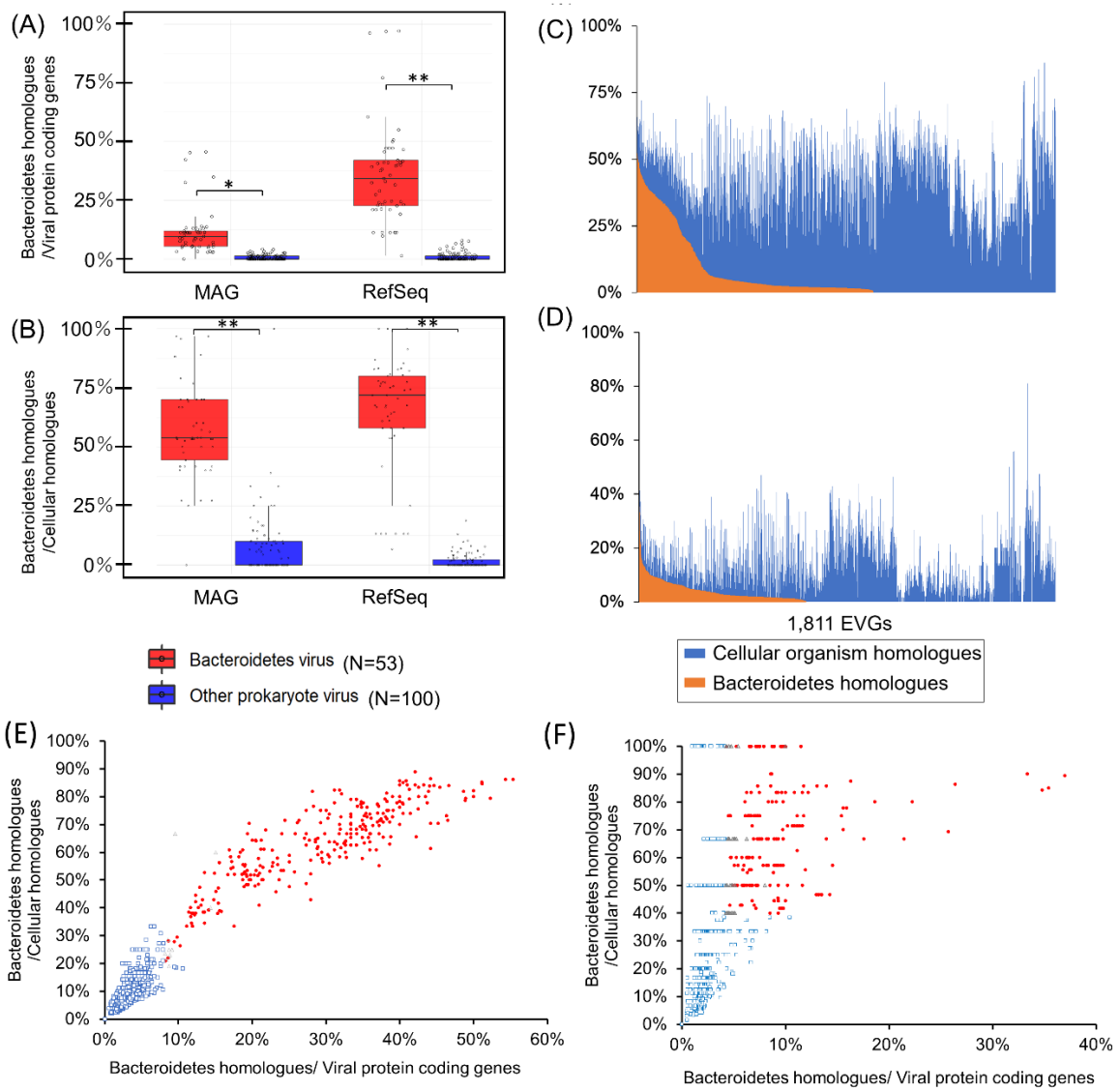

Figure 2-1. Proportion of the Bacteroidetes homologs in viral genomes.

Figure 2-1. continued.

(A) Proportion of the Bacteroidetes homologs among protein-coding genes. (B) Proportion of the Bacteroidetes homologs among cellular organism homologs. The proportions on the right are those calculated from NCBI RefSeq and those on the left are calculated from TARA MAGs. Red and blue boxes represent cultured Bacteroidetes viruses and cultured viruses infecting other prokaryotes, respectively. The boxes represent the first quartile, median, and third quartile. Asterisks denote significance (Mann–Whitney U test , *P < 0.05, **P < 0.001). (C) Proportion of the Bacteroidetes homologs in RefSeq (orange) and other cellular organism homologs in RefSeq (blue) of the 1,811 EVGs. (D) Proportion of the Bacteroidetes homologs in MAGs (orange) and other cellular organism homologs in TARA MAGs (blue) of the 1,811 EVGs. Scatter plots showing the proportion Bacteroidetes homologs among protein-coding genes (x-axis) and among cellular organism homologs (y axis) for the comparison against RefSeq (E) and TARA-MAG (F). Viruses passing the cut off values for the prediction of Bacteroidetes EVGs are shown in red circles. Viruses passing the two criteria (i.e., (i) at least 7.9% or 4.2 % of genes should be homologs of Bacteroidetes genes in RefSeq or TARA-MAGs, respectively; (ii) the Bacteroidetes homologs should account for at least 18.8% or 38.9% of cellular homologs in RefSeq or TARA-MAGs, respectively) but have only few Bacteroidetes homologs (RefSeq: homolog < 5 genes, TARA MAGs: homolog < 3 genes) are shown in gray triangles. Other viruses that did not pass the cut off values are shown in blue squares.

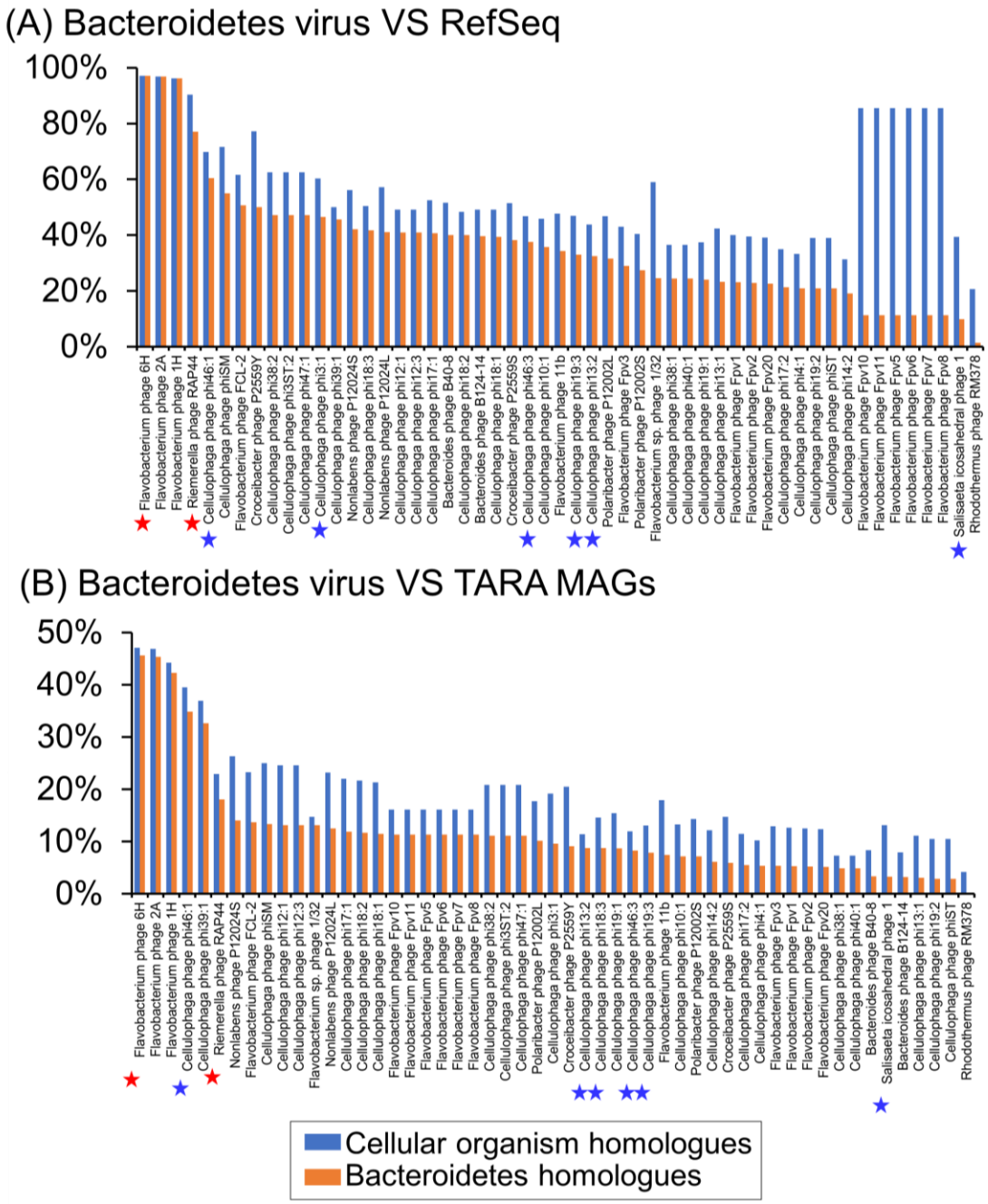

Figure 2-2. Proportion of the Bacteroidetes homologs in cultivated Bacteroidetes viruses.

(A) Proportion of the Bacteroidetes homologs in RefSeq (orange) and cellular organism homologs in RefSeq (blue) of the 53 dsDNA cultivated Bacteroidetes viruses. (B) Proportion of the Bacteroidetes homologs in TARA MAGs (orange) and cellular organism homologs in TARA MAGs (blue) of the 53 dsDNA cultivated Bacteroidetes viruses. Red and blue stars represent the viruses with a lysogenic life cycle and viruses having putative integrase homologs but not reported lysogenic life cycle, respectively.

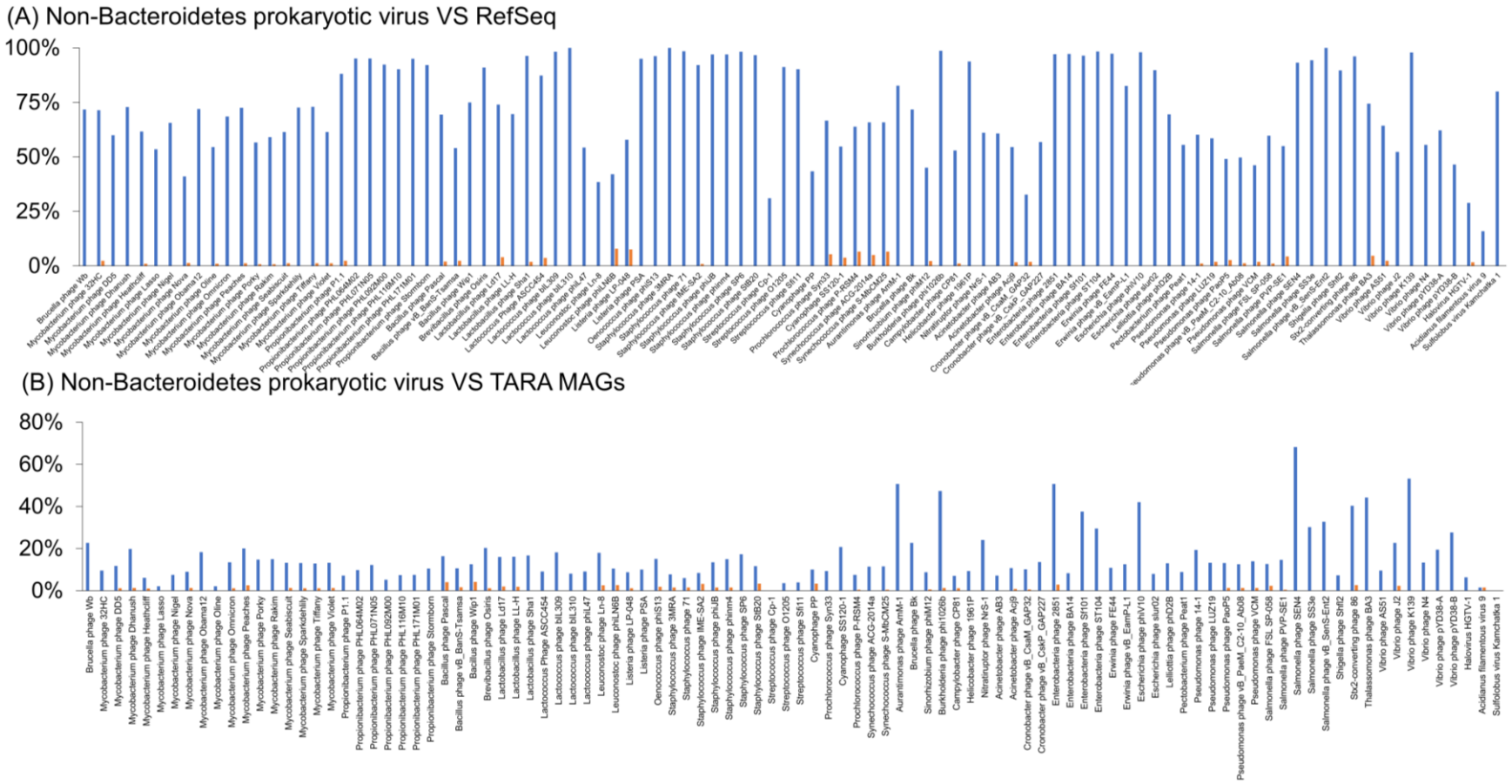

Figure 2-3. Proportion of the Bacteroidetes homologs in cultivated non-Bacteroidetes viruses.

(A) Proportion of the Bacteroidetes homologs in RefSeq (orange) and cellular organism homologs in RefSeq (blue) of the randomly selected 100 prokaryotic viruses. (B) Proportion of the Bacteroidetes homologs in TARA MAGs (orange) and cellular organism homologs in TARA MAGs (blue) of the randomly selected 100 prokaryotic viruses.

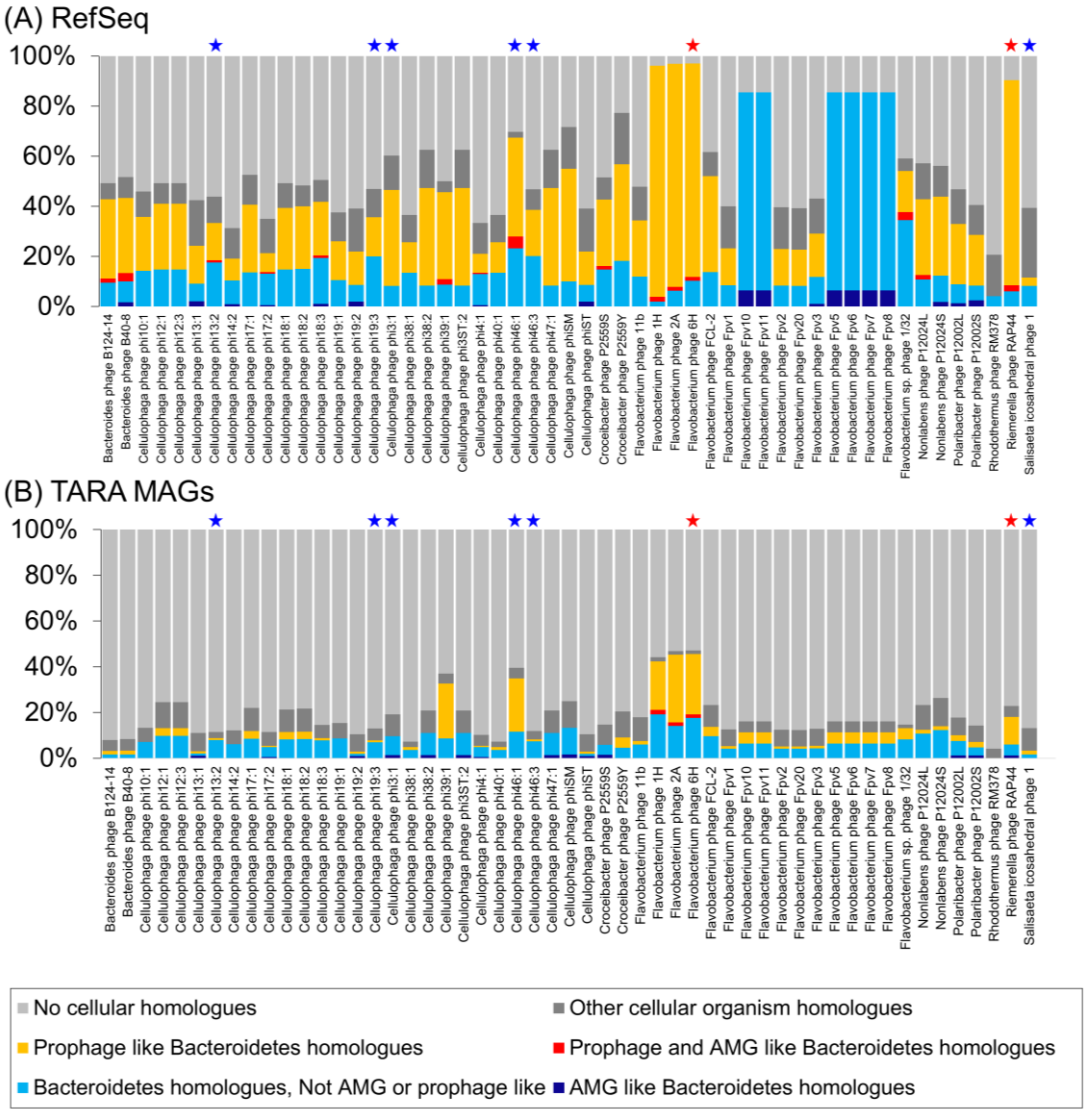

Figure 2-4. Proportion of the putative AMGs and provirus homologs in the viral Bacteroidetes homologs.

(A) Homologs of RefSeq cellular organism genomes. (B) Homologs of TARA MAGs. Proviruses were detected by VirSorter (Roux *et al*., 2015). List of Pfam domains found in putative AMGs like genes followed the lists in Roux *et al*., 2016 and Ruo *et al*., 2018. Red and blue stars represent the viruses with a lysogenic life cycle and viruses having putative integrase homologs but not a lysogenic life cycle, respectively.

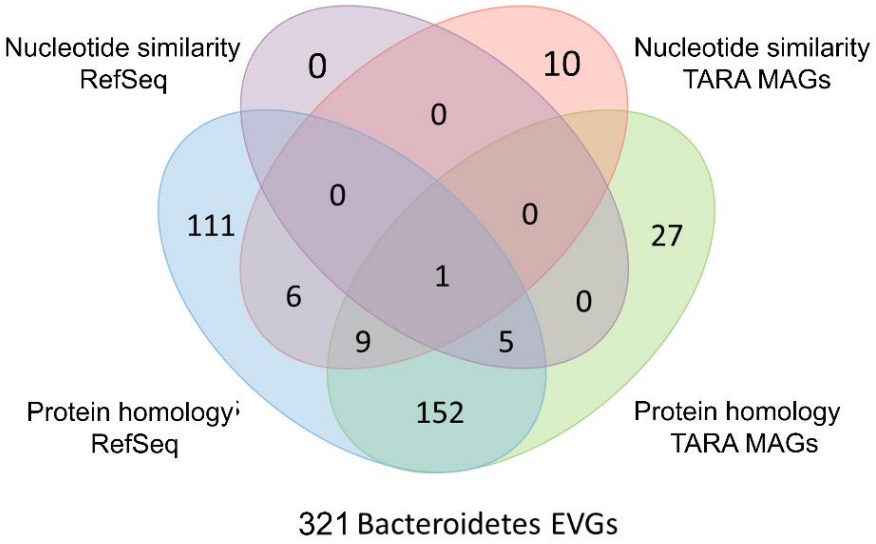

Figure 2-5. Venn diagram of the Bacteroidetes EVGs detected by the four host prediction methods

***Flavobacteriaceae* EVGs group 3**

I detected a novel group (group 3) of putative marine *Flavobacteriaceae* viral genomes (**Figure 2-6B**). This group composed of 10 EVGs classified into 6 gOTUs and 19 cultured Bacteroidetes virus genomes. The 10 EVGs ranged in size from 32 kb to 44 kb with a G+C content ranging from 32.6% to 42% (**Table 2-3**). The EVGs shared 2.8% to 30.4% of genes (two to seven genes) with the cultured members of the group 3. For example, TARA_ERS492198_N000180 (G537) and TARA_ERS490204_N000278 (G536) shared 17 and 8 genes with *Cellulophaga* siphovirus phi19:1, respectively (**Figure 2-7B**). Most of the shared genes are annotated as structural protein genes such as capsid and tail tape measure (**Figure 2-7B**). However, the EVGs rarely shared genes with *Cellulophaga* siphovirus phi10:1, which show genus level similarity with phi19:1 (**Figure 2-7B**). Similarly, within the group 3, LDNO01000008 and Flavobacterium virus 11b shared several structural protein homologs with phi10:1 but not with phi19:1 or the members of G537 and G536 (**Figure 2-7B**).

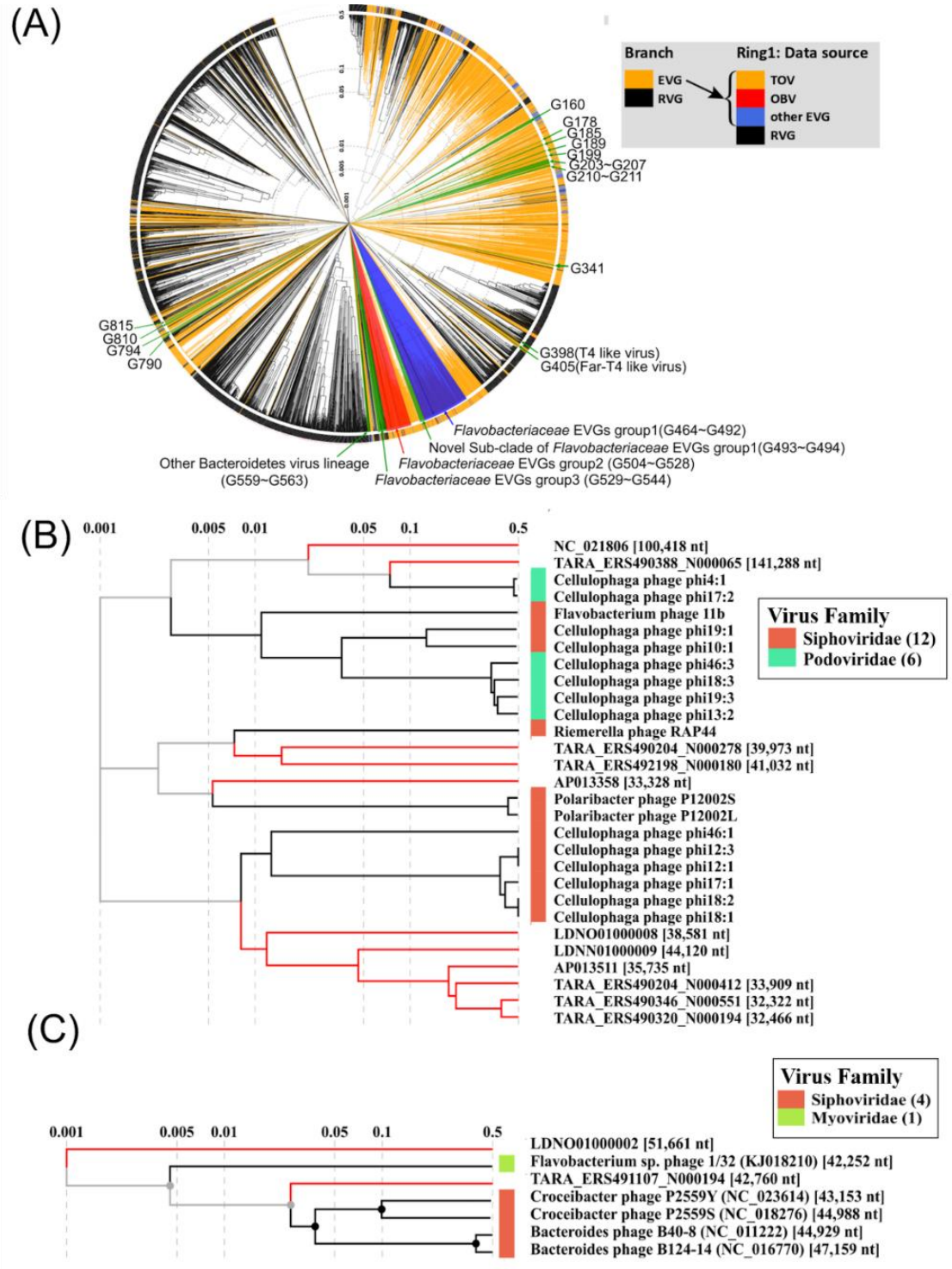

Figure 2-6. Proteomic tree representation of the Bacteroidetes EVGs with cultured Bacteroidetes viral genomes.

Figure 2-6. continued

The dendrograms represent proteome-wide similarity relationships. (A) A proteomic tree of 1,811 EVGs (orange branches) and 2,429 cultured viruses (black branches) calculated in Nishimura et al 2017a with highlighting of newly detected Bacteroidetes EVGs (green), *Flavobacteriaceae* EVGs group1 (blue), and *Flavobacteriaceae* EVGs group 2 (red). The tree is midpoint rooted. Branch lengths are indicated using a logarithmic scale. (B) A part of the proteomic tree with *Flavobacteriaceae* EVGs group 3 (red branches) and their relatives of cultured viruses (black branches). (C) A part of the proteomic tree with 2 EVGs (red branches) with cultured Bacteroides and *Flavobacteriaceae* viruses (black branches). Rings outside the dendrogram represent taxonomic groups of viral family classifications.

Other Bacteroidetes viral lineages

I identified two other EVGs (TARA_ERS491107_N000194 and LDNO01000002) positioned within a clade of the proteomic tree exclusively composed of Bacteroidetes viruses infecting members of *Flavobacteriaceae* and *Bacteroides* (**Figure 2-7C**). The two EVGs had 42 kb and 51 kb genomes with a G+C content of 32.6% and 49.2%, respectively (**Table 2-3**). TARA_ERS491107_N000194 shared a maximum 21% of the genes (17 genes) including putative capsid protein genes and phage tail protein gene with the cultured members of this group. Similarly, LDNO01000002 shared maximum 5% of the genes (two genes) such as putative terminase-like protein with the members of this group but did not share any genes with TARA_ERS491107_N000194.

Table 2-3. General genomic features of the Bacteroidetes gOTUs identified in this study.

