理學博士學位論文

Ecological and Genomic Study on Freshwater Bacteriophages

담수 생태계 내 박테리오파지의
유전체와 생태학적 연구

2017년 8월

서울대학교 대학원
생명과학부
문 기 라

Ecological and Genomic Study on Freshwater Bacteriophages

Advisor: Professor Dr. rer. nat. Sang Jong Kim

by

Kira Moon

A Thesis Submitted in Partial Fulfillment

of the Requirements for

the Degree of Doctor of Philosophy

August, 2017

School of Biological Sciences

Graduate School

Seoul National University

ABSTRACT

문기라 (Kira Moon)

생명과학부 미생물생태학

(Department of Biological Sciences, Microbial Ecology)

The Graduate School

Seoul National University

Viruses, the smallest and simplest form of life, are the most abundant biological entities on the Earth. Bacteriophages (phages) are viruses that infect of infecting bacterial cells. As bacterial cells are known to be found in almost every environment known, their predators, bacteriophages are also found in diverse environments including ocean, soil, hot spring, polar areas, and deserts. However, despite their high abundance and ability to survive under extreme conditions, environmental bacteriophages had been understudied due to limitations in isolating and culturing them in laboratory settings. As a result, number of isolated and identified bacteriophages is very low relative to their high abundance in the environment. Recently, to overcome culturability restrictions, viral metagenome, also denoted as virome, was suggested to study bacteriophage population without culturing. Therefore, based on viral metagenome technique, many large-scale marine virome projects had been performed, especially in marine settings. However, most of the virome sequences remain as un-interpreted due to the dearth of known bacteriophage genome information in the public genome databases. Furthermore, only few number of bacteriophage studies in freshwater environments including both virome and isolation of phages have been performed despite the importance of inland freshwaters as highly conserved reservoirs of diverse microbial communities. For better understanding of freshwater microbial community structure and their

ecological dynamics, this study performed both culture-independent and culture-dependent bacteriophage researches. Using viral metagenome, a culture-independent method, bacteriophage population distribution in Lake Soyang, the largest lake in South Korea, was observed. Since microbial community within a confined lake shifts as seasonal stratification takes place, bacteriophage community was also expected to change according to seasons. Therefore, 6 seasonal samples were collected from Lake Soyang and viral metagenome samples were prepared from them. When sequence similarity between 6 samples were compared, no clear seasonal variability was observed, however, gradual change of viral sequences was observed through time. When taxonomic annotation was performed using virome reads, up to 93.6% of them were not identifiable. Among those that were annotated with a taxonomic name, most of them were shown to be the phages that were isolated from marine environments. For more analysis of freshwater virome, viral contigs constructed from Lake Soyang virome data were grouped with reference viral sequences obtained from public databases, and 211 groups were found that showed no similarity with previously reported bacteriophages. In attempts to identify those unique viral groups, their putative host bacteria were predicted. Among 211 virome contig groups, 23 groups with the most viral contig sequences (976 contigs) were predicted to infect a host belonging to the phylum *Proteobacteria* and 1 group with 315 contigs was anticipated to infect a host within the phylum *Actinobacteria,* which are the two major bacterial phylum found in Lake Soyang. In spite of diverse attempts to interpret freshwater virome, inability to annotate virome reads and biasedly assigning the annotated freshwater virome reads to representative marine bacteriophage genomes indicated the under-representation and deficiency of freshwater bacteriophages. Therefore, to fill the gaps in the knowledge of freshwater bacteriophages, novel bacteriophages were physically isolated and cultured from Lake Soyang. As a result, total of 4 novel bacteriophages have been isolated from Lake Soyang. Two representative bacterial strains of the family *Comamonadaceae,*

Rhodoferax and *Curvibacter* isolates were used as hosts to screen for novel phages from Lake Soyang. Hence, two independent phages infecting *Curvibacter* sp. and one phage infecting *Rhodoferax* sp. were isolated, and they were named as P26509A and P26059B, and P26218, respectively. The bacteriophage, P19250A that infects a strain belonging to the family *Methylophilaceae,* was also isolated and revealed to be the most abundant bacteriophage in Lake Soyang in winter seasons, in which its host, LD28 clade also thrives. In the binning analyses of freshwater viromes, P19250A was the most highly-assigned freshwater phage (up to 8.7%) in several viromes of foreign countries, including five viromes from Lake Soyang that were constructed in this study. These results showed that newly isolated bacteriophages would be an essential resource for analyses of freshwater viromes. One of the major ecological roles of bacteriophages is as mediators of horizontal gene transfer (HGT) between bacterial cells. Among many bacterial protein genes, antibiotic resistance gene (ARG) is one of the ecologically and clinically important genes that are transported by bacteriophages. To observe bacteriophage community structure of the lentic environment and their roles as ARG transporters in urban area, the Han River, which flows from Lake Soyang to the Yellow sea, passing through Seoul (the capital of South Korea) was selected as the study site. When overall sequence similarity was compared, all 6 samples collected from the Han River had low dissimilarity. Also, when taxonomic assignment of virome reads were analyzed, no significant change of taxonomic assignment was observed, indicating that Han River, which flows for approximately 180 km, has stably maintained viral population. As one of the auxiliary metabolic genes carried by bacteriophages, a number of ARGs was observed within from viral metagenome reads. Among virome contigs, 7 viral contigs were shown to be carrying well conserved active antibiotic resistance genes, suggesting that those genes may be transferred to bacterial cells upon phage infection and lead to the rise of antibiotic resistance bacterial strains. Through both culture-independent and dependent methods, distribution of bacteriophage sequences and

their ecological roles in freshwater environments, both lentic and lotic, were observed. Through isolation of novel bacteriophages that are abundantly distributed in freshwater habitats, this study has provided a key information of interpreting global virome samples as well as that of Lake Soyang. Therefore, this study emphasized the need of isolation and culture of environmental bacteriophages to understand viral ecology and also viral metagenome data.

Keywords : Bacteriophage, dsDNA virus, freshwater, viral metagenome, novel bacteriophage, whole genome sequencing, antibiotic resistance gene

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFM	Artificial freshwater medium
AMG	Auxiliary metabolic genes
ARG	Antibiotic resistance genes
BATS	Bermuda Atlantic Time-series Study
BLAST	Basic local alignment search tool
BOD	Biochemical oxygen demand
CARD	Comprehensive Antibiotic Resistance Gene Database
COD	Chemical oxygen demand
DAL	Double agar layer
DO	Dissolved Oxygen
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid
HGT	Horizontal gene transfer
IMCC	Inha Microbial Culture Collection
KtW	Kill-the-winner hypothesis
MOI	Multiplicity of infection
NCBI	National Center for Biotechnology Information
NMDS	Non-metric multidimensional scaling
ORF	Open reading frame
PC	Protein clusters
PCoA	Principal coordinate analysis
PEG	Polyethylene glycol
PES	Polyethersulfone

POV	Pacific Ocean Virome
PS	Photosystem
SM buffer	Sodium-Magnesium buffer
SPOT	San Pedro Ocean Time-Series Study
SS	Suspended solids
TEM	Transmission electron microscopy
TFF	Tangential flow filtration
TN	Total Nitrogen
TP	Total Phosphate
VLP	Viral-like particle
WWTP	Wastewater treatment plant

CHAPTER 1.

General Introduction

1. Environmental bacteriophages

1.1. Bacteriophages: the most abundant biological entities on the Earth

Bacteriophages, often referred as phages, are the smallest and simplest form of life that are most widely and abundantly found on The Earth. Bacteriophages are obligate parasites that could only reproduce themselves through infection of a host bacteria. For their survival, bacteriophages can interchangeably take multiple life cycles. The most common form of bacteriophage life cycle is lytic cycle, which that once a bacteriophage particle enters the host cell, it replicates its nucleic acids and assembles new bacteriophage particles using the host cell replication machinery. Once complete particles are assembled, they burst out of the cell, leading to bacterial cell death. On the other hand, bacteriophages may take a temperate life cycle, which includes both lytic and lysogenic cycles. When bacteriophages take lysogenic life cycle, they do not actively replicate their genomes after they enter the host cell. Rather, the phage genomes integrate into the host genome and replicate along with their host genomes. Once bacteriophage genomes have been integrated into the host genome, they are called as prophages. Bacteriophages choose to take lysogenic life cycle when surrounding environments are not favorable such that bacterial cells are under stress and bacterial replication machineries are not functioning or that host bacterial densities are low and chance of further viral infection is limited (Maurice *et al.*, 2013; Payet and Suttle, 2013). When situations become favorable, those phage genomes resume replication and particle assembly and burst out of the host cell to seek for the next host bacterium. Besides these two major life cycles, bacteriophages may also be replicating and reproducing viral particles without host cell lysis, through budding or secretion from the host cell membrane (Koskella and Brockhurst, 2014; Sime-Ngando, 2014). Through highly adaptive strategies of replication, bacteriophages have survived in diverse environments along with their hosts.

As bacteria are known to inhabit in diverse and extreme environments, their predators, bacteriophages, are also known to be existing in such environments as well. Through shotgun metagenome sequencing, many researchers discovered bacteriophage genes in diverse environments such as hot springs, polar areas, and deserts (Adriaenssens *et al*., 2015; Bellas and Anesio, 2013; Breitbart *et al*., 2004b; Lopez-Bueno *et al*., 2009). Among various environments, bacteriophages are known to be most abundant in aquatic environments, where they are found as free-floating particles. Researchers have calculated that approximately 4×10^{30} viral-like particles (VLP), which includes bacteriophages, are found in ocean alone (Ignacio-Espinoza *et al*., 2013; Suttle, 2005) while 10^{31} viral particles are estimated to exist on The Earth (Cobián Güemes *et al*., 2016). Despite high number of VLPs found on the Earth, they are significantly understudied compared to bacterial cells. In open oceans, viral particles are found to be approximately 10-fold more abundant than microbial cells (Wommack *et al*., 2015). However, number of isolated and sequenced viral particles are much less than those of bacterial cells; number of bacterial genomes that have been sequenced are about 12-fold more than those of viruses.

1.2. Difficulties in environmental bacteriophage researches

Among sequenced viral particles reported on National Center for Biotechnology Information (NCBI) database, environmental viruses, mostly comprised of bacteriophages, take only 30%, which is about 2,000 sequences. Number of identified bacteriophages is significantly low compared to number of VLPs predicted to be existing in natural environments due to restrictions in culturing individual phages. One of the major limitations in culturing and isolating bacteriophages is that they are obligate parasites which require their hosts to be cultured beforehand. However only few environmental bacterial species, less than 1% of the existing bacterial population have been cultured in artificial media so far, as claimed by the ‘great plate count anomaly’ (Hugenholtz, 2002; Vartoukian *et al*.,

2010). Therefore, attempts to culture bacteriophages that infect environmental bacterial groups are also highly restricted.

Although most of the bacterial groups remain uncultured that no morphological, physiological, and genomic data are available, the 'unculturable bacteria' can be assigned with group names according to phylogenetic classification done by 16S rRNA sequences which are well conserved in bacterial and archaeal species (Yarza *et al*., 2014). However, viral genomes are known to be highly variable and subjected for frequent mutations due to horizontal gene transfers, thereby no conserved sequence was found. Lack of universally conserved sequences restricted viral particles in environments to be identified with culture-independent methods. Therefore, due to such limitations, studies on environmental viruses have been hampered, leaving bacteriophages as massive dark matters of microbial ecology.

2. Ecological roles of bacteriophages

2.1. Population control of bacterial communities by bacteriophages

As an obligate parasite, most bacteriophages lyse their hosts for their replication. Therefore, as lytic bacteriophages thrive in the environment, the density of designated hosts decreases. At different magnitudes, bacteria and their phages show alternating fluctuation in densities over time. According to a suggested model by Rodriguez-Brito and his colleagues, while both viral and microbial densities oscillate over time, viral abundance shows a peak soon after its host microbes reaches the peak abundance (Rodriguez-Brito *et al*., 2010). Such phenomenon is explained by a hypothesis, 'kill-the-winner' (KtW) (Thingstad, 2000). The KtW theory states that as a bacterial strain blooms within its habitat, its bacteriophages acquire higher chance of encountering their hosts, thereby increasing rate of infection and leading to higher rate of viral replication and decrease in host cell densities. Therefore, bacteriophages are considered as one of the significant factors

that control bacterial population. Within aquatic environment, KtW events are often observed through interactions between Cyanobacteria and cyanophages. Parsons and her colleagues illustrated roles of bacteriophages as population controllers of their hosts through observing rapid decrease of blooming Cyanobacterial population as their corresponding phages start to increase in number (Parsons *et al*., 2012).

2.2. Indirect participation of bacteriophages in geochemical cycling in freshwater environments

Bacterial cells that are abundant in aquatic environment contribute in biochemical cycling such as utilization of carbon to produce CO_2 gas (Chistoserdova, 2011) and complete nitrification (Daims *et al*., 2015). As those bacterial cells flourish in the environment, their phages also increase in number and eventually lead to bacterial cell lysis. Therefore, those bacteriophages will disrupt nutrient cyclings performed by bacterial species. As bacteriophages predate on abundant hosts, the dominant microbial population that participate in different biochemical cycling will change over time, consequently shifting biochemical components within the system. Likewise, phages indirectly participate in diverse biochemical cyclings in environments through interference and its impact is enhanced especially in enclosed aquatic systems such as lakes. Furthermore, by destroying the bacterial cells through lytic life cycles, phages also contribute in organic carbon particulate accumulation in freshwater environment (Guidi *et al*., 2016). As bacteriophages lyse host cells for release of newly produce viral particles, the host cell debris is releases to the system. Since cell debris are highly organic and rich of metabolites such as carbon, the release of cell debris leads to increase in nutrients that can be utilized by neighboring microbial communities (Pan *et al*., 2014). Phages are not only the predators for microbial cells that control bacterial abundance but also a key participant in modifying freshwater chemical parameters.

2.3. Bacteriophages as reservoirs of bacterial genes

Bacteriophages rely on their host cells' machineries for their reproduction. Therefore, the phage genomes are often replicated along with the host genomes. During process of packaging phage genomes into newly produced phage capsids, parts of bacterial genomes are occasionally mis-packaged into phage capsids as either small pieces of genes or encompassed parts of the phage genome. While phage particles carry bacterial gene within its capsid, they continue infecting other bacterial cells. At this stage, the acquired genes are either transferred to the next host cell or remain as a part of the phage genome and continue to be replicated with the phage genome. Temperate phages freely change their life cycles back and forth, from lytic to lysogenic. Therefore, when a temperate phage genome carrying acquired bacterial gene takes lysogenic life cycle, the acquired bacterial gene may be integrated into the new host's genome, enriching the genetic diversity of the host bacteria. In this process, the phage genome serves as both reservoir and transporter of the bacterial genes. This phenomenon can be found in all phages, however, temperate phages have higher chance of transferring bacterial genes from one to the other. Recent studies have revealed that it is the temperate phages that dominate in the aquatic systems rather than complete lytic phages, which was conventionally known before (Brum *et al*., 2015a; Knowles *et al*., 2016), implying more possibility of phages as agents for horizontal gene transfer (HGT) which enhance genetic diversity among bacterial groups.

Although phages are the predators of the bacteria and cause cell death through infection, some phages carry bacterial genes that benefit their host, which are called as auxiliary metabolic genes (AMG). Cyanobacteria, *Synechococcus* and *Prochlorococcus* genera are known to perform photosynthesis for their production. These Cyanobacteria were known to perform both photosystems (PS) I and II, but core genes for PS I were missing in their genomes. Interestingly, those missing PSI

core genes were found within bacteriophage genomes that infect *Cyanobacteria* (Hevroni *et al*., 2015). Upon infection, these phages insert their genome into the host cell and express the PS I core genes and missing core genes are complemented. Therefore, while the Cyanobacteria is infected by a cyanophage, both PS I and II are activated and cell production increases. Through this mechanism, the cyanophages benefit from enhanced cell production by having more efficient replication of phage particles (DeLong and Beja, 2010; Sharon *et al*., 2009). Considering that bacteriophages do not perform any metabolic pathways for their survival, carrying of an extra bacterial gene within their tight capsid may be inefficient. Therefore, presence of functional AMGs in bacteriophage genomes is a significant evidence of evolution established by HGTs.

Some phages carry *imm* genes called 'superinfection immunity' gene (Abedon, 2015). These genes are mainly carried by lysogenic phages but seldomly carried by lytic phages as well. The *imm* gene codes for plasma membrane protein which prevents superinfection, also known as coinfection of one or more phages. This is a defensive method of the phages to protect their host cell from further phage infections, which would cause competition in using host cell replication machinery. In the bacterial cell's perspective, initially infected phage genome serves as "vaccine" that protects bacterial cells against other phages. Bacteriophages are not simple predators or parasites that only cause harm to the bacterial cells, rather they are complex organisms that co-evolved with bacteria through benefiting their hosts as well. Therefore, there is a need for phages to be reconsidered as opportunistic symbiont biological entities and couriers of diverse bacterial genes.

3. Viral metagenome: culture-independent bacteriophage researches

3.1. Development of methods to prepare environmental samples for viral metagenome

With increase in awareness of roles of bacteriophages in the environment and their high abundance, numerous attempts have been made to study bacteriophages of diverse environments. Number of researchers attempted to isolate and sequence bacteriophages that infect major bacterial groups in oceans and two groups have successfully identified bacteriophages that infect SAR116 and SAR11, distinctively, the most abundant bacterial groups in the ocean (Kang *et al*., 2013; Zhao *et al*., 2013). Just as their hosts did, the bacteriophages that were identified to be infecting SAR116 and SAR11 were also found to be the most abundant bacteriophages in ocean (up to 25.3% of the viral metagenome reads analyzed) illuminating a large part of environmental bacteriophage population (Culley, 2013). However, large proportion of bacteriophage in environments, including ocean is still left as unknown.

The bacteriophages of SAR116 and SAR11 were able to be isolated because their hosts were able to be cultured in artificial media. Since most of the bacterial species still remain as uncultured, their bacteriophages are also uncultured. In order to overcome the culturability limitations, researchers have developed viral metagenome (virome) methods, specifically for viral particles in aquatic environments. Shotgun metagenome sequencing allowing access to large number of viral genetic sequences regardless of presence of marker genes. Since viral particles, especially bacteriophages, are very small in size, which are mostly less than 0.2-μm in diameter, and amount of DNA within a phage capsid is too small for metagenome library construction, collection and concentration of bacteriophage particles for

shotgun sequencing from aquatic environments required massive volume of samples. Using tangential flow filtration (TFF) system, which collects viral particles of desired size through continuous flow of water sample through filter sets, approximately 120 L of sea water is required to collect sufficient amount of viral particles for sequencing (Wommack *et al*., 2004). Recently, a method that collects viral particles through flocculation using ferric chloride ($FeCl_3$) was developed by John *et al*., 2011. The iron oxyhydroxide particles flocculates with negatively charged viral particles in water, and flocculated particles were large enough to be collected on polycarbonate filter papers. The chemical flocculation method recovered approximately 95% of the viral particles in aquatic samples while the TFF method recovered only 23% (John *et al.,* 2011), thereby allowing concentration of approximately 10L of water to collect enough bacteriophage particles for metagenome library construction. Thus, chemical flocculation provided highly efficient method to concentrate viral particles from environments using relatively small volume of samples, allowing culture-independent access to enormous bacteriophage populations in the environment with high efficiency. However, the chemical flocculation method primarily concentrates negatively charge bacteriophages with capsid, which are mostly represented by the order *Caudovirales*. Therefore, some viruses with lipid membrane may be under-represented. However, considering that most environmental bacteriophage community structures are composed of double-stranded DNA (dsDNA) phages (Roux *et al.,* 2016b), the chemical flocculation method can be accepted to prepare of viral metagenome samples that adequately represent viral populations of aquatic environments. The detailed methods of viral metagenome sample preparation using $FeCl_3$ flocculation are described in chapter 2 and 4.

3.2. Global-wide ocean viral metagenome studies

Using highly efficient viral particle concentration methods, number of research groups performed a global-wide ocean viral metagenome sampling. As of January 2017, total of 5 marine projects have been set out to collect global-wide ocean viral metagenome samples. Two expeditions, *Tara* Ocean expedition and Malaspina expedition, sailed across the Earth to collect biological water samples from 2009 to 2011. The *Tara* Ocean expedition collected surface water samples to observe biological diversity and influence of environmental factors on viral community structure in 6 oceans and revealed that marine bacteriophages are mostly comprised of non-tailed viruses and distribution of different morphology of viruses is influenced by salinity, temperature, and oxygen concentration (Brum *et al*., 2013). Meanwhile, the Malaspina expedition focused more on distribution of deep ocean viral and microbial communities (Brum and Sullivan, 2015).

The other three marine virome projects focused on more confined sampling sites with different depths and time points. The University of Southern California's Microbial observatory at San Pedro Ocean Time-series (SPOT) project surveyed dynamics of bacterial and viral population, specifically myoviruses in daily time points for approximately 3 months of time period. The SPOT project revealed that populational concentrations of bacteria and myoviruses showed fluctuations at daily intervals but when they were observed in monthly intervals, the fluctuations were rather stable and uniformly maintained with seasonal variability, indicating that viral and host relationships have been established over long time period and have predictive population shifts (Chow and Fuhrman, 2012). The Bermuda Atlantic Time-series Study (BATS) has observed viral and microbial populations over 10 years of period at the Sargasso Sea. Through long-time period observations, the BATS showed that most of the viroplanktons present in open ocean environments are predicted to be cyanophages since viroplanktons abundance fluctuated according

to abundance shifts of cyanobacteria (Parsons *et al*., 2012).

The Pacific Ocean Virome (POV) project collected sea water samples from 2009 to 2011 from different depths of the Pacific Ocean. Using the data set, a new approach of viral population identification has been proposed (Hurwitz and Sullivan, 2013), which was clustering of viral protein sequences. Due to lack of viral sequences in publicly available genome databases, most of the viral reads could not be identified using conventional BLAST search, thereby leaving large portion of viral metagenome as dark matters. Therefore, cataloging viral metagenome reads based on protein clusters (PC) that they share with other metagenome reads or reference sequences were used to classify unknown virome reads, independent from existing databases. Thus, they were able to create approximately 1.3 million PCs to assign virome reads to. Then by observing the virome sequence groups established based on shared PCs, distribution patterns of viral groups across the Pacific Ocean was observable. The POV made a conclusion that marine viral community represents the seed-bank hypothesis. A viral population may dominate in regional areas, infecting their hosts. While they remain in the local area, the viruses are the "banks" of viral genes that are easily influenced by variable environmental conditions (due to small size of the viruses). Such local viral "banks" will start influencing bacterial cells through infection. Since the microbial cells are motile, they may travel to a neighboring niche, and this time, they may be carrying bacteriophage genes within their cells. Therefore, the bacterial cells with viral genetic components will serve as the "seed" that will spread the viral particles or genetic materials to distantly located niche. Thereby, locally dominating viral groups may participate in shaping overall ocean microbial ecology (Brum *et al*., 2015b).

To expand virome studies from taxonomical identification and distribution of viral reads to functional survey of bacteriophages, auxiliary metabolic genes (AMGs) coded by viral reads were focused and studied in depth as well. Viral

metagenome data from two ocean virome projects, POV and Malaspina expedition, were analyzed for protein coding genes of the viral reads. Within viral contigs assembled from virome reads, 243 AMGs were identified, where 148 of them were newly found (Roux *et al*., 2016a). Although it was found at low frequency, *amoC* gene, which encodes the C subunit of ammonia monooxygenase that is involved in ammonia oxidation, was found in viral contigs for the first time. Therefore, it became clearer that bacteriophage may play key roles in nutrient cycling in marine environments through manipulation of their host metabolic genes.

3.3. Viral metagenome, a casket filled with novel sequences

With help of high throughput sequencing methods, billions of viral metagenome reads have been collected through many studies and most of them still remain as unknown bacteriophage genomes, waiting to be identified. Based on these viral metagenome data, attempts to fish out putative bacteriophage genomes of a bacteria of interest were made. Since bacteriophages are known to be highly susceptible of HGTs as they infect their host cells, those with specific hosts are predicted to be carrying a portion of their host bacteria genome. Especially when the host bacteria are known to carry signature genes, their bacteriophages are highly likely to be carrying those genes as well. One of the most abundant but uncultured freshwater bacterial groups is Actinobacteria. Within the Actinobacterial clade, a subgroup called acI are known to be the most widely distributed in freshwater lakes. The acI clade is known to be carrying a signature gene, *whiB* transcription factors (Ghylin *et al*., 2014; Warnecke *et al*., 2004). Hinted from this, Ghai and his colleagues searched for viral contigs assembled from viral metagenome that carry *whiB* gene (Ghai *et al*., 2017). From two viral metagenome samples prepared from a freshwater reservoir Amadorio, located in Alicante, Spain, 8 contigs that carry *whiB* genes were identified. Although putative bacteriophage sequences are obtained, morphological or physical characteristics were not obtainable, thereby leaving these

phage contigs as putative candidate phages of acI clade, not as defined bacteriophages.

4. Significance of freshwater microbial ecology

Inland freshwaters, including lakes, reservoirs, streams, and rivers, occupy approximately 3,536,000 km^2, which is about 2.60 % of the Earth's surface. Within inland freshwaters, lakes and reservoirs occupy 85%, serving as water resources for diverse human activities. Although inland freshwaters take less area than oceans, it is estimated that approximately 50% of the CO_2 gas emission of the Earth is from the world's largest lakes, one of the largest contributor being the Caspian Sea, an enclosed inland water with about 1.2% salinity (Raymond *et al*., 2013). The freshwater lakes receive large amount of dissolved organic carbon from surrounding soils and accumulates dissolved CO_2 in water system. With large surface area and gas transfer velocity of 3-4 m per day, inland freshwater participates in carbon cycling on the Earth's atmosphere (Raymond *et al*., 2013). Furthermore, CO_2 emission from inland freshwaters will participate in regulation of global climate and participate in permafrost thaw as well (Tranvik *et al*., 2009).

Although different oceans have different currents and environmental conditions, eventually, it is a connected system. However, each freshwater lake is isolated from each other, having independent systems from each other despite locations or climates. Therefore, each lake is considered to have a unique ecosystem that has developed independently. Interestingly, despite ecological niche differences, key players of the microbial community are often very similar among lakes (Glöckner *et al*., 2000; Newton *et al*., 2011). Throughout different lakes around the world, the most dominant bacterial group is known to be those of the phylum *Actinobacteria*, followed by the phylum *Proteobacteria*. Although they may appear in different proportions, freshwater *Actinobacteria* and *Proteobacteria* are almost universally found in lake environments. Surface waters, including streams and rivers

also contain relatively stable microbial structure. However, since surface running waters flow through diverse environments, they are heavily influenced by agricultural, industrial, and urban activities. Also, inland waters, both lakes and rivers are sensitive to climate and environmental changes (Tseng *et al*., 2013), providing valuable study sites and samples for seasonal and climate-dependent microbial researches (Eiler *et al*., 2014; Hahn *et al*., 2015; Niño-García *et al*., 2016).

5. Purposes and scope of the study

The general purposes of this study are to observe and understand freshwater bacteriophage diversity and distribution within inland lake and river through culture-independent and dependent methods. Through viral metagenome, a culture-independent method, non-specific and broad range of bacteriophage genomes were able to be collected for analysis. Also, through isolation and culturing bacteriophages from freshwater lake, not only that host-bacteriophage relationship was observed, but also contribution in interpretation of viral metagenome prepared from the identical site was possible. The detailed purposes of this research are as follow.

1. Survey of bacteriophage population and discovery of unique bacteriophage sequences in an oligotrophic freshwater lake, Lake Soyang:

   Freshwater lake bacterial community are known to have seasonal variability, thereby, their predators, bacteriophages were also expected to have seasonal differences. Therefore, viral metagenome samples were prepared from surface water of Lake Soyang for different seasons. Also, in attempts to identify novel and unique viral sequences that are present in Lake Soyang, viral contigs were assembled from virome reads and analyzed.

2. Culture and whole genome sequencing of novel bacteriophages isolated from Lake Soyang:

   In order to discover bacteriophages that infect dominant heterotrophic bacteria in lake, novel bacteriophages were screened using bacterial strains isolated from Lake Soyang. Isolated bacteriophages were subjected for whole genome sequencing and those sequences were used to interpret viral metagenomes.

3. Survey of bacteriophage population distribution along the Han River body as well as detection of antibiotic resistance genes carried by the bacteriophage sequences through viral metagenome:

Through viral metagenome samples prepared from different points of Han River, that flows from pristine upstream to urbanized downstream, changes of bacteriophage population along the river flow were studied. Also, to evaluate the roles of bacteriophage as bacterial gene reservoirs and transporters, antibiotic resistance genes within viral metagenome data were studied.

Therefore, through these perspectives, this study will provide an insight on freshwater bacteriophage population in both lentic and lotic freshwater systems through viral metagenome analysis. Discovery of novel bacteriophages that thrive in freshwater lake allowed more extensive understanding of bacteriophage community structure through contribution of its genome data to interpret viral metagenome prepared from diverse lakes, including Lake Soyang.

CHAPTER 2.

Seasonal Freshwater Bacteriophage Survey in a Freshwater Lake using Viral Metagenome

ABSTRACT

Inland freshwaters, which occupy approximately 2.6% of the Earth's surface, are valuable sources of microbial diversity. Especially, freshwater lakes are confined systems that each of them represents a unique biome with their own microbial community. Although each lake is independent from each other, surprisingly, major bacterial components are similar with each other. Therefore, understanding microbial diversity of inland lakes provides universally applied knowledge. Bacterial community has been studied in diverse lakes using 16S rRNA amplicon sequencing. However, viral community, which plays a major role in bacterial population shifts, has been understudied due to limited methods to observe them. Recently, viral metagenome method has been developed that allowed marine bacteriophage population studies. However, viral community of diverse freshwater systems has not been studied much, leaving them as *terra incognita*. Therefore, to understand viral community of an oligotrophic lake, Lake Soyang of South Korea was subjected for viral metagenome sequencing. For survey of change in bacteriophage community over different seasons, 6 different samples were collected from surface water of Lake Soyang. However, viral metagenome sequences did not show seasonal variability, but rather implicated gradual changes in virome sequences over time. Within virome reads, only 6.40-12.16% of them were shown to have a similar match in an existing database, being able to identify the bacteriophage community. For further analysis of the virome, metagenome contigs were assembled and protein-coding genes were predicted. Then, based on sequence similarity, protein clusters were constructed and compared to reference viral sequences. As a result, total of 693 clusters were created and among those, 211 of them were identified to be newly found from Lake Soyang virome. For identification of bacteriophage contigs, their putative hosts were predicted through manual curation of each open reading frames found within complete and circularized contigs of Lake Soyang. Hence, 23 groups with 976 contigs were predicted to have a host belong to the phylum *Proteobacteria* and 1 group with 315 contigs to have a host within the phylum *Actinobacteria*.