Group (gOTU)	No of EVGs	No. of EVGs predicted as Bacteroidetes EVG	Ave. length (bp)	Ave. GC%	Ave. RefSeq Bacteroidetes homologue in EVG (%)	Ave. MAG Bacteroidetes homologue in EVG (%)	Classified group
G160	13	9	37,551	38.5	2.2	11.1	-
G178	1	1	40,754	32.6	11.4	0	-
G185	4	2	54,812	31.7	1.9	0.5	-
G189	3	1	58,769	35.4	5.3	3.7	-
G199	2	1	36,245	35.8	5.9	2.5	-
G203	5	2	31,173	30.7	5.7	7.0	-
G204	3	3	32,490	32.4	4.4	6.4	-
G205	2	2	27,613	33.8	3.5	15.1	-
G206	4	3	27,672	35.8	3.2	7.4	-
G207	3	1	31,013	33.2	4.7	4.7	-
G210	8	5	34,852	38.2	7.5	4.6	-
G211	1	1	34,002	34.9	7.8	9.8	-
G341	1	1	39,514	39.3	10.2	10.2	-
G398	1	1	179,949	32.0	6.3	0.4	T4 like
G405	1	1	143,709	33.4	8.5	7.3	Far-T4 like

Table 2-3. Continued

G493	21	21	32,686	33.5	31.8	5.3	Novel sub-clade of *Flavobacteriaceae* group 1
G494	3	3	31,174	31.7	22.8	8.0	
G535	1	1	33,328	30.5	36.0	4.0	
G536	1	1	39,973	35.3	28.6	7.1	
G537	1	1	41,032	42.0	55.4	21.4	*Flavobacteriaceae* EVGs group 3
G541	4	4	33,608	40.6	43.6	35.1	
G542	1	1	44,120	33.1	36.1	22.2	
G544	1	1	38,581	32.6	44.1	8.5	
G561	1	1	42,760	32.6	25.8	1.6	Bacteroidetes viral lineage
G563	1	1	51,661	49.2	3.3	4.9	
G790	1	1	58,364	33.9	34.7	26.4	
G794	9	7	12,003	31.2	0	0	-
G810	3	1	43,470	46.8	2.0	2.3	-
G815	3	3	32,908	39.6	28.6	30.8	-

(A)

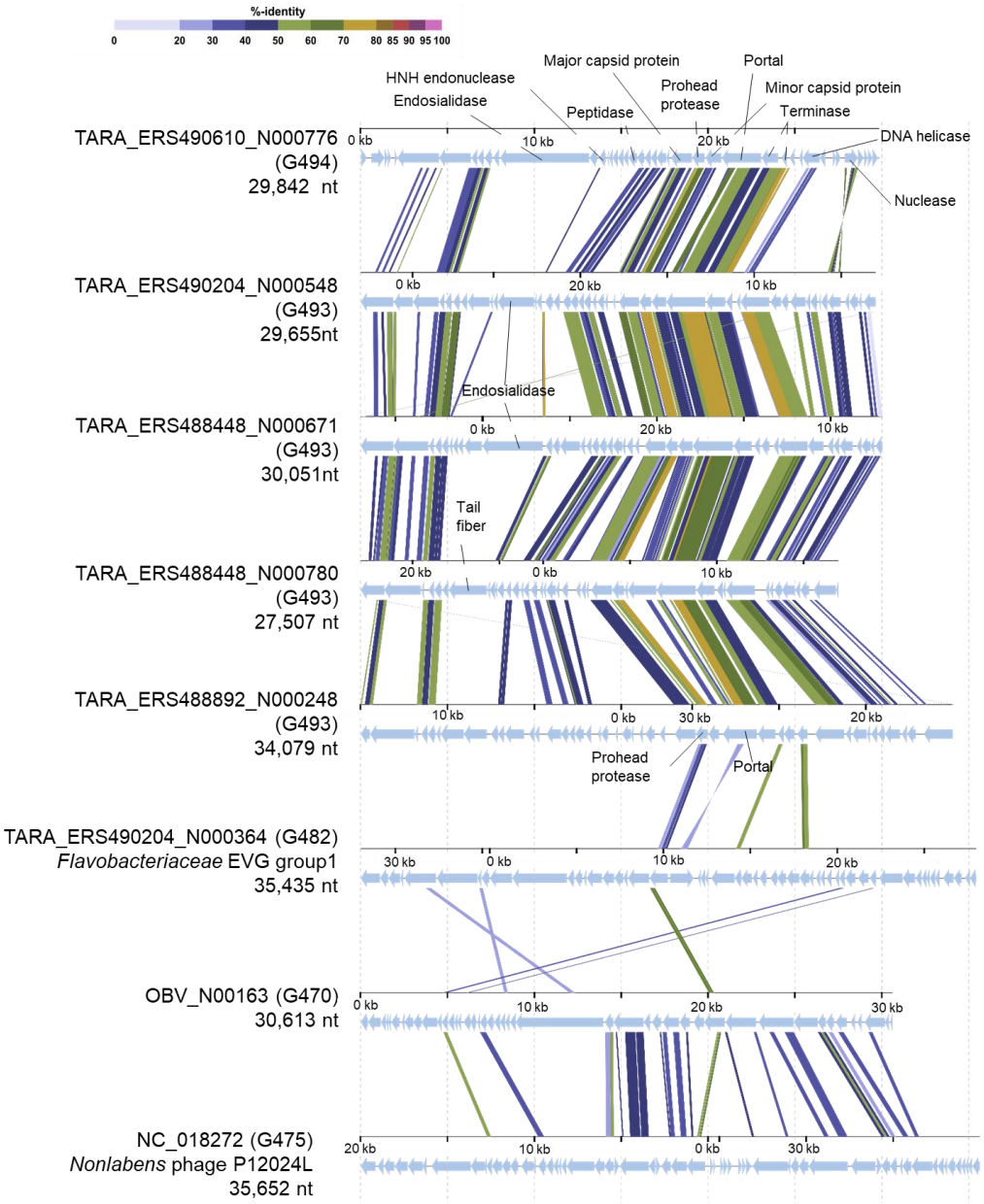

Figure 2-7. Bacteroidetes EVGs shared genomic features with cultured Bacteroidetes genomes.

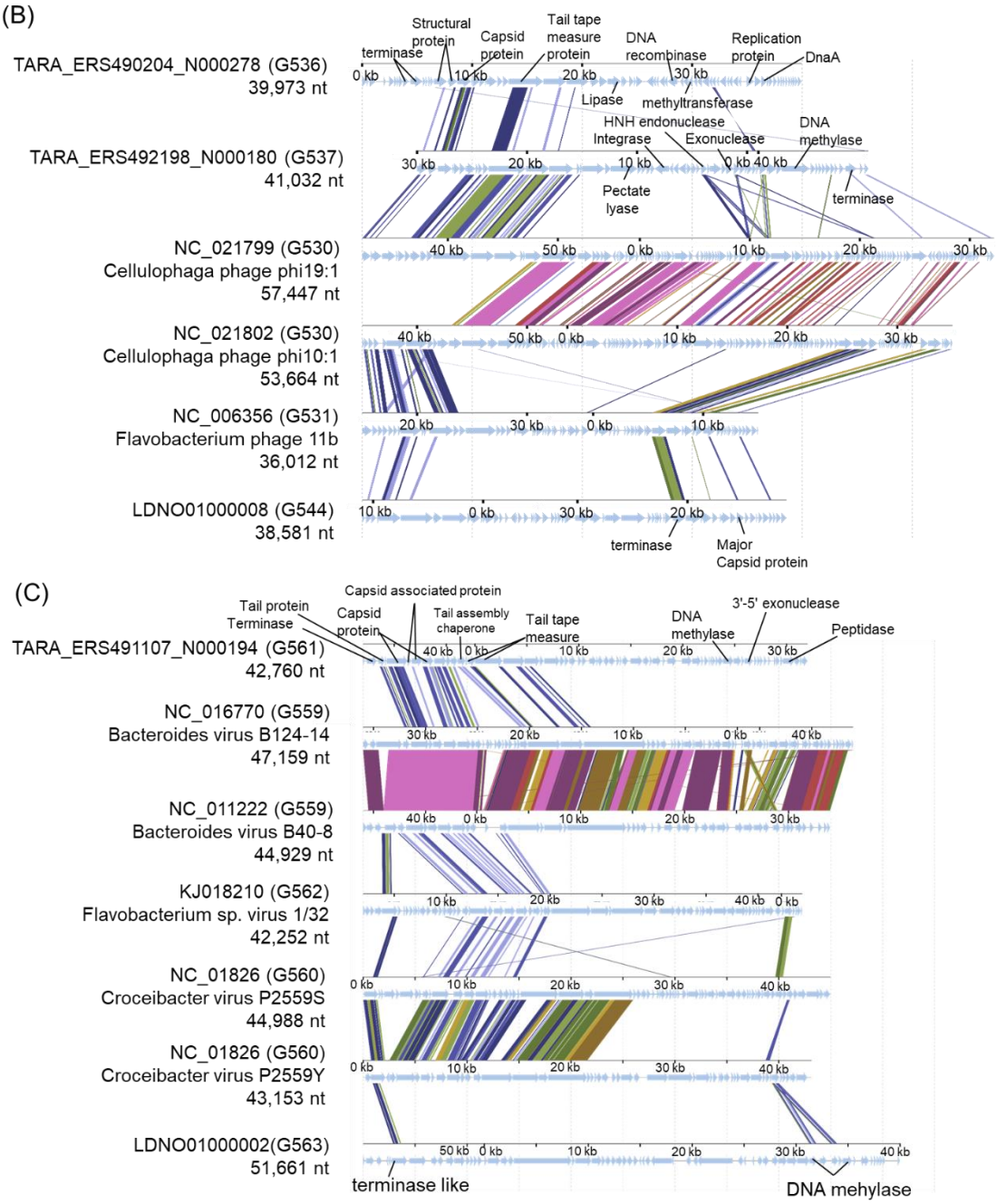

Figure 2-7. continued.

(A) A genome map of members of the G493 and G494 with other members of the *Flavobacteriaceae* EVGs group 1. (B) A genome map of members of *Flavobacteriaceae* EVGs group 3. (C) A genome map of members of G561 and G563 with *Bacteroides* and *Flavobacteriaceae* viruses. The sequences are circularly permuted and/or reversed. The sequences are circularly permuted and/or reversed for clarity. Putative gene functions are indicated. All tBLASTx alignments are represented as colored lines between the two genomes. The color scale represents tBLASTx percent identity.

T4 like viruses

I identified two EVGs which exhibited genome characteristics of the T4-like superfamily (*Tevenvirinae*), which is one of the most widespread, abundant, and extensively studied viral groups. This is the first report of T4-like viruses infecting marine Bacteroidetes excluding a virus infecting thermophilic Bacteroidetes *Rhodothermus marinus*. *Tevenvirinae* appears to be comprised of several subgroups including (i) the "true" T-evens represented by T4 and closely related viruses infecting Enterobacteria, (ii) the Pseudo and Schizo T-evens (including Aeromonas and Vibrio viruses), (iii) the Exo T-evens (including cyano- and SAR11 viruses), and Far-T4-like virus, which includes the sole isolate RM378 infecting a thermophilic Bacteroidetes *Rhodothermus marinus* (Petrov *et al.*, 2010). TARA_ERS490346_N000037 (G405), 143 kb in size with a G+C content of 33.4% (**Table 2-3**) was found to be most similar to the Far-T4-like virus RM378 among the cultured viruses as they shared 26 genes (**Figure 2-9A**). Phylogenetic tree of the major capsid protein (T4 phage gene 23) suggests that TARA_ERS490346_N000037 is a novel member of Far-T4 like viruses (**Figure 2-8**). This EVG is the first representative of complete genomes from environmental Far-T4 like virus with *in silico* identification of putative host groups. The EVGs have up to 66 genes mostly annotated as structural proteins and replication proteins shared with the Far-T4 genome fragments assembled from the freshwater viromes (**Figure 2-9A**) (Roux, François Enault, *et al.*, 2015).

TARA_ERS488589_N000003 (G398) was observed to be most similar to the marine Exo-T4 like viruses infecting *Pelagibacter* and unicellular Cyanobacteria (**Figure 2-9B**). This EVG has a 180 kb genome and G+C content is slightly lower (32%) than the known T4-like viruses (**Table 2-3**). Twenty two of the 38 core genes conserved in the

T4-like virus genomes as shown in a previous comparative genomics study (**Figure 2-9B**) (Sullivan *et al.*, 2010).

As reported in the other T4 like viruses, these T4 like EVGs encoded putative auxiliary metabolic genes (**Table 2-4**). For example, the TARA_ERS488589_N000003 has queuosine (Que) biosynthesis pathway genes (gene109 (*queF*), gene162 (*queE*), gene164 (*queD*), and gene66 (GTP cyclohydrolase)). Que biosynthesis genes were reported in two cultured *Cellulophaga* viruses (Holmfeldt *et al.*, 2013) and I found them in members of the *Flavobacteriaceae* group 1 and group 2 (**Table 2-4**). Similarly, both EVGs encode proteins putatively related to carbohydrate metabolism (**Table 2-4**). Additionally, I found that the TARA_ERS488589_N000003 encodes proteins putatively related to two cell-surface adhesion systems (curli biosynthesis (gene_61: *csrA*, gene_62: *csrG*, and gene_63; *csrF*)) and ubiquitous surface proteins (gene_52 and gene_70, **Table 2-4**) mostly found in pathogenic bacteria (Barnhart and Chapman, 2006; Tan *et al.*, 2006).

Figure 2-8. An approximately maximum likelihood phylogenetic tree computed from the multiple alignment of Gp23 (major capsid protein) of TARA_ERS490346_N000037 (G405) and T4-like superfamily viruses.

The protein sequences were collected from RefSeq and Lake Pavin viromes (Roux, François Enault, *et al.*, 2015). Circles indicate nodes with bootstraps higher than 0.9.

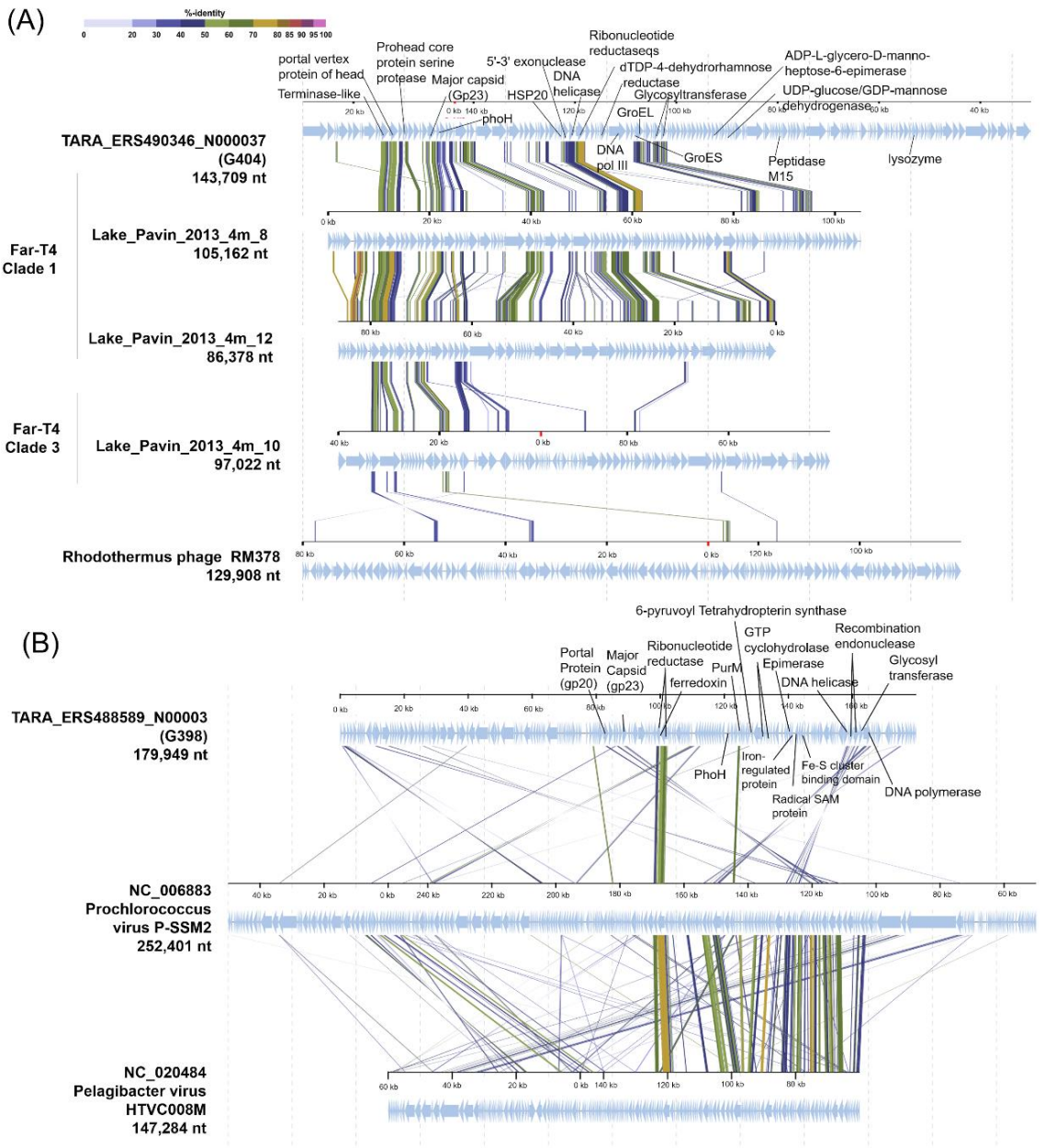

Figure 2-9. Bacteroidetes EVGs shared genomic features with T4-like super family. (A) A genome map of TARA_ERS490346_N000037 (G405), Far-T4 contigs assembled in Roux et al. 2015 and *Rhodothermus marinus* virus RM378 (B) A genome map of TARA_ERS488589_N000003 (G398) and Exo-T4 viruses. The sequences are circularly permuted and/or reversed for clarity. Putative gene functions are indicated. All tBLASTx alignments are represented as colored lines between the two genomes. The color scale represents tBLASTx percent identity.

Table 2-4. List of eggNOG and PFAM domains annotation of the putative AMGs found in Bacteroidetes EVGs

EVGs	Viral group (gOTU)	Gene	Putative function	PFAM ID	Homologues with RefSeq Bacteroidetes genomes	Homologues with Bacteroidetes MAGs
TARA_ERS488701_N000192	341	gene_10	UDP-N-acetylglucosamine 2-epimerase	PF02350.19	Yes	Yes
		gene_33	sulfate reduction	PF06508.13	Yes	-
		gene_7	D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain	PF02826.19 PF02737.18 PF03446.15	Yes	-
TARA_ERS488589_N000003	398	gene_1	RmlD substrate binding domain	PF01370.21 PF04321.17	-	-
		gene_109	GTP cyclohydrolase I family	PF14489.6 PF14819.6	Yes	-
		gene_138	Cytidylyltransferase-like	PF01467.26	-	-
		gene_139		PF03016.15	-	-
		gene_140	Male sterility protein	PF01073.19 PF01370.21 PF16363.5 PF04321.17	-	-
		gene_145	Exostosin family	PF03016.15	-	-

Table 2-4. Continued

TARA_ERS488589_N000003	398	gene_159	spermidine synthase activity	PF03602.15	-	-
		gene_160	S-adenosylmethionine decarboxylase	PF02675.15	-	-
		gene_161	Catalyzes the decarboxylation of S-adenosylmethionine to S-adenosylmethioninamine (dcAdoMet)the propylamine donor required for the synthesis of the polyamines spermine and spermidine from the diamine putrescine	PF02675.15	-	-
		gene_164	6-pyruvoyl tetrahydropterin synthase	PF01242.19	-	-
		gene_166	GTP cyclohydrolase activity	PF02649.14	-	-
		gene_176	GlcNAc-PI de-N-acetylase	PF02585.17	-	-
		gene_177	Epimerase dehydratase	PF01073.19 PF02737.18 PF01370.21 PF16363.5 PF04321.17 PF03721.14 PF00106.25	-	-

Table 2-4. Continued

TARA_ERS488589_N000003	398	gene_178	2OG-Fe(II) oxygenase superfamily	PF13640.6	-	-
		gene_182	Aminotransferase class I and II	PF00155.21	-	-
		gene_207	Catalyzes the synthesis of activated sulfate	PF01467.26	-	-
		gene_4	Belongs to the NAD(P)-dependent epimerase dehydratase family. dTDP-glucose dehydratase subfamily	PF01073.19 PF01370.21 PF16363.5 PF04321.17	-	-
		gene_51	Sulfotransferase domain	PF00685.27 PF13469.6	-	-
		gene_66	Protein of unknown function (DUF3307)	PF05138.12	-	-
TARA_ERS490346_N000037	405	gene_105	GDP-mannose dehydrogenase	PF00984.19 PF03721.14	-	-
		gene_106	biosynthetic process	PF01467.26	-	-
		gene_110	PFAM NAD-dependent epimerase dehydratase	PF01073.19 PF01370.21 PF16363.5 PF04321.17	-	-
		gene_112	Oxidoreductase family, NAD-binding Rossmann fold	PF01408.22	-	-
		gene_120	Spore maturation protein CgeB	PF01118.24	-	-

Table 2-4. Continued

TARA_ERS490346_N000037	405	gene_124	Glycosyltransferase like family 2	PF00534.20	-	-
		gene_125	PFAM Glycosyl transferase, group 1	PF00534.20	-	-
		gene_136	Catalyzes the reduction of dTDP-6-deoxy-L-lyxo-4- hexulose to yield dTDP-L-rhamnose	PF01370.21 PF16363.5 PF04321.17	Yes	
TARA_ERS488929_N000326	464	gene_19	GTP cyclohydrolase I activity	PF01227.22 PF14489.6	-	-
TARA_ERS488929_N000326	464	gene_21	synthase	PF01242.19	-	-
TARA_ERS489943_N000539	464	gene_27	synthase	PF01242.19	-	-
TARA_ERS489943_N000539	464	gene_30	Queuosine biosynthesis protein QueC	PF00733.21 PF02540.17 PF06508.13	Yes	Yes
		gene_31	Catalyzes the NADPH-dependent reduction of 7-cyano-7- deazaguanine (preQ0) to 7-aminomethyl-7-deazaguanine (preQ1)	PF01227.22 PF14489.6 PF14819.6	-	-
TARA_ERS491107_N000346	465	gene_12	PFAM GTP cyclohydrolase I	PF01227.22 PF14489.6	-	-
		gene_14	synthase	PF01242.19	-	-

Table 2-4. Continued

TARA_ERS491107_N000346	465	gene_15	Queuosine biosynthesis protein QueC	PF00733.21 PF02540.17 PF06508.13	Yes	Yes
TARA_ERS490320_N000240	468	gene_44	GDP-mannose 4,6 dehydratase	PF01073.19 PF01370.21 PF16363.5 PF04321.17 PF00106.25	-	-
TARA_ERS490953_N000263	468	gene_27		PF01126.20	Yes	-
TARA_ERS478007_N000306	469	gene_28	Psort location Cytoplasmic, score 8.96	PF01073.19 PF01073.19 PF01370.21 PF16363.5 PF04321.17	Yes	-
TARA_ERS490953_N000223	469	gene_42	Phosphoadenosine phosphosulfate reductase	PF02540.17 PF06508.13	-	-
AP013515	470	gene_22		PF00116.20	Yes	-
TARA_ERS478007_N000181	470	gene_21	Phosphoadenosine phosphosulfate reductase	PF02540.17 PF06508.13 PF01227.22 PF14489.6	-	-

Table 2-4. Continued

TARA_ERS478052_N000461	470	gene_19	protein of Mannheimia haemolytica PHL213 UniRef RepID A7JSI2_PASHA	PF01503.17	-	-
TARA_ERS488558_N000828	470	gene_28		PF00116.20	Yes	-
TARA_ERS488558_N000869	470	gene_10	Putative phage serine protease XkdF	PF02668.16	Yes	-
TARA_ERS488813_N000314	470	gene_15	protein disulfide oxidoreductase activity	PF01503.17	-	-
TARA_ERS488836_N000239	470	gene_42	protein disulfide oxidoreductase activity	PF01503.17	-	-
TARA_ERS488929_N000303	470	gene_33	Prolyl 4-hydroxylase alpha subunit homologues.	PF13640.6	-	-
TARA_ERS490953_N000167	470	gene_43	synthase	PF01242.19	-	-
TARA_ERS490953_N000167	470	gene_45	GTP cyclohydrolase I activity	PF01227.22 PF14489.6	-	-
TARA_ERS490346_N000483	472	gene_45	Pyridoxal-phosphate dependent enzyme	PF00291.25	-	-
TARA_ERS488340_N000863	474	gene_26	von willebrand factor, type A	PF09206.11	Yes	Yes
TARA_ERS488813_N000313	478	gene_30	phosphoadenosine phosphosulfate	PF06508.13	-	-
TARA_ERS488448_N000694	481	gene_5	Prolyl 4-hydroxylase alpha subunit homologues.	PF03171.20 PF13640.6	-	-

Table 2-4. Continued

TARA_ERS488836_N000172	482	gene_11	6-pyruvoyl tetrahydropterin synthase	PF01242.19	-	-
		gene_12	Queuosine biosynthesis protein QueC	PF02540.17 PF06508.13	-	-
		gene_9	gtp cyclohydrolase	PF01227.22	-	-
TARA_ERS490204_N000364	482	gene_18	catalyzes the formation of formate and 2-amino-4-hydroxy-6-(erythro-1,2, 3-trihydroxypropyl) dihydropteridine triphosphate from GTP and water	PF01227.22 PF14489.6	-	-
		gene_21	synthase	PF01242.19	-	-
TARA_ERS490204_N000364	482	gene_31		PF10014.9	-	-
TARA_ERS488448_N000283	483	gene_5	GTP cyclohydrolase I activity	PF01227.22 PF14489.6	-	-
		gene_7	6-pyruvoyl tetrahydrobiopterin synthase	PF01242.19	Yes	Yes
		gene_8	Catalyzes the ATP-dependent conversion of 7-carboxy-7- deazaguanine (CDG) to 7-cyano-7-deazaguanine (preQ(0))	PF00733.21 PF02540.17 PF06508.13	Yes	

Table 2-4. Continued

OBV_N00135	484	gene_2	GTP cyclohydrolase	PF01227.22 PF14489.6	-	-
OBV_N00135	483	gene_43	synthase	PF01242.19	Yes	-
TARA_ERS490346_N000441	487	gene_25	GTP cyclohydrolase I activity	PF01053.20 PF01227.22 PF14489.6	-	-
TARA_ERS490346_N000441	487	gene_28	synthase	PF01242.19	-	-
TARA_ERS490285_N000146	490	gene_16	Nitrogen regulatory protein P-II	PF00543.22	Yes	Yes
TARA_ERS490320_N000023	490	gene_46	Nitrogen regulatory protein P-II	PF00543.22	Yes	Yes
TARA_ERS490142_N000102	491	gene_50	Belongs to the P(II) protein family	PF00543.22	Yes	Yes
TARA_ERS490053_N000309	492	gene_48	Belongs to the P(II) protein family	PF00543.22	Yes	Yes
TARA_ERS488448_N000780	493	gene_25	domain protein	PF00122.20	Yes	-
TARA_ERS488448_N000127	504	gene_20	COG1705 Muramidase (flagellum-specific)	PF01832.20	Yes	Yes
TARA_ERS488892_N000066	504	gene_11		PF00016.20	Yes	-
TARA_ERS492198_N000080	504	gene_66		PF00016.20	Yes	-
OBV_N00025	506	gene_64		PF14489.6	Yes	Yes

Table 2-4. Continued

TARA_ERS488757_N000031	509	gene_39	Ribonucleotide reductase, small chain	PF02915.17	Yes	Yes
TARA_ERS488340_N000358	510	gene_5		PF13640.6 PF13661.6	-	-
		gene_62	5-dioxygenase	PF00534.20	-	-
TARA_ERS488354_N000081	510	gene_5		PF13640.6 PF13661.6	-	-
		gene_62	5-dioxygenase	PF00534.20	-	-
TARA_ERS489285_N000130	514	gene_42		PF13640.6 PF13661.6	-	-
LDNP01000001	519	gene_61	mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	PF01832.20	Yes	-
TARA_ERS490346_N000066	523	gene_14	Ribonucleotide Reductase	PF02915.17	Yes	Yes
TARA_ERS492160_N000078	523	gene_24	Ribonucleotide Reductase	PF02915.17	-	Yes
TARA_ERS488813_N000030	524	gene_46	NAD-dependent epimerase dehydratase	PF01073.19 PF01370.21 PF16363.5 PF04321.17 PF00106.25	-	-

Other new lineages distant from the cultured viruses

The remaining 44 EVGs classified into 17 gOTUs were 12–59 kb in size with a G+C content of 31% to 47% (**Table 2-3**). They were distributed in twelve clades in the viral proteomic tree exclusively composed of EVGs (Nishimura, Watai, *et al.*, 2017). Following the previous classification of 2,429 cultured prokaryotic viral genomes, gOTUs classification based on genomic similarity reflected the phylum-level host taxonomy with only two exceptions (Nishimura, Watai, *et al.*, 2017). This suggests that the 22 EVGs, which were not predicted as Bacteroidetes viruses by the *in-silico* virus-host prediction employed in the present study but were classified into the same gOTUs as Bacteroidetes EVGs, are also likely to be candidates of Bacteroidetes viruses (**Table 2-3**). Most of the predicted genes (71–94%) of these uncultured clades were functionally annotated as hypothetical proteins, as is common for environmental viruses (Seguritan *et al.*, 2003; Borriss *et al.*, 2007). The predicted functions/categories of the annotated genes were DNA metabolism (48%, the values provided here are averages), viral structural genes (21%), and host lysis (15%).