1. INTRODUCTION

Viruses are known to be capable of infecting every known organism, including themselves. Among them, bacteriophages, also known as phages, are viruses that infect bacteria of various environments. Bacteriophages are obligate parasites that completely depends on bacterial cell replication machinery for their replication. After bacteriophages utilize bacterial cell machinery for genome replication and assembly of proteins needed for phage particles, most of the times, they lyse the host cell membrane and release themselves to the environment. Hence, bacteriophage infection often leads to bacterial cell death, leading to bacterial population control. While bacteriophages assemble newly produced genome into bacterial capsids, bacterial gene fragments are occasionally incorporated into capsids as well. Those bacterial gene fragments can be delivered to another bacterial cell as bacteriophages infect the next host, exhibiting horizontal gene transfer (HGT). Bacteriophages also have critical ecological importance for their high abundance and distribution. They are estimated to be making 10^{24} productive infections in every second, establishing 10^{31} viral particles on the Earth (Hendrix, 2010), which includes soil, marine, lakes, hot spring, and polar areas (Adriaenssens *et al*., 2015; Parsons *et al*., 2012; Zawar-Reza *et al*., 2014). However, despite their significance, most of these phages are unknown of their identity nor hosts because there are restrictions in culturing and isolating bacteriophages. Because bacteriophages are obligate parasites of bacteria, culturing bacteriophage requires host bacteria that are grown in artificial media. However, despite development of diverse culturing technologies, more than 99% of bacterial population is known to be still uncultured (Vartoukian *et al*., 2010). In spite of unculturability, bacterial groups can be named and phylogenetically classified with 16S rRNA sequence diversity, which is highly conserved in bacterial and archaeal strains. However, viral particles do not have any conserved sequence that could be used to detect and classify them. To overcome

methodological limitations in culturing and identifying abundant and ecologically function bacteriophages, viral metagenome (virome) was suggested which collects genetic information on environmental bacteriophages without culturing.

Since the very first viral metagenome performed in 2002 (Breitbart *et al*., 2002), number of viral metagenome studies has been done in diverse environments, discovering high number of viral genes that were not reported before. From 2009 to 2011, the Pacific Ocean Virome (POV) performed large scale viral metagenome studies in ocean with different depths, providing great insight to marine bacteriophages. From POV, Hurwitz and her colleagues discovered that bacteriophage genes are niche-specifically distributed, especially from photic and aphotic zones (Hurwitz *et al*., 2015). The niche-specific bacteriophage contig sequences contained auxiliary metabolic genes (AMGs) that modify bacterial metabolic processes, specific for their habitats. The viral sequences found in aphotic zone contained flagellar genes, *flaB* and *motA*, that may improve bacterial cell motility in the deep sea for better nutrient acquisition.

With a gigantic pool of unique sequences, viral metagenome data are like caskets of unfound bacteriophage genomes and these data can be utilized to search for novel bacteriophages. When bacterial hosts of interest are known to be carrying a signature gene sequence, it is very likely that its bacteriophage would be carrying the sequence as well, as a result of horizontal gene transfer between parasites and hosts. Therefore, in theory, putative genomes of bacteriophages that are not yet identified can be searched within viral metagenome data using signature gene of the bacterial strain of interest. Ghai and his colleagues attempted to search for genomes of bacteriophages that are predicted to be infecting a freshwater bacterial group. The uncultured freshwater bacteria lineage, Actinobacteria, known as one of the most abundant freshwater bacteria and acI, a clade of Actinobacteria, is known to be carrying *whiB* transcription factors. Therefore, Ghai and his colleagues searched for

viral contigs with *whiB* sequences within freshwater lake viral metagenome data and were able to suggest 6 contigs that were suspected to be infecting Actinobacterial group, acI (Ghai *et al*., 2016). This study not only found putative phage genomes that could infect uncultured bacterial groups, but also showed that environmental virome data are reservoirs of novel bacteriophage genomes that need to be sequenced and identified.

Multiple research expeditions set out to collect environmental microbial and viral samples and produced large amount of metagenome data (Hingamp *et al*., 2013; Hurwitz and Sullivan, 2013). However, freshwater viral metagenome studies are relatively scarce, compared to large number of different lakes with different characteristics that exist across the continents. According to a review article by Bruder and her colleagues, only 13 studies on freshwater virome have been published by 2016, severely underrepresenting freshwater microbial ecology (Bruder *et al*., 2016). Surface freshwaters occupy approximately 2.6% of the Earth's surface (Raymond *et al.,* 2013) and 91.3% of surface waters are comprised of lakes. Surface waters are spread across all the continents serving as essential geographic water resources for all biological entities on the Earth. Also, freshwater lakes are isolated from each other and that each has independent ecosystem that functions as unique reservoirs of biological organisms. Interestingly, although all lakes are independent from each other with distinctive properties, overall composition of microbial communities are similar with each other. Therefore, to inspect viral community structure in an oligotrophic lake and provide data sets for freshwater microbial ecology, viral metagenome study has been performed in Lake Soyang.

Lake Soyang is the largest and oldest conserved freshwater lake located in South Korea. As an oligotrophic lake, Lake Soyang encompasses diverse freshwater microbial organisms including diverse uncultured bacterial species, such as Actinobacterial groups acI, acIV, and LD28 clade, thereby imposing possibility of

having unknown bacteriophage communities that have not been reported before. In this study, 6 surface water samples were collected from Lake Soyang, at different seasons to observe viral population distribution along the seasonally varying bacterial population. Also, as numerous uncultured bacteriophages are expected to exist in Lake Soyang, novel bacteriophage sequences retrieved form viral metagenome were analyzed.

2. MATERIALS AND METHODS

2.1. Seasonal sampling of surface water of Lake Soyang

From October 2014 to May 2016, total of 6 surface water samples, representing different seasons were collected from Lake Soyang, located in Gangwon province, South Korea (37.947421 N, 127.818872 E) (Fig. 2-1). For viral metagenome analysis, approximately 10 L of surface water samples were collected. The environmental data, such as water temperature, pH, and dissolved oxygen (DO), total nitrogen (TN), total phosphate (TP), biochemical oxygen demand (BOD), chemical oxygen demand (COD), and suspended solids (SS) levels were provided by the Water Information System of the Ministry of Environment, of South Korea (http://water.nier.go.kr) (Table 2-1). Immediately after sampling of the lake water, approximately 30 ml of lake water was fixed with 2.5% glutaraldehyde solution. Then 100 μl of the fixed samples were stained with SYBR Gold (Invitrogen, Waltham, MA, USA) and viewed under epifluorescence microscopy for viral particle enumeration.

2.2. Viral metagenome sample preparation and metagenome sequencing

The collected water samples were brought to the lab in 4°C. Upon arrival to the lab, the samples were filtered through a 0.2-μm Supor® PES Membrane filter (Pall Corporation, New York, USA) using a filter tower to remove bacterial-like particles. To the filtered water samples, 0.01 g of $FeCl_3 \cdot 6H_2O$ were added per 10 L of sample to flocculate viral particles within the sample. The samples were vigorously shaken to promote flocculation of viral particles with $FeCl_3$ ions and they were incubated at room temperature for 1 hr to 12 hrs. The flocculated viral particles were then collected on a 0.8-μm Isopore polycarbonate filter (Merck Millipore,

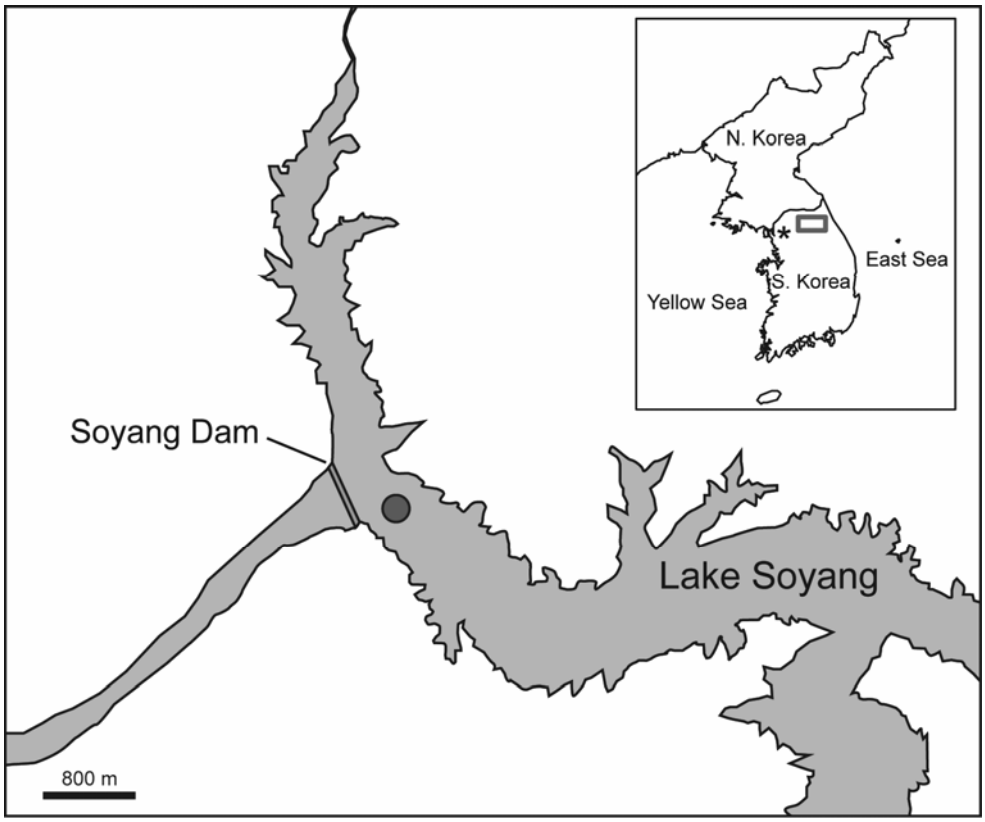

Figure 2-1. A map displaying the sampling site of Lake Soyang. The circle shows the sampling site.

Table 2-1. Environmental parameters of each sampling dates of Lake Soyang.

Sample[a]	Season	Viral particles per ml	Temp. (˚C)	pH	DO (mg/L)	BOD (mg/L)	COD (mg/L)	TN (mg/L)	TP (mg/L)	SS (mg/L)
’14 Oct.	Fall	3.08×10^7	15.70	8.30	9.50	0.60	2.80	1.635	0.006	1.300
’15 Jan.	Winter	4.07×10^7	6.00	7.80	11.80	0.50	2.10	1.625	0.006	0.800
’15 Sept.	Summer	6.13×10^7	14.10	8.30	10.70	0.30	2.40	1.985	0.010	1.700
’15 Nov.	Fall	2.97×10^7	13.90	8.50	10.40	0.30	3.10	2.047	0.008	1.500
’16 Feb.	Winter	-	4.00	7.90	13.40	0.30	2.20	1.719	0.002	0.900
’16 May	Spring	4.05×10^7	6.10	8.10	13.50	0.70	2.20	1.777	0.004	0.600

[a] Environmental data were provided by the Water Information System of the Ministry of Environment of South Korea.

Darmstadt, Germany) (John *et al*., 2011). The polycarbonate filters were placed in a conical tube and stored in 4˚C in dark until further treatment.

The polycarbonate filters with flocculated viral particles were dissolved in 0.1 M EDTA-0.2 M $MgCl_2$-0.2 M ascorbate acid buffer to chelate iron particles and suspend viral particles. Then the samples were treated with DNase I and RNase A at final concentrations of 10 U/ml and 1 U/ml (Sigma-aldrich, St. Louis, MO, USA), respectively to remove any possible external nucleic acids. After one hour of incubation with both enzymes in 20˚C, DNase and RNase were deactivated by adding 100 mM of EDTA and EGTA (Hurwitz *et al*., 2013). The viral particles within the sample were purified through cesium chloride (CsCl) step-gradient ultracentrifugation (Thurber *et al*., 2009). To a centrifuge tube, different densities of CsCl were stacked in following order; 1.7, 1.5, 1.35, and 1.2 g/cm^2, from bottom to top. Then above the top layer, approximately 15 ml of prepared sample was added. The samples were centrifuged at 24,000 rpm for 4 hrs at 4˚C in a Beckman Coulter L-90K ultracentrifuge with a SW32 Ti swing bucket. After centrifugation, the density fraction between 1.5 and 1.35 g/cm^2, which corresponds to density of double-stranded DNA (dsDNA) phages, were retrieved with a syringe. Buffer exchange of the sample with SM buffer (50 mM Tris-HCl, pH 7.5; 100 mM NaCl; 10 mM MgSO4·7H2O; 0.01% gelatin) was performed to remove CsCl remaining in the sample. Then, for sterilization, the samples were filtered through a 0.2-μm pore size Acrodisc® Syringe filter (Pall Corporation). The viral DNA was extracted from the filtrates using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). The DNA samples were then used for TruSeq library construction. Sequencing was performed using Illumina MiSeq platform, with 2 × 300-bp paired-end reads at ChunLab Inc. (Seoul, South Korea). The overall scheme of the viral metagenome preparation steps is shown in the figure 2-2.

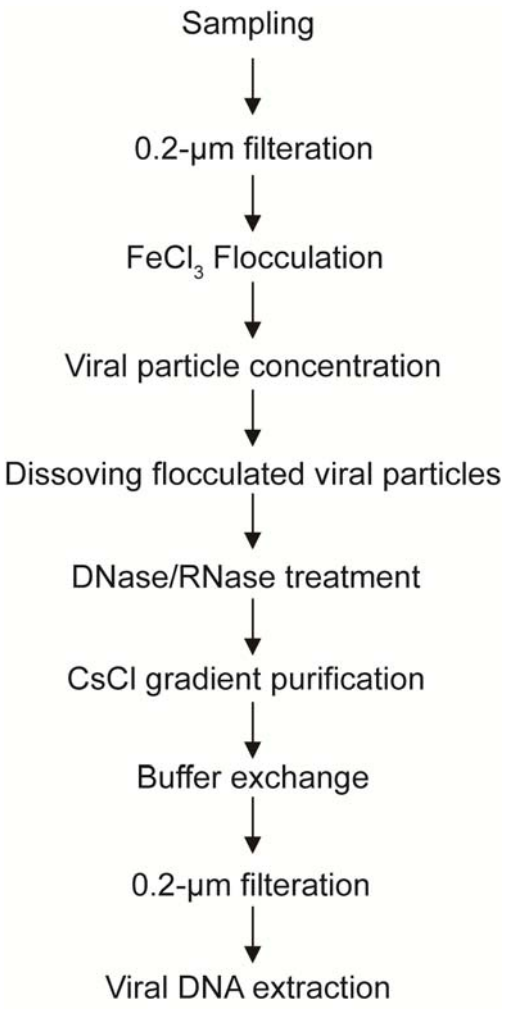

Figure 2-2. Flow chart of viral metagenome sample processing steps.

2.3. Quality trimming, assembly, and analysis of viral metagenome reads and contigs

Using the CLC Genomics Workbench (Qiagen), the raw whole genome sequencing data were mapped to the phiX174 genome for removal of sequencing control reads, followed by trimming of low quality reads using Trimmomatic program (Bolger *et al*., 2014). Although bacteria-like particles were removed, after series of bacterial cell removal processes and sterilization during viral metagenome sample process, presence of bacterial contamination was further investigated using MeTaxa program (Bengtsson-Palme *et al*., 2015). Within quality trimmed metagenome reads, presence of bacterial 16S small subunit ribosomal RNA (SSU rRNA) sequences were screened using default parameters. The trimmed reads were then assembled using SPAdes version 3.5.0 (for ’14 Oct. and ’15 Jan. samples) and 3.8.2 (for all the other samples) (Bankevich *et al*., 2012). The dissimilarity/distances between Lake Soyang virome reads were calculated using MASH algorithm (Ondov *et al*., 2016) and Non-metric multidimensional scaling (NMDS) and Principal coordinate analysis (PCoA) plots based on the dissimilarity/distance were constructed using the Vegan package and hclust provided in R (Oksanen *et al*., 2007).

All the assembled contigs from Lake Soyang were used as an input to VirSorter algorithm (Roux *et al*., 2015) to screen for contigs that are viral or prophage origin (http://de.cyverse.org/de/). The VirSorter identified viral or prophage contigs by searching for viral protein within the submitted contigs. Then based on the number of viral protein coding genes found, the submitted contigs were classified into three categories, ‘pretty sure,’ ‘quite sure,’ and ‘not so sure.’ For further analysis, only the contigs that were classified as ‘pretty sure’ and ‘quite sure’ categories were accepted and used for further analysis.

2.4. 16S rRNA amplicon sequencing for bacterial community analysis

Analysis of the bacterial community structure of Lake Soyang was performed in parallel with the virome sequencing. Along with the samples for the viral metagenome, 1 L of lake samples were also collected for bacterial 16S rRNA amplicon sequencing analysis. Six surface water samples from Lake Soyang were filtered with a mixed cellulose ester membrane filter (3.0 μm; Advantec MFS, Tokyo, Japan) to remove large-sized planktons. Then, bacterial cells were collected on a PES membrane filter (0.2-μm; Pall Corporation). The bacterial DNA was extracted from the prepared filter using Qiagen DNeasy Blood and Tissue Kit (Qiagen). The extracted DNA samples were used to amplify the V3-V4 regions of the 16S rRNA genes. Sequencing of the 16S rRNA gene amplicons was performed using the Illumina MiSeq platform at ChunLab, Inc. The sequencing data were analyzed using CLcommunity program (ChunLab).

In order to observe sequence relationship between viral metagenome sequences and 16S rRNA amplicon sequences prepared in parallel, distance matrices were constructed and analyzed. The distance matrix of viral metagenome data were prepared using MASH algorithm as described above. That of 16S rRNA amplicon sequences was constructed through OTU clustering using MOTHUR software (Schloss *et al*., 2009). Statistical analysis of correlation between distance matrices were performed using Mantel statistics test provided by the Vegan package in R (Minchin *et al*., 2015).

2.5. Phylogenetic and functional annotation of virome reads using metagenome analysis pipeline

The viral metagenome reads were analyzed using MG-RAST annotation server pipeline (Glass *et al*., 2010). Since the MG-RAST carries its own quality control processes, the viral metagenome reads that removed phiX174 mapped genes were uploaded. The MG-RAST provided taxonomic and functional annotation for each read submitted, based on RefSeq, Genbank, IMG, SEED, and Swissprot databases. The assembled contigs were uploaded to the IMG/M ER webserver (Markowitz *et al*., 2012). The IMG/M ER provided taxonomical and functional annotations of the predicted ORFs of each contigs. The assembled contigs of Lake Soyang virome are available on IMG/M ER webserver with following accession numbers: Gp0127957, Gp0127956, Gp0173525, Gp0173524, Gp0173523, and Gp013522.

2.6. Prediction of putative host bacteria of bacteriophage sequences acquired from viral metagenome

From the contigs that were assembled from Lake Soyang viral metagenome, total of 260 contigs were predicted to be viral and completely circularized. These contigs were considered as candidate bacteriophage genomes present in Lake Soyang thereby their identity and hosts were predicted. Since most of the virome contigs collected from Lake Soyang were unique and no similar viral sequences have been reported before, identification of these sequences were restricted. Therefore, protein-coding sequences of virome contigs were groups with reference viral protein sequences from NCBI RefSeq databases based on sequence similarity. The program, vContact, a program implanted in iVirus (Bolduc *et al*., 2016), allowed clustering of protein sequences based on similarity and produced protein clusters (PC) for further analysis. Then the program assigned an input contig seqeunces into different groups based on the presence of shared protein clusters, allowing assignment of unknown

bacteriophage contigs retrieved from viral metagenome to a genetically characterized group.

The putative hosts of the virome contigs were predicted by manual curation of each taxonomic annotation of ORFs within the contigs. The IMG/M ER webserver provided organism names of the best protein BLAST match to all the predicted ORFs of the virome contigs submitted. Among the virome ORFs, those predicted to be coding for bacterial proteins were considered to be acquired from their hosts. Therefore, the host of the contig was predicted when more than 40% of the taxonomic annotation results had a consensus bacterial organism.

3. RESULTS

3.1. Seasonal distribution of viral metagenome reads in Lake Soyang

From October 2014 to May 2016, six samples were collected from Lake Soyang, each representing different seasons. Two samples, ’14 Oct. and ’15 Nov. represent fall and samples, ’15 Jan. and ’15 Feb. represent winter. The samples ’15 Sept. and ’16 May each represents summer and spring, respectively. For each sample, about 10 L of lake water were used to concentrate viral particles to be sequenced. As a result, 5.5 million to 9.6 million reads were obtained through Illumina MiSeq sequencing (Table 2-2). The raw sequences were firstly quality controlled by removing reads that were mapped to phiX174 genome, which were used as sequencing control, resulting approximately 5.0 million reads to 9.5 million reads. To observe similarity of virome data, the dissimilarity distance matrix was calculated using quality controlled virome reads. Then, the distance matrix was used to construct a dendrogram and the samples appeared to be grouped into three groups according to sampling periods (Fig. 2-3). Lake Soyang virome samples of ’14 Oct. and ’15 Jan., which were the first two samples collected, were branched together, while ’15 Sept. and ’15 Nov. samples, and ’16 Feb. and ’16 May samples were grouped together. The dendrogram was expanded by calculating dissimilarities between Lake Soyang virome and those of foreign lakes and ocean, which were collected from NCBI and MetaVir server (http:/metavir-meb.univ-bpclermont.fr), a viral metagenome-specific analysis webpage. The dendrogram branches appeared to be clustering according to different lakes and oceans. Total of 26 viral metagenome data from 8 different virome projects were prepared through different methods. Out of 8 virome projects analyzed, including that of Lake Soyang, 4 of the projects used tangential flow filtration system (TFF) to concentrate viral particles from environmental samples (viromes collected from Taiwan, Michigan, US, UK, and France), 3 projects utilized Polyethylene glycol (PEG) (Virginia, US, and Canada),

Table 2-2. Viral metagenome data statistics after each quality control step

Sample	Raw sequence	ΦX 174 adaptor sequence removed		Quality trimmed		Assembled contigs	
	No.	No.	% surviving	No.	% surviving	No.	Avg. length (bp)
’14 Oct.	5,388,212	5,311,306	98.57	4,738,638	89.22	78,169	1,232
’15 Jan.	5,450,553	5,376,087	98.63	4,850,217	90.22	89,763	1,110
’15 Sept.	9,614,776	9,474,078	98.54	6,953,569	73.40	121,633	1,084
’15 Nov.	6,213,676	6,114,072	98.40	5,561,065	90.96	214,755	937
’16 Feb.	5,413,492	5,111,569	94.42	4,570,797	89.42	164,680	938
’16 May	5,193,349	4,958,959	95.49	4,321,595	87.15	140,964	1,040

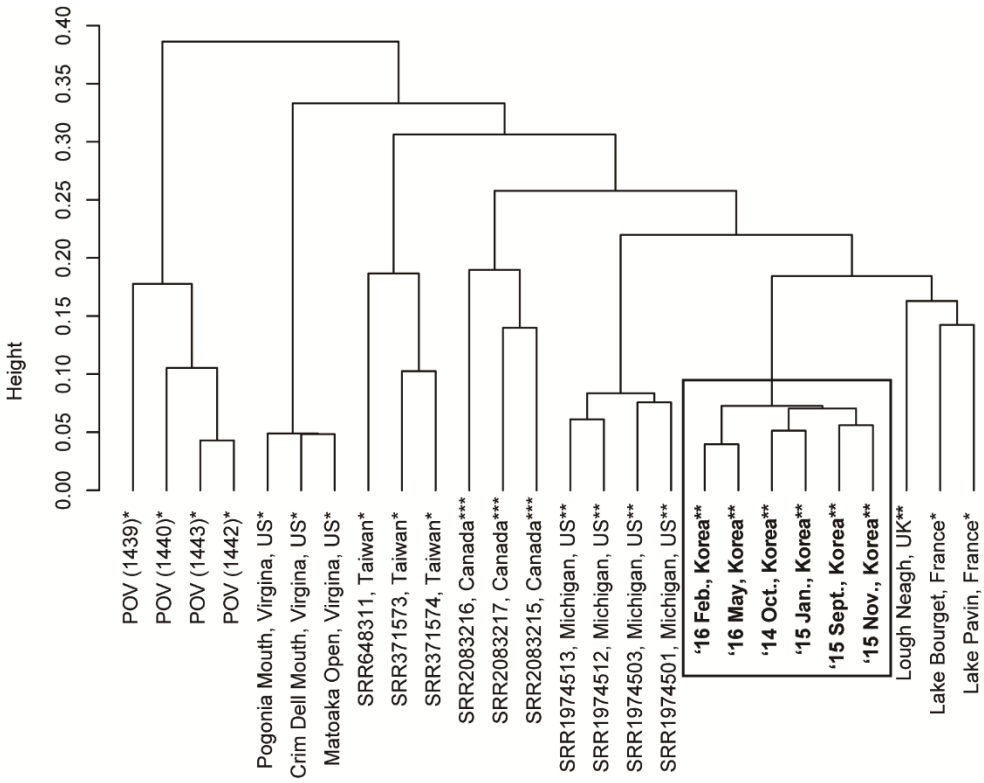

Figure 2-3. Dendrogram showing the clustering pattern of viral metagenomes prepared from aquatic environments, including Lake Soyang. The virome sequences for Pacific Ocean Virome (MetaVir ID: 1439, 1440, 1442, and 1443), Lake Matoaka, US (MetaVir ID: 2718, 2719 and 2720), Lough Neagh, UK (MetaVir ID:4925; SRR2174000), and Lake Bourget, and Lake Pavin, France (MetaVir ID: 4 and 6) were collected from the MetaVir website, while the remaining virome sequences were collected from the NCBI database. An asterisk represents metagenome sequencing done using 454 Pyrosequencing, double asterisks represent those done with Illumina MiSeq, and triple asterisks represent those done with Illumina HiSeq platform.

and 2 projects used $FeCl_3$ (POV and S. Korea), while samples prepared from France used both TFF and PEG. Also, 3 projects (Virginia, US, Taiwan, and South Korea) further concentrated viral particles specifically targeting for dsDNA phages through CsCl gradient centrifugation. Sequencing platform for all 8 projects were also different from each other, having either one of the three sequencing platforms: 454 pyrosequencing, Illumina MiSeq, and Illumina HiSeq. However, diversity of sample preparation methods or sequencing platforms appeared to cause no bias in sequence diversity among different projects – dendrogram branches were constructed irrelevant of either factors and rather, they were grouped according to sample types; freshwater or saline water, revealing that viral metagenome sequences are highly specific to their original environments.

The environmental metadata vectors were plotted onto NMDS and PCoA plots constructed based on the distance matrix, using envfit function of the Vegan package (Fig. 2-4 and 2-5). The similarity between virome sequences appeared to be significantly correlated with total nitrogen (TN) and suspended solids (SS) concentrations (Table 2-3), although influence of nitrogen and suspended solids on viral particles in environments are unclear. Overall, unlike bacterial community that are known to be highly influenced by seasonal changes due to water stratifications that occur during seasonal shifts, viral sequence distribution showed no significant differences according to seasonal changes.

Prior to further analysis, the viral metagenome sequences were analyzed for presence of bacterial sequences, which determines whether the samples were prepared properly without bacterial contamination. Using MeTaxa program, bacterial 16S SSU rRNA sequences were screened within the virome reads. Within virome samples analyzed, only 0.0001% to 0.0035% of the base pairs were appeared to be those of bacterial 16S rRNA, permitting confident neglection of possibility of

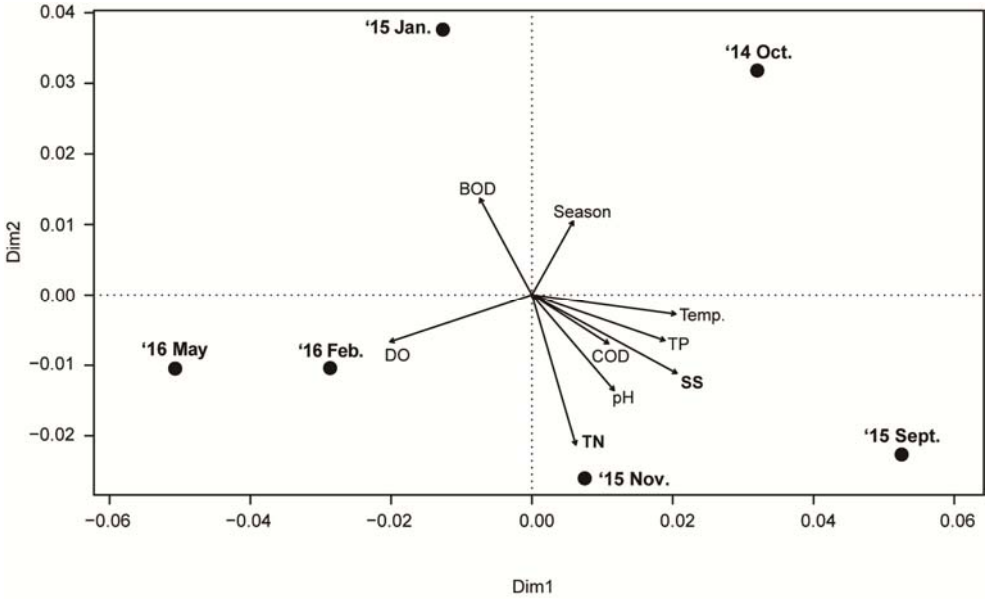

Figure 2-4. Principal coordinate analysis (PCoA) plot of six virome samples collected from Lake Soyang. The distance was calculated based on raw virome reads using MASH algorithm. Environmental vectors were added to the PCoA plot and they are depicted in arrows.

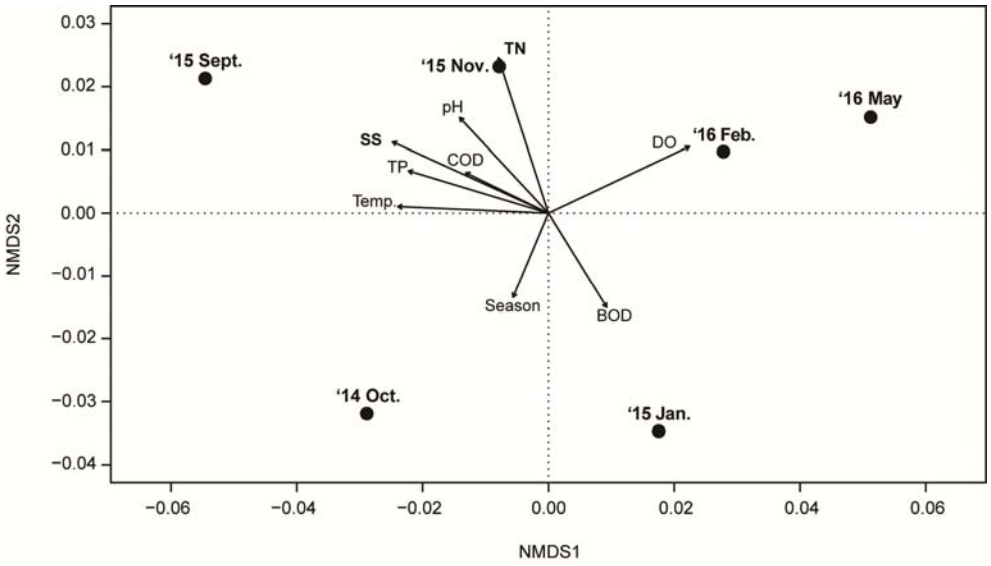

Figure 2-5. Non-metric multidimensional scaling (NMDS) plot of six virome samples collected from Lake Soyang. The NMDS distance was calculated based on raw virome reads using MASH algorithm. Environmental vectors were added to the NMDS plot and they are depicted in arrows.

Table 2-3. Envfit results of environmental data used for analysis of Lake Soyang viromes

	NMDS1	**NMDS2**	**R^2**	***p*-value**
Temp. (˚C)	-0.99905	0.04362	0.73600	0.17083
pH	-0.68245	0.73093	0.54800	0.32500
DO	0.90493	0.42556	0.78710	0.13056
BOD	0.52552	-0.85078	0.39830	0.49167
COD	-0.89890	0.43816	0.27450	0.60556
TN	**-0.30779**	**0.95145**	**0.85700**	**0.09167**
TP	-0.95755	0.28828	0.69210	0.18333
SS	**-0.91048**	**0.41356**	**0.95060**	**0.01250**
Season	-0.38688	-0.92213	0.27100	0.59444

bacterial contamination on viral metagenome sequences (Table 2-4). Therefore, the viral metagenome reads were submitted onto MG-RAST webserver for further viral metagenome analysis. According to the MG-RAST metagenome sequence quality control algorithm, the length of the metagenome reads was uniformly distributed that phylogenetic and functional assignment of the viral metagenome reads were assumed to be adequately performed without sequence length bias (Table 2-5).

3.2. Distribution of bacteriophage populations and viral protein genes in Lake Soyang

Through MG-RAST analysis server, in which that raw viral metagenome samples were uploaded, performed taxonomic assignment based on the predicted protein features of the each read submitted was performed based on BLAST algorithm (Wilke *et al*., 2012). From virome reads, 2.3 million to 4.5 million virome reads were predicted with protein coding sequences. However, only 6.40 to 12.16 % of those were identified with a known function, leaving the rest as unknown (Table 2-6). When overall known taxonomic assignment was observed at the organism level, the bacterial groups occupied more than 70% of the annotated reads, except in ’15 Sept. sample, where bacterial annotation reads occupied the total annotation reads by approximately 40% (Fig. 2-6a). Dominance of bacterial annotation in viral metagenome samples are common when general nonredundant databases, such as MG-RAST M5nr and NCBI RefSeq, are used for annotation. Because most of the environmental bacteriophages still remain undiscovered and public databases cannot represent them, environmental viral reads are often falsely assigned to bacterial taxonomy, causing limitations in viral metagenome data interpretation. Despite such biases, ’15 Sept. sample had approximately 50% of its reads assigned to viral taxonomy. Of the virus assigned reads in ’15 Sept. sample, about 85% were identified to belong to the *Myoviridae* family (Fig. 2-6b), which were mostly consist of

Table 2-4. Percent of 16S rRNA bacterial SSU sequences in Lake Soyang viral metagenome data

Site	Total bp in virome	Total 16S rRNA bp in virome	% of 16S rRNA seq. in virome
‘14 Oct.	2,222,290,309	8,806	0.00040%
‘15 Jan.	2,083,564,492	2,411	0.00012%
‘15 Sept.	3,684,907,065	9,914	0.00027%
‘15 Nov.	2,787,349,768	2,845	0.00010%
‘16 Feb.	2,405,978,956	12,080	0.00050%
‘16 May	2,354,370,573	81,912	0.00348%

Table 2-5. Average read lengths of viral metagenome collected from Lake Soyang before and after quality control (QC) of metagenome reads by a metagenome analysis server

Site	Mean seq. length (before QC)	Mean seq. length (after QC)
‘14 Oct.	296 ± 23	222 ± 61
‘15 Jan.	297 ± 20	215 ± 63
‘15 Sept.	288 ± 32	202 ± 71
‘15 Nov.	297 ± 21	240 ± 60
‘16 Feb.	298 ± 18	237 ± 61
‘16 May	295 ± 28	240 ± 61

Table 2-6. Ratio of predicted and identified protein features of Lake Soyang virome, calculated a metagenome analysis server

Samples	Predicted protein features	Identified protein features	Percentage
’14 Oct.	2,345,903	209,551	8.93
’15 Jan.	2,654,483	169,840	6.40
’15 Sept.	4,111,637	313,281	7.62
’15 Nov.	4,501,168	410,673	9.12
’16 Feb.	3,600,573	279,516	7.76
’16 May	2,411,170	293,192	12.16

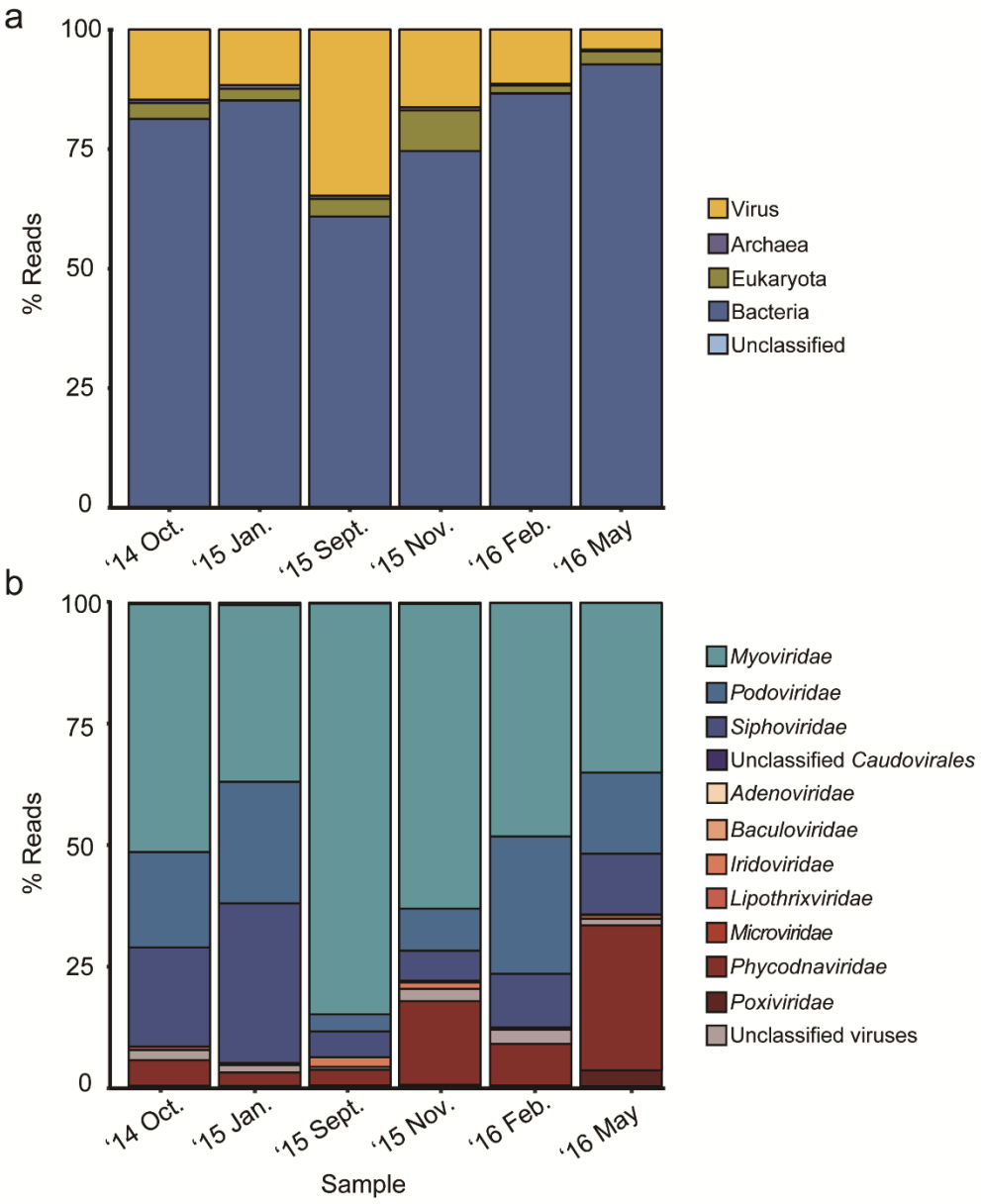

Figure 2-6. Taxonomic annotation of Lake Soyang virome samples by a metagenome analysis server. (a) Proportion of different domains that were annotated from Lake Soyang virome raw reads are depicted. (b) Proportion of different families of virus are shown.

cyanophages (Table 2-9). This phenomenon corresponded with temporal cyanobacterial bloom that took place in the same period (Fig. 2-7). Although cyanobacteria showed peak in the summer season and appeared at low abundance in others, the cyanophages were consistently thriving in Lake Soyang throughout different seasons (Fig. 2-8 and Table 2-7 to 12), making presence in the lake environment independent of their hosts. Other than cyanobacteria and cyanophages, no relationship between bacterial species and viral groups were able to be identified.

Due to deficiency in taxonomic assignment of viral metagenome reads, observation of relationship between bacteriophage and bacteria was restricted. Therefore, to be independent from limitations caused by shortness in viral genome databases, raw sequences of viral metagenome and bacterial 16S rRNA amplicon sequencing were compared through distance matrix-matrix analysis. Using the Mantel test provided by the Vegan package of R, three statistic tests were performed; Kendall, Pearson, and Spearman. All three correlation tests presented that viral metagenome and 16S rRNA amplicon sequencing had significant positive correlation, indicating that compositions of two sequences are varying together (Table 2-13).

Along with the taxonomic annotations, protein functional group annotation based on predicted protein coding genes of the virome reads were obtained through the MG-RAST analysis pipeline. The functional proteins were grouped according to subsystems of the SEED database (Overbeek *et al*., 2014). Overall, the protein groups assigned to bacteriophage-related proteins were dominating, occupying up to 55.50% of the total reads assigned to a protein functional group. However, in two samples, ’15 Nov. and ’16 May, virome reads that were assigned to bacteriophage-related protein groups appeared to be exceptionally low (Fig. 2-9 and Table 2-14). Although number of viral particles present in the ’16 May sample was covering approximately 30% of annotated reads similar to those of other samples

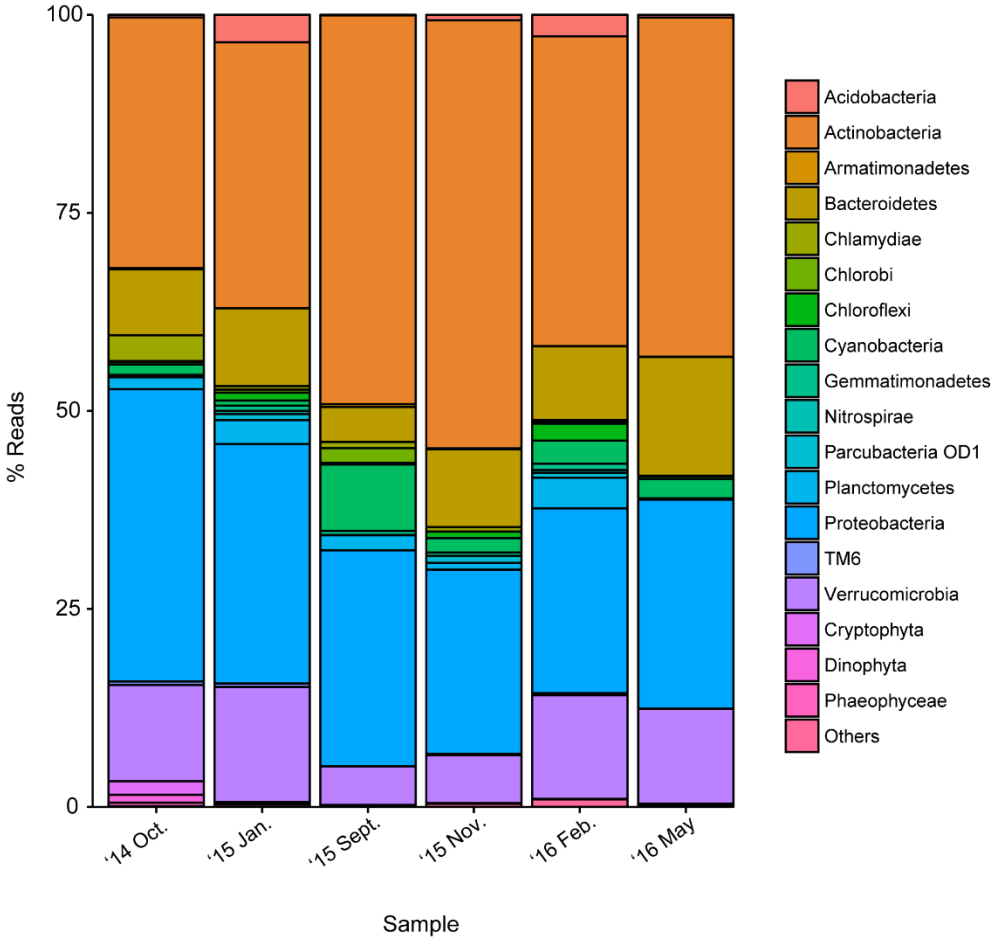

Figure 2-7. Taxonomic assignments and distribution of 16S rRNA sequences obtained from Lake Soyang. The 16S rRNA amplicon sequencing was performed using Illumina MiSeq platform and their taxonomic assignments were performed using CLcommunity program

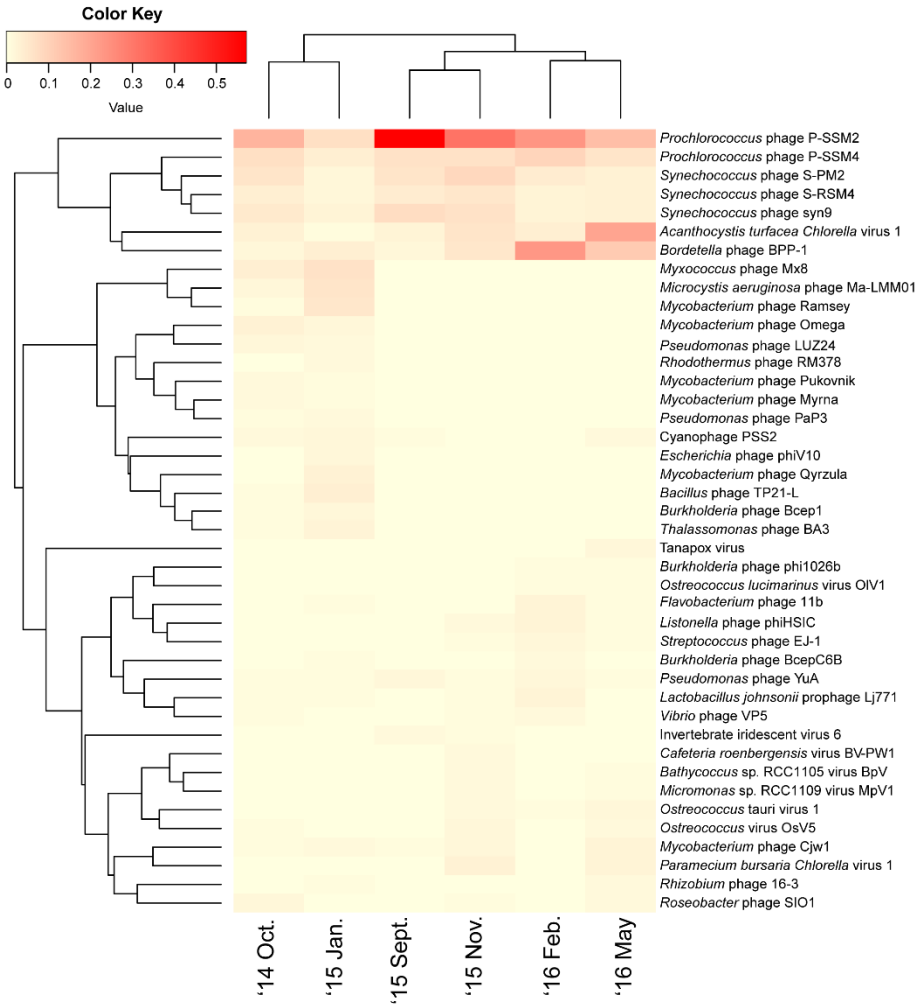

Figure 2-8. Heatmap generated by comparison of annotated viral reads from Lake Soyang. The virus annotation was performed by MG-RAST annotation server and only annotated viral species that had relative abundance of more than 1% in at least 1 sample were shown here.

Table 2-7. List of 15 viruses that were most frequently detected within the viral metagenome reads of '14 Oct. sample collected from Lake Soyang

SY – '14 Oct.			
Species name	**%**	**Norm%[a]**	**Origin of isolation**
Prochlorococcus phage P-SSM2	16.61	1.13	Marine
Prochlorococcus phage P-SSM4	7.13	0.69	Marine
Synechococcus phage S-PM2	6.20	5.44	Marine
Synechococcus phage syn9	4.69	45.58	Marine
Myxococcus phage Mx8	3.94	1.37	Soil
Synechococcus phage S-RSM4	3.89	0.34	Marine
Acanthocystis turfacea Chlorella virus 1	3.25	0.19	Freshwater
Mycobacterium phage Omega	2.99	0.46	Unknown
Pseudomonas phage LUZ24	2.13	0.80	Hospital sewage
Bordetella phage BPP-1	2.06	0.84	Animal lung
Microcystis aeruginosa phage Ma-LMM01	2.03	0.22	Freshwater
Roseobacter phage SIO1	1.87	0.81	Marine
Cyanophage PSS2	1.49	2.39	Marine
Mycobacterium phage Pukovnik	1.33	0.43	Unknown
Mycobacterium phage Myrna	1.22	0.13	Unknown

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-8. List of 15 viruses that were most frequently detected within the viral metagenome reads of '15 Jan. sample collected from Lake Soyang

SY – '15 Jan.			
Species name	**%**	**Norm%[a]**	**Origin of isolation**
Prochlorococcus phage P-SSM2	7.31	0.63	Marine
Myxococcus phage Mx8	6.59	2.90	Soil
Microcystis aeruginosa phage Ma-LMM01	5.92	0.79	Freshwater
Mycobacterium phage Ramsey	5.19	1.93	Unknown
Bordetella phage BPP-1	3.99	2.04	Animal lung
Prochlorococcus phage P-SSM4	3.95	0.48	Marine
Bacillus phage TP21-L	3.74	2.17	Unknown
Mycobacterium phage Qyrzula	3.13	1.01	Unknown
Synechococcus phage syn9	2.84	34.87	Marine
Thalassomonas phage BA3	2.79	1.63	Coral
Burkholderia phage Bcep1	2.14	0.97	Plant root
Cyanophage PSS2	2.12	4.29	Marine
Mycobacterium phage Omega	1.95	0.38	Unknown
Escherichia phage phiV10	1.89	1.05	Unknown
Synechococcus phage S-RSM4	1.80	0.20	Marine

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-9. List of 15 viruses that were most frequently detected within the viral metagenome reads of '15 Sept. sample collected from Lake Soyang

SY – '15 Sept.			
Species name	**%**	**Norm%[a]**	**Origin of isolation**
Prochlorococcus phage P-SSM2	57.08	3.86	Marine
Synechococcus phage syn9	7.47	71.94	Marine
Prochlorococcus phage P-SSM4	6.48	0.62	Marine
Synechococcus phage S-PM2	6.04	5.25	Marine
Synechococcus phage S-RSM4	4.32	0.38	Marine
Acanthocystis turfacea Chlorella virus 1	2.82	0.17	Freshwater
Pseudomonas phage YuA	1.85	0.54	Hospital sewage
Bordetella phage BPP-1	1.74	0.70	Animal lung
Invertebrate iridescent virus 6	1.37	0.11	*Drosophila*
Mycobacterium phage Cjw1	1.00	0.22	Unknown
Mycobacterium phage D29	0.94	0.33	Unknown
Synechococcus phage Syn5	0.71	0.26	Marine
Invertebrate iridescent virus 3	0.59	5.27	*Drosophila*
Cyanophage PSS2	5.59	8.88	Marine
Enterobacteria phage T4	0.59	0.06	Unknown

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-10. List of 15 viruses that were most frequently detected within the viral metagenome reads of '15 Nov. sample collected from Lake Soyang

SY – '15 Nov.			
Species name	**%**	**Norm%[a]**	**Origin of isolation**
Prochlorococcus phage P-SSM2	31.17	2.29	Marine
Synechococcus phage S-PM2	8.58	8.12	Marine
Prochlorococcus phage P-SSM4	6.57	0.68	Marine
Synechococcus phage syn9	6.38	66.87	Marine
Acanthocystis turfacea Chlorella virus 1	5.91	0.38	Freshwater
Synechococcus phage S-RSM4	5.63	0.54	Marine
Bordetella phage BPP-1	5.32	2.33	Animal lung
Paramecium bursaria Chlorella virus 1	3.12	0.18	Freshwater
Mycobacterium phage Cjw1	1.91	0.47	Unknown
Ostreococcus virus OsV5	1.72	0.17	Marine
Micromonas sp. RCC1109 virus MpV1	1.67	0.17	Marine
Listonella phage phiHSIC	1.50	0.73	Marine
Cafeteria roenbergensis virus BV-PW1	1.29	0.04	Marine
Ostreococcus tauri virus 1	1.28	0.12	Marine
Bathycoccus sp. RCC1105 virus BpV1	1.16	0.11	Marine

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-11. List of 15 viruses that were most frequently detected within the viral metagenome reads of ’16 Feb. sample collected from Lake Soyang

SY – ’16 Feb.			
Species name	**%**	**Norm%[a]**	**Origin of isolation**
Prochlorococcus phage P-SSM2	23.97	2.96	Marine
Bordetella phage BPP-1	23.22	17.01	Animal lung
Prochlorococcus phage P-SSM4	9.33	1.63	Marine
Synechococcus phage S-PM2	4.36	6.92	Marine
Acanthocystis turfacea Chlorella virus 1	3.86	0.42	Freshwater
Synechococcus phage S-RSM4	2.60	0.42	Marine
Flavobacterium phage 11b	2.56	2.21	Sea-ice
Synechococcus phage syn9	2.56	44.96	Marine
Lactobacillus johnsonii prophage Lj771	2.48	1.89	Human intestine
Listonella phage phiHSIC	2.44	2.00	Marine
Pseudomonas phage YuA	2.12	1.13	Marine
Streptococcus phage EJ-1	1.78	1.29	Unknown
Vibrio phage VP5	1.51	1.18	Wastewater
Burkholderia phage BcepC6B	1.25	0.92	Plant root
Ostreococcus lucimarinus virus OlV1	1.06	0.17	Marine

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-12. List of 15 viruses that were most frequently detected within the viral metagenome reads of ’16 May sample collected from Lake Soyang

SY – ’16 May			
Name	**%**	**Norm%[a]**	**Origin of isolation**
Acanthocystis turfacea Chlorella virus 1	20.14	1.86	Freshwater
Prochlorococcus phage P-SSM2	14.77	1.55	Marine
Bordetella phage BPP-1	1.93	1.21	Animal lung
Prochlorococcus phage P-SSM4	6.27	0.93	Marine
Synechococcus phage S-RSM4	3.35	0.46	Marine
Synechococcus phage syn9	3.33	49.90	Marine
Synechococcus phage S-PM2	3.17	4.29	Marine
Mycobacterium phage Cjw1	2.71	0.95	Unknown
Paramecium bursaria Chlorella virus 1	2.30	0.18	Freshwater
Ostreococcus tauri virus 1	1.86	0.26	Marine
Rhizobium phage 16-3	1.59	0.70	Plant root
Roseobacter phage SIO1	1.51	1.01	Marine
Yaba-like disease virus	1.32	0.24	Animal skin
Ostreococcus virus OsV5	1.16	0.17	Marine
Cyanophage PSS2	1.16	2.87	Marine

[a] Proportion of viral metagenome reads assigned to a taxonomic nomenclature was normalized by genome length of the assigned reference genome.

Table 2-13. Distance matrix-matrix correlation between bacterial 16S rRNA amplicon sequences and viral metagenome sequences collected from Lake Soyang

Correlation test	***r***	**Significance**
Kendall	0.5048	0.0042
Pearson	0.6805	0.0042
Spearman	0.6929	0.0028

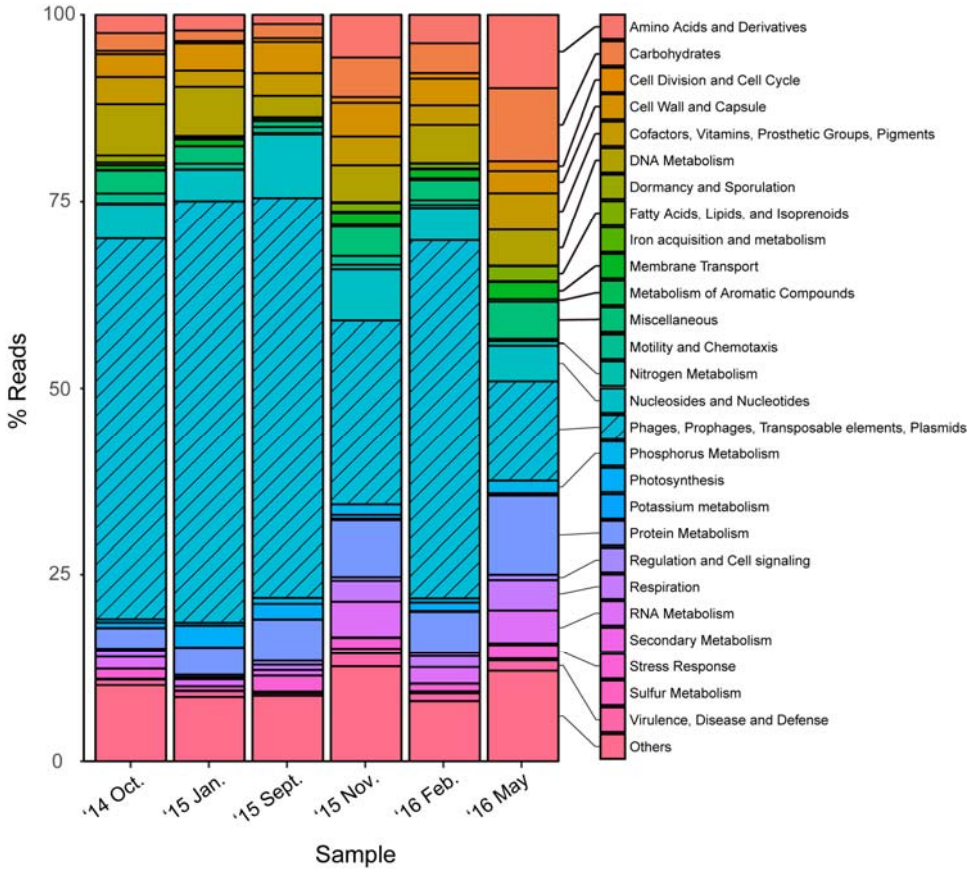

Figure 2-9. Functional gene annotation of Lake Soyang virome samples. The functional gene annotation was performed by MG-RAST annotation server. The annotation was done based on the SEED subsystem database and annotation matches with e-value threshold of 1.00E-5 were taken into consideration for data analysis.

Table 2-14. Proportion of each Lake Soyang virome reads that were assigned to function annotation categories by a metagenome analysis server

Function category	'14 Oct. (%)	'15 Jan. (%)	'15 Sept. (%)	'15 Nov. (%)	'16 Feb. (%)	'16 May (%)
Amino Acids and Derivatives	2.45	2.11	1.23	5.72	3.81	9.85
Carbohydrates	2.38	1.41	1.90	5.32	4.02	9.73
Cell Division and Cell Cycle	0.44	0.30	0.51	0.78	0.74	1.31
Cell Wall and Capsule	3.07	3.66	4.20	4.44	3.58	2.99
Clustering-based subsystems	11.01	9.51	8.83	12.78	8.88	13.12
Cofactors, Vitamins, Prosthetic Groups, Pigments	3.65	2.18	3.03	3.87	2.55	4.93
DNA Metabolism	6.80	6.53	2.78	4.96	5.14	4.85
Dormancy and Sporulation	0.95	0.22	0.01	0.13	0.06	0.09
Fatty Acids, Lipids, and Isoprenoids	0.34	0.21	0.08	1.16	0.65	1.98
Iron acquisition and metabolism	0.04	0.02	0.01	0.17	0.12	0.12
Membrane Transport	0.62	0.90	0.32	1.60	1.20	2.31
Metabolism of Aromatic Compounds	0.10	0.03	0.11	0.25	0.24	0.33
Miscellaneous	3.08	2.32	0.80	3.96	2.72	4.89
Motility and Chemotaxis	1.33	0.76	0.81	1.20	0.67	0.29
Nitrogen Metabolism	0.16	0.07	0.18	0.60	0.41	0.63
Nucleosides and Nucleotides	4.58	4.26	8.58	6.84	4.32	4.77
Phages, Prophages, Plasmids, and Transposable elements	50.11	55.50	53.45	24.50	47.08	12.29
Phosphorus Metabolism	0.45	0.39	0.80	1.41	0.56	1.69
Photosynthesis	0.71	2.96	2.09	0.52	1.13	0.16
Potassium metabolism	0.04	0.03	0.03	0.19	0.13	0.16
Protein Metabolism	2.77	3.64	5.57	7.59	5.54	10.57
RNA Metabolism	1.59	0.95	0.74	4.80	2.18	4.43
Regulation and Cell signaling	0.29	0.33	0.54	0.51	0.43	0.72
Respiration	0.82	0.29	0.71	2.79	1.50	4.08
Secondary Metabolism	0.03	0.02	0.02	0.09	0.05	0.26
Stress Response	1.35	0.55	2.13	1.46	1.02	1.80
Sulfur Metabolism	0.12	0.08	0.22	0.64	0.25	0.26
Virulence, Disease and Defense	0.73	0.78	0.32	1.75	1.05	1.39

(4.05×10^7 particles per ml, Table 2-1), number of reads that were taxonomically assigned to viruses were also the lowest in ’16 May. Such disproportionality hints at the abundance of novel bacteriophages in the ’16 May sample with unique genome sequences that were not able to be identified with current databases.