Abundance and distribution of the Bacteroidetes EVGs

Abundance and distribution of the Bacteroidetes EVGs in the global ocean were investigated by read recruitment of the *Tara* Oceans viromes, which consist of 43 viromes representing 26 oceanic locations (Brum *et al.*, 2015). Relative abundance of Bacteroidetes EVGs among the 1,811 EVGs ranged from 2.2 to 34.6% (average: 13.9%). Members of the *Flavobacteriaceae* EVGs group 1 were abundant along with the *Flavobacteriaceae* EVGs group 2, which includes phi38:1 belonging to one of the most abundant viral candidate genera in the global oceans (Roux *et al.*, 2016). Most of the newly detected Bacteroidetes EVGs were less abundant (average: 2.8%) than

Flavobacteriaceae EVG group 1 or 2 (average: 11%). However, members of the G493 were ubiquitous and fourth most abundant genus among the Bacteroidetes EVGs (**Figures 2-10**, **2-11**). Additionally, TARA_ERS491107_N000194 (G561) was rarely recruited reads from most samples but found to be locally abundant (up to 20% of the relative abundance) in the Chile-Peru Current Coastal Province deep chlorophyll maximum sample (**Figures 2-10**, **2-11**).

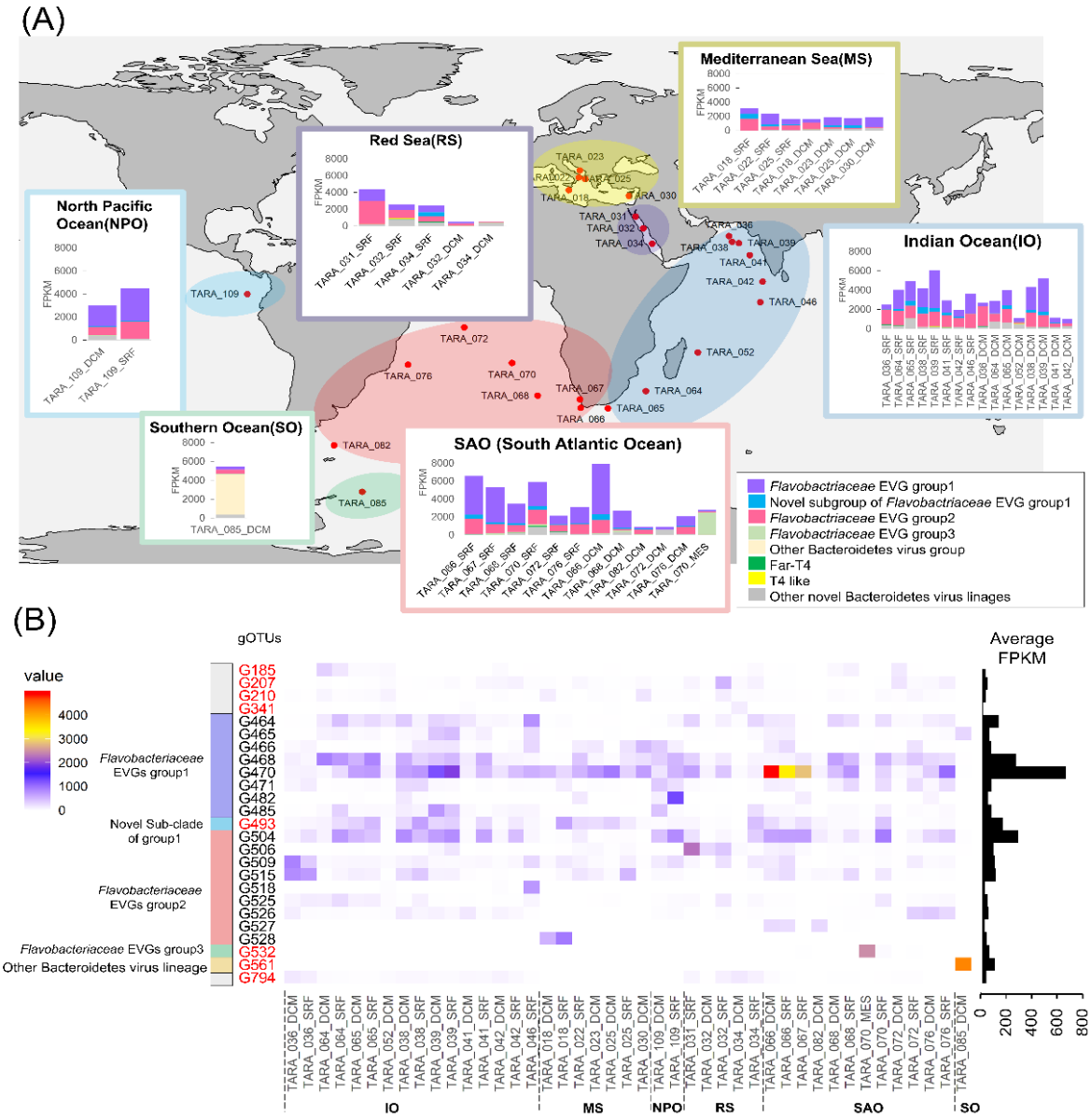

Figure 2-10. Abundance of the Bacteroidetes EVGs in Global Ocean surface waters.

(A)Virome fragment recruitments of Bacteroidetes EVG groups in each oceanic region. Sampling sites of *TARA* ocean expedition used for analysis are shown in red circle. Bar graphs represents normalized virome FPKM (fragments per kilobase per mapped million reads) of each Bacteroidetes EVG group at the site. (B) A heatmap shows normalized virome FPKM of abundant Bacteroidetes EVGs (i.e. gOTUs passing average relative abundance >0.1% and/or relative abundance >15% at least a site within Bacteroidetes EVGs). The scale bar on the left side represents FPKM value. Average FPKM values are shown in the right panel. Novel Bacteroidetes EVGs detected in this study are highlighted in red text. Oceanic region in the map in panel (A) are shown under x-axis.

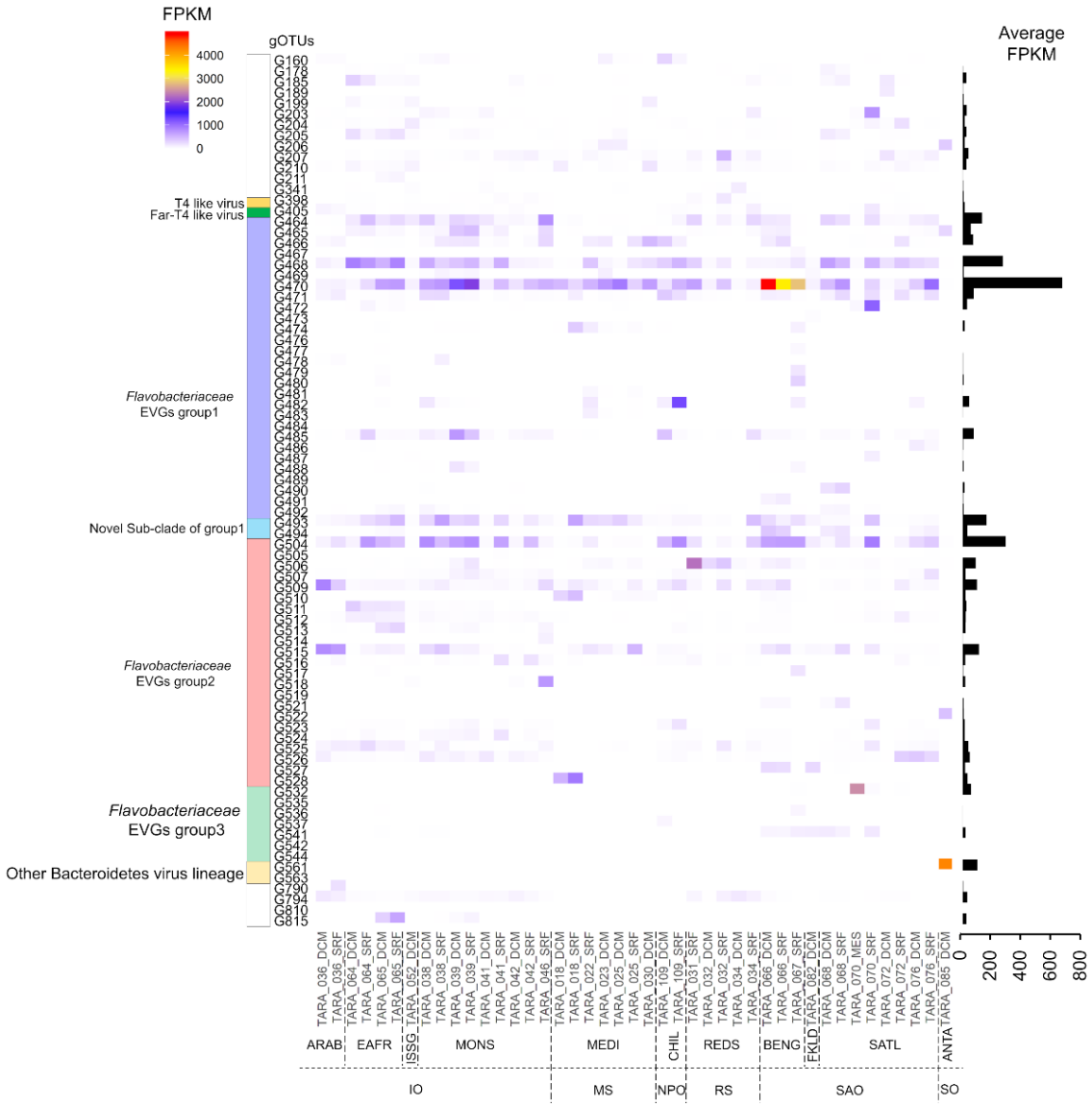

Figure 2-11. Abundance of the dominant Bacteroidetes EVGs in Global Ocean surface waters.

A heatmap shows normalized virome FPKM (fragments per kilobase per mapped million reads) of the 83 genus-level OTUs including 322 Bacteroidetes EVGs. The scale bar on the left side represents FPKM value. Average FPKM values are shown in the right panel. X-axis represents sampling sites of TARA ocean expedition and oceanic region are abbreviated as: IO (Indian Ocean), MS (Mediterranean sea), NPO (North pacific ocean), RS (Red sea), SAO (South Atlantic Ocean), and SO (Southern Ocean).

Discussion

As only limited lineages of marine Bacteroidetes can be cultivated (Alonso *et al.*, 2007), most viruses infecting marine Bacteroidetes have not been characterized. The objective of the study was to expand the knowledge of the diversity of the viruses likely infecting marine Bacteroidetes species by nucleotide/protein similarity-based approaches using MAGs as well as isolated bacterial genomes.

Firstly, I showed that Bacteroidetes MAGs from *Tara* Oceans data serve as more sensitive references for the host prediction of the uncultured marine Bacteroidetes viruses as compared to the genomes in the public database mostly derived from cultured bacteria (**Table 2-3**). This high sensitivity of MAGs obtained from simultaneously sampled metagenomes with EVGs supports ecosystem specific interactions of Bacteroidetes and these viruses. Taxonomic assignment of the Bacteroidetes MAGs suggests that these are representative genomes of previously uncultured marine Bacteroidetes lineages (**Table 2-2**). It strengthened our hypothesis that viruses of unknown hosts interact with uncultured bacteria and MAGs enabled us to detect potential interactions by overcoming the cultivation bias. However, it should be noted that MAGs likely include several contaminations of contigs from other taxa or viruses. Therefore, it is important to be careful of the pre-filtering steps such as removal of virus-like contigs and contaminated contigs of other taxa. Moreover, not only the MAGs, I identified several virus-like sequence contaminations from the reported Bacteroidetes genomes in NCBI RefSeq database. For example, I found that an 18.8 kb of circular contig from *Nonlabens* sp. 1Q3 (Accession: NZ_RMVE00000000) shows 98.7% nucleotide identity to Cyanophage P-TIM40 across 98% of the region. Pre-filtering by viral detection tools such as VirSorter

(Roux, Francois Enault, *et al.*, 2015) was also important for the accurate host prediction of viruses using cultivated bacterial genomes.

Secondly, I developed a protein homology-based host prediction approach. The approach achieved significant improvement of the detection of Bacteroidetes viruses compared to the nucleotide similarity-based approaches. High proportion of the host homologs likely derived from proviruses suggest that the methods mainly rely on the viral lysogeny (**Figure 2-2**). The observation that most of the viral genomes of cultured Bacteroidetes have a number of provirus homologs implies that lysogeny may be a widespread feature in Bacteroidetes viruses and these proviruses are maintained in host genomes. This feature might be related to a copiotrophic and r-strategist lifestyle of cultivated species of coastal Bacteroidetes (Lauro et al., 2009). Relatively large host genomes are capable of maintaining proviruses because of the weak selective pressure from genome streamlining (Lauro et al., 2009). Additionally, the viral lysogenic potential might be adaptive to respond to the multifold change of host abundance during and after phytoplankton bloom (Teeling *et al.*, 2012). The fact that lysogens are widespread (25–50% of the microbial genomes) in marine environments (Howard-Varona *et al.*, 2017) suggests that the homolog-based approach may be applicable not only for Bacteroidetes viruses but also for the environmental viruses infecting other prokaryotes. Indeed, the possession of many host-related homologs was also reported in uncultured viruses potentially infecting the marine group II (MGII) euryarchaeota (Nishimura et al., 2017a). However, viruses infecting extremophile Bacteroidetes have fewer Bacteroidetes homologs than the other Bacteroidetes viruses (Rhodothermus virus RM378: 1.4%, Salisaeta icosahedral virus: 6.6%). One possibility is that the shortage of genomes of the extremophile microorganisms due to sampling bias caused fewer matches with the host

like homologs in their viruses. Expansion of microbial genomes could assist in more precise and sensitive host prediction of uncultured viruses by the homolog proportion-based method.

The Bacteroidetes EVGs identified by these new approaches may provide useful genetic markers for studying viral importance in the ecological study of marine Bacteroidetes, such as viral roles in the rapid succession of various Bacteroidetes species during bloom(Teeling *et al.*, 2012; Needham *et al.*, 2016). For example, G493 is the fourth most abundant marine Bacteroidetes virus in the genus-level and might have a large impact on the dynamics of the uncultured marine Bacteroidetes populations. Among these newly identified Bacteroidetes EVGs, I identified not only the relatives of cultured marine Bacteroidetes viruses, but also marine viral lineages phylogenetically distinct from the cultured marine Bacteroidetes viruses.

I detected potential virus-host interactions between marine Bacteroidetes and Far-T4 viruses. They were previously reported to be common in aquatic environments but data on their complete genomes are unavailable and they are not linked with their hosts (Roux, François Enault, *et al.*, 2015). As members of Bacteroidetes are also common in aquatic environments (Kirchman, 2002; Pommier *et al.*, 2006), they are reasonable hosts of the uncultured Far-T4 lineages. These findings may provide important insights into the unknown ecology of Far-T4 viruses. Among the Far-T4 Bacteroidetes EVGs, I found several previously reported AMGs putatively related to carbohydrate metabolism, sulfur metabolism, and queuosine synthesis (**Figure 2-5**). Among them, queuosine synthesis genes were widely found in Bacteroidetes EVGs (T4 like Bacteroidetes EVG, member of *Flavobacteriaceae* EVGs group 1 and 2, **Table 2-4**). Queuosine is a hypermodified guanosine derivative in tRNAs specific for Asp, Asn, His,

or Tyr. One of the predicted roles of queuosine is the improvement of translation efficiency (El Yacoubi *et al.*, 2012) and a study suggested that it acts as a quantity control mechanism of viral structural gene products (Sabri *et al.*, 2011). Other studies suggest queuosine modification of viral DNA provides a protection mechanism against host endonucleases (Kulikov *et al.*, 2014; Thiaville *et al.*, 2016; Sazinas *et al.*, 2018). The biological role of queuosine modification is still controversial (Vinayak and Pathak, 2009); however, the prevalence of queuosine synthesis potential in marine Bacteroidetes EVGs suggests its advantage to the viruses during infection in marine Bacteroidetes.

Additionally, I found two systems putatively related to cell adhesion (curli production and ubiquitous cell surface proteins) in an EVG (**Table 2-4**). Curli amyloid fiber is a major proteinaceous component of the extracellular matrix produced mainly by Enterobacteriaceae (Barnhart and Chapman, 2006) and was also reported in Bacteroidetes genomes by bioinformatic analysis (Dueholm *et al.*, 2012). The ubiquitous surface proteins are essential for the attachment of pathogenic *Moraxella* (Lafontaine et al., 2000; Tan et al., 2006). The genes might promote the attachment of infected host cells near the uninfected host cells during infection. Such aggregation during infection was recently reported in Tupanvirus infecting amoebas and thought to promote progeny production (Oliveira *et al.*, 2019). Further studies are needed to clarify the role of these proteins in the life cycle of the EVGs.

Conclusions

From the analysis of the host prediction of 1,811 circular complete genomes, I detected 321 viral genomes that most likely correspond to Bacteroidetes dsDNA viruses. Microbial MAGs have advantages in the computational detection of uncultured marine Bacteroidetes viruses compared with the microbial genomes in the current public

databases. I also developed a sensitive method for predicting Bacteroidetes viruses based on bacterial homolog detection in viral genomes. This enhanced prediction approach using MAGs and homolog detection tested on the marine Bacteroidetes-virus systems might be applicable for the host prediction of diverse uncultured viral genomes and might also expand the realm of characterized viruses in various environments. The newly identified Bacteroidetes EVGs expanded our knowledge of the marine Bacteroidetes viruses such as identification of interactions between aquatic ubiquitous viral group Far-T4 and marine Bacteroidetes. They may serve as useful genetic markers for the future studies on the interactions between Bacteroidetes and their viruses.

Chapter 3

Prevailed viral frequency dependent selection toward coastal marine prokaryotes revealed by monthly time-series virome analysis

Abstract

Viruses infecting marine prokaryotes have large impacts on the diversity and dynamics of their host. In model systems, it has been argued that the viral infection rate is frequency-dependent, where rising cell densities drive increased virus-host encounters. However, it is unclear whether the frequency-dependent viral infection occurred in the natural prokaryotic community. Here, I examined the prevalence of viral infection in abundant prokaryotes through the comparison of prokaryotic and viral diversity by 16S rRNA amplicon sequencing and virome sequencing of samples collected monthly for two years at a Japanese coastal site, Osaka bay. Maximum community similarity between samples occurred at 12-month intervals in both prokaryotic and viral communities, suggesting the seasonality of the viral community was shaped by the seasonality of the prokaryotic community via the host-specific infection. To support this, the composition of viral putative hosts determined by *in silico* prediction (covering 62 % of the viral community) was similar to the taxonomic composition of the prokaryotic community. To test whether each virus increased according to its specific host abundance, co-occurrence network analysis between the viruses and abundant prokaryotic populations was performed. In total, 6,423 co-occurring virus-host pairs were determined and increasing of viruses in respond to their host abundance was observed between these pairs. Persistently abundant populations such as the most abundant populations of *Synechococcus* and SAR11 had few co-occurring viruses. However, faster temporal change and weak annual periodicity of viral community suggest dominant species of

viruses infecting these populations changed during observation. Altogether, the results suggest the prevalence of frequency-dependent viral infection in coastal marine prokaryotes.

Introduction

Marine prokaryotes are ubiquitous in the ocean and play key roles in biogeochemical processes such as carbon cycling (Falkowski *et al.*, 2008). The diversity analysis of the marine prokaryotic community based on sequencing of the 16S rRNA gene has revealed over 35,000 species-level operational taxonomic units (OTUs, based on 97% sequence identity) (Sunagawa *et al.*, 2015). The most of the observed prokaryotic species fall into 13 major phyla (class for proteobacteria) such as α-proteobacteria (e.g. SAR11 clade, SAR116 clade, and *Roseobacter* clade), γ-proteobacteria (e.g. SAR86 clade and SAR92 clade), Bacteroides (e.g. members of *Flavobacteriaceae*), and Cyanobacteria (e.g. *Synechococcus* and *Prochlorococcus*) (Pommier *et al.*, 2006; Sunagawa *et al.*, 2015). In spite of the divergence of metabolic capacity and physiologies among these species, each species often can be largely divided into either one of two growth strategist based on its potential growth rate and temporal dynamics: (i) *K*-strategist (slow-growing and persistently dominant, e.g. SAR11) and (ii) *r*-strategist (fast-growing and opportunistic, e.g. members of *Flavobacteriaceae*) (Suttle, 2007). However, recent high-frequency sampling schemes (e.g. daily) have given extended insights into the temporal dynamics of each species (Teeling *et al.*, 2012; Bunse and Pinhassi, 2017). For example, a OTU of Marine group II euryarchaeota (MGII), which had not been recognized as *r*-strategist species, showed drastic fluctuation following a spring phytoplankton bloom (Needham *et al.*, 2016). Further, taxonomic classification based on the single-nucleotide variation within 16S rRNA and internal transcribed spacer (ITS) sequences have revealed finely resolved populations (genotypes or strains) within a species-level OTUs (Eren *et al.*, 2013, 2015; Tikhonov *et al.*, 2015). Such populations often showed distinct temporal dynamics, indicating species which described as *K*-

strategist also can show frequent fluctuation under the finely resolved taxonomical resolution (Needham *et al.*, 2017; Chafee *et al.*, 2018).

Viruses infecting prokaryotes are ubiquitously and abundantly present in the ocean (Suttle, 2005, 2007; Breitbart *et al.*, 2018). The viruses are estimated to lyse 20–40% of the prokaryotic cells each day (Suttle, 2005, 2007; Breitbart *et al.*, 2018). Basically, viruses are believed to infect their specific hosts (often restricted to strains within a species) in a frequency-dependent manner according to the encounter rate between viruses and their hosts (Fuhrman and Suttle, 1993; Winter *et al.*, 2010). In particular, a study suggests 10^4 cells/ml is a host cell density threshold for rapid propagation of viral infection (Wiggins and Alexander, 1985). Thus, viruses are predicted to affect host diversity via frequency-dependent selection, in which viruses infect host population (species or strain) that become relatively abundant in environment and frequencies of host and viruses oscillate over time, maintaining host diversity (Thingstad, 2000; Rodriguez-Valera *et al.*, 2009).

Mathematical models of viral and host abundance have demonstrated that a prokaryotic species (or lineage) with faster growth rate than others can be susceptible to viral infection (Thingstad *et al.*, 1993; Thingstad, 2000). This trait allows the slow-growing *K*-strategist such as SAR11 to reach a higher abundance than the fast-growing *r*-strategist such as members of *Flavobacteriaceae* because of decrease of viral propagation (Suttle, 2007). However, the discovery of SAR11 viruses as the most abundant viruses raise a question to the prediction (Zhao *et al.*, 2013; Zhang *et al.*, 2020). Thus, it is still unclear whether *K*-strategist also suffer from viral infection and whether viral infection is prevalent in abundant prokaryotic populations according to their abundance.