3.3. Novel bacteriophage contigs recovered from viral metagenome

Using the SPAdes assembler, contigs were assembled from viral metagenome reads prepared from Lake Soyang. Among assembled contigs, only those with 10 kbp or longer were used for further analysis to assure that the contig is a viable bacteriophage genome candidate. The obtained contigs were determined whether they are viral origin or not using the VirSorter algorithm. Thus, only 1.28 to 4.02% of the assembled contigs were confidently identified as viral origin (Table 2-15). Among them, total of 260 circularized viral contigs were identified for all 6 samples combined. These contigs were considered as complete bacteriophage sequences. As seen from analysis of viral metagenome reads, taxonomic annotation of metagenome reads based on BLAST with public genome database is highly limited due to insufficient number of viral sequences in databases and lack of universal viral marker genes that could allow classification of the sequences. Therefore, identification of virome contigs were attempted through clustering of protein coding sequences based on similarity of protein-coding sequences, alone. Using the VirSorter algorithm, protein-coding sequences for all viral-predicted contigs were computed. Therefore, those protein sequences were identified as viral origin and were used to build protein clusters along with viral protein sequences collected from NCBI RefSeq database (release 79), using vContact program based on sequence similarity. As a result, total of 693 groups with similar sequences were constructed. Among those, 211 groups were consisted of contigs retrieved from viral metagenome only, representing unreported viral sequences that cannot be found on public databases. Of all, 28 largest groups that consisted of 50 or more viral

Table 2-15. Number of viral metagenome contigs that were identified as virus or prophage by VirSorter

Sample	Contigs	Viruses		Prophages
		Contigs	Complete contigs	Contigs
’14 Oct.	78,169	2,043	53	8
’15 Jan.	89,763	2,512	58	6
’15 Sept.	121,633	2,743	13	10
’15 Nov.	214,755	3,265	49	13
’16 Feb.	164,680	3,141	45	10
’16 May	140,964	2,868	42	6

sequences covered 32.9% of total number of sequences that were clustered and most of them were comprised or either reference sequences of virome contigs only. Groups consisted of references sequences only often showed grouping of viral proteins that belonged to an identical viral or bacterial taxonomic groups thereby providing evidence that the protein clusters are reliable to predict taxonomic groups for each ORF group. For example, groups 4, 10, and 15 were comprised of both virome contigs and reference sequences, that taxonomic prediction of virome contigs were possible (Fig. 2-10). Contigs clustered to groups 4 and 10 were predicted to be those of cyanophages or enteric bacteria phages, respectively, according to the reference sequences found within these groups. However, the reference seqeunces of group 15 showed no consensus characteristics that 14 virome contigs that were assigned to the group were not able to be identified. Including those of group 15, the remaining Lake Soyang virome contigs (2,030 ORFs) remained unclassified, thus concluded as unreported viral protein sequences found from Lake Soyang.

For more detailed analysis, 260 circularized and complete viral contigs were used for taxonomic assignment using public databases. Each ORF within the circularized contigs were predicted with a taxonomic assignment through BLAST with NCBI RefSeq and IMG protein database. However, most of the ORFs of circularized contigs were not found within either NCBI or IMG databases. ORFs that were able to be identified by the public databases were mostly assigned to bacterial groups. Thereby, assuming that bacterial protein sequences within a bacteriophage genome are the results of HGT during infection, number of ORFs that were assigned to a bacterial taxonomy were counted to predict their putative hosts. When more than 40% of the ORFs with a taxonomic assignment had a consensus, then the consensus bacterial group was predicted as a putative host (Table 2-16). Out of 260 contigs, most of the them were predicted to have hosts belonging to the phylum *Proteobacteria* (66.54%) and about half of those were not able to be identified with hosts with lower taxonomic level because they did not have consensus bacterial

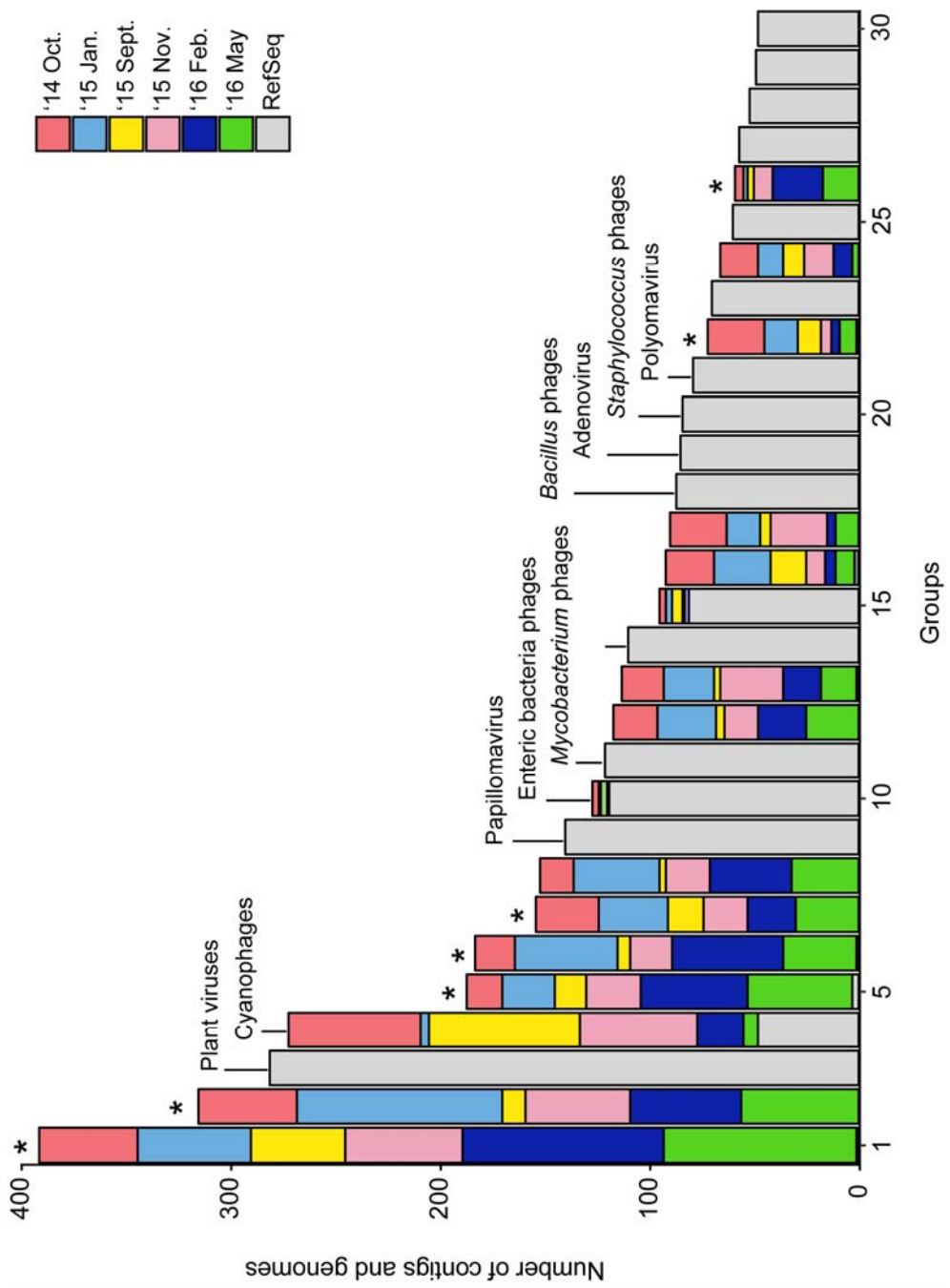
'14 Oct.
'15 Jan.
'15 Sept.
'15 Nov.
'16 Feb.
'16 May
RefSeq
Number of contigs and genomes
0
100
200
300
400
Groups
1
5
10
15
20
25
30
Plant viruses
Cyanophages
Papillomavirus
Enteric bacteria phages
Mycobacterium phages
Bacillus phages
Adenovirus
Staphylococcus phages
Polyomavirus
*

Figure 2-10. Viral sequence groups constructed based on shared protein clusters between viral metagenome contigs and references sequences collected from the RefSeq database. The taxonomic assignment of each group was made based on the reference sequences that were included in the groups. For the groups with asterisks, the hosts of the virome contigs were predicted based on IMG taxonomic annotations. The predicted host of the groups 1, 5, 6, 7, and 26 was predicted to belong to the phylum *Proteobacteria*, those of group 2 was predicted to belong to the phylum *Actinobacteria*, and that of group 22 was predicted to belong to the phylum *Bacteroidetes*.

Table 2-16. Predicted hosts of complete bacteriophage contigs retrieved from viral metagenome samples.

	'14 Oct.	**'15 Jan.**	**'15 Sept.**	**'15 Nov.**	**'16 Feb.**	**'16 May**	**Total**
*Proteobacteria*_p	26	14	6	18	18	12	94
*Alphaproteobacteria*_c	4	7	1	5	5	4	26
*Rhizobiales*_o	1	6	0	2	2	0	11
*Betaproteobacteria*_c	1	6	1	3	2	2	15
*Burkholderiales*_o	4	3	0	1	6	2	16
*Gammaproteobacteria*_c	2	4	2	2	1	0	11
*Actinobacteria*_p	4	5	2	3	0	3	17
*Bacteroidetes*_p	1	2	0	2	5	6	16
*Cyanobacteria*_p	1	0	0	0	0	0	1
*Firmicutes*_p	0	1	0	2	1	2	6
*Planctomycetaceae*_f	0	0	0	1	0	0	1
Unknown	9	10	1	10	5	11	46
Total	53	58	13	49	45	42	260

taxonomic assignments. Within the phylum *Proteobacteria, Alphaproteobacter* was shown to be the most prevalently infected bacterial group. Using this method, the furthest taxonomic level of bacterial host that can be predicted was the family level, which was *Planctomycetaceae* (’15 Nov.). No more detailed prediction on putative bacterial host was able to be made because the virome contigs were compsed of too diversely originated ORFs. Interestingly, although virome reads’ taxonomic assignment showed that ’15 Sept. sample were dominated by cyanophages (Fig. 2-6b, Table 2-9), no complete contig from the sample was shown to belong to cyanophages. Rather, only one contig, ’14 Oct.-25, showed high resemblance with a *Synechococcus* phage S-CBS4 (70% identity). Since bacteriophage genomes are usually composed of ORFs of multiple bacteria, prediction of their hosts was restrictive. Thereby among the complete contigs, 46 of them were not able to be predicted on their hosts nor their identity (Table 2-16). Complete viral sequences with predicted hosts were reciprocally tracked within the viral protein-clustering groups that were constructed earlier in attempts to reveal the identity of both complete sequences and viral groups. As a result, 15 complete viral contigs that had the phylum *Proteobacteria* as predicted host group belonged to the viral sequence group 1 and eight complete contigs that had the phylum *Actinobacteria* as predicted host group were found within the viral sequence group 2, allowing indirect host prediction of 706 contigs belonging to groups 1 and 2 (Fig. 2-10). In the same context, viral sequence groups 5, 6, 7, and 26 were also revealed to have the phylum *Proteobacteria* as their host group and group 22 appeared to have the phylum *Bacteroidetes* as their putative host group. Although the host prediction using protein clustering and protein sequence annotation were not able to provide delicate and detailed prediction on virome contigs, 1,872 viral contigs among 2,142 virome sequences that are distinctive from existing databases were able to be predicted with their putative host groups at the phylum level, providing the first steppingstones for isolation of unreported freshwater bacteriophages.

4. DISCUSSION

In this study, freshwater bacteriophage population distribution was observed using metagenome approach in Lake Soyang, the largest artificial lake in South Korea. To observe seasonal changes of the viral population, 6 samples were collected from 2014 to 2016, at different seasons. After collecting and concentrating viral particles only, the viral DNA was sequenced for each metagenomic analysis. Using the raw sequences obtained, the dissimilarity distance index was calculated to observe the sequence similarities without any bias caused by known sequence databases. The viral metagenome sequences of all 6 samples appeared to be highly similar with each other that no significant difference between seasons was observed. A dendrogram constructed based on distance index of metagenome sequences showed that viral samples can be grouped into three branches, not by seasons but by sampling periods. Samples collected on October 2014 and January 2015 formed a branch, while those collected on September 2015 and November 2015 formed another and those of February 2016 and May 2016 formed one (Fig. 2-3). This indicated that viral population is shifting gradually over time, rather than having seasonal cycles. Besides the temporal changes, environmental influence on virome read distribution was observed through plotting NMDS and PCoA plots. Although TN and SS appeared to be having positive correlation with viral sequence compositions (Table 2-3). It is widely known that as lytic bacteriophage infects their hosts and lysis the bacterial cell for release of newly produced bacteriophage particles, organic carbon and nitrogen that were constituting the bacterial cell are released to the surroundings. It has been calculated that viral lysis could release 4-40 nM of nitrogen per day when it is assumed that 2-20% of bacterioplankton are lysed by bacteriophages per day (Hewson and Fuhrman, 2008). Therefore, the positive correlation of TN concentration in Lake Soyang and viral metagenome reads may indicate bacteriophages are actively lysing bacterial cells for their reproduction.

When the viral population changes were observed after the sequenced reads were assigned to taxonomic groups, no seasonal pattern was observed as well, except for cyanophage bloom that occurred in the summer season. In the ’15 Sept. sample, the number of reads that were assigned to the family *Myoviridae* was significantly high, occupying 85% of all the reads that were assigned to viruses. The corresponding 16S rRNA amplicon sequencing that were performed in parallel with viral metagenome, showed increased proportion of cyanobacteria in ’15 Sept. sample. The correlation between cyanobacteria and myovirus population has been observed before in many different environments, especially in oceans (Chow and Fuhrman, 2012). Hence, paired seasonal abundance of cyanobacteria and myovirus population seems to be a universal phenomenon that could be observed in both marine and freshwater environments. Virome reads assignment at the species level also showed high abundance of cyanophages in ’15 Sept., the summer sample (Fig. 3-8). Especially, *Prochlorococcus* phage P-SSM2, which showed high dominance in viral groups of ’15 Sept. sample, showed gradual decrease in its abundance after the bloom. When virome reads were annotated with taxonomic assignments, cyanophage bloom was most observable. This may be due to actual Cyanobacteria-cyanophage bloom that took place, but also may be a biased observation due to relatively well represented and identified cyanophages. Among environmental bacteriophages, cyanophages are the most identified and well-studied group. Therefore, public databases contain relatively high number of cyanophages genomes, which may have led to biased interpretation in viral metagenome reads. Seasonal variation was also not visible when predicted functional genes were annotated from virome reads. Clear seasonal pattern of viral population distribution may not have been observable due to small sample size. However, low resolution of viral population annotation, which was caused by limited size of the viral database with low representation of viral population may have caused limitations in interpretation of viral metagenome data as well.

From the MG-RAST annotation server, approximately 87.84% to 93.60% of the virome reads that were submitted were predicted to having protein sequences but not able to be identified, indicating that Lake Soyang virome is mostly consisted of unreported viral sequences (Table 2-6). Therefore, those sequences were subjected for further analysis after construction of viral contigs from viral metagenome reads. Using virome contigs that were 10 kbp or longer, protein sequence clusters were constructed based on sequence similarity with protein sequences collected from NCBI RefSeq viral database. As a result, total of 693 groups were established and among those, 211 groups were composed of Lake Soyang virome contigs only. Taxonomic prediction for only 16 groups that consisted of both Lake Soyang virome contigs and reference sequences were able to be made based on the reference sequences that they were grouped with.

From the viral groups that were only consisted of Lake Soyang virome contigs, circularized contigs were collected for manual curation of their host prediction. Through IMG/M ER annotation server and RefSeq database, ORFs of the circularized contigs were predicted and taxonomic assignment was made for each ORF. The taxonomic predictions made on all ORFs carried by a single contig were summarized and when more than 40% of the taxonomic predictions had consensus to a specific bacterial organism, it was accepted as a predicted host. Most viral genomes are composed of diverse protein coding genes of bacterial origin, those are suspected to be obtained from horizontal gene transfer, during infection. Therefore, host prediction was made according to bacterial taxonomic assignments made on each ORFs. As a result, 7 major groups that were unique to Lake Soyang virome were able to be predicted with their putative hosts. Within 28 largest groups that consisted of 50 contigs or more, 5 Lake Soyang groups (groups 1, 5, 6, 7, and 26) were identified to have putative hosts within the phylum *Proteobacteria* and 2 groups were each predicted to have hosts within in the phylum *Actinobacteria* (group 2) and *Bacteroidetes* (group 22). Hence, total of 976 contigs were newly identified as

bacteriophages that could possibly infect hosts within the phylum *Proteobacteria*, 315 as those infecting a member of the phylum *Actinobacteria*, and 59 as those infecting those of the phylum *Bacteroidetes*.

In the era of massive viral metagenome data, interpretation of those data is often being hampered by limited number of viral sequences that have been identified and sequenced before. Although large number of unique viral contigs were found in viral metagenome study performed in Lake Soyang, most of them were not able to be identified expanding more dark matter to the environmental viral genome database. Therefore, although viral metagenome approach was suggested to overcome the limitations of the bacteriophage cultures, it is the experimental bacteriophage culture and isolation that could provide information for viral metagenome analysis. Therefore, two methods, viral metagenome and bacteriophage isolation, must be accompanied together for better understanding of bacteriophage ecology in diverse environments.

CHAPTER 3.

Genomic and Ecological Study on Novel Bacteriophages Isolated from Lake Soyang

ABSTRACT

Recently, through number of viral metagenome studies performed in diverse environments, thousands of putative novel bacteriophage genomes became known. However, most of those genomes remain as unknown or unclassified due to dearth of environmental bacteriophage genome databases that could be used to identify viral sequences retrieved from viral metagenome. Therefore, physical isolation and genome sequencing of individual bacteriophages are necessary not only for expanding our knowledge on environmental microbial ecology but also for better interpretation of viral metagenome data. From Lake Soyang, the largest conserved freshwater lake in South Korea, 4 new bacteriophages were isolated using 3 different bacterial strains that were also isolated from Lake Soyang; P19250A infecting a strain belonging to the LD28 group, P26059A and P26059B infecting *Curvibacter* sp., and P26218 infecting *Rhodoferax* sp. strain. The bacteriophages P19250A and P26059A appeared to be members of the *Siphoviridae* family and P26218 and P26059B was those of the *Podoviridae* family. Through Illumina MiSeq platform, whole genome of all four phages have been sequenced. Using obtained sequences, the binning analyses were performed on freshwater viromes and it was shown that the phage P19250A was the most highly-assigned freshwater phage (up to 8.7%) in Lake Soyang. Also, the proportion of P19250A-assigned reads fluctuated following the seasonal abundance of LD28 clade in Lake Soyang, which indicated host-dependent bacteriophage population shifts. The phages P19250A and P26218 showed seasonal preference in winter. Meanwhile, P26059A showed weak seasonality in summer season, reflecting bacterial host abundance according to seasonal changes. These results showed that novel bacteriophages isolated from Lake Soyang and their genomes would be essential resources for understanding freshwater bacteriophage community and also suggest that phages of other abundant freshwater bacteria need to be isolated as well.

1. INTRODUCTION

Bacteria, which are present in diverse environments, are known to thrive in lake waters at 10^6 cells per ml and, their predators, bacteriophages (phages), are calculated to be present as 10-times more abundant than their hosts, being the most abundant biological entities on the Earth (Ignacio-Espinoza *et al*., 2013). Phages in the environment often take lytic life cycle, which they aggressively reproduce through host cell lysis, and thereby actively control the bacterial populations. The phages also participate in nutrient cycle in aquatic environments. Not only that they interfere with various nutrient cycles performed by bacterial cells by predation, but they also contribute in increase of the dissolved carbon source in the environment through lysis of bacterial cells and release of cell debris into the surrounding system (Guidi *et al*., 2016). Phages also participate in proliferation of bacterial gene diversity by mediating horizontal gene transfer through unintentional carriage of host gene fragments while infecting one host after the other (Yu *et al*., 2016). In process of carrying their host genomes within the phage capsid, some phages acquired bacterial functional genes that could benefit both the host and itself. Those bacterial functional genes carried by phages are called auxiliary metabolic genes (AMG) and those genes are known to be involved in photosystem, glycolysis, and phosphorous, sulfur, and nitrogen cycling (Adriaenssens and Cowan, 2014; Hurwitz and U'Ren, 2016; Sharon *et al*., 2009). The AMGs enter the bacterial cell and are expressed to enhance cell metabolism upon phage infection and eventually benefit phage reproduction. Despite diverse functions of the phages in the environment, recognition and appreciation of environmental phages are highly lacking. In the NCBI Genome database, as of March 2017, total of 91,075 prokaryote genomes are available while only 7,140 viral genomes are accessible. Among those, bacteriophage genomes are even less – 2,101 genomes sequenced and those of environmental phages are expected to be lesser. Despite high number of phage

particles found in the environments, number of identified bacteriophages is relatively low due to hardships in culturing individual phages. One of the major limitations in culturing phages is that they require their hosts to be cultured before phage screening to begin. Only few environmental bacterial species have been cultured in laboratory settings so far, and even if they are successfully cultured, maintaining those strains in artificial media are challenging, which limits the attempts to culture their bacteriophages.

In order to overcome the culturability restrictions, viral metagenome approach, also known as virome, was suggested to discover unknown environmental bacteriophage sequences (Edwards and Rohwer, 2005). Without the need of culturing both bacteria and phages, virome analysis allowed access to large amount of bacteriophage genomes in diverse environments such as ocean (Angly *et al*., 2006; Brum *et al*., 2015b; Hurwitz and Sullivan, 2013), freshwater (Roux *et al*., 2012; Skvortsov *et al*., 2016), Antarctic freshwater (de Cárcer *et al*., 2015; Lopez-Bueno *et al*., 2009), hot spring (Breitbart *et al*., 2004b), and soil (Reavy *et al*., 2015; Srinivasiah *et al*., 2013). Through assembling environment virome data, many studies identified novel and abundant bacteriophage genomes in various environments (Brum *et al*., 2015b; Hurwitz and Sullivan, 2013; Lopez-Bueno *et al*., 2009). Also, some studies were able to propose putative phage genomes from virome data sets using marker genes conserved in specific viral groups or host genomes (Ghai *et al*., 2016; Zawar-Reza *et al*., 2014). However, there still were limitations in analysis of virome data that most of the assembled phage genomes were not able to find a close match within established genome databases. Furthermore, phage genome sequences alone was not sufficient to provide information on their hosts or morphology, which are required basic information for phage classification. As mentioned above, viral genome database is considerably small compared to that of bacterial genome. Number of novel environmental phage genomes acquired from

viral metagenomes outnumbers the existing viral genome groups and those genomes are not being able to be classified or group to an existing phage/viral genome groups and they are currently named as simply “unclassified.” Also, because viral genomes are too diverse and they do not carry any universally conserved sequences, construction of taxonomic trees to classify viral particles with only sequence information was not achievable. Furthermore, despite numerous and extensive viral metagenome studies performed recently, lack of representative freshwater phages has hampered proper taxonomic and functional interpretation of freshwater viromes (Bruder *et al*., 2016), and has resulted in many freshwater virome reads being assigned to marine phages (Green *et al*., 2015; Skvortsov *et al*., 2016) which are relatively studied more. Therefore, virome approaches to study environmental phages must be accompanied with individual phage cultures and experimental observations.

A number of studies on marine bacteriophages have been done, including isolation of the most abundant bacteriophages in the ocean (Kang *et al*., 2013; Zhao *et al*., 2013) and survey of marine viral population through metagenome (Hurwitz and Sullivan, 2013; Roux *et al*., 2016a). In depth studies on specific bacteriophages were also performed and discovered marine cyanophages with photosystem genes as AMGs to assist their host metabolism and enhance phage reproductivity (Sharon *et al*., 2009). However, all these extensive studies on bacteriophage were confined to marine environments, still leaving un-pioneered spaces of bacteriophages in other biospheres, such as freshwater lakes (Cobián Güemes *et al*., 2016). Inland waters, including lakes, reservoirs, streams, and rivers, play important roles in global biogeochemical cycles and climate change (Raymond *et al*., 2013; Tranvik *et al*., 2009). There are large number of freshwater lakes across the continents that are diverse in size and characteristics with large ecological values. Since each lake is enclosed and isolated from each other, despite how similar climate or environment

they have, each of them have unique and independent systems. Meanwhile, major bacterial composition is very similar across different lakes (Glöckner *et al*., 2000; Newton *et al*., 2011; Salcher, 2013), providing interesting aspects in freshwater microbial evolution and ecology. Also, inland lake microbial communities react more sensitively to climate and environmental changes compared to those of oceans due to smaller area (Tseng *et al*., 2013), providing valuable study sites for seasonal and climate-dependent microbial researches. In this regard, there are numerous studies on freshwater microbes with diverse aspects, such as community structures influenced by salinity gradient, climate changes and water chemistry (Eiler *et al*., 2014; Hahn *et al*., 2015; Niño-García *et al*., 2016). Also, in-depth studies on individual bacterial strains that inhabit in freshwater lakes have been done by many researchers (Hahn *et al*., 2016; Jezbera *et al*., 2013; Salcher *et al*., 2015). Yet, many of the freshwater bacterial strains still remain uncultured, along with their phages. Until today, no phage has been isolated that infects major freshwater heterotrophic bacterial groups, such as acI, acIV, LD12, *Limnohabitans*, *Polynucleobacter*, and LD28. Hence, isolation and culturing of freshwater phages using freshwater bacteria are necessary to understand the freshwater virosphere.

Lake Soyang, located in South Korea, is the largest and oldest artificial lake in Korea that serves as tap water reservoir for Seoul metropolitan area. As well as other conserved oligotrophic lakes, Lake Soyang inhabits diverse bacterial lineages and phage groups. To lead the study on the freshwater microbial population and dynamics, number of bacterial strains and bacteriophages have been isolated and studied from the site. For this study, two of representative families of the class *Betaproteobacteria* were selected to isolate their bacteriophages; the families *Methylophilaceae* and *Comamonadaceae* within the class *Betaproteobacteria*. Among the diverse freshwater bacterial groups, the class *Betaproteobacteria* is often the most abundant group in freshwater environments, though less abundant in marine

environments (Cottrell *et al*., 2005; Zwart *et al*., 2002). Thereby the freshwater *Betaproteobacteria* is the best-studied and the most cultured bacterial group (Newton *et al*., 2011).

One of the major heterotrophic bacterial groups in freshwater is the methylotrophs belonging to the *Betaproteobacteria* group, who are responsible for single-carbon (C1) utilization and contribute in carbon cycle of its inhabiting environment (Beck *et al*., 2014; Chistoserdova, 2015; Halsey *et al*., 2012; Hanson, 1998). By participating in C1 compound metabolism, methylotrophic bacteria are expected to play important roles in the control of the emission of greenhouse gases such as methane and carbon dioxide. While methylotrophic bacteria are distributed among diverse phylogenetic groups (Chistoserdova and Lidstrom, 2013) with various metabolic pathways, a few phylogenetically related clades in the family *Methylophilaceae* have been described as a major methylotrophic group in water column of marine and freshwater environments. In marine habitats, the OM43 clade of the *Methylophilaceae* was found to be a major methylotrophic group by several studies (Gifford *et al*., 2013; Rappe *et al*., 1997; Sowell *et al*., 2011). Isolation and genome sequencing of HTCC2181, a coastal strain of the OM43 clade, showed the ability of C1 compound utilization. In freshwater habitats, the LD28 and PRD01a001B groups are known to be frequently found in pelagic freshwater. Especially, the LD28 clade, a close relative of the OM43 clade, was found to be widespread and abundant (Newton *et al*., 2011; Salcher *et al*., 2011). Recently, Salcher *et al.,* (2015) successfully isolated bacterial strains affiliated with the LD28 and PRD01a001B clades and described the isolates as type strains of two novel species of the *Methylopumilus*, a novel *Candidatus* genus within the *Methylophilaceae*. The genome sequences of the two strains showed the existence of methylotrophic pathway, and methanol was revealed to enhance the growth of strain MMS-2-53, a LD28 isolate. Considering the recurrent seasonal variation of

the LD28 clade (Salcher *et al*., 2015; Salcher *et al*., 2011), studies on phages infecting the LD28 clade bacteria are expected to contribute in better understanding of the dynamics of the LD28 clade.

Metagenomic studies on several freshwater bacteria revealed that the family *Comamonadaceae*, arbitrarily named betI (Zwart *et al*., 2002), is the most frequently found family (Kaden *et al*., 2014) within the class *Betaproteobacteria*. The genus *Rhodoferax* (Newton *et al*., 2011), belonging to the family *Comamonadaceae*, is found in diverse habitats including ditch water, activated sludge, Antarctic microbial mats, and water reservoirs (Cottrell *et al*., 2005; Hiraishi *et al*., 1991; Madigan *et al*., 2000; Newton *et al*., 2011). The *Curvibacter* genus is a member of the family *Comamonadaceae*, a representing family of the class *Betaproteobacteria* (Willems, 2014), which is one of the dominating bacterial group in freshwater environments (Newton *et al*., 2011). Therefore, understanding the ecology of the genera *Rhodoferax* and *Curvibacter* and their lytic phage will contribute to the understanding of freshwater microbial dynamics and help in further freshwater phage genomic studies.

Therefore, in this study, three bacterial strains isolated from Lake Soyang, that belong to the genus *Betaproteobacteria*, were selected for their phage isolation; one belonging to the family *Methylophilaceae* (IMCC19250) and two strains belonging to the family *Comamonadaceae* (IMCC26218 and IMCC26059). Thus, total of four bacteriophages were successfully isolated and sequenced. The phage P26218 was isolated using *Rhodoferax saidenbachensis* strain IMCC26218, and the phages P26059A and P26059B were isolated using a strain IMCC26059, a strain belonging to *Curvibacter* species. Lastly, P19250A, a lytic phage was isolated and shown to be infecting IMCC19250, a strain belonging to LD28 clade. After whole genomes of the isolated phages were obtained, their genomic distribution in Lake Soyang were observed through competitive binning analysis of viral metagenome

data prepared from the identical lake. The binning analysis revealed that P19250A was the most abundant bacteriophage found in winter seasons while other three phages showed relatively low appearances in Lake Soyang. Through sequencing of novel bacteriophages isolated from Lake Soyang, not only that our knowledge on freshwater virosphere was extended, but also was able to provide enhanced interpretation on unknown parts of the viral metagenome studies.

2. MATERIALS AND METHODS

2.1. Isolation and purification of freshwater bacteriophages

2.1.1. Isolation and cultivation of the host strains from Lake Soyang

On October 2011, a freshwater sample was collected at a depth of 30 m from Lake Soyang, located in Gangwon province of South Korea (37.947421 N, 127.818872 E), and was transported to the laboratory. For media preparation, 2 L of the water sample was filtered through a 0.2-μm pore-size polyethersulfone (PES) membrane filter (Pall Corporation, New York, USA), autoclaved (2 h), cooled, and aerated (4 h). Then, 10 μM NH_4Cl, 10 μM KH_2PO_4, 50 μM pyruvic acid, 5 μM D-glucose, 5 μM *N*-acetyl-D-glucosamine, 5 μM acetic acid, 1 μM $FeCl_3$, 1 μM methionine, 1 μM glycine, 1 μM cysteine, and a vitamin mixture (Cho and Giovannoni, 2004) were added to the treated water to be used as culture media. A small volume of untreated water was diluted to a microbial cell density of 10 cells ml^{-1} using the media prepared as above, and aliquoted into 48-well microtiter plates (1 ml per well). The plates were incubated at 15°C in the dark for 6 weeks. After incubation, the cell density in each well was measured using a Guava® EasyCyte™ Plus Flow Cytometry System (Merck Millipore), and the growth-positive wells were harvested for phylogenetic analysis based on 16S rRNA gene sequences as described by Yang *et al.* (2016).

Among the bacterial strains that were initially cultivated, a bacterial strain IMCC19250, which was classified to the LD28 clade was purified by a subsequent dilution culturing and was selected as a host strain for the isolation of bacteriophages. The IMCC19250 strain was grown in artificial freshwater media (AFM) using methanol as a sole carbon source and it did not form colonies on agar medium. Thereby, all experiments, including phage isolation, were performed using AFM.