Measuring the abundance of viruses infecting each prokaryotic population in the environment is difficult because of the enormous diversity of marine viruses (Brum and Sullivan, 2015). Since the vast majority of marine prokaryotes could not be cultivated using standard techniques (Rappé and Giovannoni, 2003), 50% of class to genus-level taxonomic groups still remain uncultivated (Lloyd *et al.*, 2018). Thus, the viruses infecting marine prokaryotes have been rarely cultivated except for well-studied marine *Synechococcus* and *Prochlorococcus* virus–host systems (Suttle and Chan, 1993; Waterbury and Valois, 1993; Sullivan *et al.*, 2003, 2005) and several isolates infects other taxa such as SAR11 (Zhao *et al.*, 2013). Although viruses lack a universally conserved gene such as 16S rRNA gene of prokaryotes (Edwards and Rohwer, 2005), viral metagenomes (viromes) recently became a powerful tool to characterize the diversity of the uncultivated viruses (Brum and Sullivan, 2015; Nishimura, Watai, *et al.*, 2017; Breitbart *et al.*, 2018). However, the majority of uncultured viruses derived from viromes had no cultured relatives and had not yet been connected with their hosts (Brum and Sullivan, 2015; Breitbart *et al.*, 2018). For example, a study reported that 78.4% (1,420 genomes) of 1,811 circular viral genomes from the marine viromes were not connected with their host (Nishimura, Watai, *et al.*, 2017). On the other hand, to overcome such limitation in prediction of virus–host pairs, *in silico* host prediction approaches using viral and microbial genomes have been developed (Edwards *et al.*, 2016; Ahlgren *et al.*, 2017; Tominaga *et al.*, 2020).

Monthly observation of microbial community is the most common interval in oceanic time-series studies (Bunse and Pibynhassi, 2017; Chow and Fuhrman, 2012; Fuhrman et al., 2015; Ignacio-Espinoza et al., 2020; Xia et al., 2011). These studies uncovered that seasonal oceanographic features, such as temperature, salinity, and

nutrient concentrations have a strong influence on prokaryotic dynamics (Fuhrman *et al.*, 2015; Bunse and Pinhassi, 2017). Dynamics of viral community also known to show seasonal variability (Chow and Fuhrman, 2012; Pagarete *et al.*, 2013; Ignacio-Espinoza *et al.*, 2020). Since viruses are obligate parasites, the seasonal dynamics of viruses can be shaped by the dynamics of its hosts and prevalence of viral infection in abundant prokaryotic populations can be addressed by comparison of virus and host dynamics. However, the viral seasonality was often discussed independently with its host dynamics except for few prokaryotic-virus pairs such as *Synechococcus/Prochlorococcus* and T4-like viruses (Xia *et al.*, 2011; Chow and Fuhrman, 2012; Ahlgren *et al.*, 2019; Ignacio-Espinoza *et al.*, 2020).

In this study, I aimed to solve the above two fundamental questions whether viral infection is prevalent among abundant prokaryotic populations and whether viral infection manners is different among populations according to taxa and/or its growth strategy. I conducted a two-years monthly observations of prokaryotic and viral diversity at the eutrophic coastal site, Osaka bay, Japan. I performed *in silico* host prediction analysis of the viral genomes obtained from the time series samples. According to the prediction, dynamics of viruses and their putative hosts were compared and potential virus-host pairs were determined by their co-occurrence dynamics. I examined whether viruses were abundant when their putative hosts were abundant in the determined virus-host pairs. Interactions between *K*-strategist prokaryotes and their viruses were further discussed based on the difference of their temporal dynamics.

Materials and Methods

Sampling and processing

Seawater samples (4L) were collected from a 5 m depth at the entrance of Osaka Bay (34°19′28″N, 135°7′15″E), Japan, within 3h from before or after high tide, between March, 2015 to November 2016, at a monthly resolution. Seawater was filtered through a 142-mm-diameter (3.0-μm-pore-size) polycarbonate membrane (Millipore, Billerica, MA) and then through a sequentially through 0.22 μm-pore Sterivex filtration units (SVGV010RS, EMD Millipore). After filtration, the 0.22 μm filtration units were directly transferred to −80 °C (for subsequent DNA extraction). The filtrates were stored at 4°C prior to treatments. Water temperature and salinity were monitored by fixed water intake systems of the research institute of environment agriculture and fisheries, Osaka prefecture. Nutrient concentrations (NO_3-N, NO_2-N, NH_4-N, PO_4-P, and SiO_2-Si) were measured by continuous flow analysis (BL TEC K.K., Japan.).

rRNA gene amplicon sequencing analysis

For prokaryotic community analysis, DNA was extracted from the stored filtration units using previously described protocol (Yoshida *et al.*, 2018; Takebe *et al.*, 2020). Total 16 S rDNA was amplified using a primer set based on the V3–V4 hypervariable region of prokaryotic 16 S rRNA genes (Takahashi *et al.*, 2014) with added overhang adapter sequences at each 5′-end according to the sample preparation guide (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf). Amplicons were sequenced using MiSeq sequencing system and MiSeq V3 (2 × 300 bp) reagent kits (Illumina, San Diego, CA).

Paired-end sequences of the 16S rDNA amplicon were merged using VSEARCH with option "-M 1000" (Rognes *et al.*, 2016). Merged reads containing ambiguous nucleotides (i.e., "N") were discarded. The remaining merged reads were clustered using VSEARCH to form operational taxonomic units (OTUs) with a sequence identity threshold of 99%. Singleton OTUs were discarded. The representative sequences of the remaining OTUs were searched against the SILVA ribosomal RNA gene database (release 138) (Quast *et al.*, 2013) to taxonomically annotate OTUs using SINA (Pruesse *et al.*, 2012) with a threshold of 99% sequence identity. Abundant OTUs were defined as OTUs exceeding 1 % relative abundance according to the reported minimum host cell density for effective viral infection ($\fallingdotseq 10^4$ cells/ml) (Wiggins and Alexander, 1985) and typical coastal marine prokaryotic cell density ($\fallingdotseq 10^6$ cells/ml) (Whitman *et al.*, 1998).

To identify statistically relevant variants within abundant OTUs, we applied minimum entropy decomposition (MED) (Eren *et al.*, 2015) as following previous study (Needham *et al.*, 2017). All of the sequences from each of these 99% OTU were aligned with MAFFT v7.123b (-retree 1 -maxiterate 0 -nofft -parttree) (Katoh *et al.*, 2002). The alignments of sequences that had entropy at sites >0.25 were decomposed based on those positions, and decomposition continued until all positions had entropy <0.25. The minimum number of the most abundant sequence within each amplicon sequence variants (ASVs) must exceed 50 and if ASVs did not exceed 1% of the parent 99% OTU's composition on average, they were removed from analysis as previously reported (Needham *et al.*, 2017).

Virome sequencing, assembly, classification, and calculation of relative abundance

The viruses in the filtrate were concentrated by $FeCl_3$ precipitation (John *et al.*, 2011) and purified using DNase and a CsCl density centrifugation step (Hurwitz *et al.*,

2013). The DNA was then extracted as previously described (Kimura *et al.*, 2012). A sample (February, 2016) generated insufficient amount of DNA for virome sequencing, it was removed from the analysis. Libraries were prepared using a Nextera XT DNA sample preparation kit (Illumina, San Diego, CA) according to the manufacturer's protocol, except that I used 0.25 ng viral DNA. Samples were sequenced with a MiSeq sequencing system and MiSeq V3 (2 × 300 bp) reagent kits (Illumina, San Diego, CA).

Viromes were individually assembled using SPAdes 3.9.1 with default *k*-mer lengths (Bankevich *et al.*, 2012). Additionally, I used scaffolds of these assemblies (hereafter referred as contigs for simplicity). Circular contigs were determined as previously described (Nishimura, Watai, *et al.*, 2017). The contig sequences were clustered at 95% global average nucleotide identity (ANI) with cd-hit-est (options: -c 0.95 -G 1 -n 10 -mask NX) (Li and Godzik, 2006). Total 5,226 mts-OBV contigs (monthly time series Osaka bay viral contigs, >10 kb, 62~926 contigs/samples, including 202 circular ones) were obtained.

In addition, this assembly generated 181,131 short contigs (i.e., longer than 1 kb but not longer than 10 kb). The abundance of these contigs was assessed based on the relative abundance of terminase large subunit genes (*terL*) as previously described (Yoshida *et al.*, 2018). As a result, 4,666 genes were detected as putative *terL* genes (i.e., genes with the best hit to PF03354.14, PF04466.12, PF03237.14, and PF05876.11). The FPKM (fragments per kilobase per mapped million reads) for putative *terL* genes were calculated by in-house ruby scripts.

The mts-OBV contigs with complete viral genomic sequence set collected in a previous study (Nishimura, Watai, *et al.*, 2017) were used for viral abundance estimation from read mapping. The complete viral genomic sequence belong to one of the following

two categories; (i) 1,811 environmental viral genomes (EVGs; all are circularly assembled genomes, 45 are assembled in Osaka bay in previous study) derived from marine virome studies, and (ii) 2,429 reference viral genomes (RVGs) of cultured dsDNA viruses. Genus-level genomic OTUs (gOTUs) were previously assigned for complete genomes based on genomic similarity score by ViPTree (Nishimura, Yoshida, *et al.*, 2017). For the mts-OBV contigs, if the sequence showed similarity to one of the complete genomes (with genomic similarity SG > 0.15), the sequence was assigned to the gOTU of the most similar circular genome as previously described (Nishimura, Watai, *et al.*, 2017; Yoshida *et al.*, 2018). Quality controlled virome reads were obtained through quality control steps as previously described (Nishimura, Watai, *et al.*, 2017). These reads were mapped against the viral genomic sequence set using Bowtie2 software with a parameter "--score-min L,0,-0.3" (Langmead and Salzberg, 2012). FPKM values were calculated by in-house ruby scripts.

Viral host prediction

First, I assigned putative host groups based on the genomic similarity with viral genomic sequence set collected in a previous study (Nishimura, Watai, *et al.*, 2017). If a mts-OBV contigs were classified into the same gOTU with the viruses which previously assigned host group by cultivation or predicted by genomic contents (Nishimura, Watai, *et al.*, 2017), the host groups were applied to the contigs. I also compared similarity with mts-OBV contigs and the viral genomes deposited in virus host database (as of October 2018) and recently reported isolates (Mihara *et al.*, 2016; Zhang *et al.*, 2020).

In addition, for the viruses not assigned host group by genomic similarity, we performed *in silico* host prediction based on the nucleotide sequence similarity between viruses and prokaryotes as previously described (Paez-Espino *et al.*, 2016; Roux *et al.*,

2016; Tominaga *et al.*, 2020). First, total 220,103 viral genomes/contigs derived from marine viromes were collected used for the analysis (Mizuno *et al.*, 2016; Luo *et al.*, 2017; Nishimura, Watai, *et al.*, 2017; Gregory *et al.*, 2019; Ignacio-Espinoza *et al.*, 2020) (**Table 3-2**). For the putative host genomes, I collected total 8,016 MAGs/SAGs from marine metagenomic or single cell genomic studies (Tully *et al.*, 2017, 2018; Delmont *et al.*, 2018; Krüger *et al.*, 2019; Pachiadaki *et al.*, 2019). From Pachiadaki *et al*, I only used 1,040 high quality SAG assemblies ≥ 80% completion (Pachiadaki *et al.*, 2019). To remove the contamination of virus-like contigs from the MAGs/SAGs, 14,967 contigs classified as viral-like sequences by VirSorter (category 1, 2, and 3) (Roux, Francois Enault, *et al.*, 2015) were discarded (**Table 3-1**). Details of each methods are reviewed elsewhere (Edwards *et al.*, 2016). (i) CRISPR spacers match: CRISPR spacer sequences were predicted by CRISPR Recognition Tool (Bland *et al.*, 2007) then total 13,305 sequences were extracted. Detected spacer sequences and spacers sequences deposited in CIRSPRdb (Grissa *et al.*, 2007) were queried against viral genomes using the BLASTn-short function (Camacho *et al.*, 2009) with these parameters: at least 95% identity over the whole spacer length and only 1–2 SNPs at the 5′end of the sequence was allowed. (ii) tRNA match: tRNAs were recovered from MAGs/SAGs and viral genomes by ARAGORN with '-t' option (Laslett and Canback, 2004). Total 213,939 and 31,439 tRNAs were recovered from the MAGs/SAGs and viral genomes, respectively. The recovered prokaryotic and viral tRNAs with 111,385 tRNAs deposited in GtRNAdb (Chan and Lowe, 2016) were compared by BLASTn (Camacho *et al.*, 2009) and only a perfect match (100% length and 100% sequence identity) were considered as indicative of putative host-virus pairs. (iii) Nucleotide sequences homology of prokaryotic and viral genomes: viral genomes/contigs were queried against prokaryotic MAGs/SAGs and

prokaryotic genomes in NCBI RefSeq (as of December 2019) using BLASTn (Camacho *et al.*, 2009). Only the best hits above 80% identity across alignment with the length ≥1500 bp were considered as indicative of host-virus pairs. For the prediction based on the contigs of MAG/SAGs, I performed taxonomic validation of the matched contigs in MAG/SAGs as previously described (Tominaga *et al.*, 2020). Viruses belonging to the same gOTU were assigned to have consistent host groups according to the previous study (Nishimura, Watai, *et al.*, 2017), with three exceptional gOTUs (G404, G405, and G495) which annotated multiple host lineages. For the contigs assigned into the three gOTUs, genomic similarity among the same gOTU members were calculated and potential host of each contigs were assigned based on the most similar genomes/contigs which annotated by host prediction

Statistical analyses

Before statistical analyses, amplicon reads were rarefied using the "vegan" package in R (Dixon, 2003). To examine within-sample alpha-diversity (Shannon diversity, evenness, and richness) and beta-diversity (Bray–Curtis similarity: 1 - Bray–Curtis dissimilarity, for all of the possible pairwise combinations among all of the sampling points) using the vegan package in R (Dixon, 2003). Mantel tests were performed in R via the vegan package (Dixon, 2003) on only fully overlapping set of data. Pairwise correlations between estimated abundance of prokaryotic ASVs and viral contigs (having putative host information and exceeding FPKM >10 at least a month, 2,735 contigs) on fully overlapping set of data were then determined via Spearman correlations ($P<0.01$, $Q<0.05$) as implemented in the local similarity analysis program. (Xia *et al.*, 2011, 2013). Network visualizations of correlation matrices were generated in Cytoscape_v3.8.0 (Shannon *et al.*, 2003).

Table 3-1. Basic statistics of microbial metagenome-assembled genomes used for the host prediction analysis.

Dataset	Number of genomes	Number of contigs	Contigs removed by VirSorter	Number. of CRISPR spacer sequences	Number. of tRNAs	Reference
GORG	1040	34,175	182	82	29,187	Pachiadaki *et al,* 2019
TOBG	2,631	214,181	7,317	8,379	75,981	Tully *et al.*, 2018
TMED	290	18,380	1,046	95	7,416	Tully *et al.*, 2017
TARA_MAG	957	323,552	3,164	2,954	25,047	Delmont *et al.*, 2018
NS_MAG	3098	557,045	3,258	1,795	76,308	Krüger *et al.*, 2019
Total	8,016	1,147,333	14,967	13,305	213,939	

Table 3-2. Basic statistics of viral metagenome-assembled genomes used for the host prediction analysis.

Dataset	Number of contigs	Number of of tRNAs	Reference
SPOT	19,907	1,592	Ignacio-Espinoza *et al.* 2020
EVG+RVG	4,240	6,322	Nishimura *et al.* 2017
GOV2	195,728	23,473	Gregory *et al.* 2019
Deep Ocean	99	19	Mizuno *et al.* 2017
ALOHA	129	33	Luo *et al,* 2017
Total	220,103	31,439	

Estimation of growth strategy of ASVs

I established indexes for the approximation of the *r* (intrinsic rate of natural increase) and *K* (carrying capacity) of each ASVs from their monthly dynamics. For the approximation of the *r* of each ASVs, maximum increasing of normalized relative rank (0-1) per month was applied. Similarly, for the approximation of *K* of each ASVs, length of continuously dominant month (>0.1% relative abundance, 1-18 months) of each ASVs was applied.

Detection of SNPs

Reads were mapped to the viral contigs using Bowtie2 with a parameter "--score-min L,0,-0.3" as above (Langmead and Salzberg, 2012) and the resulting alignment files were converted to BAM format sorted using samtools (Li, 2011). Average genome entropy of the contigs which exceeded more than 10 coverage at each month were computed using the DiversiTools (http://josephhughes.github.io/DiversiTools/).

Results and Discussion

Overview of prokaryotic and viral community in Osaka bay

The prokaryotic community composition was determined by sequencing amplicon libraries of the 16S rRNA gene V3-V4 region derived from 0.22 to 3 μm size fractions. Total 2.8 M paired-end reads (24,168 to 846,565 reads per sample) were obtained from the 18 samples and the sequences were clustered into 35,191 OTUs (1,462 to 18,268 OTUs per month) with a sequence identity threshold of 99% (hereafter referred to as species-level populations, **Table 3-3**). The prokaryotic community was dominated by α-proteobacteria (41%), γ-proteobacteria (21%), Bacteroidetes (19%), and Cyanobacteria (7%) at the phylum level (class level only for Proteobacteria).

To explore viral community composition, I obtained a total 60 M paired-end reads of viromes (929,884 to 8,124,354 sequence) which generated from the virus enriched <0.22 μm size fraction of 17 samples that were concomitantly collected with the prokaryotic size fractions (**Table 3-3**). After decontamination of prokaryotic sequences, 5,226 virus-like large contigs (> 10kb, monthly time series Osaka bay viral contigs: mts-OBV contigs) including 202 circularly assembled viral genomes were obtained (**Table 3-3**). In this study, I refer to each contig as a species-level viral population, according to the proposal in viral ecology (Roux *et al.*, 2019). The majority (~75%) of these mts-OBV contigs showed high genomic similarity (Genomic similarity score; $S_G > 0.15$) with one of the complete (circular) viral genomes reported previously (Nishimura, Watai, *et al.*, 2017) or the 202 circular genomes assembled in this study. Based on the S_G, these mts-OBV contigs were classified into 314 gOTUs (**Table 3-3)**. Average 40% of virome reads (29 to 53% per sample) were mapped to the mts-OBV contigs or previously reported viral genomes (Nishimura, Watai, *et al.*, 2017). The mts-OBV contigs occupied average 96% relative abundance at each samples (based on the FPKM values calculated from read counts). To confirm that each mts-OBV contig represents an abundant viral population, I estimated their relative abundance using the whole set of contigs (>1kb) and relative abundance of terminase large subunit genes (*terL*) as previously described (Yoshida *et al.*, 2018). Although the relative abundance of the mts-OBV contigs range widely on average (0.0013% to 1.26%, **Figure 3-1**), all mts-OBV contigs ranked in the top >30% of whole community at least a month (the lowest of maximum relative abundance was 0.0115 % , 16Jan_NODE_472).

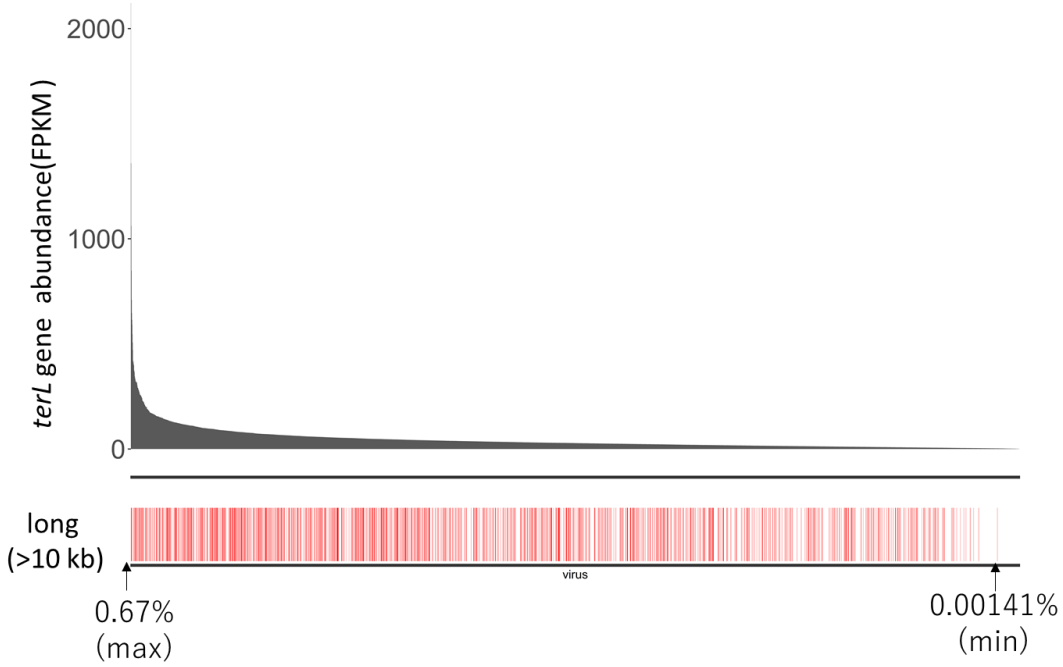

Figure 3-1. Virome abundance of OBV long contigs as assessed by putative *terL* genes. Abundance of 1,078 mts-OBV long contigs (indicated by red ticks) was assessed by the abundance of putative terL genes (from 4,666 contigs in total). x-axis represents rank of the contigs. y-axis represents the percentage of *terL* FPKM (average of 17 samples).

Alpha diversity (Shannon index) of the viral community was significantly higher than the prokaryotic community ($p < 0.001$, **Figures 3-2A, 3-2B**). Both richness and evenness were also significantly higher in the viral community than the prokaryotic community (**Figures 3-2C, 3-2D, 3-2E, 3-2F,** $p < 0.001$). The lower evenness of prokaryotic community indicate a steeply declining rank–abundance curve as often observed in the marine environment (Pedrós-Alió, 2006). One of the possible explanation for the higher viral diversity is that a prokaryotic species (OTU) can simultaneously be infected by multiple viral species at each time point (discussed below). Changes in alpha diversity of the viral community showed a weak inverse correlation with that of the prokaryotic community (**Figure 3-2B**). Especially, inverse correlation was apparent in species richness ($p < 0.05$, **Figure 3-2D**).

Table 3-3. 16S rRNA amplicon and virome read sequences in each time series samples.

Date	16S rRNA(V3-V4)					Virome					
	Raw reads	Analyzed reads	OTUs (identity 99%)	Abundant OTUs	Abundant ASVs	Raw reads	Contigs (>10kb)	Number. of observed gOTU	Circular	Used for read mapping	Mapped reads
26/5/2015	584,310	383,233	12,108	16	77	2,406,326	216	58	8	1,935,576	38%
21/6/2015	169,680	96,012	6,908	16	74	3,169,845	215	56	5	2,664,889	42%
22/7/2015	179,210	101,327	8,270	13	73	2,969,901	205	45	8	2,524,397	47%
18/8/2015	205,748	122,268	8,230	10	66	3,038,377	237	54	11	2,506,479	37%
8/9/2015	196,590	93,356	7,727	20	65	2,872,112	188	39	6	2,400,621	53%
30/10/2015	846,565	476,567	18,268	17	72	2,853,038	132	31	4	2,304,810	34%
26/11/2015	130,132	101,127	6,553	16	72	3,500,833	106	34	3	3,170,941	50%
15/12/2015	45,041	38,431	3,185	17	69	2,966,567	336	62	7	2,682,291	44%
14/1/2016	32,124	26,130	1,589	17	71	8,124,354	1,143	149	41	6,695,026	46%
9/2/2016	47,849	42,071	1,462	18	70	na	na	na	na	na	na
10/3/2016	59,464	52,042	2,576	18	73	5,451,541	698	116	28	4,116,774	46%
27/4/2016	45,725	40,408	1,466	18	68	1,508,014	131	42	1	1,424,427	41%
25/5/2016	72,102	63,239	2,481	15	72	929,884	84	30	7	846,803	35%
24/6/2016	94,774	83,776	3,987	17	73	1,821,038	145	55	4	1,699,251	31%
20/7/2016	48,386	42,265	2,134	18	69	1,565,957	64	35	2	1,470,533	29%
22/8/2016	24,168	20,803	1,778	20	63	2,327,061	194	59	2	2,148,345	29%
26/9/2016	27,118	21,828	2,284	17	54	7,824,235	926	180	32	6,975,203	31%
1/11/2016	67,236	45,444	3,362	18	57	7,651,460	755	133	38	6,623,871	36%

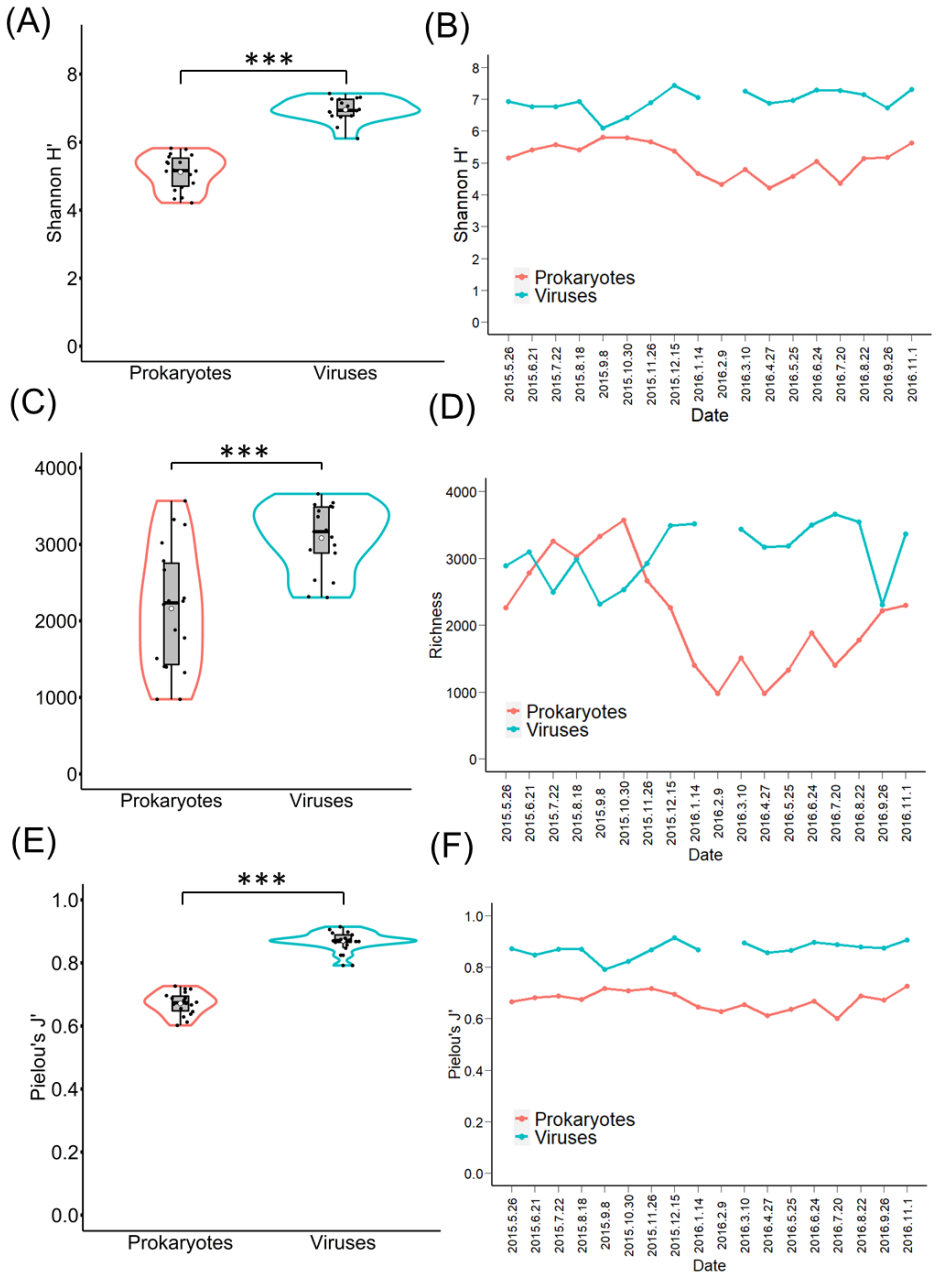

Figure 3-2. Alpha diversity profiles of prokaryotic and viral communities in Osaka bay during observation.

Average of Shannon H' (A), richness (number of OTUs or contigs, C), and evenness (Pielou's j: Shannon diversity divided by log richness, E) were calculated from normalized abundances of prokaryotic OTUs based on rarefied reads and viral contigs from fragments per kilobase of per million reads mapped (FPKM) value. The boxes represent the first quartile, median, and third quartile. Asterisks denote significance (Student's *t*-test, ***$p < 0.001$). The change of Shannon H' (B), richness (D), and evenness (F) of prokaryotic and viral communities of the time-series were plotted.

Seasonal dynamics of the prokaryotic and viral communities

Water temperature was higher in summer and lower in winter during observation (**Figure 3-3**).The concentration of SiO_2, PO_4-P, and inorganic nitrogen was increased in summer presumably because of the river inflow increasing during rainy season (**Figure 3-3**).

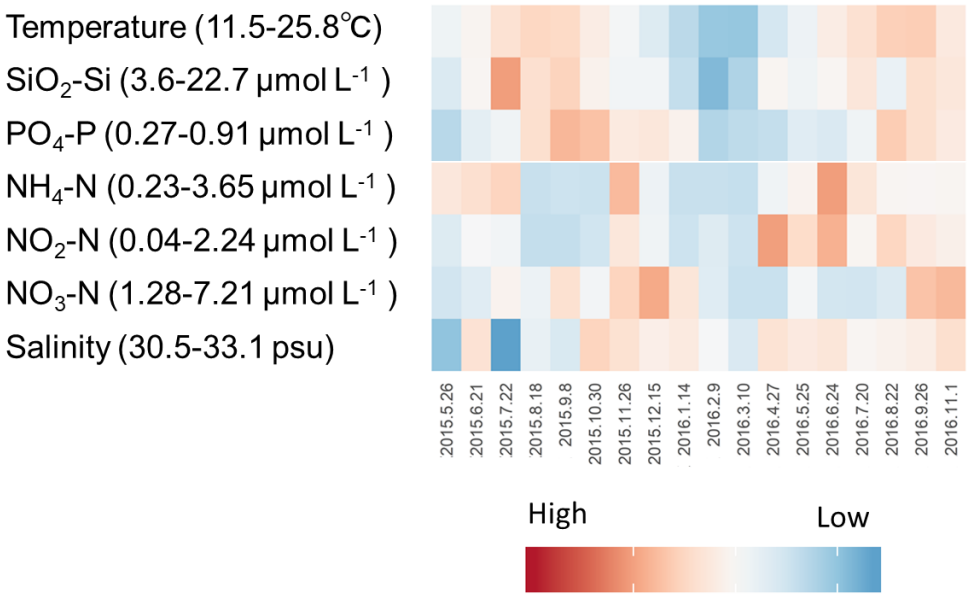

Figure 3-3. Changes in environmental parameters at the Osaka Bay (OB).
Heatmap represents z-score transformed value of measured environmental parameters.

I investigated seasonal dynamics of the prokaryotic and viral communities inferred from Bray-Curtis similarity for all of the pairwise combinations (136 pairs, 1 to 17-month intervals). Both prokaryotic and viral community showed clear seasonal patterns, with peaks of maximum average similarity at intervals of around 12 months, representing the same seasons, and minimum average similarity at 6 months intervals, representing opposite seasons (**Figure 3-4**). Although the similarity between samples were constantly lower in viral community than prokaryotic community (**Figure 3-4,** discussed below), the viral community composition was significantly correlated with the prokaryotic community composition (Mantel $rho = 0.51$, $p < 0.01$). The seasonality of

viruses following prokaryotes was consistent with the assumption that the viral community was shaped by the prokaryotic community because each virus only can propagate in its specific host.

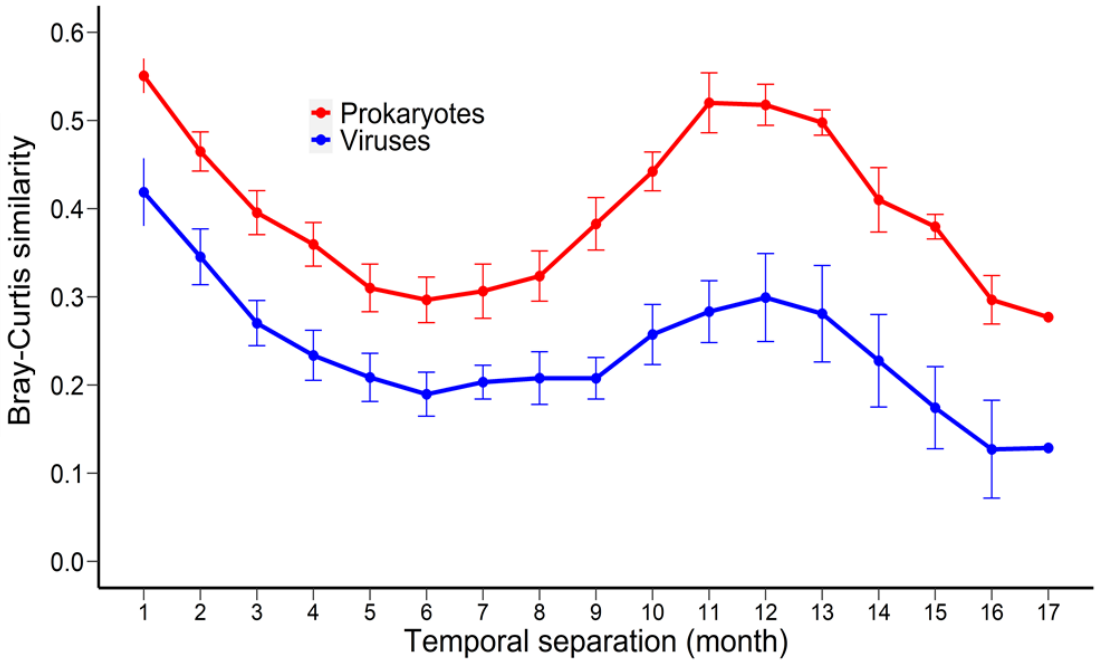

Figure 3-4. Seasonality of the prokaryotes and viruses at the Osaka Bay (OB) during observation.

The Bray–Curtis community similarity index was calculated among all of the possible sample pairs from normalized abundances of prokaryotic OTUs and OBV contigs and plotted as a function of the number of months separating their sampling.

If the viral community composition was shaped by prokaryotic community composition, abundance of each virus might reflect the abundance of its host. To test the hypothesis, composition of the prokaryotic community and viral community were compared based on the viral putative hosts (mostly host phylum- or class level composition). The putative host groups of viruses were predicted by four genome-based *in silico* prediction methods (Similarity with known viruses, CRISPR-spacer match, tRNA match, and genome homology). First, based on the similarity with cultured viruses, putative host groups of 951 mts-OBV contigs (22 gOTUs) were predicted (182 contigs (6

gOTUs), *Synechococcu/Prochlorococcus*; 501 contigs (8 gOTUs), SAR11; 214 contigs (2 gOTUs), SAR116; 31 contigs (1 gOTUs), *Roseobacter*; 23 contigs (5gOTUs), Others). Similarly, putative host groups of 504 mts-OBV contigs (39 gOTUs) were predicted based on the similarity with uncultured viruses which previously assigned their putative host (468 contigs (31 gOTUs), Bacteroidetes; 36 contigs (4 gOTUs), MGII, predicted by Nishimura et al., 2017a and Tominaga et al., 2020). In other 1,460 mts-OBV contigs (35 gOTUs), putative host groups were predicted by the sequence similarity (CRISPR-spacer match, tRNA match, and genome homology) with metagenome assembled genomes of marine prokaryotes or the genomes in the public database mostly derived from cultured prokaryotes (621 contigs (14 gOTUs), α-proteobacteria; 80 contigs (5 gOTUs), Bacteroidetes; 236 contigs (5 gOTUs), γ-proteobacteria; 326 contigs (2 gOTUs), δ-proteobacteria; 53 contigs (8 gOTUs), Others, **Table 3-4, 3-5**). Finally, I assigned potential host groups for the 2,844 mts-OBV contigs. Note that the host prediction based on genome analysis is mostly phylum or class level except for contigs showing similarity with cultured viruses such as *Synechococcus/Prochlorococcus* cyanoviruses. Thus, the host prediction could not completely reflect species (or strain)-specific virus-host pairs.

Table 3-4. General genomic features of the host assigned viral genus-level group (gOTUs) inferred from host prediction analysis based on the sequence similarity with microbial genomes.

gOTU (Defined in Nishimura *et al*, 2017)	Predicted host taxa	Ave. genome size of circular genomes (kbp)	Ave. G+C (%)	No. of OBV contigs	Ave. relative abundance (%)
G6	γ-proteobacteria	40.3	41.3	14	0.23
G39	α-proteobacteria	52.7	37.6	26	0.49
G46	δ-proteobacteria	35.5	42.4	134	3.46

Table 3-4. Continued

G52	α-proteobacteria	46.1	44.3	70	2.64
G79	α-proteobacteria	37.2	42.4	37	0.73
G92	γ-proteobacteria	43.4	41.1	15	0.21
G102	α-proteobacteria	56.7	43.5	205	3.54
G106	α-proteobacteria	55.1	37.6	20	0.39
G107	γ-proteobacteria	57.8	43.7	143	2.95
G108	Bacteroidetes	58.4	46.8	9	0.10
G112	γ-proteobacteria	61.4	41.4	63	0.93
G113	α-proteobacteria	52.7	36.6	17	0.23
G114	α-proteobacteria	57.2	36.5	48	1.07
G117	γ-proteobacteria	58.5	43.0	7	0.07
G125	α-proteobacteria	36.0	46.8	30	0.42
G131	Bacteroidetes	37.0	33.4	3	0.07
G179	Bacteroidetes	42.9	35.2	7	0.10
G190	Verrucomicrobia	58.2	32.2	4	0.06
G198	Verrucomicrobia	34.6	32.5	6	0.21
G204	Verrucomicrobia	32.5	32.4	4	0.05
G264	Bacteroidetes	31.7	35.2	2	0.01
G266	Verrucomicrobia	41.5	49.8	1	0.01
G317	α-proteobacteria	35.3	51.0	2	0.02
G389	α-proteobacteria	145.8	37.7	6	0.04
G398	Verrucomicrobia	179.9	32.0	11	0.08
G404	α-proteobacteria or δ-proteobacteria	180.7	40.7	241	6.76
G405	Bacteroidetes or Marinimicrobia	143.7	33.4	66	1.43
G410	Verrucomicrobia	111.8	33.8	13	0.28
G495	α-proteobacteria or Bacteroidetes	43.9	40.9	33	0.55
G865	α-proteobacteria	35.6	46.3	12	0.13
G1072	α-proteobacteria	32.5	36.1	66	1.50
G1078	γ-proteobacteria	39.4	58.3	1	0.04

Table 3-5. Numbers of host-virus pairs between viral genomes and prokaryotic MAGs or genomes in reference databases detected by three host prediction methods.

gOTUs (Defined in Nishimura *et al,* 2017)	BLASTn	CRISPR	tRNA
G6	-	1	-
G39	-	-	1
G46	-	2	-
G52	1	-	-
G79	21	-	-
G92	8	-	-
G102	23	-	-
G106	20	-	-
G107	16	-	-
G108	-	1	-
G112	15	-	-
G113	5	-	-
G114	-	-	1
G117	-	1	-
G125	1	-	-
G131	6	-	-
G179	2	-	-
G190	1	-	-
G198	2	-	-
G204	1	-	-
G264	-	4	-
G266	-	1	-
G317	6	-	-
G389	1	-	13
G398	-	-	49
G404	-	-	9
G405	5	-	505
G410	-	1	4
G495	17	-	-
G865	-	1	-
G1072	2	-	-
G1078	14	-	-

Major phyla (or class level only for proteobacteria) in the prokaryotic community were not changed drastically but relative abundance of several phyla (class) exhibit a remarkable seasonal dynamics (**Figure 3-5**). The composition and seasonal dynamics of viral community generally followed composition and dynamics of their putative hosts (**Figure 3-5**). For example, Cyanobacteria (79% of reads were assigned to OTU_8, *Synechococcus*) dominated in summer (up to 9.6% and 22.6 % of community at June 2015 and July 2016, respectively, **Figure 3-5A**) and *Synechococcus* viruses also increased summer (up to 5.3 % and 12.1% of community at Augest 2015 and Augest 2016, respectively, **Figure 3-5B**). Similarly, relative abundance of Bacteroidetes increased from winter to spring (up to 33.7% of community at May 2016, **Figure 3-5A**) and Bacteroidetes viruses also increased during spring (up to 30.2% of community at May 2016, **Figure 3-5B**). Both SAR11 (from 5% to 47% of community, **Figure 3-5A**) and SAR11 viruses (from 9% to 22% of community, **Figure 3-5B**) were always abundant throughout the observation.

If the viral infection increased in association with host density, the viral composition might be predictable by the composition of abundant prokaryotes. To examine the hypothesis, I calculated the ratio of viruses potentially infect abundant OTUs (73 OTUs, 52~42 OTUs/month, **Table 3-6**) which was selected based on the minimum cell density for effective viral propagation in cultured viruses (Wiggins and Alexander, 1985). The majority (78 ~100 %) of the putative host taxa of viruses matched with taxa of the abundant OTUs of each month (**Figure 3-6**). The result was in agreement with the assumption that viral infection increased with host density and frequently occurred in abundant prokaryotes.

However, viral abundance did not always match with their putative host abundance (**Figure 3-5**). For example, the proportion of putative γ-proteobacteria viruses was lower comparing with that of γ-proteobacteria and the proportion of putative δ-proteobacteria viruses was much higher comparing with that of δ-proteobacteria (**Figure 3-5**). Therefore, viral abundance did not correlated with host abundance except for a few pairs such as Bacteroidetes and their viruses (**Figure 3-7**). As mentioned above, the host taxa which were predicted for each viral contig were phylum- or class-level but not species- or population- level in most cases. Thus, the lack of a tight correlation between viral and host abundance may not be surprising. Further, still nearly 40% of contigs were not assigned their host and it may cause the underestimation of viruses infecting some taxa. Difference of burst sizes among viruses, which have been estimated to range 6 to 300 in marine environment (Parada *et al.*, 2006), also can influence on the estimation of viral abundance. Therefore, to investigate whether viruses increased according to the its specific host abundance, I next statistically examined associations (i.e. co-occurrence) between nearly strain-level populations extracted from the abundant 73 prokaryotic OTUs and the viruses.

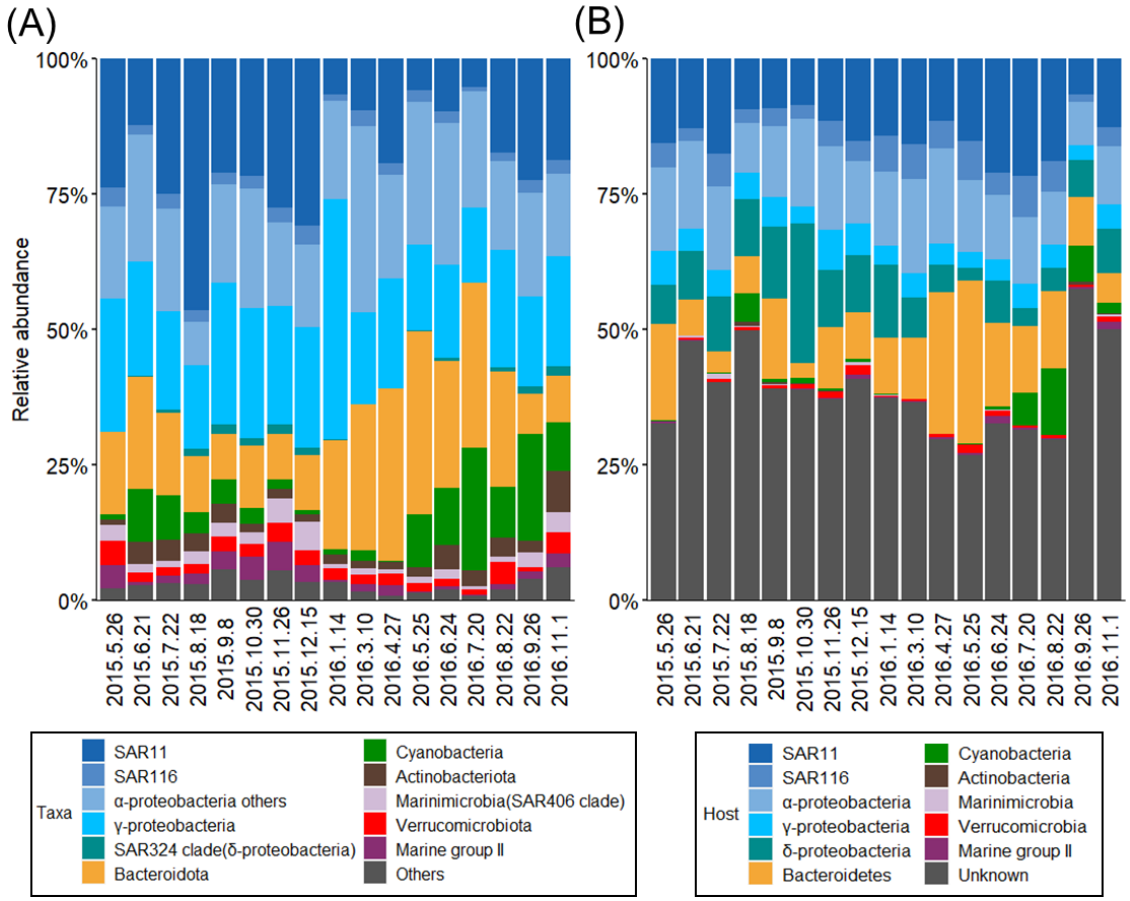

Figure 3-5. Comparison of prokaryotic and viral taxonomic community composition based on the host prediction.

(A) Relative abundance of phylogenetic groups of prokaryotic communities. Quality-controlled reads were clustered into OTUs with sequence identity of 99% using VSEARCH (Rognes et al., 2016). These OTUs were classified at the phylum level (class level for Proteobacteria) using SINA (Pruesse et al., 2012).

(B) Relative abundance of viruses based on their putative hosts assigned by host prediction. Normalized abundances of viral contigs were calculated from fragments per kilobase of per million reads mapped (FPKM) value.

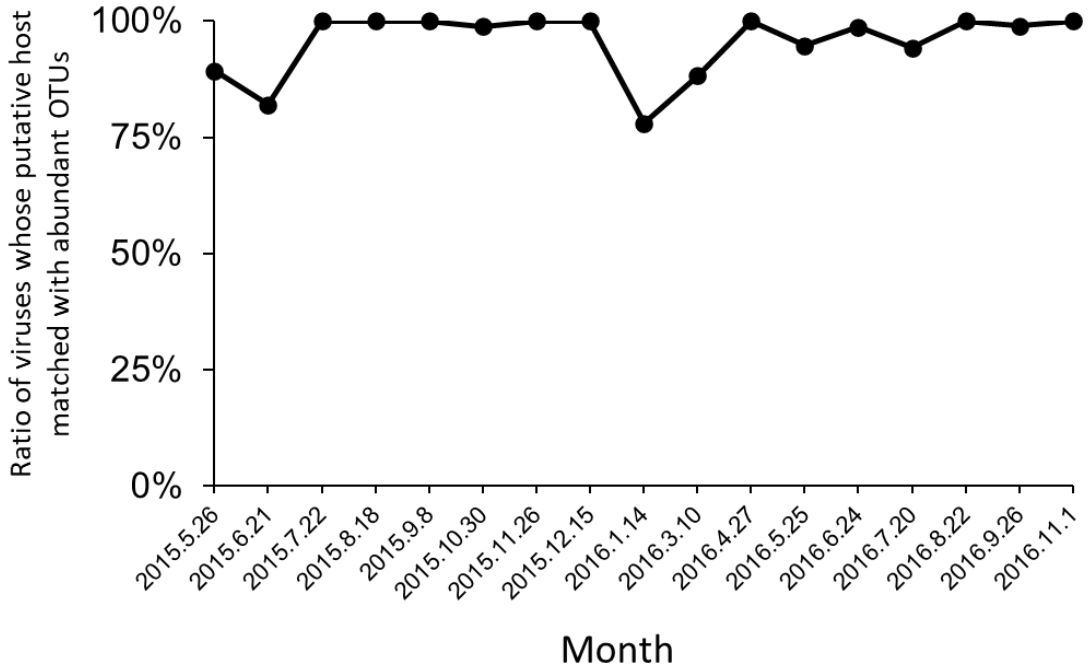

Figure 3-6. Ratio of viruses possibly infects abundant OTUs based on the assigned host group.

The plot showing the sum of the relative abundance of viral populations which possibly infects abundant OTUs of the month (i.e. viruses whose host group matched with the taxa of abundant OTUs) among the viruses which assigned their putative hosts.

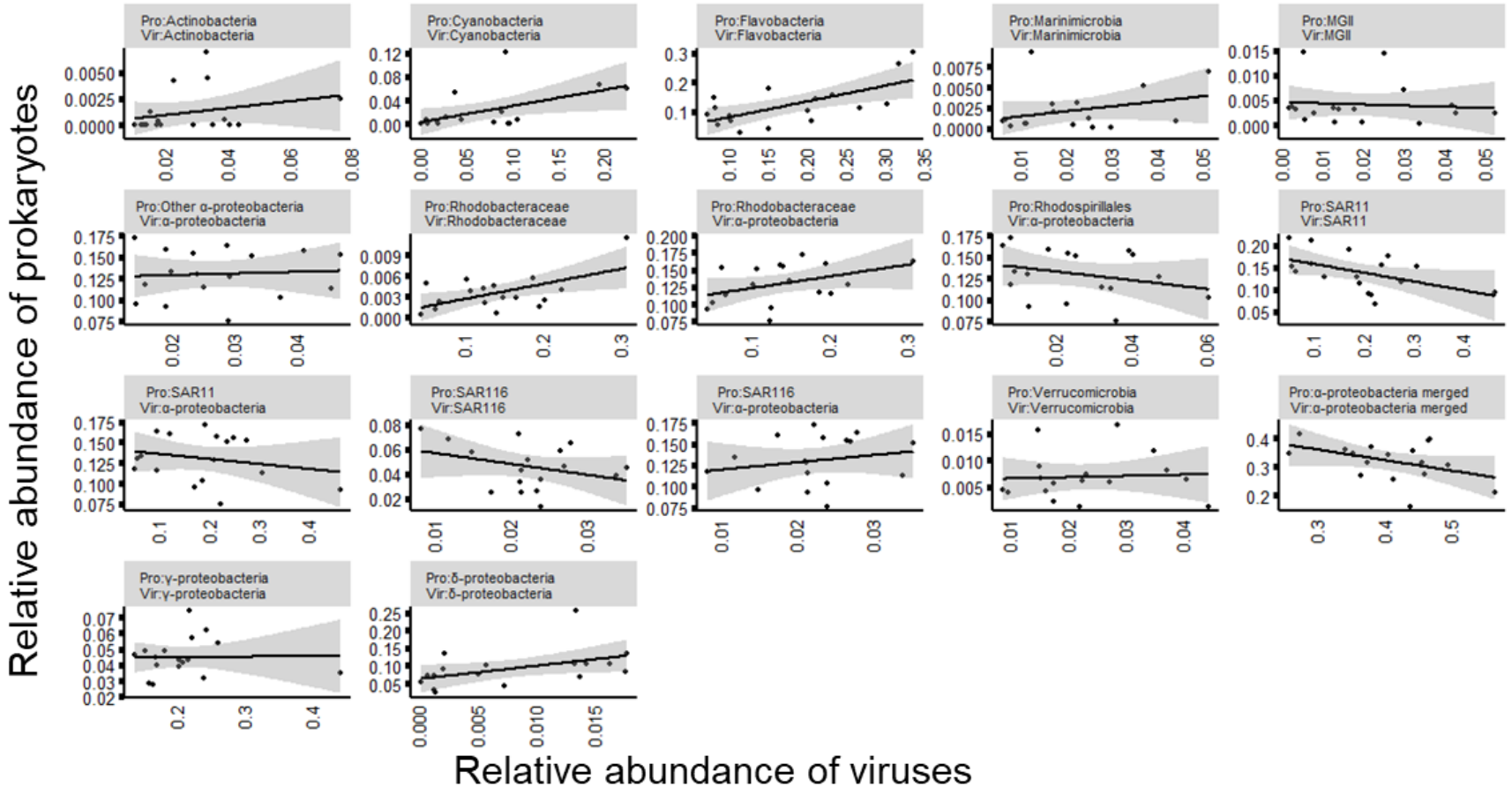

Figure 3-7. Relationship of relative abundance of prokaryotic taxa and viruses predicted to be infect to the corresponding prokaryotic taxa.

x-axis indicate relative abundance of viruses at each month. y-axis indicate relative abundance of prokaryotes at corresponding month. Pro indicate the prokaryotic taxa and Vir indicate putative host of the viruses.

Table 3-6. Table of taxonomic descriptions of abundant prokaryotic OTUs .

OTU_ID	Pylum	Class	Order	Family	Genus
OTU_7	Thermoplasmatota	Thermoplasmata	Marine Group II		
OTU_3	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	HIMB11
OTU_5	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Ascidiaceihabitans
OTU_6	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Planktomarina
OTU_13	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	
OTU_11	Proteobacteria	Alphaproteobacteria	Rhodospirillales	AEGEAN-169marine group	
OTU_1	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade I	Clade Ia
OTU_4	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade II	
OTU_184	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade I	Clade Ia
OTU_5825	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade I	Clade Ia
OTU_8783	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade I	Clade Ia
OTU_100823	Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade I	Clade Ia
OTU_33	Proteobacteria	Gammaproteobacteria	Cellvibrionales	Porticoccaceae	SAR92 clade
OTU_2	Proteobacteria	Gammaproteobacteria	SAR86 clade		
OTU_12	Proteobacteria	Gammaproteobacteria	SAR86 clade		
OTU_28	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio
OTU_15	Marinimicrobia (SAR406 clade)				
OTU_8	Cyanobacteria	Cyanobacteriia	Synechococcales	Cyanobiaceae	Synechococcus CC9902

Table 3-6. Continued.