Table 3-1. Composition of the artificial freshwater medium (AFM) used to culture an LD28 strain, IMCC19250

Chemicals	**Final concentration**
KH_2PO_4	200μM
$(NH_4)_2SO_4$	300μM
$MgSO_4$ $7H_2O$	300μM
KCl	200μM
$CaCl_2$ $2H_2O$	500μM
$NaHCO_3$	300μM
CH_3OH	200μM
Vitamin mixture[a]	1 ×
Trace metal[b]	1 ×
Na_2CO_3	to pH 7.2

[a]See Stingl *et al*. (2008) for detailed composition

[b]See Carini *et al*. (2013) for detailed composition

The recipe for AFM is shown in table 3-1. AFM was prepared by adding salts to MilliQ water, followed by autoclaving (1.5 h), aeration ($\geq$ 12 h), and the addition of a vitamin mixture and trace metals.

On April 2014, a freshwater sample was collected from the identical site, at depth of 1 m. Using the identical method described above, bacterial strains that each belong to the genera *Rhodoferax* (IMCC26218) and *Curvibacter* (IMCC26059) were isolated. Based on a comparative 16S rRNA gene sequence analyses, strain IMCC26218 was found to belong to the genus *Rhodoferax* with 98.7% sequence similarity to *R. saidenbachensis* ED16^T. The IMCC26059 strain showed 98.00% 16S rRNA similarity with *Curvibacter delicatus*, and, it also showed close relatedness to *Curvibacter fontanus* when Neighbor-joining phylogenetic tree was constructed using 16S rRNA sequences, making ambiguous phylogenetic identification of IMCC26059 and leaving the strain as putative novel species (Fig. 3-1). Both strains were able to form colonies on R2A agar (Becton, Dickenson and Company, Franklin Lakes, NJ, USA) at 20˚C (Quast *et al*., 2012).

2.1.2. Isolation of a bacteriophage infecting IMCC19250, a non-colony former

The surface water sample collected on April 2014 was filtered through a 0.2-μm PES membrane filter (Merck Millipore, Darmstadt, Germany) (Brum *et al*., 2015b) to remove large particles and retain only those smaller than 0.2-μm in diameter, which was mostly comprised of viral particles, and 200 μM methanol and 1 × vitamin mixture were added. Strain IMCC19250 grown in AFM with 200 μM methanol was inoculated into 800 ml of the lake water sample processed as described above, at the density of 5×10^4 cells ml^{-1}, and incubated at 20°C for 3 weeks to enrich the bacteriophages present in the lake water that could infect the host strain. During incubation, 10 ml of the enrichment culture was sub-sampled every week. Collected

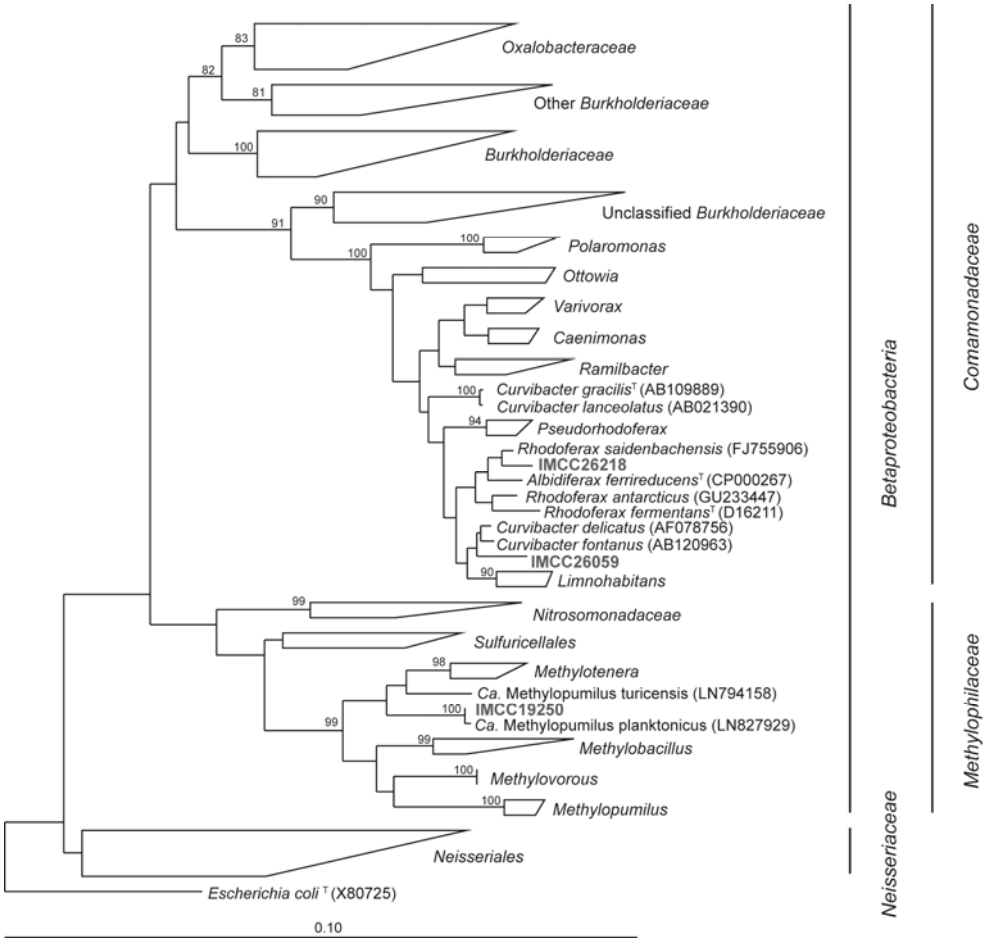

Figure 3-1. 16S rRNA neighbor-joining phylogenetic tree of representative bacterial strains of the phylum *Betaproteobacteria*. Bacterial strains, IMCC26218, IMCC26059, and IMCC19250, that were isolated from Lake Soyang and used as hosts for phage screening are marked in bold. The tree was constructed with bootstrap value of 1,000 based on the SILVA ribosomal RNA gene database (Quast *et al*., 2012).

sub-samples were mixed with 2 ml of chloroform, vortexed, and centrifuged at 3,000 × *g* for an hour to remove bacterial cells. The aqueous phase, which contained putatively retained bacteriophage particles, was collected and stored at 4°C until analysis.

To screen for phages infecting IMCC19250, an exponentially growing culture of IMCC19250 was diluted to approximately 10^4 cells ml^{-1} using AFM, and 1 ml of the bacterial culture was aliquoted into 48-well plates, and 10 μl of the enrichment culture samples prepared as above were inoculated. After 2 weeks of incubation at 20°C, the cell density of each well was measured with a Guava® EasyCyte™ Plus Flow Cytometry System (Merck Millipore), and compared to the densities of uninoculated control wells (with host only). Several wells that showed much lower cell densities than the control wells were selected for further experiments. Samples were collected from each of the selected wells, treated with 200 μl of chloroform, and inoculated into a 20 ml of IMCC19250 culture containing about 10^4 cells ml^{-1}. During incubation, the cell density of the cultures was monitored by flow cytometry. Growth retardation and/or cell lysis was observed in many cultures, suggesting phage infection. The presence of phage particles in those cultures was confirmed by epifluorescence microscopy according to the method of Patel *et al.,* 2007*,* with slight modifications. Samples (10–100 μl) of the cultures were filtered onto 0.02-μm Whatman Anodisc 25 filters (Sigma-Aldrich, St. Louis, MO, USA), stained with SYBR Gold (Invitrogen, Waltham, MA, USA), and examined under a Nikon Eclipse 80i epifluorescence microscope (Nikon Instruments, Melville, NY, USA). The cultures that were confirmed to contain phage particles were stored in either 7% DMSO (Sigma-Aldrich) or 10% glycerol (Sigma-Aldrich) at -80°C. A phage strain was established from one selected sample through co-culture with IMCC19250, and was named P19250A, after its bacterial host.

2.1.3. Isolation of bacteriophages infecting colony-forming bacterial strains

Bacterial strains IMCC26218, a strain belonging to *Rhodoferax* species and IMCC26059, a strain belonging to *Curvibacter* species were isolated from Lake Soyang and both were able to form colonies on R2A agar (Becton, Dickenson and Company, Franklin Lakes, NJ, USA). Therefore bacteriophages infecting these strains were isolated using R2A media. To screen for putative bacteriophages infecting *Rhodoferax* sp. IMCC26218 and *Curvibacter* sp. IMCC26059, surface waters from Lake Soyang were collected and brought to lab at 4˚C on October 2014 and May 2015, respectively. The water samples were filtered through a 0.2 μm PES membrane filter (Merck Millipore) (Brum *et al*., 2015b) immediately after the samples were brought to lab. To 400 ml of treated water sample, 100 ml of 5 × R2A broth (MB Cell, Los Angeles, CA, USA) and liquid culture of IMCC26218 and IMCC26059 were each added and incubated at 20°C for 2 weeks for enrichment of bacteriophages infecting target hosts. During the incubation period, 10 ml of the enrichment culture was sub-sampled for 5 times at a 3-day interval. Each sub-sample was treated with approximately 3 ml of chloroform to inactivate the bacterial cells. The treated samples were used for spot-double agar layer (DAL) plaque assay on its designated host lawn plates for phage screening via appearance of plaques (Grabow, 2004), resulting in the isolation of phage P26218 and P26059A. The DAL plates were prepared with 1.5 × R2A agar as the bottom layer and 0.7 × R2A agar with bacterial liquid culture as the top agar.

On June 2016, another surface water sample was collected from the identical site. After filtering the water sample through 0.2 μm PES membrane filter (Merck Millipore), 1 L of water sample was concentrated to approximately 12 ml using 50 kDa Centrifugal Device (Pall Corporation). The samples were filtered through a 0.2 μm Acrodisc® Syringe Filter (Pall Corporation) for sterilization. Ten

μl of concentrated samples were spotted on IMCC26059 bacterial lawn plate and plaques were obtained from the spotted regions. The plaque was retrieved and purified through series of DAL plating for purification and obtained phage particle was named as P26059B.

2.2. Growth curves of isolated bacteriophages

2.2.1. Co-culture growth curve of host and its bacteriophage

For phage P19250A, its growth curve was constructed through co-culture analysis with its host, IMCC19250. An exponentially-growing culture of strain IMCC19250 was inoculated into six culture flasks that contained 30 ml of fresh AFM, at an initial cell density of 10^4 cells ml^{-1}. Subsequently, P19250A was added to three flasks at a multiplicity of infection (MOI) of 18. Another 3 flasks, without phage, were used as controls. The cultures were incubated at 20°C, and the growth of the host strain was measured every day with a Guava Flow Cytometer (Merck Millipore). At the same time, a 1 ml of sub-sample was taken from each culture flask and analyzed to enumerate the phage particles by epifluorescence microscopy after staining with SYBR Gold (Patel *et al*., 2007).

2.2.2. One-step growth curves of bacteriophages P26059A and P26059B

One-step growth curves for P26059A and P26059B were constructed using exponentially growing IMCC26059 liquid culture and their phage stocks (2.43×10^8 PFU/ml and 7.98×10^8 PFU/ml, respectively) prepared in SM buffer. The phage stocks were each inoculated to the host liquid culture at MOI of 0.33 (P26059A) and 6.29 (P26059B). The mixtures were incubated in a shaking incubator at 20˚C and 100 rpm for 10 min. The incubated samples were serially diluted to 10^{-4} fold and was placed in shaking incubator for 3 hours. During incubation, the liquid culture was withdrawn every 20 min and plated on DAL plate in triplicate. The plaques were

counted after two days of incubation in 20˚C and enumerated plaque numbers were used to draw one-step growth curves.

2.3. Enrichment and concentration of bacteriophage particles

Bacteriophages P19250A, P26218, and P26059A were enriched and concentrated through preparing 800 ml of lysate solutions. P19250A particles were amplified by co-culture with IMCC19250 in 800 ml of AFM and P26218 and P26059A particles were enriched in co-culture with their hosts, IMCC26218 and IMCC26059, respectively, in R2A Broth.

Lysate solution of P26059B was prepared differently from other phages. For the phage P26059B, 10 confluent DAL plates with propagated phages were prepared. To extract phage particles from the plaques, 5 ml of SM buffer were added to each plate and they were incubated on a gyratory shaker in 4˚C. After an overnight incubation, the SM buffer was retrieved. Ten ml of chloroform was added to approximately 50 ml of SM buffer with harvested phage particles was harvested and vigorously vortexted for 5 min for removal of bacterial cells. Then the sample was centrifuged at 3,000 × g for 30 min. and only the top aqueous layer was collected for further procedures.

All the lysates prepared were collected and concentrated according to the methods in “Molecular Cloning: A Laboratory Manual” (Green and Sambrook, 2012) with minor modifications. After treatment with 1 μg ml^{-1} DNase I and RNase (Sigma-Aldrich) and the addition of 1 M NaCl and 10% (w/v) polyethylene glycol (PEG) 8000, the lysate was incubated on ice for overnight, and then centrifuged at 11,000 × g for 40 min to precipitate phage particles. The pellet was soaked in SM buffer and resuspended. An equal volume of chloroform was added to the resuspended pellet, vortexed, and centrifuged at 3,000 rpm for 30 min at 4°C. The top aqueous phase was collected and ultracentrifuged for 2 h at 240,000 × g in a Beckman Coulter

L-90K ultracentrifuge with a SW 55 Ti swinging-bucket rotor. The pellet was resuspended in 100 μl of SM buffer for further analysis.

2.4. Morphological analysis of isolated phages using transmission electron microscopy

For morphology analysis of the phages, copper grid samples were prepared to be observed under a transmission electron microscope, TEM (CM200; Phillips, Amsterdam, Netherlands). Ten μl of the phage concentrates were adsorbed onto formvar and carbon-coated copper grids. The grids were negatively stained using 2% uranyl acetate by two short stainings followed by 45 sec of a final staining step (Ackermann and Heldal, 2010). After observation, taxonomic classification of phages was made based on its morphology (King *et al*., 2012).

2.5. Whole genome sequencing of phages and quality control

From the prepared phage concentrates, genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Whole genome sequencing for the phage genomes was performed by ChunLab, Inc. (Seoul, South Korea). The sequencing library was constructed using the TruSeq DNA sample preparation kit (Illumina, San Diego, US) and the samples were sequenced using an Illumina MiSeq system with 2 × 300 bp paired-end reads. The sequences for phages were each assembled using SPAdes-assembler (Bankevich *et al*., 2012). Assembled contigs were checked for their completeness followed by gene prediction by the RAST server (Aziz *et al*., 2008). Each predicted gene was analyzed by BLAST against NCBI's nr and env-nr protein databases (Lavigne *et al*., 2008) for their function prediction and only the results with e-values less than 0.001 were accepted. The protein coding genes that did not have a predicted function were further analyzed with BLAST upon UniProt database (Apweiler *et al*., 2004), Pfam database (Finn *et al*., 2013), and Conserved Domain

Database (CDD) (Marchler-Bauer *et al*., 2011). Then their functions were predicted based on the protein domain found with the e-value threshold of 0.001. For further analysis of the phage genomes, tRNAs were searched using tRNAscan-SE v. 2.0 (Lowe and Eddy, 1997) and ARAGORN v. 1.2.38 (Laslett and Canback, 2004), which were available on-line.

2.6. Competitive binning analysis of sequenced phage genomes within virome data

To observe the abundance and distribution of bacteriophage genomes that have been obtained, six viromes of Lake Soyang and other publicly available freshwater viromes were used for the analyses. Public viromes were downloaded from NCBI and MetaVir (http://metavir-meb.univ-bpclermont.fr/). Viral metagenome data from Lake Soyang were collected from October 2014 to May 2016. Total of 6 surface water samples were collected and viral particles were concentrated from approximately 10 L of lake water using $FeCl_3$ (John *et al*., 2011). Then collected viral particles were sequenced using Illumina MiSeq sequencing platform. Virome sequences of Lake Soyang were trimmed using Trimmomatic based on quality score and length (Bolger *et al*., 2014), and phiX174 control sequences were removed by discarding sequences that were mapped to the phiX174 genome in the read mapping using CLC Genomics Workbench (Qiagen). In the binning analysis, each read of the viromes was assigned to the best-matching protein in a custom-made search database by the DIAMOND algorithm (Buchfink *et al*., 2015) with a bitscore cutoff of 40. The search database was constructed by adding annotated protein sequences of isolated phages to all the viral proteins and non-redundant bacterial proteins of RefSeq (release 72 (Nov. 5, 2015) for analysis of the phage P19250A and release 79 (Nov. 2016) for analysis of the phages P26218, P26059A and P26059B). The binning results were summarized by calculating the number of virome reads assigned to each viral and bacteriophage genome.

Along with viral metagenome, bacterial 16S rRNA amplicon sequencing was performed using identical samples. From 1-2 L of surface lake waters, bacterial DNA was extracted as described in Yang *et al*. (2016). The V3-V4 regions of the 16S rRNA genes were amplified using Illumina MiSeq platform at ChunLab, Inc. Taxonomic classification of processed sequences was performed with the RDP classifier in MOTHUR, using a custom-made database that was based on the SSURef NR database of Silva (Release 123; available at https://www.arb-silva.de/). The abundances of the host bacterial strains were calculated by dividing the number of sequencing reads assigned to the target bacteria by the total sequencing reads.

To search for virome contigs that showed synteny to the phage genomes, reads from selected viromes including those of Lake Soyang were assembled using SPAdes (Bankevich *et al*., 2012). Contigs ($\geq$10 kb) assembled from each virome were compared by local tBLASTx to a custom-made search database that included the genomes of the phages obtained in this study in addition to the all viral genomes in RefSeq (release 72 or 79) to search for contigs that showed high similarity to bacteriophage genomes of interest. The BLAST results were summarized to calculate the total bitscore between all pairs of viral genomes and virome contigs. Virome contigs, for which the total bitscore with the phage of interest was higher than the bitscore with any other viral genome, were picked for further analyses. Selected contigs were further analyzed by tBLASTx, provided at the NCBI website, against both the Reference genomic sequences database ("refseq_genomic") for further confirmation. The total bitscore between the selected virome contigs with target phage genome was compared to that of the best hit of selected virome contig in "refseq_genomic." Then, only contigs that had a higher bitscore with the target phage genome compared to the best match found from the existing database were used for the synteny analysis.

3. RESULTS

3.1. Physical characteristics of bacteriophages isolated from Lake Soyang

3.1.1. Morphology, growth curve, and host range of the phage P19250A

In attempts to isolate phages infecting major freshwater bacterial groups, bacteriophage P19250A, which infects strain IMCC19250 of the LD28 clade (Fig. 3-1), was isolated from Lake Soyang, where the host strain was previously isolated as well. Morphological characterization by TEM revealed that P19250A belonged to the family *Siphoviridae*, with an icosahedral shaped head (approximately 51 nm in diameter) and a long non-contractile tail (approximately 95 nm in length; Fig. 3-2a). P19250A showed a lytic life cycle when co-cultured with its host. Concentration of P19250A particles increased exponentially from 8.04×10^3 per ml (immediately after inoculation) to 3.04×10^8 particles per ml within 5 days (Fig. 3-2b). Concurrently, the number of host cells started to decrease after 2 days of incubation and reached 9.55×10^4 cells per ml after 5 days, while the host cultures not inoculated with the phage entered latent period with a cell density of 7.96×10^6 cells per ml. After confirming the lytic ability toward the original host strain, the host range of P19250A was tested using phylogenetically related bacterial strains isolated from the same lake. The P19250A was able to infect all the four tested isolates of the LD28 group that showed 99.85–100% sequence similarity of 16S rRNA gene. However, IMCC30193, a strain that belongs to the PRD001a001B group with 96.31% 16S rRNA sequence similarity to IMCC19250, was not infected by P19250A (Fig. 3-3). Considering that the LD28 and PRD01a001B clades were suggested to form two different species within a same genus (Salcher *et al*., 2015), our results showed that the host range of P19250A is restricted to only those within the same species.

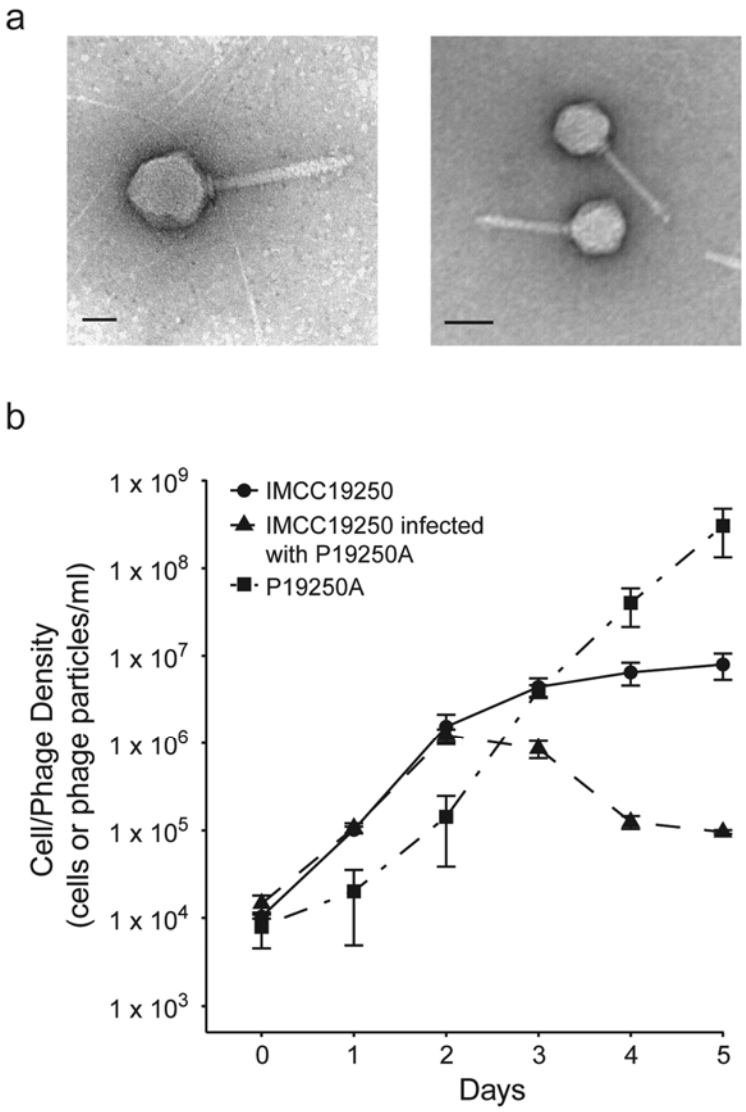

Figure 3-2. General characteristics of the phage P19250A. (a) Transmission electron micrographs of P19250A particles showing icosahedral capsids and long non-contractile tails. The scale bars represent 20 nm (left) and 50 nm (right). (b) Lysis of host strain IMCC19250 by P19250A during co-culture. For comparison, IMCC19250 growth was also measured in the absence of P19250A. Error bars represent standard error (n = 3).

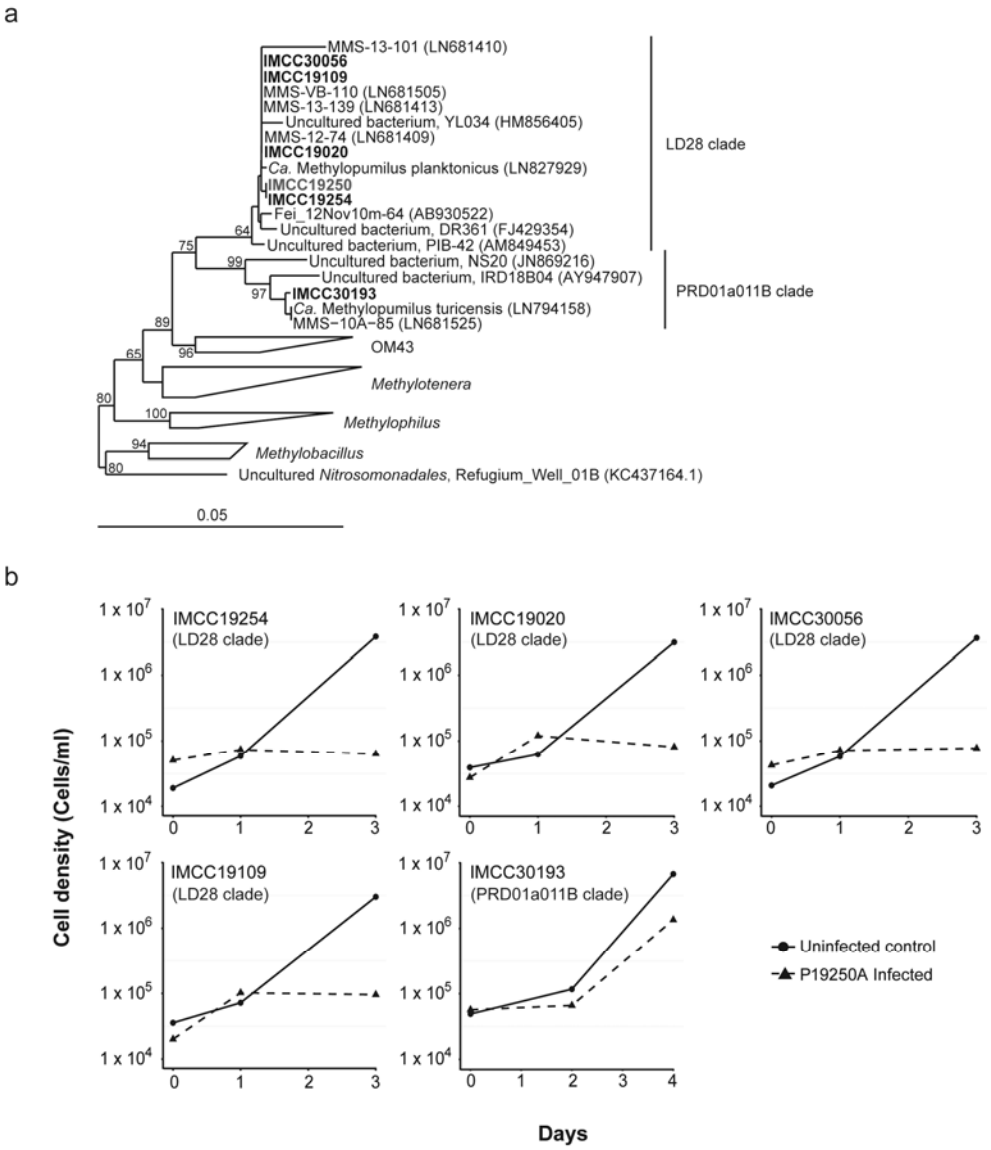

Figure 3-3. Phylogenetic position of the host strain, IMCC19250, among related strains in the family *Methylophilaceae*, and determination of the host range of the phage P19250A. (a) Neighbor-joining phylogenetic tree based on 16S rRNA gene sequences showing the phylogenetic position of the strains that were used for host range determination. Four strains of the LD28 clade and one strain of the PRD01a011B clade used for the experiment are shown in bold. Bootstrap values are shown at the nodes. (b) Growth curves of the bacterial strains used for determination of the host range of P19250A. Bacterial cells co-cultured with P19250A (closed triangles) and uninfected cultures (closed circles) are shown.

3.1.2. Physical characteristics of the phages P26218, P26059A, and P26059B

Phage P26218 is a lytic phage that forms plaques of 1 to 2 mm in diameter, on *Rhodoferax* sp. IMCC26218 culture plates. TEM of purified phage particles revealed its icosahedral-shaped head (52.1 nm in diameter) with a short tail for 9.4 nm in length (Fig. 3-4), classifying the P26218 as a member of the family *Podoviridae* of the order *Caudovirales* (King *et al*., 2012).

Two phages, P26059A and P26059B of *Curvibacter* sp. IMCC26059 were independently isolated from Lake Soyang. The phage P26059A was isolated on May 2015 using phage-enrichment method and the phage P26059B was isolated on April 2016 through concentration of phage particles present in the lake water. Both phages formed plaques on the bacterial lawn plate, indicating active lytic cycle of both phages. The plaque size for P26059A was approximately 1 mm in diameter and that of P26059B was 5 mm in diameter. When their morphology was observed under TEM, the two phages revealed to have different morphologies as well. P26059A belonged to the family *Siphoviridae* with a long tail of 153.14 nm in length with 62.20 nm head in diameter. Meanwhile, the phage P26059B appeared to be a member of the family *Podoviridae* with a short tail (9.00 nm) and an icosahedral shaped head (58.86 nm in diameter) (Fig. 3-5a). Furthermore, one-step growth curves for both phages were constructed to observe their life cycles. The latent periods for P26059A and P26059B were 120 min and 80 min each and their burst sizes were approximately 15 and 58, respectively (Fig. 3-5b).

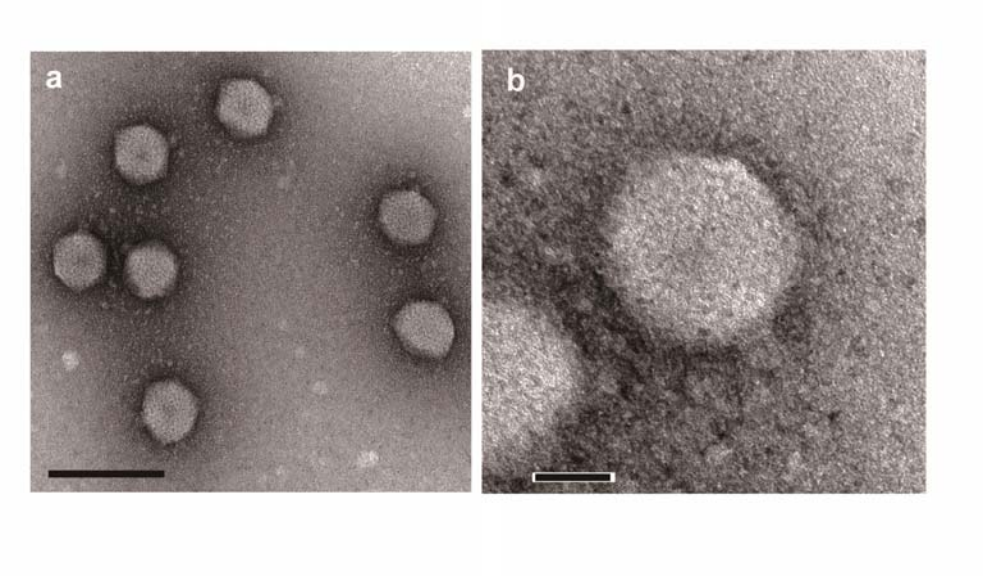

Figure 3-4. Transmission electron micrographs of the phage P26218 particles infecting *Rhodoferax* sp. IMCC26218. The TEM images were obtained using Philips CM200 electron microscope. Scale bars represent 100 nm in (A) and 20 nm in (B).