OTU_51	Cyanobacteria	Cyanobacteriia	Synechococcales	Cyanobiaceae	Synechococcus CC9902
OTU_9	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Formosa
OTU_14	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS4 marine group
OTU_20	Bacteroidota	Bacteroidia	Flavobacteriales	Cryomorphaceae	uncultured
OTU_23	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	
OTU_35	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Ulvibacter
OTU_32	Actinobacteriota	Acidimicrobiia	Actinomarinales	Actinomarinaceae	Candidatus Actinomarina
OTU_29	Verrucomicrobiota	Verrucomicrobiae	Verrucomicrobiales	Rubritaleaceae	Persicirhabdus
OTU_21	SAR324 clade(Marine group B)				
OTU_18	Proteobacteria	Alphaproteobacteria	Puniceispirillales	SAR116 clade	
OTU_16	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	uncultured
OTU_19	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	uncultured
OTU_22	Proteobacteria	Gammaproteobacteria	Burkholderiales	Methylophilaceae	OM43 clade
OTU_53	Proteobacteria	Gammaproteobacteria	Cellvibrionales	Halieaceae	OM60(NOR5) clade
OTU_68	Proteobacteria	Gammaproteobacteria	SAR86 clade		
OTU_6445	Proteobacteria	Gammaproteobacteria	Thiomicrospirales	Thioglobaceae	SUP05 cluster
OTU_10	Proteobacteria	Gammaproteobacteria	Thiomicrospirales	Thioglobaceae	SUP05 cluster
OTU_58	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio
OTU_102	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio
OTU_83	Cyanobacteria	Cyanobacteriia	Synechococcales	Cyanobiaceae	Cyanobium PCC-6307
OTU_24	Bacteroidota	Bacteroidia	Flavobacteriales	NS9 marine group	

Table 3-6. Continued.

OTU_25	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Polaribacter
OTU_40	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS5 marine group
OTU_42	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS5 marine group
OTU_44	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS4 marine group
OTU_45	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	
OTU_50	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS5 marine group
OTU_133	Verrucomicrobiota	Verrucomicrobiae	Opitutales	Puniceicoccaceae	MB11C04 marine group
OTU_17	Thermoplasmatota	Thermoplasmata	Marine Group II		
OTU_27	Proteobacteria	Alphaproteobacteria	Parvibaculales	PS1 clade	
OTU_62	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	
OTU_69	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Ascidiaceihabitans
OTU_357	Proteobacteria	Alphaproteobacteria	Rhodospirillales	AEGEAN-169 marine group	
OTU_127	Proteobacteria	Gammaproteobacteria	Burkholderiales	Methylophilaceae	OM43 clade
OTU_71	Proteobacteria	Gammaproteobacteria	Cellvibrionales	Porticoccaceae	SAR92 clade
OTU_93	Proteobacteria	Gammaproteobacteria	Cellvibrionales	Halieaceae	OM60(NOR5) clade
OTU_89	Proteobacteria	Gammaproteobacteria	Thiomicrospirales	Thioglobaceae	SUP05 cluster
OTU_126	Cyanobacteria	Cyanobacteriia	Synechococcales	Cyanobiaceae	Cyanobium PCC-6307
OTU_34	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Formosa
OTU_57	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Polaribacter
OTU_60	Bacteroidota	Bacteroidia	Flavobacteriales	Cryomorphaceae	uncultured
OTU_75	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS5 marine group
OTU_78	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS4 marine group

Table 3-6. Continued.

OTU_101	Bacteroidota	Bacteroidia	Flavobacteriales	Cryomorphaceae	Uncultured
OTU_112	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Winogradskyella
OTU_114	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS3a marine group
OTU_148	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	
OTU_168	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	uncultured
OTU_211	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS4 marine group
OTU_213	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Formosa
OTU_112397	Bacteroidota	Bacteroidia	Flavobacteriales	Flavobacteriaceae	NS4 marine group
OTU_46	Bacteroidota	Bacteroidia	Sphingobacteriales	NS11-12 marine group	
OTU_55	Actinobacteriota	Acidimicrobiia	Actinomarinales	Actinomarinaceae	Candidatus Actinomarina
OTU_137	Actinobacteriota	Acidimicrobiia	Actinomarinales	Actinomarinaceae	Candidatus Actinomarina
OTU_41	Actinobacteriota	Acidimicrobiia	Microtrichales	Microtrichaceae	Sva0996 marine group

Co-occurrence network analysis between the abundant prokaryotes and viruses

To examine the dynamics of closely related (nearly strain-level) prokaryotic populations, 114 intraspecies-level populations (ASVs, 1~4 ASVs per OTU, **Figure 3-8**) were extracted from the abundant 73 OTUs by minimum entropy decomposition method (Eren *et al.*, 2013, 2015; Needham *et al.*, 2017). Then, pairwise correlations (co-occurrence network) between the 114 ASVs and the viral contigs which was predicted to infect the ASV by host prediction (e.g. 37 Bacteroidetes ASVs and 548 mts-OBV contigs predicted as Bacteroidetes virus) were determined via Spearman's correlations. In total, 6,423 significant correlation comprised 104 ASV and 1,366 mts-OBV contigs (**Figure 3-9**). The majority (88.6%) of ASVs correlated with at least an mts-OBV contig. The number of co-occurring viral contigs ranged from 0 contig (13 ASVs) to 359 contigs (ASV6-1, classified into *Planktomarina*) and the median value was 16 contigs.

Using the detected 6,423 putative virus-host pairs, I examined whether the viruses were abundant when its putative host was abundant. Firstly, cyanobacterial 4 ASVs and co-occurring 130 cyanoviral contigs were examined for this issue. Since substantial numbers of *Synechococcus*/*Prochlorococcus*-viruses pairs have been reported on the culture-based studies (Suttle and Chan, 1993; Waterbury and Valois, 1993; Sullivan *et al.*, 2003, 2005), host prediction for cyanoviruses most likely to be reliable. Relative rank (from 0 to 1) of the cyanobacterial ASVs in prokaryotic community and relative rank of the co-occurring cyanoviral contigs in viral community were compared to each other at the months each ASV was the most abundant and the least abundant (**Figure 3-10**). When cyanobacteria ASVs were the most abundant (average 0.98), their co-occurring viruses dominated in viral community (average 0.77) (**Figure 3-10**). In contrast, at the month the cyanobacteria ASVs were the least abundant (average 0.019),

the co-occurring viruses were less abundant (average 0.16, **Figure 3-10**). This viral increase with host abundance were observed in 98 other ASVs and their co-occurring viruses (**Figure 3-10**). The results clearly indicated frequency-dependent viral infection is prevailed in abundant prokaryotic populations at least between the detected virus-host pairs.

Characterization of virus-host interaction manner by host taxa and host growth strategy

As observed in higher alpha-diversity in viral community than prokaryotic community (**Figure 3-2**), the co-occurring pairs were often observed between a host population with multiple viral populations (median 16 viral contigs per ASV). This suggests an abundant prokaryotic population can interact with multiple viral contigs. Note that the numbers of co-occurring viruses were overestimated since each contig might be a partial genome fragment derived from a same viral genome. However, the viruses classified into different genera (average 8 gOTUs, up to 24 gOTUs) often co-occurred with an ASV. Next, I characterized the "one to many" virus-host interaction manner (i.e. how many viruses co-occurred with each ASV) by their host taxa and host growth strategy.

Number of co-occurring contigs with each ASV was dependent on the predicted number of their viruses by host prediction (**Figure 3-11**). For example, Bacteroidetes viruses (548 contigs) were the second most frequently observed virus and average 71.5 viruses co-occurred with a Bacteroidetes ASVs (1-208 viruses per ASV, between 37 Bacteroidetes ASVs and 339 Bacteroidetes viral contigs). In contrast, the taxa whose viruses were less frequently detected (e.g. MGII, 38 viruses) had a smaller number of co-occurring contigs (0 to 3 contigs per ASVs, **Figure 3-11**). Thus, numbers of co-occurring viruses might be underestimated in such taxa because of the limitation of host prediction.

Exceptionally, SAR11 had relatively few co-occurring contigs even though there were more than 500 putative SAR11 viral contigs (**Figure 3-11**). SAR11 is often regarded as a *K*-strategist which has been believed to be resistant to viral infection (Suttle, 2007) and the growth strategy may influence the co-occurrence dynamics with viruses. Next, I examined the number of co-occurring viruses among ASVs classified in the same taxa by growth strategy to solve this issue.

The growth strategy of each ASV was evaluated by the approximation indexes for *r* and *K* (see methods). According to the indexes, 13 ASVs were determined as *K*-strategist like populations (i.e. *K*-index>12, *r*-index< 0.1). Among the 13 ASVs, 7 ASVs were classified into SAR11 (**Figure 3-12**). Twenty two of 57 ASVs belonging to the taxa previously predicted as r-strategist (e.g. *Flavobacteriaceae*, *Rhodobacteraceae*, *Vibrio*, and Marine Group II) were classified into the *r*-strategist like ASVs (*K*-index<3, *r*-index>0.5, total 33 ASVs) (**Figure 3-12**). Generally, *r*-strategist like ASVs such as members of Bacteroidetes often had large number of co-occurring viruses (**Figure 3-13**). The ASVs of Bacteroidetes and *Rhodobacteraceae* showing relatively *K*-strategist like dynamics also often accompanied with large number of co-occurring viruses (e.g. ASV9-1 classified into *Formosa* had 165 co-occurring contigs, *K*-index=15 and *r*-index = 0.139, **Figure 3-12**). On the other hands, *K*-strategist like population of *Synechococcus* and SAR11 had relatively few co-occurring viruses (**Figure 3-12**). The most abundant ASVs of *Synechococcus* (ASV8-1, occupied 76.7% of whole cyanobacterial reads) and SAR11 (ASV1-1, occupied 7-64% of whole SAR11 reads of each month) had 7 and 16 co-occurring viruses, respectively, even though there were 183 cyanoviruses and 500 SAR11 viruses during observation **(**c.f. maximum 100 co-occurring viruses with ASV-83-1 classified into *Synechococcus*, **Figure 3-11)**.

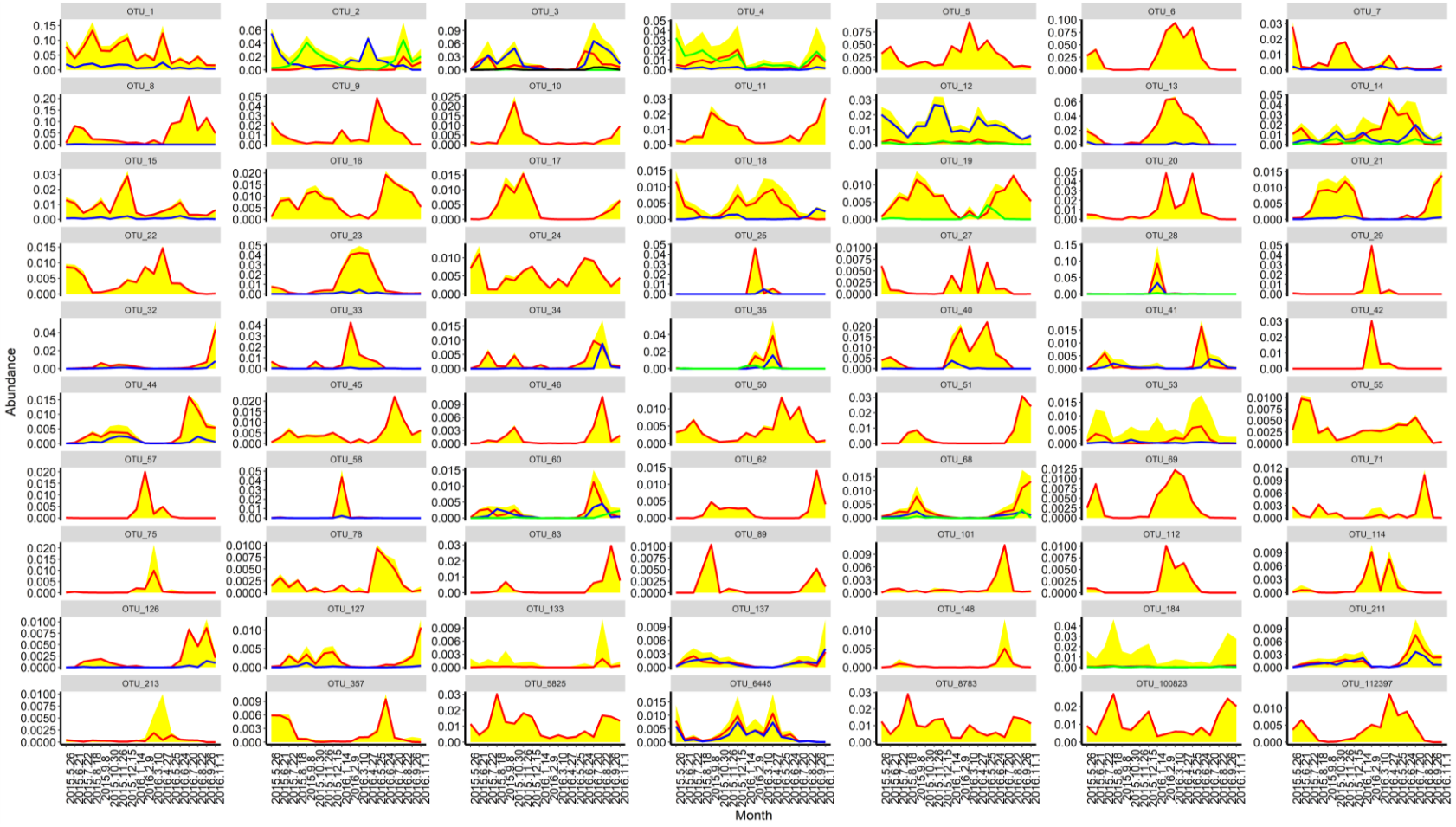

Figure 3-8. Dynamics of abundant prokaryotic OTUs and its decomposed ASVs. The yellow area-graph represents the relative abundance over time of each abundant OTU as a proportion of the whole community. The colored lines are the estimated relative abundance of each ASV (only >0.1% in abundance among whole community are shown) as a proportion of the whole community of prokaryotic sequences.

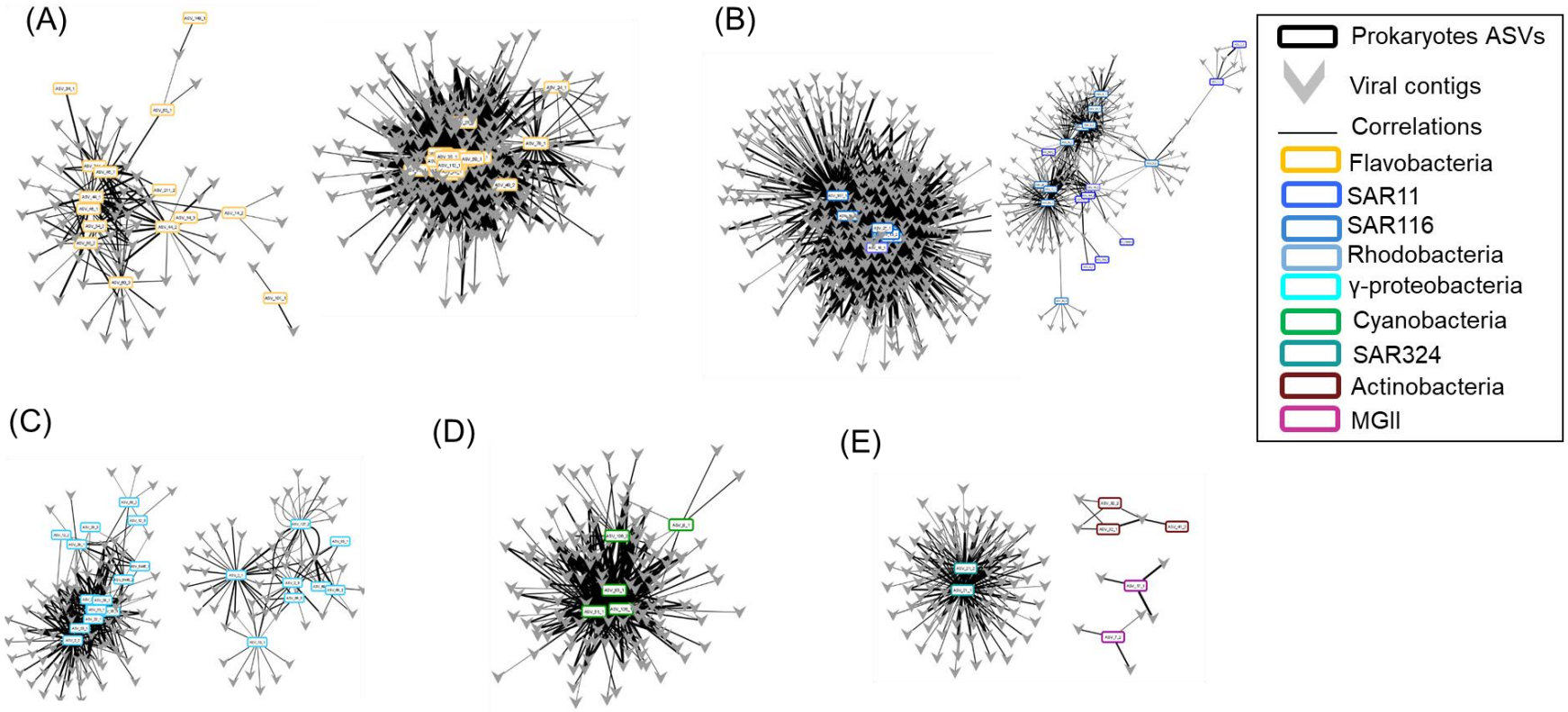

Figure 3-9. Broad overview of detected positive correlations between prokaryotic ASVs and viral populations which potentially infect each prokaryotic taxa based on host prediction analysis.

(A) Flavobacteria and their viruses. (B) α-proteobacteria and viruses. (C) γ-proteobacteria and their viruses. (D) Cyanobacteria and their viruses. (E) Other major groups (SAR324, Marine group II, and Actinobacteria) and their viruses. Prokaryotic nodes are circles and viral node are v-shapes. Node color indicates prokaryotic taxa. Solid lines are positive correlations.

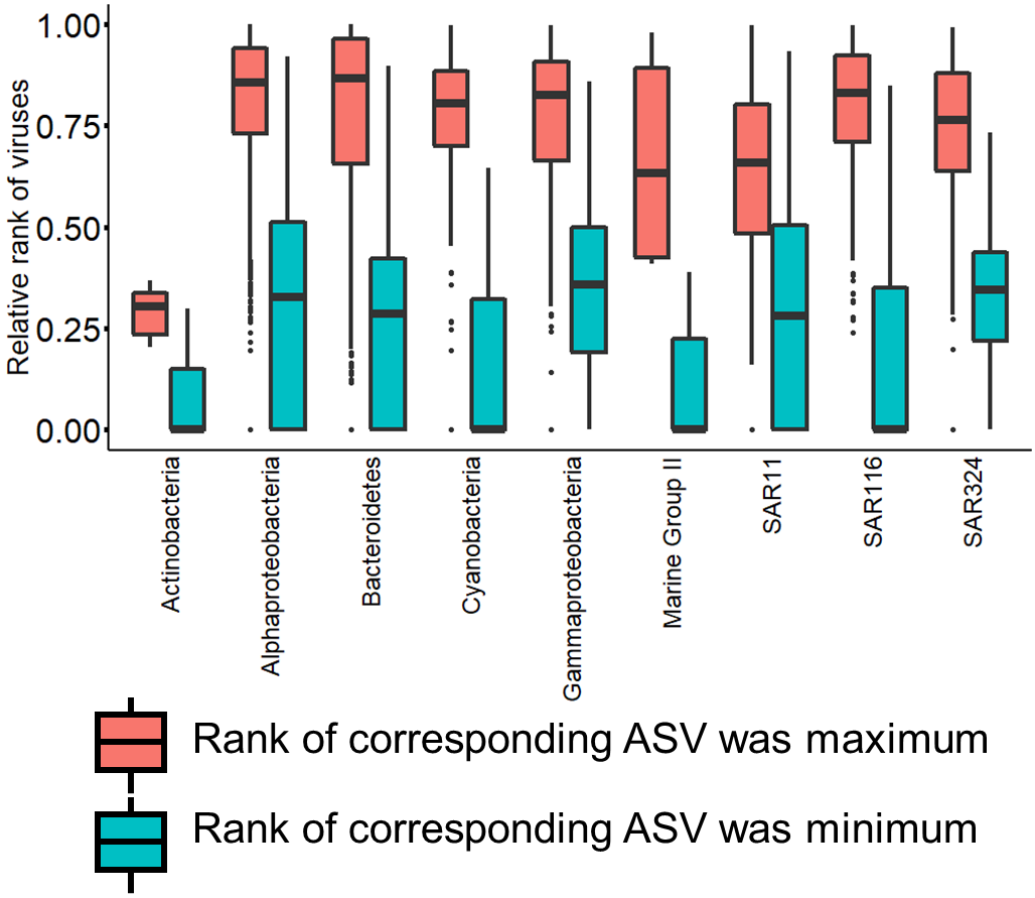

Figure 3-10. **Comparison of relative rank of co-occurring host-virus pairs when the host was the most abundant and the least abundant during observation.**
Relative rank of viruses during the month when the corresponding prokaryotic ASVs ranked maximum (red) or minimum (blue) during the sampling periods. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles; outliers are displayed as points.

If the temporal switching of virus-ASV pairs occurred, some pairs were omitted by the co-occurrence analysis. Therefore, I compared dynamics of these two ASVs and viruses which did not co-occur with other ASVs. Among the 53 viruses that did not co-occurred with any cyanobacterial ASV, 41 viruses were classified into 2 gOTUs (G14; T7-like cyanosiphovirus, and G386; T4-like cyanomyovirus) known to infect subcluster 5.1a (e.g. *Synechococcus* sp. WH 8103, clade II). Representative sequence of ASV8-1

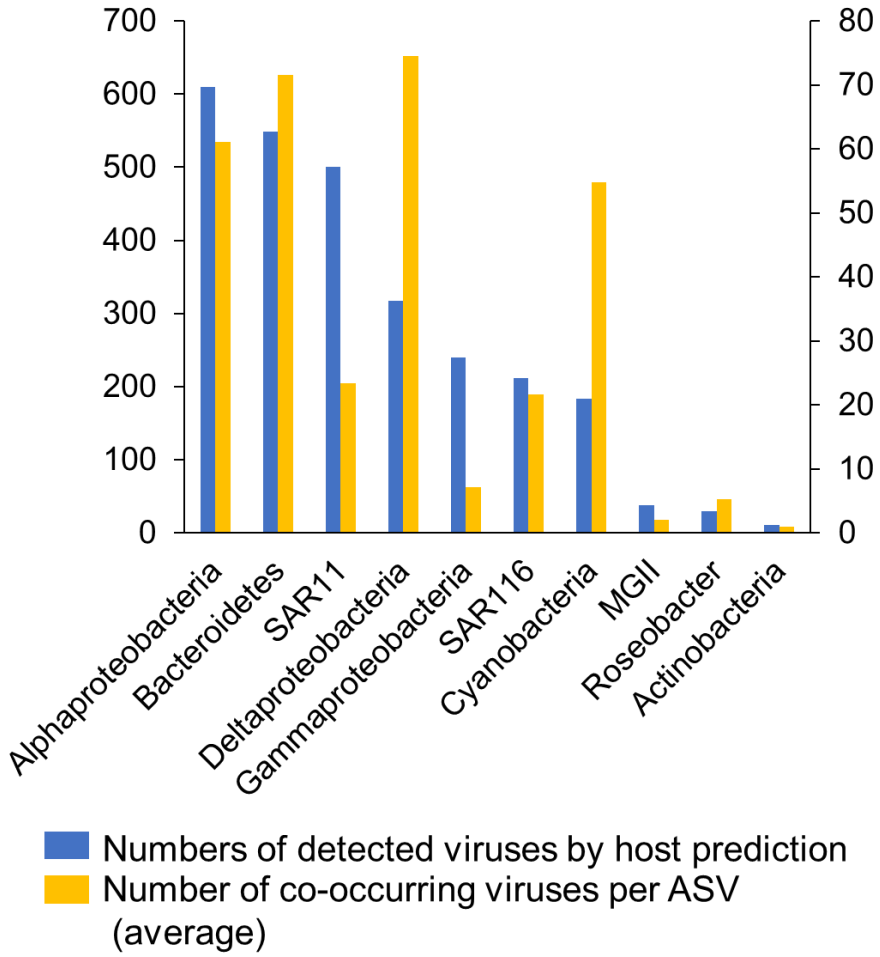

Figure 3-11. **Number of virus-host co-occurring pairs by taxa.**
Number of detected viruses by host prediction of each host taxa were shown as blue (first y-axis) and number of co-occurring viruses per an ASV (on average) by host taxa were show as yellow (second y-axis).

matched with the members of *Synechococcus* subcluster 5.1a with 100% of identity, suggesting interaction between ASV8-1 and these viruses. ASV8-1 especially dominated during summer (maximum 8% and 21% of prokaryotic community at June 2015 and July 2016, respectively, **Figure 3-14A**). Of these 53 cyanoviruses, which also increased in summer, 4 viruses were only abundant in 2015 (from 5 to > 170 times more abundant in 2015 than 2016) and other 38 viruses were only abundant in 2016 (from 5 to >300 times more abundant in 2016 than 2015) **(Figure 3-14B)**. Similarly, ASV1-1 of SAR11was

always abundant (**Figure 3-15A**) and SAR11 viruses occupied major fraction of the viral community. However, abundant member of SAR11 viruses were replaced in relatively short time-period (a few month) (**Figure 3-15B**). These results suggest that the host-virus interaction might have been underestimated by the co-occurrence analysis and *K*-strategist also interact with multiple viruses based on the their cell density.

Finally, I investigated whether the observed viruses including ones co-occurring with hosts (i.e. 53 cyanoviruses and 309 SAR11 viruses) were produced via increased contact frequency with hosts. A previous study suggested that the majority of viral genomes detected in viromes were derived from virions that were daily produced through local viral–host interactions (Yoshida *et al.*, 2018). Thus, if variants (i.e. single-nucleotide variants) within the abundant viruses were observed, it likely indicate multiple infection events (DNA replication) via increased contact frequency of host-viruses accompanying stochastic mutation, rather than the mixture of persistently existed virions. Therefore, to examine the variants, single-nucleotide polymorphisms (SNPs) from mts-OBV contigs with more than 10 coverage (2,356 contigs) were calculated. I observed that increase of intrapopulation genetic diversity (SNPs quantified by average genomic entropy) as a function of overall population abundance regardless of their hosts (**Figure 3-16**). This suggests that occurrence of frequent reproduction (replication) in abundant viruses, thus the increase of contact frequency with hosts and abundant viruses regardless of whether the viruses showed co-occurrence with the ASVs. The result was in agreement with previous observations: rapid diversification of freshwater cyanoviruses via increased contact frequency of host and viruses (Kimura *et al.*, 2013) and increase of SNPs in abundant populations of marine viruses in other coastal site (Ignacio-Espinoza *et al.*, 2020).

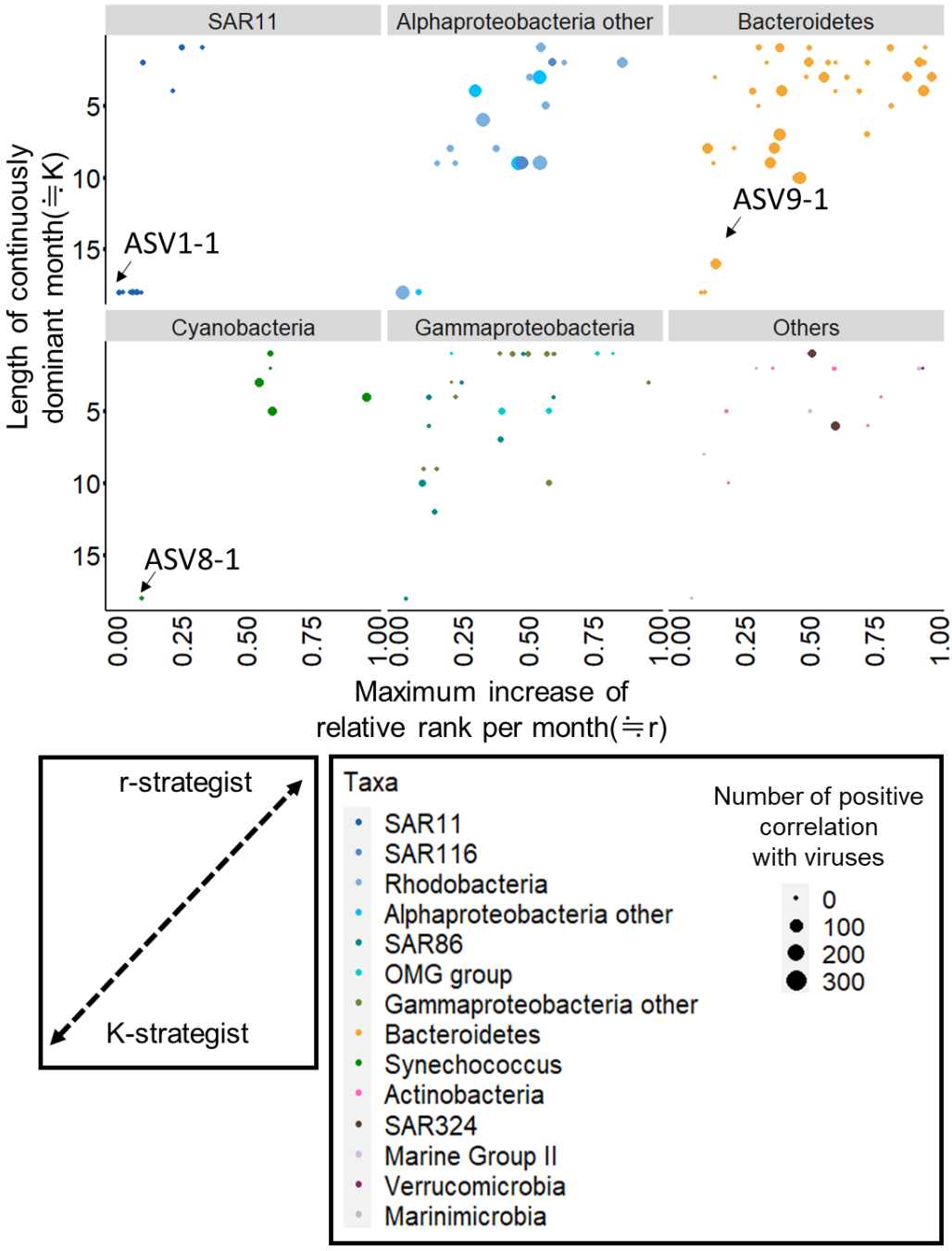

Figure 3-12. Distribution of the number of co-occurring viruses among prokaryotic ASVs based on their growth strategy inferred from approximated index of carrying capacity (*K*) and intrinsic rate of natural increase (*r*) based on their dynamics.
x-axis indicates approximation index of *r* and y-axis indicates approximation index of *K*. Size of the circles represents the number of co-occurring viruses with each ASV. Color of the circles indicate the taxa of each ASV.

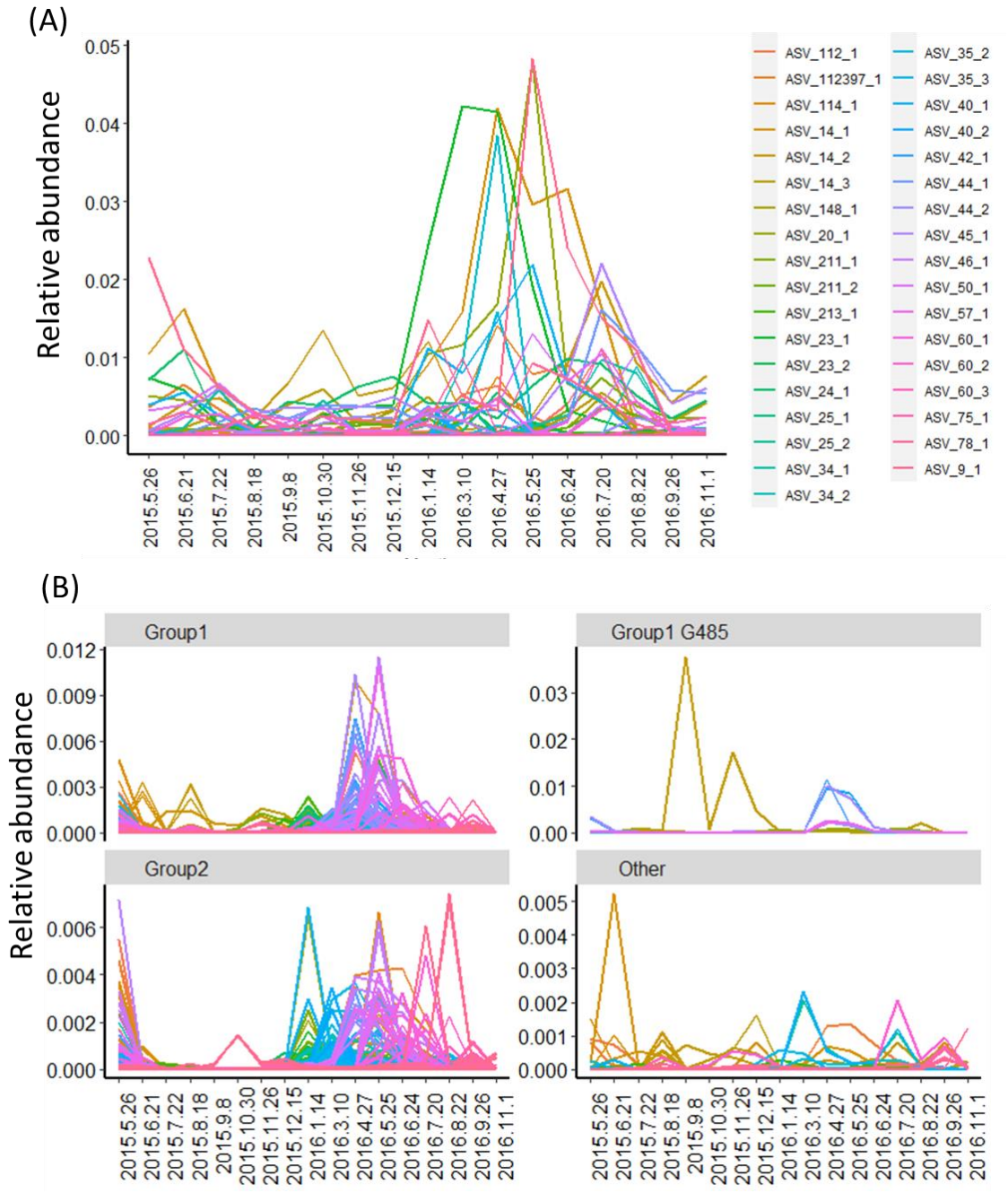

Figure 3-13. Dynamics of Bacteroidetes ASVs and co-occurring putative Bacteroidetes viruses.

(A) Dynamics of Bacteroidetes 37 ASVs and (B) Co-occurring 339 Bacteroidetes viruses (37 gOTUs). Lines represents relative abundance of the ASVs or the contigs over time. The panels of viruses were separated into four based on the classification by Nishimura *et al.* 2017 (i.e., Group 1, Group 2, and others predicted by this study, members of G485 of Group1 were separated since a contigs of G485 was highly abundant than other members of Group 1).

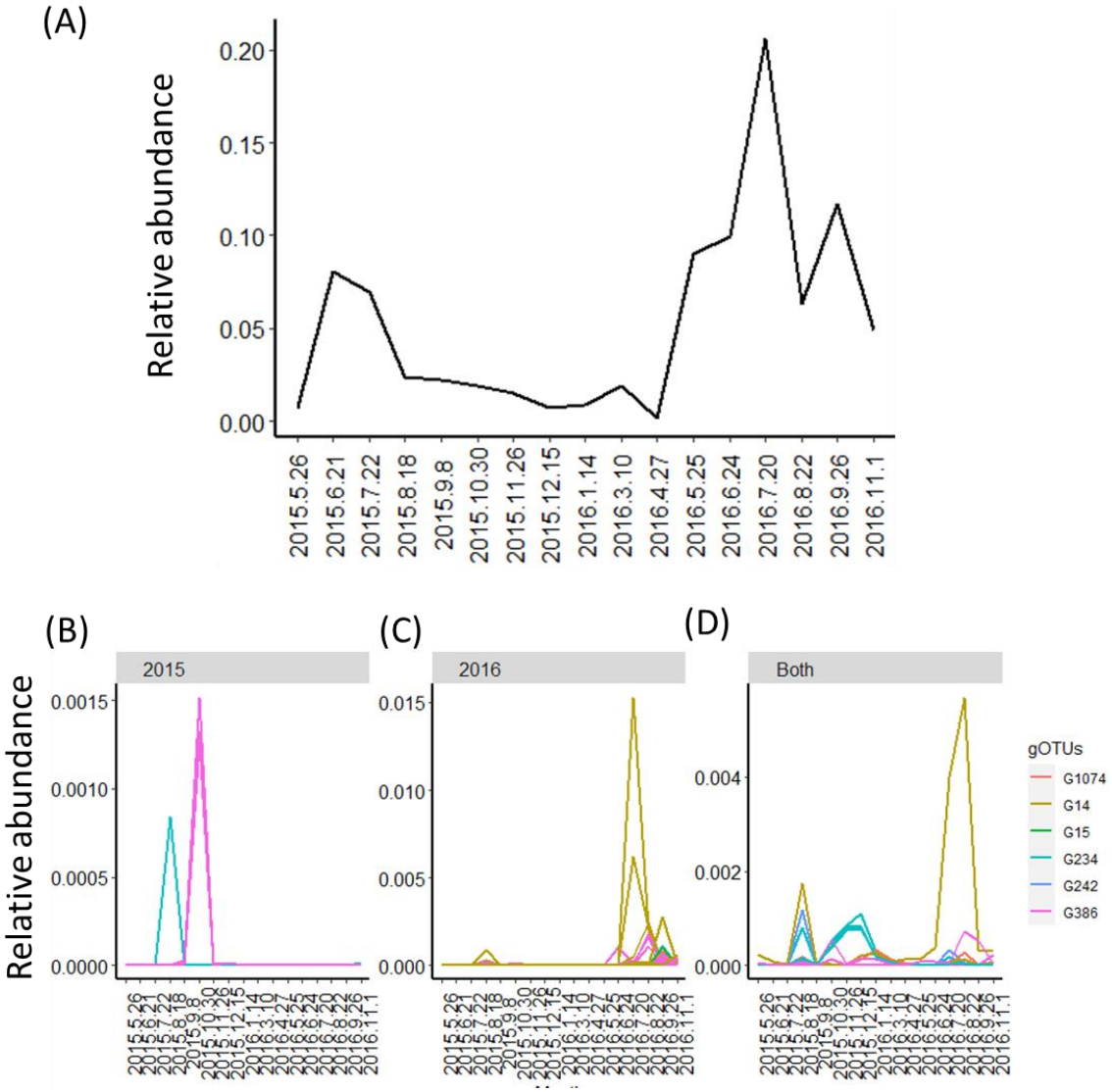

Figure 3-14. Dynamics of the most dominant *Synechococcus* population (ASV8-1) with *Synechococcus* viruses which did not co-occurred with ASVs.

(A) Dynamics of ASV8-1. (B) Dynamics of 53 cyanoviruses which did not co-occurred with ASVs. Lines represents relative abundance of the ASVs or the contigs over time. The panels of viruses were separated by their annual pattern (2015 type, 2016 type, and both years, if the virus was more than five times abundant in one year comparing with another year, the virus was defined as year-specific virus). Colors represent gOTU of the virus.

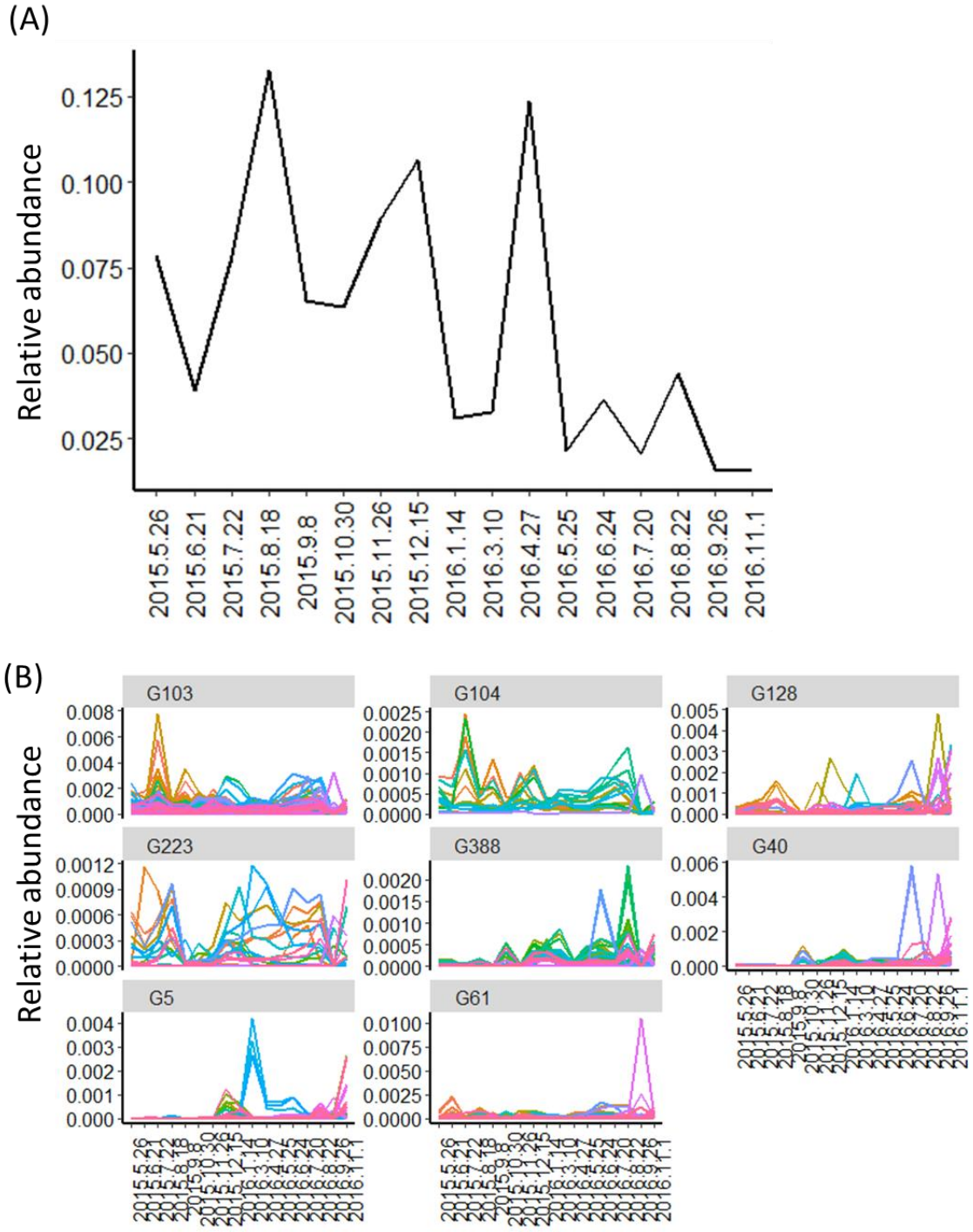

Figure 3-15. Dynamics of the most dominant SAR11 ASV(ASV1-1) with SAR11 viruses which did not co-occurred with ASVs.

(A) Dynamics of ASV1-1. (B) Dynamics of 309 putative SAR11 viruses co-occurred with ASVs. Lines represents relative abundance of the ASVs or the contigs over time. The panel were separated based on the classified gOTUs of each virus.

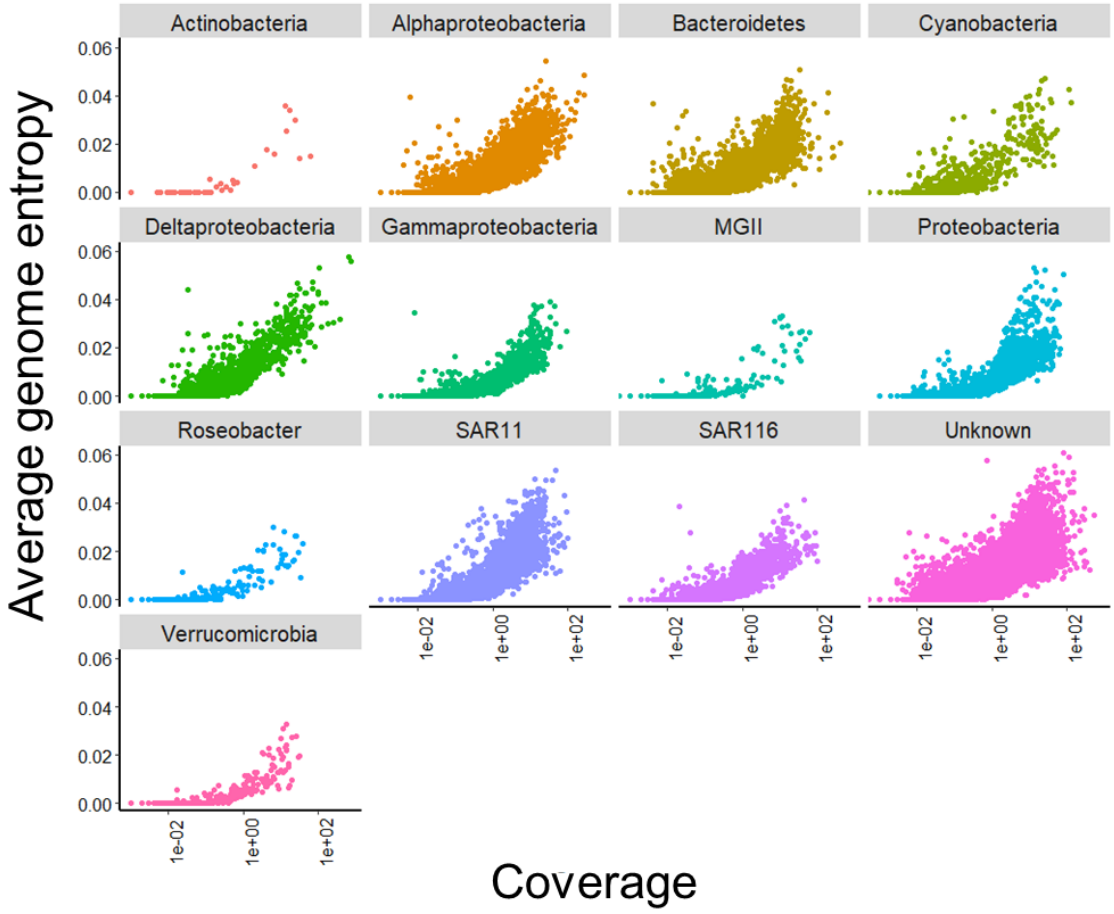

Figure 3-16. Correlation of genome average entropy and abundance of OBV contigs calculated from SNP profiles.

The graphs show the average genomic entropy of mts-OBV contigs and read coverage of the mts-OBV contigs at given time-series samples. The panel were separated based on the predicted hosts of the mts-OBV contigs.

There are three possible mechanisms of the temporal switching of virus-host pairs . First, it can be interpreted as a result of founder effects following the host fluctuation via genetic drift (Cohan and Perry, 2007). Seasonal fluctuating of host population cause bottleneck, and thereby founder effects following bottleneck enable that several viral species have an equal chance of increasing. This was suggested as a mechanism of incomplete selective sweep in the fresh water cyanobacteria populations

having different CRISPR-spacer genotype (Kimura *et al.*, 2018). The scenario is more plausible between ASV8-1 and their viruses because the ASV8-1 experienced clear seasonal fluctuation (**Figure 3-14A**). Second, the temporal acquisition of host resistance or viral counter-resistance as observed in culture model systems (Koskella and Brockhurst, 2014) may cause the switching the dominant viral species. Third, more closely related populations which undiscriminated by the polymorphism of 16S rRNA genes might have been included in each ASV. Previous studies focusing the polymorphism of ITS sequences (ITS-ASV) in SAR11 and cyanobacteria reported that dynamics of ITS-ASVs more correlated with viral dynamics which inferred from T4-like viral marker genes than dynamics of 16S-ASVs of these taxa (Needham *et al.*, 2017; Ahlgren *et al.*, 2019). Therefore, dynamics of more closely related populations within these 16S-ASVs (e.g. ITS-ASVs or whole genome sequence based-populations) also might have synchronized with observed viral dynamics. The idea in agreement with the revised version of Kill the winner hypothesis, which assumes strain level diversity in a host species and overall viral abundance infecting the host species were regarded as the sum of the viruses interacting each strains (Thingstad *et al.*, 2014).

Altogether, I revealed that the frequency dependent infection was occurred in abundant prokaryotic population according to the cell density with “one-to-many” manner regardless of host growth strategy. The “one-to-many” manner may suggest a prokaryotic cell attacked by multiple viruses having different infection strategy (e.g. different cell surface targets). This can cause the difficulty to establish complete resistance toward multiple co-existing viruses and sustain continuous virus-host interaction in environments, Therefore, the “one-to-many” manner may be a potential

mechanism for the prevailed frequency-dependent selection on abundant marine prokaryotes.

Conclusions

Comparison of seasonal dynamics between abundant prokaryotes and their viruses revealed that abundant prokaryotes were exposed by frequent viral infection regardless of their taxa or growth strategy. This suggests that lysis of the abundant prokaryotes via viral infection have a considerable contribution to the biogeochemical cycling and maintenance of prokaryotic community diversity. Further, these abundant prokaryotic populations should reflect actively growing members of community since they were able to become dominant even though they suffer frequent loss by viral lysis.

Chapter 4

Integration and outlook

In the oceans, viruses infecting prokaryotes affect the marine biogeochemical cycle through host lysis. Since viral infection is believed to be dependent on host cell density, viral infection should have larger influence on the abundant prokaryotic populations. However, it was unclear whether the host frequency-dependent viral infection occurred in complex prokaryotic community because of the enormous diversity of marine viruses.

In chapter 2, I focused on the interactions between marine Bacteroidetes and their virus as a model systems of abundant prokaryotes-virus pairs. I developed efficient methods for the prediction of viruses infecting Bacteroidetes from a thousand uncultured viral genomes by using recently reported metagenome assembled genomes of marine Bacteroidetes. I successfully identified novel 81 viral species from 26 genera, including the marine dominant viral linage Far-T4. The methods enhanced the existing knowledge on the diversity of Bacteroidetes viruses and their potential interaction with their hosts in marine environments.

In chapter 3, I applied the host prediction methods developed in chapter 2 to other prokaryotic taxa. To examine whether the frequency-dependent viral infection occurred in natural community, I compared the seasonal dynamics of the abundant prokaryotes and their viruses in Osaka Bay. Increasing of viral abundance in response to their host abundance was observed between more than 6,000 of putative virus-host pairs. Further, the faster temporal change of the viral community than the prokaryotic community suggested that the viruses interacting with continuously dominant prokaryotic

population might have changed temporally. These results revealed that abundant prokaryotes were infected by the viruses with frequency-dependent manner regardless of their taxa and survival strategy.

In these studies, I revealed a general trend that viral frequency-dependent infection is prevailed in abundant prokaryotes. The finding supports that frequency-dependent viral infection maintains the diversity of the prokaryotic community and the active recycling of organic matters in marine environment. Future works focusing on their co-evolutional dynamics from the observation of host and viral genome co-diversification will also help us to understand the underlying mechanism to establish the continuance interactions between abundant prokaryotes with their viruses.

Acknowledgements

First of all, I am profoundly grateful to Professor Takashi Yoshida for giving me a chance to study in this field, insightful comments, suggestions, and encouragement. I would like to express my gratitude to Honorary Professor Yoshihiko Sako and Associate Professor Ryoma Kamikawa for spending a lot of time to discuss with me and supporting me patiently. I am also grateful to Professor Shigeki Sawayama for thoughtful words and reviewing this dissertation.

I would like to express my deep appreciate to Professor Hiroyuki Ogata, Professor Keizo Nagasaki, and Dr. Hisashi Endo for their many grateful suggestions and comments. Furthermore, I would like to show my grateful appreciate to Dr. Daichi Morimoto, Dr. Yosuke Nishimura, Dr. Sigitas Sulcius, Mr. Florian Prodinger and Mr. Yoshiaki Sato who is one of the respectable researchers and great collaborators. Their approaches and attitudes on the research works encouraged me greatly and changed my way of thinking about the studies.