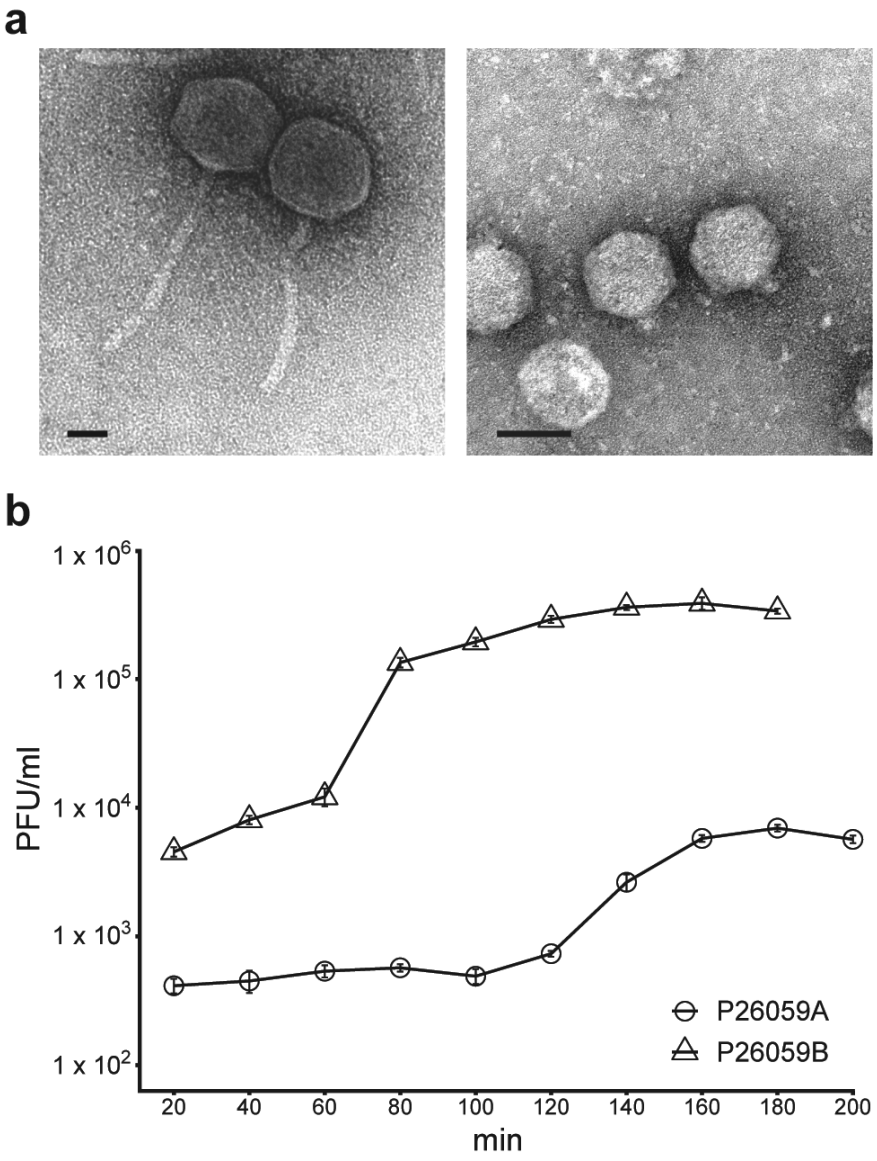

Figure 3-5. Transmission electron microscopy images and one-step growth curves of the phages P26059A and P26059B. (a) The phage P26059A is shown in the left panel and the phage P26059B is shown in the right panel. The scale bars each represent 20 nm and 50 nm, respectively. (b) One-step growth curves of the phages P26059A and P26059B are shown with standard error bars (n = 3).

3.2. Genomic characteristics of bacteriophages isolated from Lake Soyang

3.2.1. Genome features of the phage P19250A

Complete genome of the phage P19250A was obtained through Illumina MiSeq sequencing platform with 694 × fold coverage. The total size of the genome was 38,562 bp in length with 35.40% of G+C content (Table 3-2). A total of 58 genes were predicted in the genome through annotation by the RAST annotation service (Aziz *et al*., 2008), GeneMark.hmm (Lukashin and Borodovsky, 1998), and GLIMMER (Delcher *et al*., 1999). Each annotated protein coding genes were analyzed with BLAST against the NCBI non-redundant protein database and only the BLAST match with e-value threshold of 0.001 or less were accepted as putative function of the gene. Among 58 predicted protein coding genes, 20 of them were functionally annotated and they encoded proteins typically found in phages in a modular architecture (Table 3-3).

Within the phage genome, terminase small and large subunits (ORFs 1 and 2) and a portal protein (ORF 3) constituted a phage genome packaging module. Capsid-related proteins were clustered together (ORFs 4-6) followed by tail-related proteins (ORFs 7, 15, 19, and 21), forming a structure module together. ORF 23 was annotated to code for the collagen triple helix repeat-containing protein which has been first identified in a giant mimivirus, *Acanthamoeba Polyphaga mimivirus* (Colson *et al*., 2011; La Scola *et al*., 2008). Although collagen proteins are known to be dominantly found in mammals, they are found, albeit rarely, in prokaryotes and viruses and known to function as structural proteins (Rasmussen *et al*., 2003). Within the GenBank database, only few of phages belonging to the order *Caudovirales* carry the collagen-like protein and among the family *Siphoviridae*, only *Synechococcus* phage S-CSB2 and *Bacillus* phage PM1 were shown to carry it.

Table 3-2. Sequencing information of the phage P19250A genome

Features	**P19250A**
Length	38,562 bp
G+C content	35.40%
Number of contigs	1
Number of annotated genes	58
Gene coding content	93.25%
Sequencing platform	Illumina MiSeq (2 × 300 bp)
Library used	TruSeq (shotgun)
NCBI Accession number	KX815270

Table 3-3. Genome annotation of the phage P19250A. Only the ORFs with assigned function are shown.

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
1	63	536	+	*Anoxybacillus* sp. DT3-1	Phage terminase, small subunit, P27 family (M5QS93)	5.00E-05	UniProtKB
2	632	2 290	+	*Psychrobacter arcticus*	Phage terminase large subunit (WP_011279763.1)	6.00E-93	GenBank nr
3	2,300	3,709	+	*Geobacillus* phage GBK2	Portal protein (W8ECU7)	3.70E-19	UniProt_Viruses
4	3,812	4,621	+	*Enterobacteria* phage P27	Putative prohead protease (Q8W628)	3.30E-08	UniProt_Viruses
5	4,625	6,013	+	*Alpha proteobacterium* L41A	Phage capsid protein (WP_017505173.1)	6.00E-14	GenBank nr
6	6,091	7,149	+	*Sphingopyxis fribergensis*	Phage major capsid protein, HK97 (A0A0A7PBY3)	3.50E-06	UniProtKB
7	7,153	7,818	+	*Rhizobium leguminosarum*	Hypothetical protein (WP_025416011.1)	4.00E-06	GenBank nr
12	9,365	9,919	+	*Ruegeria mobilis*	Uncharacterized protein (A0A0F4RQV6)	8.40E-04	UniProtKB
15	10,743	12,527	+	*Ralstonia syzygii* R24	Putative phage hk97 tail length tape measure-related protein (G3A6M6)	6.20E-21	UniProtKB
19	14,268	17,567	+	*Yersinia* phage phiR201	Phage tail length tape-measure protein (I7K2R8)	7.20E-14	UniProt_Viruses
20	17,577	18,047	+	*Nitratireductor indicus* C115	Uncharacterized protein (K2N9B3)	7.40E-28	UniProtKB

Table 3-3. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
21	18,070	20,847	+	*Synechococcus* phage S-CBS4	Tail fiber protein (H2BCT9)	5.90E-36	UniProt_Viruses
23	21,415	25,020	+	*Acanthamoeba castellanii* mamavirus	Collagen triple helix repeat-containing protein (B5LMG2)	4.50E-149	UniProt_Viruses
24	25,035	25,664	+	*Methanosaeta concilii*	Glycosyl transferase sugar-binding domain protein (H2EDX2)	4.10E-14	UniProtKB
25	25,657	26,625	+	*Acinetobacter baumannii*	Glycosyl transferase 2 family protein (A0A009KR96)	2.30E-04	UniProtKB
27	27,252	27,413	+	*Synechococcus* phage S-CBS4	Uncharacterized protein (R9TF71)	2.30E-09	UniProt_Viruses
28	27,382	27,774	+	*Synechococcus* phage S-CBS4	Srd putative anti-sigma factor (H2BCV1)	9.10E-25	UniProt_Viruses
30	28,493	28,266	-	*Delftia acidovorans*	Hypothetical protein (WP_012205941.1)	2.00E-06	GenBank nr
31	29,133	28,477	-	*Synechococcus* phage S-CBS4	Exonuclease (AGN30409.1)	2.00E-54	GenBank nr
32	29,563	29,126	-	*Delftia acidovorans*	Genome assembly B1459, Plasmid (IA0A0C7KFE5)	7.30E-05	UniProtKB
34	30,272	29,730	-	*Pseudoalteromonas* phage pq0	Single-strand DNA binding protein (A0A0H4A6R9)	2.70E-33	UniProt_Viruses
36	31,041	30,691	-	*Bradyrhizobium* sp. CCBAU 15544	Hypothetical protein (WP_038950364.1)	5.00E-07	GenBank nr

Table 3-3. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
39	31,781	32,071	+	*Weeksella virosa*	VRR-NUC domain-containing protein (ADX66977.1)	2.00E-20	GenBank nr
40	32,123	32,788	+	*Pseudomonas fluorescens*	Uncharacterized protein (A0A0B7DB19)	3.10E-33	UniProtKB
41	32,797	33,153	+	*Paracoccus zeaxanthinifaciens*	Cell wall hydrolase (WP_022706699.1)	5.00E-18	GenBank nr
50	34,921	35,319	+	*Herminiimonas* sp. CN	Hypothetical protein (WP_025915816.1)	6.00E-04	GenBank nr
51	35,595	35,834	+	No significant similarity found			
52	35,827	36,117	+	*Enterobacter aerogenes*	Hypothetical protein (KLE87140.1)	2.00E-29	GenBank nr
53	36,181	36,585	+	Yellowstone lake phycodnavirus 2	Hypothetical protein (YP_009174670.1)	2.00E-09	GenBank nr
54	36,585	37,022	+	*Reyranella massiliensis*	Peptidase M15 (WP_020697714.1)	1.00E-36	GenBank nr
56	37,327	37,719	+	*Methylobacterium* sp. EUR3 AL-11	Hypothetical protein (WP_043342770.1)	4.00E-23	GenBank nr
57	37,716	38,150	+	*Ruminococcus flavefaciens*	HNH endonuclease (WP_037298081.1)	9.00E-15	GenBank nr
58	38,202	38,522	+	*Paracoccus* sp. J39	Hypothetical protein (WP_036698142.1)	6.00E-20	GenBank nr

Several genes were predicted to encode proteins related to nucleic acid metabolism. ORF31 encoded an exonuclease, and ORF 39 was annotated to encode a VRR-NUC (virus-type replication-repair nuclease) domain-containing protein. The VRR-NUC domains were known to help segregation of phage genomes and repair of double strand breaks (Kinch *et al*., 2005). Recently, crystal structure of VRR-NUC domain was identified, which showed that the domain has similar structure as Holliday junction resolvase and further revealed that VRR-NUC dimer functions as one (Pennell *et al*., 2014). The ORFs 28 to 53 are mostly comprised of hypothetical or unknown proteins, indicating that these genes are highly specific and unique to P19250A genome.

P19250A's phylogenetic location was explored using its terminase large subunit, capsid protein, and phage tail tape-measuring like protein (ORFs 2, 5, and 15, respectively). Related protein sequences were collected from Pfam database (Finn *et al*., 2013) and GenBank nr database, and they were used for construction of maximum-likelihood phylogenetic trees. Three different phylogenic trees were constructed; however, no consistency was observed among all trees (Fig. 3-7). Different marker genes of P19250A were placed in separated branches from other bacteriophages, not being able to be classified into existing phage groups and remained as unique and novel unclassified group of phages.

3.2.2. Genome features of the phage P26218

The capsid of the phage P26218 encapsulated a linear dsDNA with length of 36,315 bp with 56.7% G+C content (Table 3-4). Although the phage P26218 showed a morphology of a typical *Podoviridae* family, when its genomic characteristics were considered, no similar genomic architecture to those of a known phage was found among the known viral genera, leaving P26218 without an assigned genus. The amino acid sequence of DNA polymerase I (encoded by *polA*) of P26218,

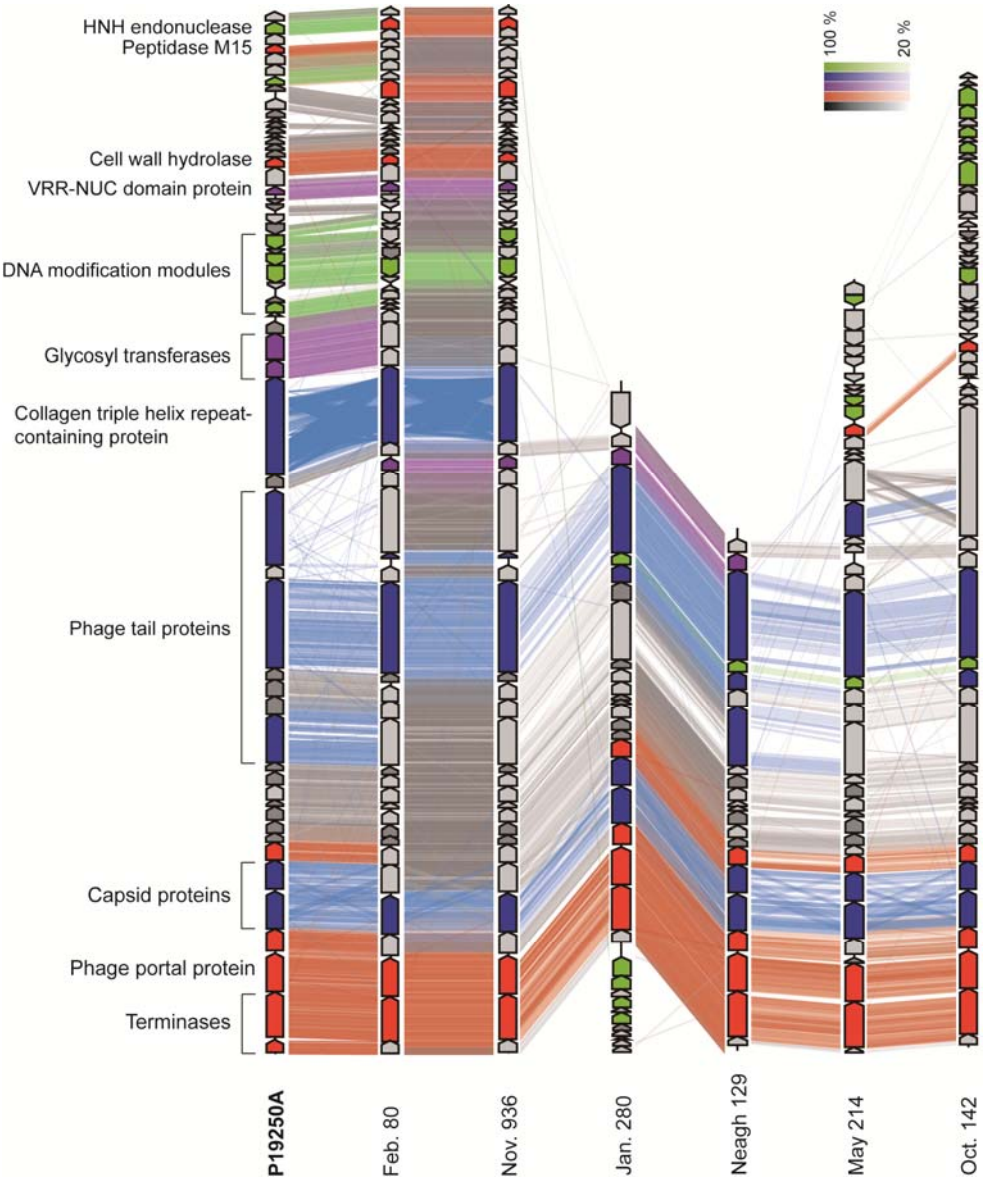
HNH endonuclease
Peptidase M15
Cell wall hydrolase
VRR-NUC domain protein
DNA modification modules
Glycosyl transferases
Collagen triple helix repeat-containing protein
Phage tail proteins
Capsid proteins
Phage portal protein
Terminases
100 %
20 %
P19250A
Feb. 80
Nov. 936
Jan. 280
Neagh 129
May 214
Oct. 142

Figure 3-6. Genome map of the phage P19250A and its synteny contigs recovered from viral metagenomes. Within the genome map, structural genes are shown in blue, genes related to DNA replication, recombination, and modification are shown in green, genes related to cell lysis and packaging are shown in red, and genes involved in auxiliary functions are shown in purple. Synteny between the P19250A genome and contigs assembled from Lake Soyang and Lough Neagh viromes are shown below. The sequence comparison was performed with tBLASTx, and similar regions are connected by rectangles.

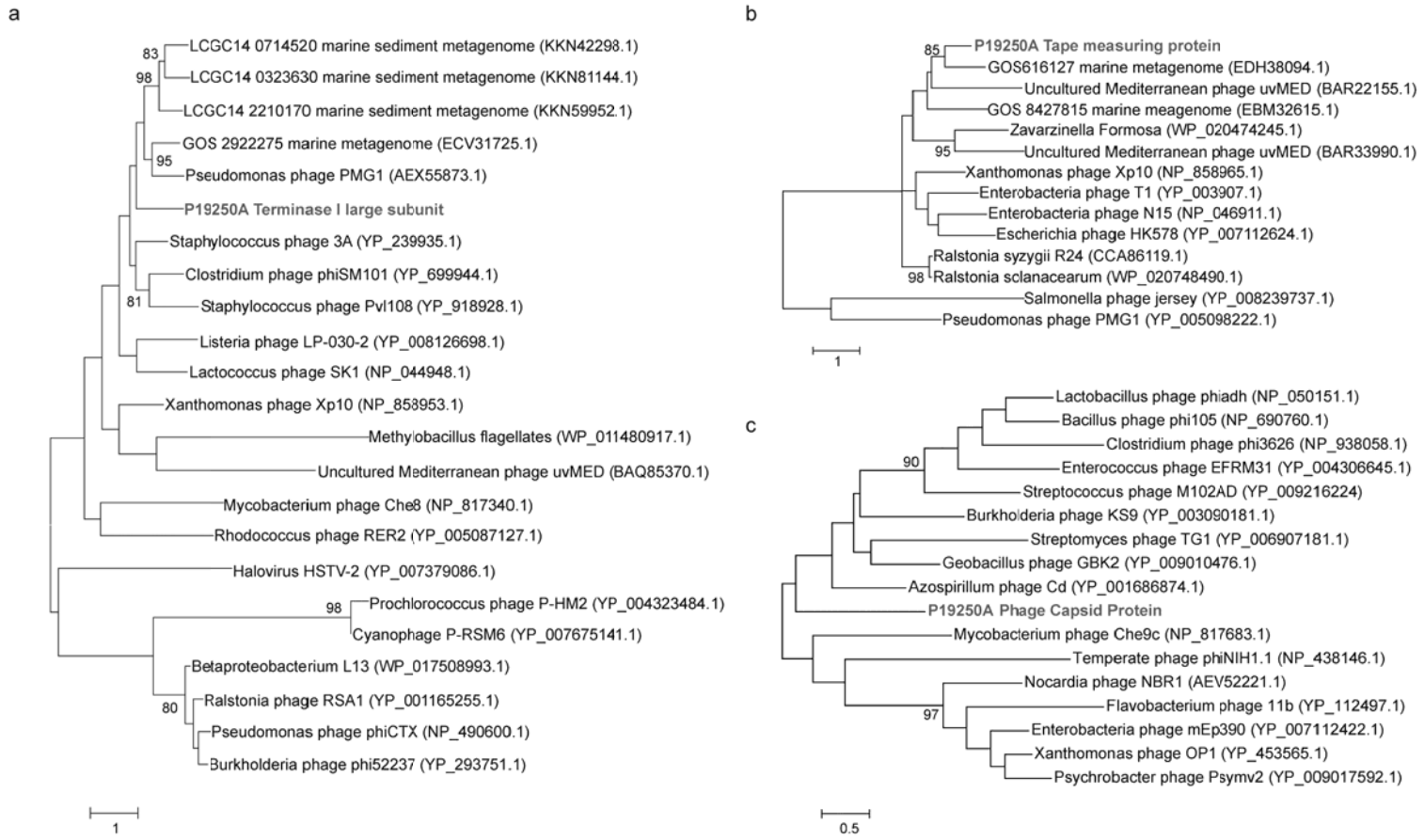
a
LCGC14 0714520 marine sediment metagenome (KKN42298.1)
LCGC14 0323630 marine sediment metagenome (KKN81144.1)
LCGC14 2210170 marine sediment metagenome (KKN59952.1)
GOS 2922275 marine metagenome (ECV31725.1)
Pseudomonas phage PMG1 (AEX55873.1)
P19250A Terminase I large subunit
Staphylococcus phage 3A (YP_239935.1)
Clostridium phage phiSM101 (YP_699944.1)
Staphylococcus phage Pvl108 (YP_918928.1)
Listeria phage LP-030-2 (YP_008126698.1)
Lactococcus phage SK1 (NP_044948.1)
Xanthomonas phage Xp10 (NP_858953.1)
Methylobacillus flagellates (WP_011480917.1)
Uncultured Mediterranean phage uvMED (BAQ85370.1)
Mycobacterium phage Che8 (NP_817340.1)
Rhodococcus phage RER2 (YP_005087127.1)
Halovirus HSTV-2 (YP_007379086.1)
Prochlorococcus phage P-HM2 (YP_004323484.1)
Cyanophage P-RSM6 (YP_007675141.1)
Betaproteobacterium L13 (WP_017508993.1)
Ralstonia phage RSA1 (YP_001165255.1)
Pseudomonas phage phiCTX (NP_490600.1)
Burkholderia phage phi52237 (YP_293751.1)
83
98
95
81
80
1
b
P19250A Tape measuring protein
GOS616127 marine metagenome (EDH38094.1)
Uncultured Mediterranean phage uvMED (BAR22155.1)
GOS 8427815 marine meagenome (EBM32615.1)
Zavarzinella Formosa (WP_020474245.1)
Uncultured Mediterranean phage uvMED (BAR33990.1)
Xanthomonas phage Xp10 (NP_858965.1)
Enterobacteria phage T1 (YP_003907.1)
Enterobacteria phage N15 (NP_046911.1)
Escherichia phage HK578 (YP_007112624.1)
Ralstonia syzygii R24 (CCA86119.1)
Ralstonia sclanacearum (WP_020748490.1)
Salmonella phage jersey (YP_008239737.1)
Pseudomonas phage PMG1 (YP_005098222.1)
85
95
98
1
c
Lactobacillus phage phiadh (NP_050151.1)
Bacillus phage phi105 (NP_690760.1)
Clostridium phage phi3626 (NP_938058.1)
Enterococcus phage EFRM31 (YP_004306645.1)
Streptococcus phage M102AD (YP_009216224)
Burkholderia phage KS9 (YP_003090181.1)
Streptomyces phage TG1 (YP_006907181.1)
Geobacillus phage GBK2 (YP_009010476.1)
Azospirillum phage Cd (YP_001686874.1)
P19250A Phage Capsid Protein
Mycobacterium phage Che9c (NP_817683.1)
Temperate phage phiNIH1.1 (NP_438146.1)
Nocardia phage NBR1 (AEV52221.1)
Flavobacterium phage 11b (YP_112497.1)
Enterobacteria phage mEp390 (YP_007112422.1)
Xanthomonas phage OP1 (YP_453565.1)
Psychrobacter phage Psymv2 (YP_009017592.1)
90
97
0.5

Figure 3-7. Phylogenetic trees of the phage P19250A, constructed using maximum likely method with bootstrap of 100, provided by the MEGA6. The sequences were collected from Pfam database and NCBI website. Phylogenetic tree in panel (a) was constructed using terminase I, large subunit protein sequences, that in panel (b) was constructed with bacteriophage tape measuring proteins, and that in panel (c) was constructed with bacteriophage capsid proteins.

Table 3-4. Sequencing information of the phage P26218

Features	**P26218**
Length	36,315 bp
G+C content	56.70%
Number of contigs	1
Number of annotated genes	44
Gene coding content	93.18%
Sequencing platform	Illumina MiSeq (2 × 300 bp)
Library used	TruSeq (shotgun)
Assemblers	SPAdes version 3.1.1
GenBank ID	KP792623

one of the widely used viral phylogenetic markers (Adriaenssens and Cowan, 2014; Breitbart *et al*., 2004a), was aligned with that of representative strains of the families *Podoviridae* and *Siphoviridae* and the aligned sequences were used for phylogenetic analysis. The phylogenetic tree based on DNA polymerase I revealed that P26218 formed a clade with a marine metagenome sequence, parted from previously known type species, confirming limitations in its assignment to a known genus (Fig. 3-8).

Out of 44 predicted ORFs, only 15 (34%) were assigned with a known function. Four ORFs were predicted to be related to DNA replication, 2 to DNA metabolism, 5 to packaging and structural functions, and 4 to other known functions (Fig. 3-9, Table 3-5). BLASTp analyses showed that each ORF with an identified function was homologous to ORFs from different phages belonging to different viral families. All ORFs encoding viral packaging function were closely related to those of other viruses in the family *Podoviridae*. The ORFs encoding DNA polymerase I, ATPase component, thymidylate synthase, and hydrolase-like protein were similar to those of the family *Siphoviridae*, while the genes for DnaB-like ATP-dependent helicase and ParB-like nuclease domain showed a high degree of homology to those of the family *Myoviridae.* This genomic architecture of P26218 confirmed the mosaic genome structure, known to be a result of lateral gene transfer usually predicted in viral genomes in attempts to enhance their genetic diversity (Swanson *et al*., 2012; Yoshida *et al*., 2015) and often observed in species of the order *Caudovirales* such as phages P22 and lambda.

3.2.3. Genome features of the phages P26059A and P26059B

The genome size of P26069A was 84,008 bp with 43.60% G+C content. The genome coded for 124 genes and contained two tRNA genes. The P26059B genome was 41,471 bp long with 54.30% G+C content. For P26059B, a total of 46 genes were predicted by the RAST annotation server (Table 3-6). After protein

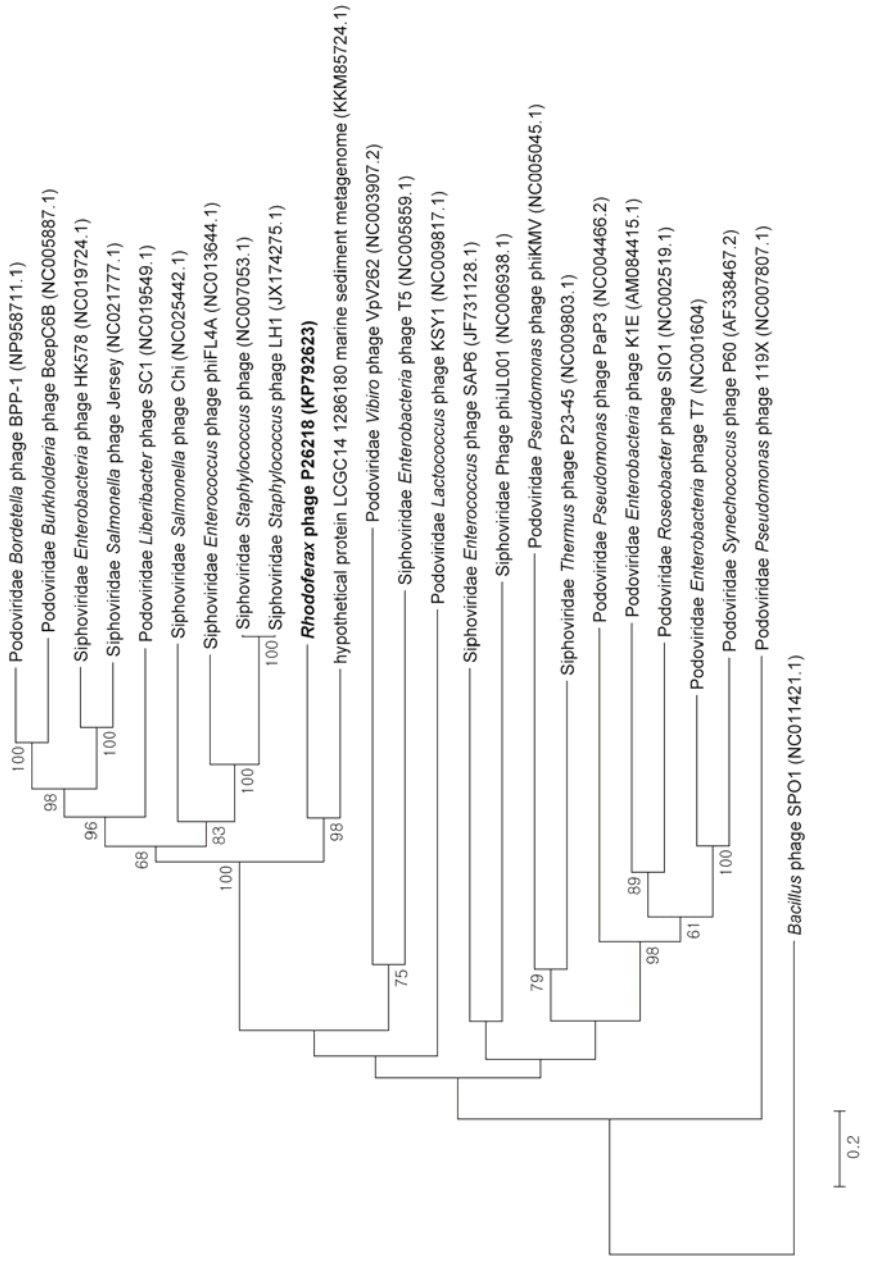
Podoviridae *Bordetella* phage BPP-1 (NP958711.1)
Podoviridae *Burkholderia* phage BcepC6B (NC005887.1)
Siphoviridae *Enterobacteria* phage HK578 (NC019724.1)
Siphoviridae *Salmonella* phage Jersey (NC021777.1)
Podoviridae *Liberibacter* phage SC1 (NC019549.1)
Siphoviridae *Salmonella* phage Chi (NC025442.1)
Siphoviridae *Enterococcus* phage phiFL4A (NC013644.1)
Siphoviridae *Staphylococcus* phage (NC007053.1)
Siphoviridae *Staphylococcus* phage LH1 (JX174275.1)
***Rhodoferax* phage P26218 (KP792623)**
hypothetical protein LCGC14 1286180 marine sediment metagenome (KKM85724.1)
Podoviridae *Vibiro* phage VpV262 (NC003907.2)
Siphoviridae *Enterobacteria* phage T5 (NC005859.1)
Podoviridae *Lactococcus* phage KSY1 (NC009817.1)
Siphoviridae *Enterococcus* phage SAP6 (JF731128.1)
Siphoviridae Phage phiJL001 (NC006938.1)
Podoviridae *Pseudomonas* phage phiKMV (NC005045.1)
Siphoviridae *Thermus* phage P23-45 (NC009803.1)
Podoviridae *Pseudomonas* phage PaP3 (NC004466.2)
Podoviridae *Enterobacteria* phage K1E (AM084415.1)
Podoviridae *Roseobacter* phage SIO1 (NC002519.1)
Podoviridae *Enterobacteria* phage T7 (NC001604)
Podoviridae *Synechococcus* phage P60 (AF338467.2)
Podoviridae *Pseudomonas* phage 119X (NC007807.1)
Bacillus phage SPO1 (NC011421.1)
100
98
100
96
68
83
100
100
98
100
75
79
98
89
61
100
0.2

Figure 3-8. Phylogenetic tree highlighting the relationship of the phage P26218 infecting *Rhodoferax* sp. IMCC26218 with representatives of the families *Podoviridae* and *Siphoviridae*. Sequences of DNA polymerase I (*polA*) collected from NCBI were aligned using CLUSTALW software, with Bacillus phage B103 (X99260) and SPO1 (NC011421.1) as an outgroup. The phylogenetic tree was generated using the neighbor-joining method implemented in MEGA6. Bootstrap values representing over 60% in 1,000 replicates are shown in the tree.

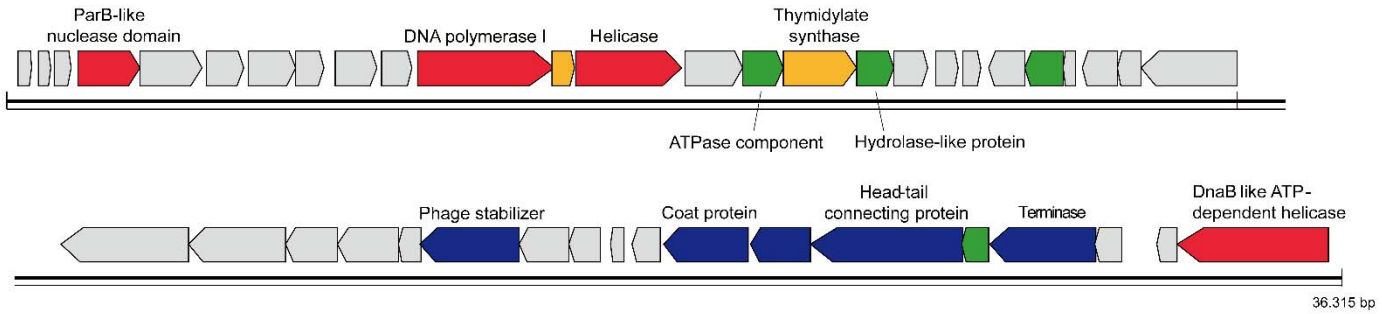

Figure 3-9. Genome map of the *Rhodoferax* phage P26218. Total length of the genome is 36,315 bp and contig functions are color-coded as follows: light grey represents hypothetical proteins, yellow represents DNA metabolism, red represents DNA replication, and blue represents structural and packaging genes and green represents other known functional genes.

Table 3-5. Genome annotation of the phage P26218. Only the ORFs with assigned function are shown.

ORF	Start	Stop	Strand	Best BLAST match (Virus family)	Function (Accession No.)	E-value	DB
4	1034	1918	+	*Leptospira phage LE1 (Myoviridae)*	ParB-like nuclease (CAE14777.1)	5.00E-04	Genbank_nr
6	2879	3424	+	*Escherichia phage N4 (Podoviridae)*	Hypothetical protein (YP_950527.1)	9.00E-18	Genbank_nr
7	3486	4172	+	*Bacillus phage PBC1 (Siphoviridae)*	Hypothetical protein (YP_006383481.1)	5.00E-17	Genbank_nr
8	4169	4570	+	Bacteriophage APSE-2 (Podoviridae)	Hypothetical protein (ABA29383.1)	4.00E-09	Genbank_nr
9	4737	5342	+	*Clostridium phage phiCT453A (UN)*	Hypothetical protein (AJA42507.1)	1.00E-19	Genbank_nr
11	5925	7889	+	*Staphylococcus phage tp310-2 (Siphoviridae)*	PolA DNA polymerase I (YP_001429916.1)	1.00E-96	Genbank_nr
12	7865	8176	+	Bacteriophage APSE-5 (Unclassified)	VRR-NUC domain (ACJ10148.1)	8.00E-14	Genbank_nr
13	8201	9730	+	*Xylella phage Xfas53 (Podoviridae)*	Helicase (YP_003344894.1)	2.00E-103	Genbank_nr
14	9773	10606	+	*Pseudomonas phage vB_PaeS_PAO1_Ab18 (Siphoviridae)*	Hypothetical protein (CEF89648.1)	2.00E-22	Genbank_nr
15	10603	11181	+	*Pseudomonas phage vB_PaeS_PAO1_Ab18 (Siphoviridae)*	Gln Q ABC-type polar amino acid transport system (CEF89651.1)	6.00E-15	Genbank_nr
16	11195	12244	+	*Pseudomonas phage MP1412 (Siphoviridae)*	Thymidylate synthase (YP_006561023.1)	8.00E-75	Genbank_nr

Table 3-5. (continued)

ORF	Start	Stop	Strand	Best BLAST match (Virus family)	Function (Accession No.)	E-value	DB
17	12241	12780	+	*Pseudomonas phage YuA (Siphoviridae)*	HD containing hydrolase-like protein (YP_001595841.1)	6.00E-18	Genbank_nr
18	12777	13265	+	*Pseudomonas phage M6 (Siphoviridae)*	Hypothetical protein (YP_001294569.1)	2.00E-15	Genbank_nr
21	14668	14138	-	*Thalassomonas phage BA3 (Podoviridae)*	Hypothetical protein (YP_001552992.1)	4.00E-15	Genbank_nr
22	15213	14668	-	*Acinetobacter phage (Podoviridae)*	zliS Lysozyme family protein (YP_007010632.1)	8.00E-63	Genbank_nr
26	17721	16342	-	*Thalassomonas phage BA3 (Podoviridae)*	Hypothetical protein (YP_001552270.1)	4.00E-14	Genbank_nr
28	21106	19718	-	EBPR Podoviridaevirus 1 (Podoviridae)	Hypothetical protein (AEI70866.1)	7.00E-30	Genbank_nr
29	21860	21120	-	*Vibrio phage VvAW1 (Podoviridae)*	Phage protein (YP_007518345.1)	1.00E-35	Genbank_nr
30	22732	21857	-	*Ralstonia phage RSK1 (Podoviridae)*	Hypothetical protein (YP_008853798.1)	2.00E-20	Genbank_nr
32	24478	23051	-	*Vibrio phage VvAW1 (Podoviridae)*	Phage stabilisation protein (YP_007518349.1)	3.00E-111	Genbank_nr
33	25184	24480	-	EBPR Podoviridaevirus 1 (Podoviridae)	Hypothetical protein (AEI70872.1)	6.00E-41	Genbank_nr
37	27771	26560	-	EBPR Podoviridaevirus 1 (Podoviridae)	P22 Coat Protein (AEI70875.1)	0.00E+00	Genbank_nr

Table 3-5. (continued)

ORF	Start	Stop	Strand	Best BLAST match (Virus family)	Function (Accession No.)	E-value	DB
38	28658	27801	-	EBPR Podoviridaevirus 1 (Podoviridae)	Phage-scaffold protein (AEI70876.1)	2.00E-44	Genbank_nr
39	30847	28655	-	EBPR Podoviridaevirus 1 (Podoviridae)	Head-tail connecting protein (AEI70877.1)	0.00E+00	Genbank_nr
40	31236	30844	-	EBPR Podoviridaevirus 1 (Podoviridae)	GNAT acetyltransferase (AEI70878.1)	1.00E-08	Genbank_nr
41	32769	31252	-	*Pelagibacter phage HTVC010P (Podoviridae)*	Phage terminase, large subunit (YP_007517700.1)	3.00E-105	Genbank_nr
44	36115	33944	-	*Yersinia phage PY100 (Myoviridae)*	DnaB-like ATP-dependent helicase (CAJ28484.1)	4.00E-11	Genbank_nr

Table 3-6. Genome sequencing information of the phages P26059A and P26059B

Features	**P26059A**	**P26059B**
Sequencing library	Paired-end TruSeq library	Paired-end TruSeq library
Sequencing platform	Illumina MiSeq	Illumina MiSeq
Fold coverage	1,205 ×	4,247 ×
Genome length	84,008 bp	41,471 bp
G+C%	43.60%	54.30%
No. of coding sequences	124	46
tRNA	2	0
Assembler	SPAdes-3.5.0	SPAdes-3.8.2
Gene calling	RAST ver. 2.0	RAST ver. 2.0
GenBank ID	KY981271	KY981272

coding genes were predicted by the RAST server, each gene was analyzed using BLAST against the NCBI nr and env-nr protein database. Only the BLAST match results with e-value less than 0.001 were considered as the predicted function of the gene. For the coding genes with no predicted function from the NCBI nr and end-nr BLASTp search, their functions were predicted based on the protein domain search made against CDD and Pfam database (Fig. 3-10). For P26059A, out of 124 predicted protein genes, 63 of them had a significant match in either one of NCBI nr, env-nr, or UniProt database. Nine of the protein coding genes of the P26059A did not have a functional protein assignment, so their functions were predicted based on the conserved protein domain found in the gene. Despite the extensive search, 27 of the genes remain as unknown with its function and 15 among them were found to have no significant BLAST match at all, leaving them as unique proteins of P26059A (Table 3-7). When P26059B protein coding genes were analyzed using BLAST upon NCBI protein nr database, 31 out of 46 genes had a significant match results. The remaining genes were searched upon protein databases, but no conserved protein domain was found. Among the 31 annotated genes, 12 of them were not able to be assigned with a known function (Table 3-8).

The P26059A genome carried 17 genes related to DNA modification and replication. Among those, total of 6 endonuclease genes were found and 4 of them (ORFs 11, 15, 20, and 58) contained a GIY-YIG domain which is widely found in prokaryotes and eukaryotes. The GIY-YIG domains are known to be found in endonucleases that function as repair system of damaged DNA in prokaryotes. Within bacteriophage genome, it is known to function in cleaving the host DNA to utilize the host nucleotides in phage genome replication (Mak *et al*., 2010), leading to active replication of phage genomes. Along with endonucleases, the phage P26059A encoded for the PhoH family protein (ORF 117). The PhoH proteins are well distributed among marine bacteriophages and cyanophages and their coding

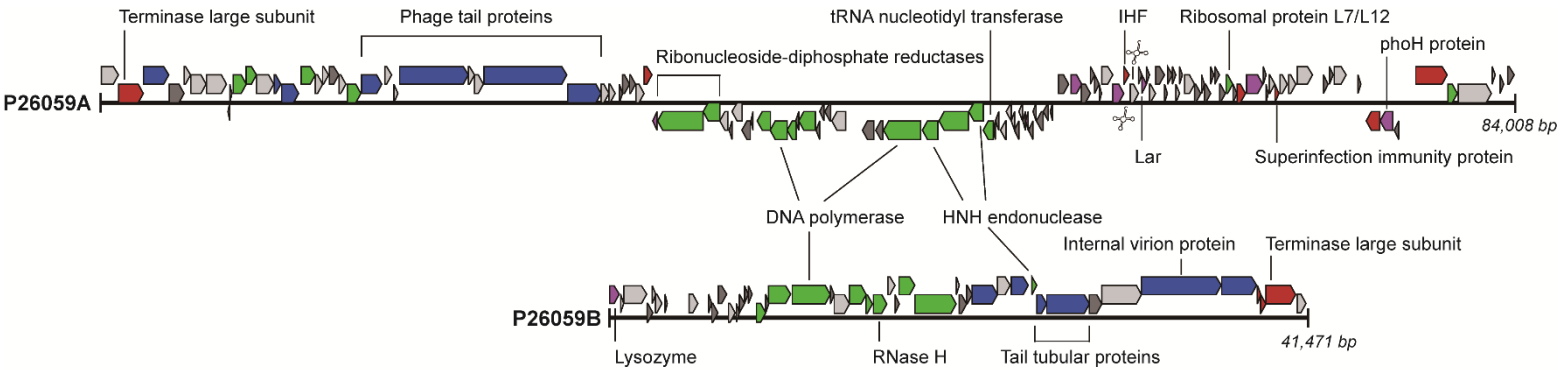

Figure 3-10. Genome map of the phages P26059A and P26059B. The blue color represents structural genes, the green represents genes related to DNA replication, recombination, and modification, the red represents those related to cell lysis and packaging, and the purple color represent auxiliary metabolic genes. The tRNAs are marked with cross signs.

Table 3-7. Genome annotation of the phage P26059A. Only the ORFs with assigned function are shown.

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
1	51	1067	+	*Vibrio owensii*	Hypothetical protein (WP_039839651.1)	3.00E-26	GenBank nr
2	1078	2550	+	*Pseudomonas* phage JG004	Terminase large subunit (YP_007002481.1)	1.00E-101	GenBank nr
3	2568	4064	+	*Pseudomonas* phage KPP10	Structural protein (YP_004849306.1)	5.00E-119	GenBank nr
5	4980	5345	+	*Sinorhizobium medicae*	Hypothetical protein (WP_011975491.1)	3.00E-20	GenBank nr
6	5369	6277	+	*Rhodobacter* sp. SW2	conserved Hypothetical protein (EEW23746.1)	3.00E-12	GenBank nr
7	6296	7480	+	*Brevundimonas* sp. Root1279	Hypothetical protein (A0A0Q7E659)	2.00E-26	Swiss-prot
10	7886	8653	+	*Pseudomonas* phage VCM	Hypothetical protein (cl24270)	1.31E-27	CDD
11	8661	9266	+	*Bacillus cereus* VD107	Hypothetical protein (cl15257)	2.47E-09	CDD
12	9263	10288	+	*Pseudomonas* phage PAK_P1	Hypothetical protein PAK_P100004 (YP_004327197.1)	5.00E-38	GenBank nr
13	10319	10717	+	*Rhodobacteraceae* bacterium HLUCCA12	Bacteriophage lambda head decoration protein D (KPQ04613.1)	1.00E-14	GenBank nr
14	10756	11769	+	*Pseudomonas* phage KPP10	Major structural protein (YP_004306755.1)	8.00E-83	GenBank nr
15	11928	12758	+	*Pseudomonas* phage phiPto-bp6g	HNH endonuclease family protein (G3KB07)	1.50E-09	Swiss-prot
16	12767	13192	+	*Pseudomonas* phage phiPsa374	Hypothetical protein (W8EIG9)	2.30E-16	Swiss-prot

Table 3-7. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
17	13194	13550	+	Marine metagenome	Hypothetical protein (EBO74624.1)	8.00E-34	env-nr
19	14165	14569	+	*Methylovorus glucosotrophus*	Hypothetical protein (WP_015830281.1)	1.00E-26	GenBank nr
20	14659	15465	+	Uncultured archaeon	Putative endonuclease (D1JFH5)	1.20E-13	Swiss-prot
21	15494	16708	+	*Pseudomonas* phage PaMx74	Putative major tail structural protein (YP_009199465.1)	4.00E-44	GenBank nr
22	16890	17300	+	*Burkholderia ubonensis*	Hypothetical protein (WP_059735913.1)	2.00E-16	GenBank nr
23	17434	17661	+	*Rhizobium tropici*	Hypothetical protein (WP_052227533.1)	2.00E-06	GenBank nr
24	17766	21848	+	*Pseudomonas syringae* pv. apii	Phage tail tape measure protein lambda (A0A0P9J5U)	2.00E-61	Swiss-prot
25	21858	22223	+	*Desulfovibrio vulgaris*	Hypothetical protein (WP_010939433.1)	2.00E-15	GenBank nr
26	22220	22741	+	*Desulfovibrio vulgaris*	Hypothetical protein (WP_014524470.1)	2.00E-39	GenBank nr
27	22767	27701	+	*Methyloceanibacter caenitepidi*	Phage-related protein, tail component (A0A0A8K5V0)	1.30E-108	Swiss-prot
28	27748	29724	+	*Sinorhizobium* phage phiN3	Putative tail fiber protein (A0A0R8UEH9)	9.00E-37	Swiss-prot
29	29734	30204	+	*Variovorax paradoxus*	Hypothetical protein (WP_057592238.1)	5.00E-20	GenBank nr

Table 3-7. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
30	30205	30567	+	*Dechlorosoma suillum* (strain ATCC BAA-33 / DSM 13638 / PS)	Hypothetical protein (G8QPT8)	2.40E-13	Swiss-prot
34	31798	32289	+	*Labrenzia alba*	Hypothetical protein (A0A0M6YUJ9)	9.10E-06	Swiss-prot
35	32286	32750	+	*Burkholderia vietnamiensis*	Peptidase M15 (WP_011886401.1)	2.00E-36	GenBank nr
36	33053	32811	-	*Halosimplex carlsbadense*	Thiol reductase thioredoxin (WP_006884752.1)	1.00E-10	GenBank nr
37	35798	33093	-	*Acinetobacter* phage YMC13/03/R2096	Ribonucleoside-diphosphate reductase (YP_009146801.1)	0.00E+00	GenBank nr
38	36780	35791	-	*Shewanella* sp. phage 1/40	RnR beta subunit (YP_009104029.1)	6.00E-112	GenBank nr
39	37253	36792	-	*Pseudomonas* phage KPP10 (Bacteriophage KPP10)	Hypothetical protein (D6RRL7)	3.90E-04	Swiss-prot
41	38112	37516	-	Marine sediment metagenome	Hypothetical protein (A0A0F9K2S2)	2.00E-14	Swiss-prot
44	39794	39000	-	*Bacillus* phage B4	Putative DNA-binding protein 2 (YP_006908397.1)	1.00E-16	GenBank nr
45	40783	39779	-	Uncultured Mediterranean phage uvMED	DNA polymerase I (BAQ92511.1)	5.00E-32	GenBank nr
46	41326	40793	-	Enterobacteria phage vB_KleM-RaK2	Recombination endonuclease VII (H6X3P3)	2.50E-13	Swiss-prot
48	42458	41481	-	*Klebsiella* phage vB_KpnM_KB57	Putative exodeoxyribonuclease (YP_009187734.1)	1.00E-44	GenBank nr

Table 3-7. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
52	44267	43443	-	*Bacillus* phage G	Gp205 (G3MBS1)	1.00E-28	Swiss-prot
55	48724	46520	-	*Vibrio* phage vB_VchM-138	Putative DNA polymerase (YP_007006391.1)	2.00E-135	GenBank nr
56	49735	48824	-	*Salmonella enterica* subsp. enterica serovar Give str. 564	HNH endonuclease (ESH10154.1)	2.00E-25	GenBank nr
57	51554	49806	-	*Pseudomonas* phage PAK_P1	Putative DNA primase/helicase (YP_004327242.1)	2.00E-141	GenBank nr
58	52425	51565	-	*Salmonella* phage Shivani	HNH homing endonuclease (A0A0A7TWE4)	2.30E-11	Swiss-prot
59	53035	52436	-	*Escherichia* phage phAPEC8	tRNA nucleotidyl transferase/ poly (A) polymerase (YP_007348551.1)	2.00E-10	GenBank nr
62	53794	53456	-	*Achromobacter xylosoxidans*	Hypothetical protein Axylo_1461 (AKP88981.1)	1.00E-16	GenBank nr
63	54099	53791	-	*Colwellia* phage 9A	Hypothetical protein (I3UMK3)	9.10E-06	Swiss-prot
64	54287	54096	-	Marine sediment metagenome	Hypothetical protein (A0A0F8XL42)	3.20E-21	Swiss-prot
66	54771	54550	-	*Synechococcus* phage S-EIVl	Hypothetical protein (A0A0C4K633)	2.00E-09	Swiss-prot
75	57650	58270	+	*Pseudomonas* phage vB_PaeM_C2-10_Ab1	Putative phosphoesterase (YP_007236845.1)	9.00E-40	GenBank nr
80	59472	60125	+	Marine sediment metagenome	Hypothetical protein (A0A0F9MA30)	1.80E-10	Swiss-prot
81	60118	60783	+	Uncultured bacterium	Thymidylate synthase complementing protein (AIA11069.1)	5.00E-76	GenBank nr

Table 3-7. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
82	60783	61100	+	-	Integration host factor (IHF) (cd13832)	8.03E-05	CDD
83	61174	61653	+	*Acidovorax* sp. Root70	Hypothetical protein (A0A0Q8LHN8)	9.60E-06	Swiss-prot
85	61853	62125	+	-	Restriction alleviation protein, Lar family (cl08047)	8.01E-04	CDD
87	62305	62664	+	Marine sediment metagenome	Hypothetical protein (A0A0F9K1M2)	3.10E-07	Swiss-prot
92	63868	64089	+	*Morganella morganii*	Hypothetical protein (WP_049246392.1)	9.00E-06	GenBank nr
94	64394	64981	+	*Burkholderia pseudomallei* (*Pseudomonas pseudomallei*)	Putative bacteriophage protein (A0A0H5L1B2)	3.50E-55	Swiss-prot
101	66866	67249	+	-	Ribosomal protein L7/L12 (pfam00542)	6.10E-04	Pfam
103	67510	68001	+	*Paracoccus sphaerophysae*	Cell wall hydrolase (WP_036720692.1)	2.00E-19	GenBank nr
104	68072	69037	+	*Chlamydia trachomatis*	Predicted acyltransferase (CRH65641.1)	9.00E-63	GenBank nr
107	69419	69751	+	*Campylobacter* sp. FOBRC14	Hypothetical protein (WP_009650654.1)	7.00E-05	GenBank nr
108	69754	69981	+	*Burkholderia* sp. RPE67	Superinfection immunity protein (pfam14373)	9.00E-12	Pfam
109	70061	70495	+	Marine metagenome	Hypothetical protein GOS_7775461 (EBQ27659.1)	3.00E-07	env-nr
110	70568	70999	+	*Hafnia alvei* FB1	Hypothetical protein (A0A097R3K3)	5.30E-17	Swiss-prot

Table 3-7. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
111	71073	71990	+	*Escherichia coli*	Phage protein (WP_021549731.1)	1.00E-30	GenBank nr
114	73284	73991	+	*Pseudomonas* phage PAK_P1	Hypothetical protein PAK_P100105 (YP_004327286.1)	8.00E-80	GenBank nr
116	75992	75186	-	*Vibrio* phage PWH3a-P1	Hypothetical protein (cd07016)	7.23E-17	CDD
117	76763	76002	-	*Escherichia* phage phAPEC8	Putative PhoH family protein (YP_007348464.1)	1.00E-37	GenBank nr
119	78123	79991	+	*Burkholderia glumae* PG1	Putative membrane-anchored cell surface protein (A0A0B6S8R5)	9.10E-05	Swiss-prot
120	80032	80592	+	Uncultured organism MedDCM-OCT-S09-C20	Putative Hypothetical protein (pfam13392)	1.93E-05	Pfam
121	80644	82602	+	*Klebsiella pneumoniae* subsp. *pneumoniae* KP5-1	Hypothetical protein (A0A0M5JR58)	1.30E-16	Swiss-prot
122	82653	82847	+	Marine metagenome	Hypothetical protein GOS_2910462 (ECV37839.1)	1.00E-14	env-nr

Table 3-8. Genome annotation table of the phage P26059B. Only the ORFs with assigned function are shown.

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
1	510	1	-	*Acinetobacter guillouiae*	Hypothetical protein (BAP37666.1)	2.00E-10	GenBank nr
2	2289	520	-	*Caulobacte*r phage Percy	Large terminase (YP_009225282.1)	0.00E+00	GenBank nr
3	2597	2286	-	*Caulobacter* phage Cd1	Putative DNA maturase A (ADD21677.1)	7.00E-07	GenBank nr
4	2778	2584	-	*Paraburkholderia diazotrophica*	Phage holin T7 family, holin superfamily II (SEI42881.1)	6.00E-13	GenBank nr
5	4900	2828	-	*Sphingobacterium deserti*	Hypothetical protein (WP_037498051.1)	3.00E-28	GenBank nr
6	9673	4940	-	*Caulobacter* phage Cd1	Internal virion protein (ADD21673.1)	9.00E-139	GenBank nr
7	12030	9682	-	*Pseudomonas* phage VSW-3	Hypothetical protein (ANH51101.1)	1.00E-26	GenBank nr
9	15296	12768	-	*Caulobacter* phage Cd1	Tail tubular protein B (ADD21670.1)	0.00E+00	GenBank nr
10	15887	15297	-	*Xanthomonas* phage f30-Xaj	Tail tubular protein A (AMM44688.1)	4.00E-35	GenBank nr
11	16183	15890	-	*Desulfosporosinus acididurans*	HNH endonuclease (WP_053006391.1)	9.00E-21	GenBank nr
12	17401	16394	-	*Ralstonia* phage RSB1	Major capsid-like protein (YP_002213721.1)	3.00E-119	GenBank nr
13	18223	17474	-	*Burkholderia thailandensis* MSMB43	Hypothetical protein (EIP87426.1)	4.00E-36	GenBank nr
14	19725	18220	-	*Caulobacte*r phage Cd1	Head-to-tail joining protein (ADD21666.1)	1.00E-154	GenBank nr

Table 3-8. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
17	20662	20492	-	*Burkholderia ubonensis*	Hypothetical protein (WP_060288674.1)	1.00E-08	GenBank nr
18	23135	20685	-	*Caulobacter* phage Percy	DNA-dependent RNA polymerase (YP_009225265.1)	0.00E+00	GenBank nr
19	24070	23132	-	*Burkholderia* phage Bp-AMP1	Putative ATP-dependent DNA ligase (CDK30097.1)	1.00E-33	GenBank nr
21	24721	24308	-	*Candidatus Accumulibacter* sp. SK-11	Hypothetical protein (EXI76492.1)	2.00E-37	GenBank nr
22	25577	24798	-	*Burkholderia* sp. BDU5	RNase H superfamily protein (WP_059473379.1)	3.00E-96	GenBank nr
23	26042	25659	-	*Xylella* phage Prado	DNA endonuclease VII (YP_008859405.1)	7.00E-43	GenBank nr
24	27006	26017	-	*Burkholderia* phage JG068	DNA exonuclease (YP_008853860.1)	1.00E-56	GenBank nr
25	27914	27006	-	*Ralstonia solanacearum*	Hypothetical protein (WP_042591591.1)	2.00E-64	GenBank nr
27	30585	28150	-	*Caulobacter* phage Cd1	DNA polymerase (ADD21653.1)	0.00E+00	GenBank nr
28	31892	30585	-	*Ralstonia* phage RSJ2	Putative DNA helicase (YP_009216554.1)	5.00E-168	GenBank nr
30	32625	32068	-	*Burkholderia cepacia*	DNA primase (WP_060050871.1)	5.00E-46	GenBank nr
35	34287	33886	-	*Vibrio vulnificus*	Hypothetical protein (KOR91322.1)	1.00E-15	GenBank nr
36	34889	34392	-	*Rhizobium leguminosarum*	Hypothetical protein (WP_027684348.1)	3.60E-01	GenBank nr

Table 3-8. (continued)

ORF	Start	Stop	Strand	Best BLAST match	Function (Accession No.)	E-value	DB
39	36622	36095	-	*Pseudomonas* phage YMC11/06/C171_PPU_BP	Hypothetical protein (YP_009275032.1)	2.00E-20	GenBank nr
41	38611	38231	-	*Bradyrhizobium* sp. Cp5.3	Hypothetical protein (WP_027554547.1)	5.00E-08	GenBank nr
42	38807	38604	-	*Vibrio nigripulchritudo* SOn1	Hypothetical protein (CCO46700.1)	3.00E-04	GenBank nr
44	40499	39141	-	*Sphingomonas adhaesiva*	Hypothetical protein (WP_066707580.1)	1.00E-65	GenBank nr
45	40693	40496	-	*Gemmatimonas* sp. SG8_17	Hypothetical protein (KPJ91942.1)	3.00E-04	GenBank nr
46	41267	40773	-	*Burkholderia* phage Bp-AMP4	Putative lysozyme (CDL65241.1)	9.00E-38	GenBank nr

genes are used to make phylogenetic classifications (Fig. 3-11) (Adriaenssens and Cowan, 2014; Goldsmith *et al*., 2011). The *phoH* gene expression within bacteria is known to be induced under phosphate starvation, which would promote uptake of phosphorous from the environment. Thereby, *phoH* gene carried by the phage genome is suspected to be expressed in the host cells to promote phosphorous uptake and lead to increase in phosphorus level within the cell to be utilized for phage genome replication. In oligotrophic environments, where phosphorous is typically limiting, such strategy would provide advantage of faster phage DNA replication with sufficient amount resources.

For its genome packaging and assembly of phage particles, P26059A carried diverse proteases with different purposes. Number of typical proteases found in phage genomes were also found in the genome of P26059A, such as serine protease XkdF (ORF 10), peptidase M15 (ORF 35), and cell wall hydrolase (ORF 103). ORF 82 encoded for an integration host factor, IHF, which functions for condensation of nucleotides in bacterial cells (Sanyal *et al*., 2014). In bacteriophage genomes, the IHFs also condenses bacteriophage genomes in order to package them into capsid proteins. ORFs 101 and 116 encode for caseinolytic proteases (Clp) which are commonly found in bacterial genes. Within many bacterial cells, ATP-activated chaperon subunit, ClpA and protease subunit, ClpP, form a chaperon-protease pair to degrade foreign proteins found in bacterial cell, which are mostly phage proteins (Gaillot *et al*., 2000). The protease subunit ClpP have been found in bacteriophage genomes as well and it was revealed that these proteases are utilized by phage particles and act as prohead proteases for packaging (Cheng *et al*., 2004). ClpS is an adaptor protein that binds to the ClpA and ClpP pair to stabilize the ClpA, thereby forming ClpAPS complex to degrade aggregated proteins and foreign proteins (Dougan *et al*., 2002). The ClpS protein is commonly found in prokaryotic cells but not in viral cells. Only few phages infecting enteric bacteria have been

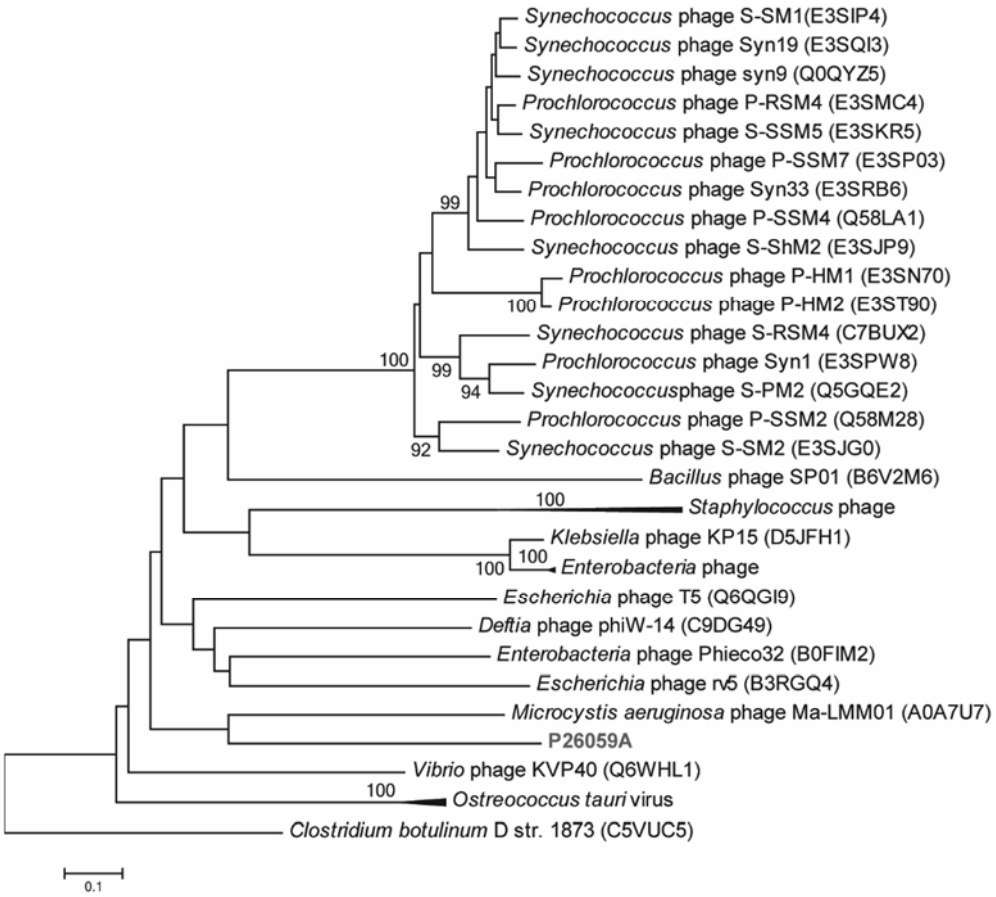

Figure 3-11. Neighbor-joining phylogenetic tree of the phage P26059A using *phoH* gene. The reference sequences were collected from Pfam database. The tree was constructed using MEGA 6 (Tamura *et al*., 2013) with bootstrap of 1,000 after performing alignment using CLUSTAL X (Thompson *et al*., 2002)

reported to be carrying *clpS* gene. In P26050A genome, ORF 116 encodes for ClpP and ORF 101 encodes for ClpS- like protein. From previous studies, it is reasonable to hypothesize that ClpP and ClpS would function as proteases to cleave phage proteins translated by the host bacteria into designated size and structure for phage packaging. However, mechanism for phage particles to utilize Cp proteases for their beneficial purposes without inactivation of their own genes must be studied further for understanding of phage genomics. While the phage P26059A carried numerous genes for active replication of its genome, it also had genes for defense mechanism. The ORF 85 encodes for a Lar family protein, a restriction alleviation protein (King and Murray, 1995), which protects phage genome from host restriction endonucleases. Also, ORF 108 encodes for a superinfection immunity protein to prevent infection of secondary phages upon infection of the first (Abedon, 2015; Lu and Henning, 1989). Presence of such defense mechanisms may have provided better survival of P26059A over other phages infecting the identical host.