I thank all my colleagues in Laboratory of Marine Microbiology for their support and encouragement. Especially, I give a special thanks to Dr. Yuto Fukuyama, Dr. Kimiho Omae, and Mr. Hiroaki Takebe for valuable discussion, technical advices and mutual help to each other. I am also deeply grateful to Dr. Shigeko Kimura, and Dr. Masao Inoue for giving me helpful comments and thoughtful attention. Many thanks for Ms. Nana Haruki, Mr. Tatsuhiro Isozaki, Mr. Kentaro Fujiwara, Mr. Kohei Nishimon, Mr. Rei Hoshino, Mr. Tomoya Suzuki, Mr. Mao Matsumoto, Mr. Kentaro Yoshikawa, Mr. Shuto Ashizawa, Mr. Ryutaro Yamada and Ms. Tomoka Ashitani for their supports. We have discussed and done several experiments together. Last, I appreciate sincerely to my parents for giving me all supports and encouragements.

My doctoral works were supported by Grants-in Aids for Scientific Research (B) (No. 17H03850) and challenging Exploratory Research (No. 26660171) from the Japan Society for the Promotion of Science (JSPS), The Canon Foundation (No. 203143100025), JSPS Scientific Research on Innovative Areas (No. 16H06437), and the Bilateral Open Partnership Joint Research Project (Japan-Lithuania Research Cooperative Program) “Research on prediction of environmental change in Baltic Sea based on comprehensive metagenomic analysis of microbial viruses”.

References

Ahlgren, N.A., Perelman, J.N., Yeh, Y., and Fuhrman, J.A. (2019) Multi-year dynamics of fine-scale marine cyanobacterial populations are more strongly explained by phage interactions than abiotic, bottom-up factors. *Environ Microbiol* **21**: 2948–2963.

Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017) Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* **45**: 39–53.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., et al. (2019) A new genomic blueprint of the human gut microbiota. *Nature* **568**: 499–504.

Alonso, C., Warnecke, F., Amann, R., and Pernthaler, J. (2007) High local and global diversity of Flavobacteria in marine plankton. *Environ Microbiol* **9**: 1253–1266.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**: 13219.

Andrews, J.H. and Harris, R.F. (1986) r- and K-Selection and Microbial Ecology. Springer, Boston, MA, pp. 99–147.

Azam, F., Fenchel, T., Field, J.G., Gray, J.S., Meyer-Reil, L.A., and Thingstad, F. (1983) The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* **10**: 257–263.

Azam, F. and Malfatti, F. (2007) Microbial structuring of marine ecosystems. *Nat Rev Microbiol* **5**: 782–91.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Barnhart, M.M. and Chapman, M.R. (2006) Curli Biogenesis and Function. *Annu Rev Microbiol* **60**: 131–147.

Bergh, Ø., BØrsheim, K.Y., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467–468.

Bhunchoth, A., Blanc-Mathieu, R., Mihara, T., Nishimura, Y., Askora, A., Phironrit, N., et al. (2016) Two asian jumbo phages, ϕRSL2 and ϕRSF1, infect Ralstonia solanacearum and show common features of ϕKZ-related phages. *Virology* **494**: 56–66.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209.

Borriss, M., Lombardot, T., Glöckner, F.O., Becher, D., Albrecht, D., and Schweder, T. (2007) Genome and proteome characterization of the psychrophilic Flavobacterium bacteriophage 11b. *Extremophiles* **11**: 95–104.

Bouvier, T. and Del Giorgio, P.A. (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ Microbiol* **9**: 287–297.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017) Minimum information about a single amplified genome

(MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725–731.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018) Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**: 754–766.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., et al. (2015) Patterns and ecological drivers of ocean viral communities. *Science (80-)* **348**: 1261498–1261498.

Brum, J.R. and Sullivan, M.B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* **13**: 147–159.

Buchan, A., LeCleir, G.R., Gulvik, C.A., and González, J.M. (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* **12**: 686–698.

Bunse, C. and Pinhassi, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends Microbiol* **25**: 494–505.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Castillo, D., Espejo, R., and Middelboe, M. (2014) Genomic structure of bacteriophage 6H and its distribution as prophage in *Flavobacterium psychrophilum* strains. *FEMS Microbiol Lett* **351**: 51–58.

Castillo, D. and Middelboe, M. (2016) Genomic diversity of bacteriophages infecting the fish pathogen *Flavobacterium psychrophilum*. *FEMS Microbiol Lett* **363**: fnw272.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., and Amann, R.I. (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* **12**: 237–252.

Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**: D184–D189.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*.

Cheng, L., Chen, H., Zheng, T., Fu, G., Shi, S., Wan, C., and Huang, Y. (2012) Complete Genomic Sequence of the Virulent Bacteriophage RAP44 of Riemerella anatipestifer. *Avian Dis* **56**: 321–327.

Chow, C.E.T. and Fuhrman, J.A. (2012) Seasonality and monthly dynamics of marine myovirus communities. *Environ Microbiol* **14**: 2171–2183.

Cohan, F.M. and Perry, E.B. (2007) A Systematics for Discovering the Fundamental Units of Bacterial Diversity. *Curr Biol* **17**: R373–R386.

Cole, J.J., Findlay, S., and Pace, M.L. (1988) Bacterial production in fresh and saltwater ecosystems: a cross-system overview. *Mar Ecol Prog Ser* **43**: 1–10.

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., et al. (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**: 15955.

Cram, J.A., Parada, A.E., and Fuhrman, J.A. (2016) Dilution reveals how viral lysis and grazing shape microbial communities. *Limnol Oceanogr* **61**: 889–905.

Dai, X., Gao, X., Zhang, X.-H., and Zhang, Z. (2015) Hyunsoonleella pacifica sp. nov., isolated from seawater of South Pacific Gyre. *Int J Syst Evol Microbiol* **65**: 1155–1159.

Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S.T., Rappé, M.S., et al. (2018) Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**: 804–813.

Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**: 927–930.

Dueholm, M.S., Albertsen, M., Otzen, D., and Nielsen, P.H. (2012) Curli Functional Amyloid Systems Are Phylogenetically Widespread and Display Large Diversity in Operon and Protein Structure. *PLoS One* **7**: e51274.

Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* **40**: 258–272.

Edwards, R.A. and Rohwer, F. (2005) Opinion: Viral metagenomics. *Nat Rev Microbiol* **3**: 504–510.

Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013) Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* **4**: 1111–1119.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**: 968–979.

Esteves, A.I.S., Hardoim, C.C.P., Xavier, J.R., Gonçalves, J.M.S., and Costa, R. (2013) Molecular richness and biotechnological potential of bacteria cultured from Irciniidae sponges in the north-east Atlantic. *FEMS Microbiol Ecol* **85**: 519–536.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**: 1034–9.

Field, C.B. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science (80-)* **281**: 237–240.

Fuhrman, J. and Suttle, C. (1993) Viruses in Marine Planktonic Systems. *Oceanography* **6**: 51–63.

Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**: 133–146.

Glöckner, F.O., Fuchs, B.M., and Amann, R. (1999) Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. *Appl Environ Microbiol* **65**: 3721–6.

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al. (2019) Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**: 1109-1123.e14.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**: 465–470.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., Verberkmoes, N.C., and Sullivan, M.B. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci U S A* **110**: 12798–803.

Howard-Varona, C., Hargreaves, K.R., Abedon, S.T., and Sullivan, M.B. (2017) Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *ISME J* **11**: 1511–1520.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**: 2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., et al. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**: D309–D314.

Hurwitz, B.L., Deng, L., Poulos, B.T., and Sullivan, M.B. (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* **15**: 1428–40.

Hurwitz, B.L. and Sullivan, M.B. (2013) The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS One* **8**: e57355.

Ignacio-Espinoza, J.C., Ahlgren, N.A., and Fuhrman, J.A. (2020) Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat Microbiol* **5**: 265–271.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K.M., Kern, S., et al. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3**: 195–202.

Jover, L.F., Effler, T.C., Buchan, A., Wilhelm, S.W., and Weitz, J.S. (2014) The elemental composition of virus particles: Implications for marine biogeochemical cycles. *Nat Rev Microbiol* **12**: 519–528.

Kang, I., Jang, H., and Cho, J.-C. (2015) Complete genome sequences of bacteriophages P12002L and P12002S, two lytic phages that infect a marine Polaribacter strain. *Stand Genomic Sci* **10**: 82.

Kang, I., Jang, H., and Cho, J.-C. (2012) Complete genome sequences of two Persicivirga bacteriophages, P12024S and P12024L. *J Virol* **86**: 8907–8.

Kang, I., Kang, D., and Cho, J.-C. (2016) Complete genome sequence of bacteriophage P2559Y, a marine phage that infects Croceibacter atlanticus HTCC2559. *Mar Genomics* **29**: 35–38.

Kang, I., Kang, D., and Cho, J.-C. (2012) Complete genome sequence of Croceibacter bacteriophage P2559S. *J Virol* **86**: 8912–3.

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 Res* **30**: 3059–3066.

Keller-Costa, T., Silva, R., Lago-Lestón, A., and Costa, R. (2016) Genomic Insights into Aquimarina sp. Strain EL33, a Bacterial Symbiont of the Gorgonian Coral Eunicella labiata. *Genome Announc* **4**: 855–871.

Kimura, S., Sako, Y., and Yoshida, T. (2013) Rapid microcystis cyanophage gene diversification revealed by long- and short-term genetic analyses of the tail sheath gene in a natural pond. *Appl Environ Microbiol* **79**: 2789–95.

Kimura, S., Uehara, M., Morimoto, D., Yamanaka, M., Sako, Y., and Yoshida, T. (2018) Incomplete selective sweeps of Microcystis population detected by the leader-end CRISPR fragment analysis in a natural pond. *Front Microbiol* **9**: 425.

Kimura, S., Yoshida, T., Hosoda, N., Honda, T., Kuno, S., Kamiji, R., et al. (2012) Diurnal infection patterns and impact of Microcystis cyanophages in a Japanese pond. *Appl Environ Microbiol* **78**: 5805–5811.

Kirchman, D.L. (2002) The ecology of Cytophaga-Flavobacteria in aquatic environments. *FEMS Microbiol Ecol* **39**: 91–100.

Koskella, B. and Brockhurst, M.A. (2014) Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev* **38**: 916–931.

Krüger, K., Chafee, M., Ben Francis, T., Glavina del Rio, T., Becher, D., Schweder, T., et al. (2019) In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *ISME J* **13**: 2800–2816.

Kulikov, E.E., Golomidova, A.K., Letarova, M.A., Kostryukova, E.S., Zelenin, A.S., Prokhorov, N.S., and Letarov, A. V. (2014) Genomic sequencing and biological characteristics of a novel Escherichia Coli bacteriophage 9g, a putative representative of a new siphoviridae genus. *Viruses* **6**: 5077–5092.

Kuno, S., Yoshida, T., Kaneko, T., and Sako, Y. (2012) Intricate interactions between the bloom-forming cyanobacterium Microcystis aeruginosa and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. *Appl Environ Microbiol* **78**: 5353–60.

Laanto, E., Bamford, J.K.H., Ravantti, J.J., and Sundberg, L.-R. (2015) The use of phage FCL-2 as an alternative to chemotherapy against columnaris disease in aquaculture. *Front Microbiol* **6**: 829.

Lafontaine, E.R., Cope, L.D., Aebi, C., Latimer, J.L., McCracken, G.H., and Hansen, E.J. (2000) The UspA1 protein and a second type of UspA2 protein mediate adherence of Moraxella catarrhalis to human epithelial cells in vitro. *J Bacteriol* **182**: 1364–73.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11–16.

Lee, Y.M., Hwang, C.Y., Lee, I., Jung, Y.-J., Cho, Y., Baek, K., et al. (2014) Lacinutrix jangbogonensis sp. nov., a psychrophilic bacterium isolated from Antarctic marine sediment and emended description of the genus Lacinutrix. *Antonie Van Leeuwenhoek* **106**: 527–533.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**: W256–W259.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

Li, W. and Godzik, A. (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**:.

Locey, K.J. and Lennon, J.T. (2016) Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* **113**: 5970–5975.

Luhtanen, A.-M., Eronen-Rasimus, E., Kaartokallio, H., Rintala, J.-M., Autio, R., and Roine, E. (2014) Isolation and characterization of phage–host systems from the Baltic Sea ice. *Extremophiles* **18**: 121–130.

Luo, E., Aylward, F.O., Mende, D.R., and DeLong, E.F. (2017) Bacteriophage Distributions and Temporal Variability in the Ocean's Interior. *MBio* **8**: e01903-17.

Mahmoudabadi, G. and Phillips, R. (2018) A comprehensive and quantitative exploration of thousands of viral genomes. *Elife* **7**:.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016) Linking Virus Genomes with Host Taxonomy. *Viruses* **8**: 66.

Miyazaki, M., Nagano, Y., Fujiwara, Y., Hatada, Y., and Nogi, Y. (2010) Aquimarina macrocephali sp. nov., isolated from sediment adjacent to sperm whale carcasses. *Int J Syst Evol Microbiol* **60**: 2298–2302.

Mizuno, C.M., Ghai, R., Saghaï, A., López-García, P., and Rodriguez-Valeraa, F. (2016) Genomes of abundant and widespread viruses from the deep ocean. *MBio* **7**:.

Morrissey, J.P., Dobson, A.D.W., Jackson, S.A., O'Gara, F., and Kennedy, J. (2015) Maribacter spongiicola sp. nov. and Maribacter vaceletii sp. nov., isolated from marine sponges, and emended description of the genus Maribacter. *Int J Syst Evol Microbiol* **65**: 2097–2103.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**: 505–510.

Nedashkovskaya, O.I., Kim, S.B., Lysenko, A.M., Frolova, G.M., Mikhailov, V. V, Lee, K.H., and Bae, K.S. (2005) Description of Aquimarina muelleri gen. nov., sp. nov., and proposal of the reclassification of [Cytophaga] latercula Lewin 1969 as Stanierella latercula gen. nov., comb. nov. *Int J Syst Evol Microbiol* **55**: 225–229.

Needham, D.M., Chow, C.-E.T., Cram, J.A., Sachdeva, R., Parada, A., and Fuhrman, J.A. (2013) Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J* **7**: 1274–1285.

Needham, D.M., Fuhrman, Jed A., Cram, J.A., Fuhrman, J. A., and Sun, F. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* **1**: 16005.

Needham, D.M., Sachdeva, R., and Fuhrman, J.A. (2017) Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* **11**: 1614–1629.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., et al. (2017) Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere* **2**: e00359-16.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017) ViPTree: the viral proteomic tree server. *Bioinformatics* **33**: 2379–2380.

Okazaki, Y., Nishimura, Y., Yoshida, T., Ogata, H., and Nakano, S. (2019) Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environ Microbiol* **21**: 4740–4754.

Oliveira, G., Silva, L., Leão, T., Mougari, S., da Fonseca, F.G., Kroon, E.G., et al. (2019) Tupanvirus-infected amoebas are induced to aggregate with uninfected cells promoting viral dissemination. *Sci Rep* **9**: 183.

Pachiadaki, M.G., Brown, J.M., Brown, J., Bezuidt, O., Berube, P.M., Biller, S.J., et al. (2019) Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**: 1623-1635.e11.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., et al. (2016) Uncovering Earth's virome. *Nature* **536**: 425–430.

Pagarete, A., Chow, C.-E.T., Johannessen, T., Fuhrman, J.A., Thingstad, T.F., and Sandaa, R.A. (2013) Strong seasonality and interannual recurrence in marine myovirus communities. *Appl Environ Microbiol* **79**: 6253–9.

Parada, V., Herndl, G.J., and Weinbauer, M.G. (2006) Viral burst size of heterotrophic prokaryotes in aquatic systems. *J Mar Biol Assoc United Kingdom* **86**: 613–621.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**: 996.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**: 1533–1542.

Parsons, R.J., Breitbart, M., Lomas, M.W., and Carlson, C.A. (2012) Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* **6**: 273–84.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**: 649-662.e20.

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.

Petrov, V.M., Ratnayaka, S., Nolan, J.M., Miller, E.S., and Karam, J.D. (2010) Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol J* **7**: 292.

Pommier, T., Canbäck, B., … L.R.-M., and 2007, U. (2006) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**: 867–880.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**: e9490.

Proctor, L.M. and Fuhrman, J.A. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60–62.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012) SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Puig, M. and Girones, R. (1999) Genomic structure of phage B40-8 of Bacteroides fragilis. *Microbiology* **145**: 1661–1670.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Rappé, M.S. and Giovannoni, S.J. (2003) The Uncultured Microbial Majority. *Annu Rev Microbiol* **57**: 369–394.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., et al. (2010) Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739–751.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009) Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7**: 828–36.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Rohwer, F. and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529–35.

Van Rossum, T., Ferretti, P., Maistrenko, O.M., and Bork, P. (2020) Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* **18**: 491–506.

Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E. V., Kropinski, A.M., Krupovic, M., et al. (2019) Minimum information about an uncultivated virus genome (MIUVIG). *Nat Biotechnol* **37**: 29–37.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689–693.

Roux, S., Enault, Francois, Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**: e985.

Roux, S., Enault, François, Ravet, V., Pereira, O., and Sullivan, M.B. (2015) Genomic characteristics and environmental distributions of the uncultivated Far-T4 phages. *Front Microbiol* **6**: 199.

Sabri, M., Hauser, R., Ouellette, M., Liu, J., Dehbi, M., Moeck, G., et al. (2011) Genome Annotation and Intraviral Interactome for the Streptococcus pneumoniae Virulent Phage Dp-1. *J Bacteriol* **193**: 551–562.

Sazinas, P., Redgwell, T., Rihtman, B., Grigonyte, A., Michniewski, S., Scanlan, D.J., et al. (2018) Comparative genomics of bacteriophage of the genus seuratvirus. *Genome Biol Evol* **10**: 72–76.

Schwalbach, M., Hewson, I., and Fuhrman, J. (2004) Viral effects on bacterial community composition in marine plankton microcosms. *Aquat Microb Ecol* **34**: 117–127.

Seguritan, V., Feng, I.-W., Rohwer, F., Swift, M., and Segall, A.M. (2003) Genome sequences of two closely related Vibrio parahaemolyticus phages, VP16T and VP16C. *J Bacteriol* **185**: 6434–47.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003) Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Sieradzki, E.T., Ignacio-Espinoza, J.C., Needham, D.M., Fichot, E.B., and Fuhrman, J.A. (2019) Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**: 1–9.

Stewart, R.D., Auffret, M.D., Warr, A., Wiser, A.H., Press, M.O., Langford, K.W., et al. (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**: 870.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations. *PLoS Biol* **3**: e144.

Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R., et al. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**: 3035–3056.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature* **424**: 1047–1051.

Sullivan, M.B., Weitz, J.S., and Wilhelm, S. (2017) Viral ecology comes of age. *Environ Microbiol Rep* **9**: 33–35.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., et al. (2015) Structure and function of the global ocean microbiome. *Science (80-)* **348**: 1261359.

Suttle, C.A. (2007) Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Suttle, C.A. (1994) The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol* **28**: 237–243.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356–361.

Suttle, C.A. and Chan, A.M. (1993) Marine cyanophages infecting oceanic and coastal strains of Synechococcus: abundance, morphology, cross-infectivity and growth characteristics. *Mar Ecol Prog Ser* **92**: 99–109.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014) Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS One* **9**:.

Takebe, H., Tominaga, K., Fujiwara, K., Yamamoto, K., and Yoshida, T. (2020) Differential Responses of a Coastal Prokaryotic Community to Phytoplanktonic Organic Matter Derived from Cellular Components and Exudates. *Microbes Environ* **35**: n/a.

Tan, T.T., Forsgren, A., and Riesbeck, K. (2006) The Respiratory Pathogen *Moraxella catarrhalis* Binds to Laminin via Ubiquitous Surface Proteins A1 and A2. *J Infect Dis* **194**: 493–497.

Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., et al. (2012) Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science (80-)* **336**: 608–611.

Thiaville, J.J., Kellner, S.M., Yuan, Y., Hutinet, G., Thiaville, P.C., Jumpathong, W., et al. (2016) Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc Natl Acad Sci U S A* **113**: E1452-9.

Thingstad, T.F. (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.

Thingstad, T.F., Heldal, M., Bratbak, G., and Dundas, I. (1993) Are viruses important partners in pelagic fend webs? *Trends Ecol Evol* **8**: 209–213.

Thingstad, T.F., Vage, S., Storesund, J.E., Sandaa, R.A., and Giske, J. (2014) A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc Natl Acad Sci U S A* **111**: 7813–7818.

Tikhonov, M., Leach, R.W., and Wingreen, N.S. (2015) Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* **9**: 68–80.

Tominaga, K., Morimoto, D., Nishimura, Y., Ogata, H., and Yoshida, T. (2020) In silico Prediction of Virus-Host Interactions for Marine Bacteroidetes With the Use of Metagenome-Assembled Genomes. *Front Microbiol* **11**: 738.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**: 170203.

Tully, B.J., Sachdeva, R., Graham, E.D., and Heidelberg, J.F. (2017) 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**: e3558.

Unfried, F., Becker, S., Robb, C.S., Hehemann, J.-H., Markert, S., Heiden, S.E., et al. (2018) Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms. *ISME J* **12**: 2894–2906.

Vinayak, M. and Pathak, C. (2009) Queuosine modification of tRNA: its divergent role in cellular machinery. *Biosci Rep* **30**: 135–148.

Waterbury, J.B. and Valois, F.W. (1993) Resistance to co-occurring phages enables marine Synechococcus communities to coexist with cyanophages abundant in seawater. *Appl Environ Microbiol* **59**: 3393–3399.

Whitman, W.B., Coleman, D.C., Wiebe, W.J., Schwalbach, M.S., Brown, M. V., Green, J.L., and Brown, J.H. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**: 6578–83.

Wiggins, B.A. and Alexander, M. (1985) Minimum bacterial density for bacteriophage replication: implications for significance of bacteriophages in natural ecosystems. *Appl Environ Microbiol* **49**: 19–23.

Wilhelm, S.W. and Suttle, C.A. (1999) Viruses and Nutrient Cycles in the Sea. *Bioscience* **49**: 781–788.

Winter, C., Bouvier, T., Weinbauer, M.G., and Thingstad, T.F. (2010) Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: the "Killing the Winner" Hypothesis Revisited. *Microbiol Mol Biol Rev* **74**: 42–57.

Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., and Keeling, P.J. (2015) Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science (80-)* **347**: 1257594.

Woyke, T., Xie, G., Copeland, A., González, J.M., Han, C., Kiss, H., et al. (2009) Assembling the Marine Metagenome, One Cell at a Time. *PLoS One* **4**: e5299.

Xia, L.C., Ai, D., Cram, J., Fuhrman, J.A., and Sun, F. (2013) Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* **29**: 230–237.

Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., et al. (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol* **5**: S15.

Xing, P., Hahnke, R.L., Unfried, F., Markert, S., Huang, S., Barbeyron, T., et al. (2015) Niches of two polysaccharide-degrading Polaribacter isolates from the North Sea during a spring diatom bloom. *ISME J* **9**: 1410–1422.

El Yacoubi, B., Bailly, M., and de Crécy-Lagard, V. (2012) Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs. *Annu Rev Genet* **46**: 69–95.

Yoshida, M., Yoshida, T., Kashima, A., Takashima, Y., Hosoda, N., Nagasaki, K., and Hiroishi, S. (2008) Ecological dynamics of the toxic bloom-forming cyanobacterium Microcystis aeruginosa and its cyanophages in freshwater. *Appl Environ Microbiol* **74**: 3269–3273.

Yoshida, T., Morimoto, D., and Kimura, S. (2019) Bacteria–Virus Interactions. In *DNA Traffic in the Environment*. Singapore: Springer Singapore, pp. 95–108.

Yoshida, T., Nishimura, Y., Watai, H., Haruki, N., Morimoto, D., Kaneko, H., et al. (2018) Locality and diel cycling of viral production revealed by a 24 h time course cross-omics analysis in a coastal region of Japan. *ISME J* **12**: 1287–1295.

Yu, T., Zhang, Z., Fan, X., Shi, X., and Zhang, X.-H. (2014) Aquimarina megaterium sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* **64**: 122–127.

Zhang, Z., Qin, F., Chen, F., Chu, X., Luo, H., Zhang, R., et al. (2020) Culturing novel and abundant pelagiphages in the ocean. *Environ Microbiol* 1462-2920.15272.

Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C., et al. (2013) Abundant SAR11 viruses in the ocean. *Nature* **494**: 357–360.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132–e132.

Zimmerman, A.E., Howard-Varona, C., Needham, D.M., John, S.G., Worden, A.Z., Sullivan, M.B., et al. (2020) Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol* **18**: 21–34.

Publication list

1. Tominaga, K., Morimoto, D., Nishimura, Y., Ogata, H., Yoshida, T. *In silico* Prediction of Virus-Host Interactions for Marine Bacteroidetes With the Use of Metagenome-Assembled Genomes. (2020). *Front. Microbiol.* 11:738.

Other publications

1. Takebe, H., Tominaga, K., Fujiwara, K., Yamamoto, K., Yoshida, T. Differential Responses of a Coastal Prokaryotic Community to Phytoplanktonic Organic Matter Derived from Cellular Components and Exudates. (2020). *Microbes Environ.* 35(3). ME20033.
2. Martinez-Hernandez, F., Luo, E., Tominaga, K., Ogata, H., Yoshida, T., DeLong, E. F., Martinez-Garcia, M. Diel Cycling of the Cosmopolitan Abundant Pelagibacter Virus 37-F6: One of the Most Abundant Viruses on Earth. (2020). *Environ Microbiol Rep.* 12(2):214-219.
3. Sato Y., Tominaga K., Aoki H., Murayama M., Oishi K., Hirooka H., Yoshida T., Kumagai H., Calcium salts of long-chain fatty acids from linseed oil decrease methane production by altering rumen microbiome *in vitro*. (2020). *PLOS ONE,* 15(11):e0242158
4. Prodinger, F., Endo, H., Gotoh, Y., Li, Y., Morimoto, D., Omae, K., Tominaga, K., Blanc-Mathieu, R., Takano, Y., Hayashi, T., Nagasaki, K., Yoshida, T., Ogata, H. An Optimized Metabarcoding Method for *Mimiviridae*. (2020). *Microorganisms*. 8(4):506.
5. Morimoto, D., Tominaga, K., Nishimura, Y., Yoshida, N., Kimura, S., Sako, Y., Yoshida, T. Cooccurrence of Broad- and Narrow-Host-Range Viruses Infecting the

Bloom-Forming Toxic Cyanobacterium *Microcystis aeruginosa*. (2019). *Appl. Environ. Microbiol.* 85(18):e01170-19.

6. Morimoto, D., Šulčius, S., Tominaga, K., Yoshida, T. Predetermined Clockwork Microbial Worlds: Current Understanding of Aquatic Microbial Diel Response from Model Systems to Complex Environments. (2020). *Adv. Appl. Microbiol.* 113:163-191