Within the genome of P26059A, the ORF 59 was annotated as tRNA nucleotidyl transferase/ poly (A) polymerase, which contributes in tRNA elongation (Table 3-7), hinting for the presence of tRNA gene. Therefore, tRNA was searched using the tRNAscan-SE 2.0 (Lowe and Eddy, 1997) and one Arg-tRNA was found between 61,021 bp and 61,096 bp. Also, when tRNA was searched further using ARAGORN engine (Laslett and Canback, 2004), Pyl-tRNA (Pyrrolysine) was also found between 62,398 and 61,396 bp of P26059A genome. The Pyl-tRNA are known to be strictly found only in bacteria and archaea while found rarely in phage genomes.

For phylogenetic analysis of the phages P26059A and P26059B, Neighbor-Joining phylogenetic tree was constructed using *terL* gene sequences that were carried by both phages (Fig. 3-12). As expected, two phages were too divergent from each other that they were not able to be classified into a monophyletic group. The phage P26059B was clustered with representative bacteriophage strains of the family

Podoviridae. However, P26059A *terL* gene was clustered with representative sequences of the family *Myoviridae*, despite that it was morphologically classified as a member of the family *Siphoviridae.*

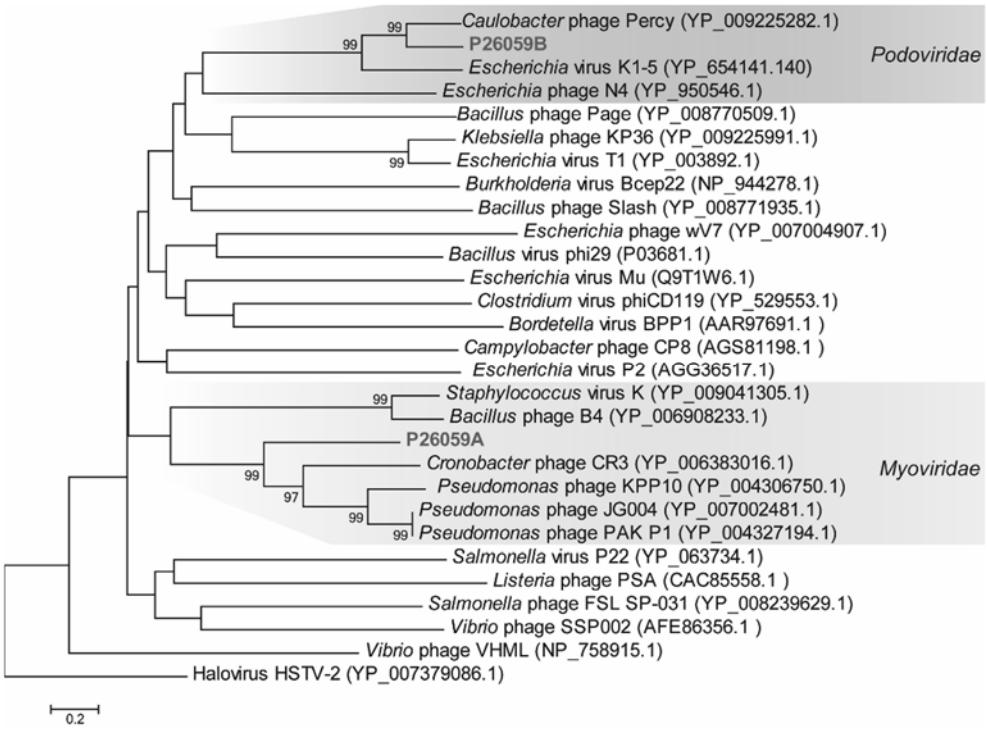

Figure 3-12. Neighbor-joining phylogenetic tree of the phages P26059A and P26059B. The tree was constructed using genes coding for terminase large subunit. The reference sequences were collected from NCBI nr database. The tree was constructed using MEGA 6 (Tamura *et al*., 2013) with bootstrap of 1,000 after performing alignment using CLUSTAL X (Thompson *et al*., 2002)

3.3. Abundance and distribution of isolated bacteriophages in freshwater lakes

Along with physical isolation and genome analysis of the phages isolated from Lake Soyang, their genomic abundance and distribution within their original habitat was observed through competitive binning analysis using the viral metagenome samples prepared from Lake Soyang. More than 9 million sequences from each virome were assigned by DIAMOND (Buchfink *et al*., 2015) to the best-matching protein in a database that included all viral proteins and all bacterial non-redundant proteins from NCBI RefSeq (release 72 or 79) in addition to the proteins annotated in the phage genomes obtained in this study. Among the reads assigned to viruses, which comprised 12.45–29.39% of all assigned reads, four bacteriophages isolated from this study occupied 2.08% to 11.26%, unveiling the identity of virome reads that were unknown. When competitive binning results for all four phages isolated from Lake Soyang were compared, P19250A appeared to be the most abundant bacteriophage, especially in winter seasons (’15 Jan and ’16 Feb; Table 3-9). Also, P19250A was the most highly-assigned freshwater phage for five out of six samples (except in ’15 Sept., Table 3-10), while most other highly-assigned viruses were from marine environments (Fig. 3-13a). Interestingly, the proportion of P19250A-assigned reads showed the same appearance as that of the LD28 clade (Fig. 3-13); the proportion of both P19250A-assigned reads and the LD28 clade showed peaks in winter (’15 Jan. and ’16 Feb.) and recorded the lowest values in summer (’15 Sept.; Fig. 3-13b), suggesting that proliferation of a phage type represented by P19250A is dependent on the presence of LD28 clade.

P19250A was also highly assigned in binning analysis of other freshwater viromes. When 40 viromes from 8 freshwater lakes and reservoirs (Green *et al*., 2015; Mohiuddin and Schellhorn, 2015; Roux *et al*., 2012; Skvortsov *et al*., 2016; Tseng *et al*., 2013; Watkins *et al*., 2015) were analyzed using the same method as that for

Table 3-9. Competitive binning results of four phages isolated from Lake Soyang

	'14 Oct.		**'15 Jan.**		**'15 Sept.**		**'15 Nov.**		**'16 Feb.**		**'16 May**	
	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank
Viral reads[a]	13.99%	-	12.45%	-	29.39%	-	24.65%	-	15.97%	-	16.04%	-
P19250A[b]	1.45%	9	4.77%	1	0.65%	25	2.19%	8	8.28%	1	3.90%	2
P26218	0.47%	45	1.37%	10	0.41%	47	0.53%	31	2.63%	3	2.73%	4
P26059A	0.76%	27	0.38%	53	1.01%	15	0.18%	104	0.34%	55	0.72%	26
P26059B	0.01%	513	0.01%	776	0.01%	673	0.01%	505	0.01%	691	0.20%	98

[a](Number of reads assigned to viruses)/(Number of reads assigned to bacteria or viruses) × 100

[b](Number of reads assigned to a bacteriophage)/(Number of reads assigned to viruses) × 100

Table 3-10. Ranks of bacteriophages within analyzed virome samples

Sample name	Rank	Name	% within viral binned reads	Origin
Lake Soyang	1	*Synechococcus* phage S-SM2	6.54	Marine
'14 Oct.	2	*Synechococcus* phage S-CBS4	3.88	Estuary
	3	*Pelagibacter* phage HTVC008M	3.64	Marine
	4	*Pelagibacter* phage HTVC010P	3.28	Marine
	5	*Persicivirga* phage P12024S	3.24	Marine
	6	*Puniceispirillum* phage HMO-2011	3.16	Marine
	7	*Synechococcus* phage S-SKS1	2.65	Marine
	8	Cyanophage KBS-S-2A	1.95	Marine
	9	*Prochlorococcus* phage P-SSM2	1.89	Marine
	10	*Synechococcus* phage S-SSM7	1.78	Marine
	11	**P19250A**	1.32	Freshwater
Lake Soyang	1	**P19250A**	5.13	Freshwater
'15 Jan.	2	*Puniceispirillum* phage HMO-2011	3.03	Marine
	3	*Rhodothermus* phage RM378	2.80	Hot spring
	4	*Pelagibacter* phage HTVC008M	2.32	Marine
	5	*Synechococcus* phage S-CBS4	2.14	Estuary
	6	*Synechococcus* phage S-SM2	1.95	Marine
	7	Cyanophage KBS-S-2A	1.82	Marine
	8	*Pelagibacter* phage HTVC010P	1.51	Marine
	9	*Microcystis* phage Ma-LMM01	1.34	Freshwater
	10	*Synechococcus* phage S-CBP3	1.33	Estuary
	11	*Synechococcus* phage S-CBS2	1.22	Estuary
Lake Soyang	1	*Synechococcus* phage S-SM2	8.31	Marine
'15 Sept.	2	*Synechococcus* phage S-SKS1	7.83	Marine
	3	*Prochlorococcus* phage P-SSM2	6.02	Marine
	4	*Pelagibacter* phage HTVC008M	5.29	Marine
	5	*Pelagibacter* phage HTVC010P	4.49	Marine
	6	*Puniceispirillum* phage HMO-2011	2.91	Marine
	7	*Synechococcus* phage ACG-2014f	2.73	Marine
	8	*Synechococcus* phage S-SSM7	2.49	Marine
	9	*Synechococcus* phage S-CBS4	1.81	Estuary
	10	*Synechococcus* phage S-PM2	1.60	Marine
	11	*Synechococcus* phage S-RIM8 A.HR1	1.33	Marine

Table 3-10. (continued)

Sample name	Rank	Name	% within viral binned reads	Origin
Lake Soyang	1	*Chrysochromulina ericina* virus	6.15	Marine
'15 Nov.	2	*Phaeocystis globosa* virus	5.46	Freshwater
	3	*Pelagibacter* phage HTVC010P	4.30	Marine
	4	*Synechococcus* phage S-SM2	3.97	Marine
	5	*Synechococcus* phage S-SSM7	3.41	Marine
	6	*Pelagibacter* phage HTVC008M	3.05	Marine
	7	*Synechococcus* phage S-SKS1	2.77	Marine
	8	**P19250A**	2.28	Freshwater
	9	*Puniceispirillum* phage HMO-2011	2.17	Marine
	10	*Prochlorococcus* phage P-SSM2	1.88	Marine
	11	*Aureococcus anophagefferens* virus	1.52	Marine
Lake Soyang	1	**P19250A**	8.70	Freshwater
'16 Feb	2	Cyanophage KBS-S-2A	4.64	Marine
	3	*Idiomarinaceae* phage 1N2-2	3.54	Marine
	4	*Salicola* phage CGphi29	3.33	Marine
	5	*Synechococcus* phage S-SM2	2.44	Marine
	6	*Synechococcus* phage S-SSM7	2.27	Marine
	7	*Puniceispirillum* phage HMO-2011	2.20	Marine
	8	*Pelagibacter* phage HTVC008M	1.81	Marine
	9	*Synechococcus* phage S-SKS1	1.79	Marine
	10	*Synechococcus* phage S-CBS4	1.51	Estuary
	11	*Synechococcus* phage S-CBS1	1.29	Marine
Lake Soyang	1	Cyanophage KBS-S-2A	7.59	Marine
'16 May	2	*Puniceispirillum* phage HMO-2011	4.72	Marine
	3	**P19250A**	4.20	Freshwater
	4	*Chrysochromulina ericina* virus	2.57	Marine
	5	*Idiomarinaceae* phage 1N2-2	2.19	Marine
	6	*Phaeocystis globosa* virus	2.11	Freshwater
	7	*Acanthocystis turfacea Chlorella virus* 1	1.92	Freshwater
	8	*Salicola* phage CGphi29	1.90	Marine
	9	*Persicivirga* phage P12024S	1.65	Marine
	10	*Pelagibacter* phage HTVC008M	1.62	Marine
	11	*Synechococcus* phage S-RIP2	1.41	Marine

Table 3-10. (continued)

Sample name	Rank	Name	% within viral binned reads	Origin
Matoaka open,	1	*Puniceispirillum* phage HMO-2011	18.59	Marine
US	2	**P19250A**	5.30	Freshwater
	3	*Acanthocystis turfacea Chlorella virus* 1	4.34	Freshwater
	4	*Persicivirga* phage P12024S	3.85	Marine
	5	*Nitrincola* phage 1M3-16	2.61	Unknown
	6	*Citrobacter* phage CVT22	2.01	Termite gut
	7	*Celeribacter* phage P12053L	1.83	Marine
	8	*Cellulophaga* phage phi38:1	1.67	Marine
	9	*Roseobacter* phage SIO1	1.34	Marine
	10	Cyanophage PP	1.24	Freshwater
	11	*Pelagibacter* phage HTVC008M	1.22	Marine
Lough Neagh,	1	**P19250A**	6.26	Freshwater
UK	2	*Idiomarinaceae* phage 1N2-2	6.04	Marine
	3	*Salicola* phage CGphi29	5.99	Marine
	4	*Persicivirga* phage P12024S	4.30	Marine
	5	*Synechococcus* phage S-CBS4	2.59	Estuary
	6	Cyanophage KBS-S-2A	2.26	Marine
	7	*Puniceispirillum* phage HMO-2011	2.05	Marine
	8	*Planktothrix* phage PaV-LD	1.89	Freshwater
	9	*Pelagibacter* phage HTVC008M	1.80	Marine
	10	*Cronobacter* phage vB CsaM GAP32	1.56	Sewage
	11	*Cellulophaga* phage phi38:1	1.46	Marine

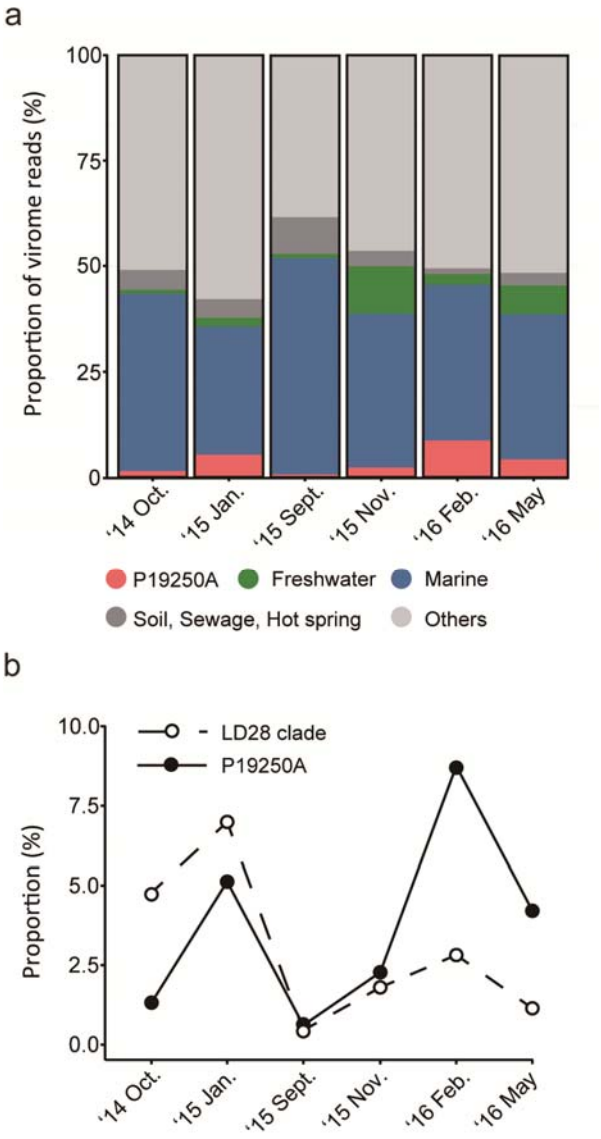
a
100
75
50
25
0
Proportion of virome reads (%)
'14 Oct.
'15 Jan.
'15 Sept.
'15 Nov.
'16 Feb.
'16 May
P19250A
Freshwater
Marine
Soil, Sewage, Hot spring
Others
b
10.0
7.5
5.0
2.5
0.0
Proportion (%)
LD28 clade
P19250A
'14 Oct.
'15 Jan.
'15 Sept.
'15 Nov.
'16 Feb.
'16 May

Figure 3-13. Binning of virome reads from Lake Soyang to reference viral genomes, including the phage P19250A genome. (a) Proportion of virome reads assigned to viral groups according to the habitat they were isolated from. The top 30 most highly-assigned viruses for each virome were grouped into three categories: Freshwater; Marine or estuary; and Soil, sewage, or hot spring), and are indicated by different colors. P19250A was not grouped with other viruses and is marked separately. All other viruses were grouped together and marked as "Others." Supporting Information Table S3 shows detailed information on the highly-assigned viruses and their proportion. Note that the proportion was calculated using the number of reads assigned to all viruses as the denominator. (b) Seasonal change in the proportion of P19250A-assigned virome reads and the relative abundance of the LD28 clade. The proportions of P19250A-assigned reads are the same as those presented in (a). The relative abundance of the LD28 clade was calculated based on the taxonomic classification of 16S rRNA gene amplicon sequences obtained from the water samples.

Lake Soyang and contribution of P19250A among the virus-assigned reads ranged from 0.06% to 6.26% (Table 3-10). P19250A-assigned reads showed the highest proportion in the virome of Lough Neagh (Skvortsov *et al*., 2016), at 6.26%, followed by the virome of Lake Matoaka (open) (Green *et al*., 2015), at 5.34% (Table 3-11). P19250A ranked the first among freshwater phages for both samples, while most other highly-assigned viruses were of marine origin, as was observed in Lake Soyang viromes.

Then, contigs that were assembled from viromes that could show synteny and similarity to the P19250A genome were searched. When contigs assembled from several viromes were compared to the P19250A genome and all viral genomes in RefSeq (release 72) using tBLASTx, 20 contigs were found to be most similar to the P19250A genome: 17 contigs from Lake Soyang and 3 from Lough Neagh. These contigs showed highly-conserved synteny to the P19250A genome and in particular, two contigs from Lake Soyang had similarity to the entire P19250A genome (Fig 3-6 and Fig. 3-14). Finding these highly syntenic contigs showed the existence of a phage type that shares genomic content with P19250A.

Compared to P19250A, the phages P26218, P26059A, and P26059B occupied less portion of the viral population in Lake Soyang (Table 3-9). When their distribution and abundance were observed in Lough Neagh and Lake Michigan, the phage P26218 showed high abundance in Lough Neagh and some of Lake Michigan samples, but P26059A and P26059B showed low abundance in all samples analyzed (Table 3-12). Albeit at low appearances, P26059A and P26218 showed slight seasonality. Within Lake Soyang virome, P26059A appeared to be more abundant in summer season (’15 Sept) while P26218 showed higher abundance in winter seasons, along with P19250A.

Table 3-11. Percentage of binned reads matching to the phage P19250A genome in viromes obtained from various freshwater lakes and reservoirs

Country	Sampling site	Accession number	Sample name	Sampling date	% of binned reads[a]	% of viruses[b]	% of P19250A[c]
UK	Lough Neagh	SRR2147000	SRR2147000	Apr. 2014	33.30	14.45	6.26
US	Lake Matoaka	4523576.3	Matoaka Open	Sept. 2013	49.11	21.62	5.34
		4523574.3	Crim Dell Mouth	Sept. 2013	47.10	23.92	3.26
		4523575.3	Pogonia Mouth	Oct. 2013	46.59	24.02	3.62
US	Lake Michigan[d]	SRR1974494	Wilmette	May 2014	76.68	2.91	2.21
		SRR1974488	Wilmette	June 2014	89.48	0.40	1.13
		SRR1974511	Wilmette	May 2014	79.56	2.15	1.23
		SRR1974497	Montrose	May 2014	78.39	2.91	2.21
		SRR1974501	Montrose	June 2014	83.16	1.85	1.57
		SRR1974512	Montrose	Aug. 2014	82.79	1.51	1.69
		SRR1974491	57th St.	May 2014	74.95	2.27	1.61
		SRR1974503	57th St.	July 2014	74.58	3.52	1.40
		SRR1974513	57th St.	Aug. 2014	83.25	1.62	1.32
Canada	Lake Erie - Long Beach C.E.	SRR2082964	SRR2082964	Aug. 2012	63.62	2.60	0.68
	Lake Erie - Nickel Beach	SRR2083213	SRR2083213	Aug. 2012	29.10	27.78	0.79
	Lake Erie - Long Beach	SRR2083214	SRR2083214	Aug. 2012	32.08	39.92	0.45
	Lake Erie - Long Beach C.E.	SRR2083218	SRR2083218	Aug. 2012	17.46	30.72	0.63
	Lake Erie - Nickel Beach_VD	SRR2083219	SRR2083219	Aug. 2012	17.16	31.48	0.52
	Lake Erie - Long Beach 4	SRR2083220	SRR2083220	Aug. 2012	35.40	6.30	1.34
	Lake Erie - Long Beach 3	SRR2083221	SRR2083221	Aug. 2012	28.87	21.53	0.30

Table 3-11. (continued)

Country	Sampling site	Accession number	Sample name	Sampling date	% of binned reads[a]	% of viruses[b]	% of P19250A[c]
	Lake Erie - Long Beach 2	SRR2083222	SRR2083222	Aug. 2012	32.54	28.18	0.31
	Lake Erie - Long Beach 1	SRR2083223	SRR2083223	Aug. 2012	16.94	35.63	0.64
	Lake Ontario - Queen's Royal	SRR2083215	SRR2083215	Aug. 2012	12.49	10.60	0.63
	Lake Ontario - Fifty Point	SRR2083216	SRR2083216	Aug. 2012	14.62	6.54	0.61
	Lake Ontario - Lakeside ED	SRR2083217	SRR2083217	Aug. 2012	25.44	19.79	0.18
	Lake Ontario - Fifty Point VD	SRR2083224	SRR2083224	Aug. 2012	30.64	14.16	0.17
	Lake Ontario - Queen's Royal VD	SRR2083224	SRR2083225	Aug. 2012	26.28	13.81	0.32
	Lake Ontario - Lakeside 4	SRR2083227	SRR2083227	Jul. 2013	38.96	4.39	0.92
	Lake Ontario - Lakeside 3	SRR2083230	SRR2083230	Aug. 2012	21.18	25.73	0.50
	Lake Ontario - Lakeside 2	SRR2083231	SRR2083231	Aug. 2012	26.14	13.85	0.28
	Lake Ontario - Lakeside 2	SRR2083504	SRR2083504	Aug. 2012	24.95	16.44	0.26
	Lake Ontario - Lakeside 1	SRR2083509	SRR2083509	Aug. 2012	24.95	16.44	0.26
France	Lake Bourget	ERP000339	Lake Bourget	Jul. 2008	54.10	20.58	1.30
	Lake Pavin	ERP000339	Lake Pavin	Jun. 2008	48.54	21.56	1.97
Taiwan	Feitsui Reservoir	SRR648314	648314	Jan. 2009	49.73	11.75	0.14
		SRR648313	648313	Aug. 2008	39.46	20.38	0.26
		SRR648312	648312	Jul. 2008	54.54	8.31	0.26
		SRR618311	648311	Jan. 2008	45.18	32.38	0.35
		SRR371574	371574	Aug. 2007	34.07	24.04	0.25
		SRR371573	371573	Jul. 2007	34.45	14.86	0.34

[a](Number of reads assigned to bacteria or viruses) × 100/(Number of total reads)

[b](Number of reads assigned to viruses) × 100/(Number of reads assigned to bacteria or viruses)

[c](Number of reads assigned to P19250A) × 100/(Number of reads assigned to viruses)

[d] Competitive binning analysis of Lake Michigan was performed using RefSeq database release 79 while all the other samples were analyzed using RefSeq database release 72.

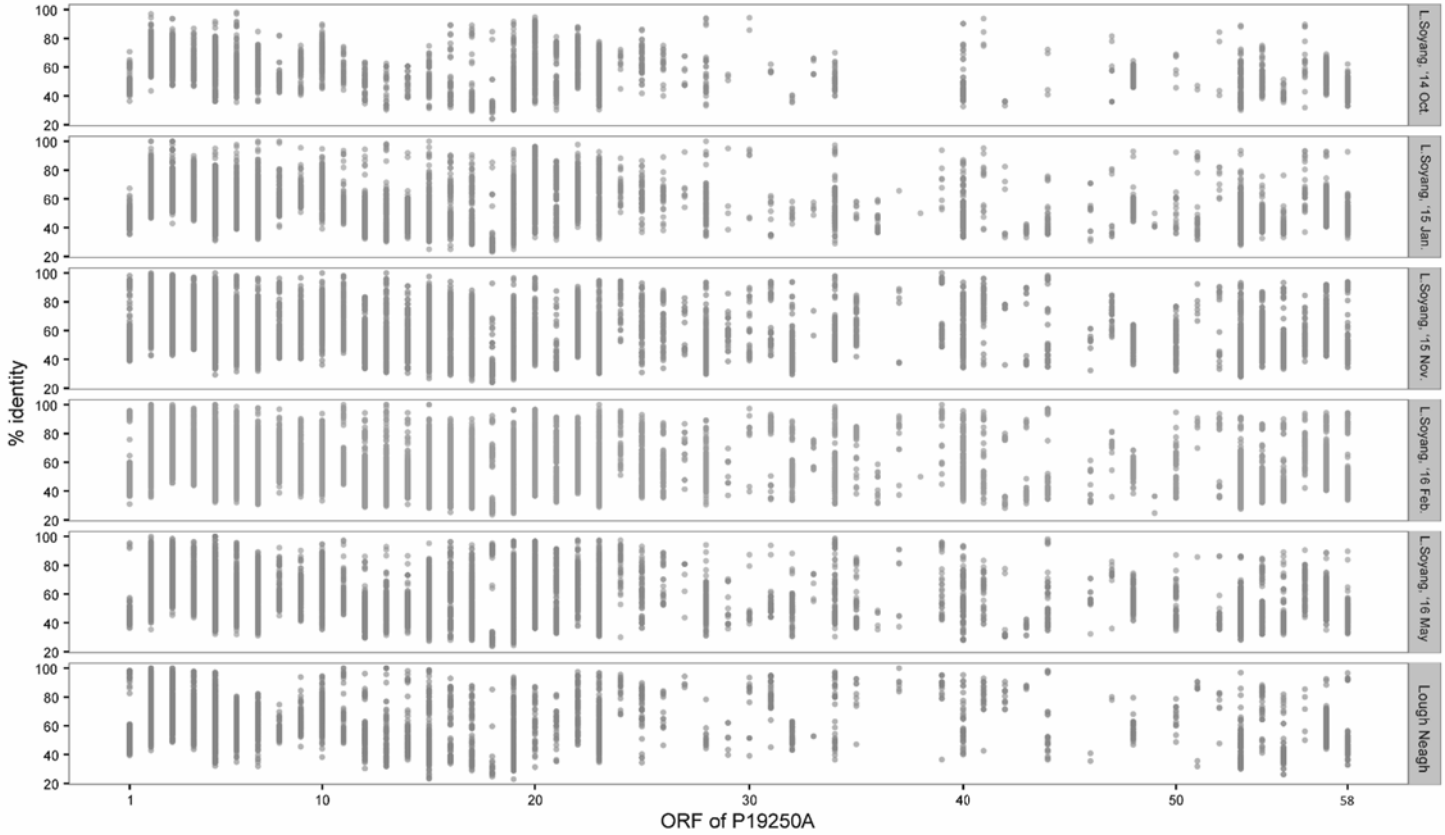

Figure 3-14. Fragment recruitment plot of the phage P19250A ORFs in six virome data; 5 of Lake Soyang virome and one Lough Neagh virome. Each dot represents a matching contig from the virome data set to the ORFs of the phage P19250A.

Table 3-12. Competitive binning results for the phage P26218, P26059A, and P26059B in Lough Neagh and Lake Michigan viromes

Virome samples	Virus / total binned	P26059A		P26059B		P26218	
	%[a]	%[b]	Rank	%	Rank	%	Rank
Lough Neagh	14.45	0.35	53	0.08	189	3.16	4
LM[c]-SRR1974494	2.91	2.20	6	0.38	51	0.11	160
LM-SRR1974488	0.40	0.97	17	0.31	65	0.30	68
LM-SRR1974511	2.15	1.01	16	0.58	35	0.05	281
LM-SRR1974497	1.95	0.77	21	1.42	9	1.55	7
LM-SRR1974501	1.85	0.64	30	0.44	46	1.21	13
LM-SRR1974512	1.51	0.79	23	0.47	50	0.17	134
LM-SRR1974491	2.27	0.44	52	0.32	75	0.09	190
LM-SRR1974513	1.62	1.27	8	0.35	68	0.18	126
LM-SRR1974503	3.52	0.54	34	0.27	74	0.11	162

[a] Proportion of virus-binned reads among all the binned reads

[b] Proportion of P26059A, P26059B, or P26218 among all the reads that were binned to viruses

[c] LM is an abbreviation for Lake Michigan

4. DISCUSSION

From Lake Soyang, an oligotrophic lake located in South Korea, 4 different bacteriophages were isolated using 3 bacterial strains that were also isolated from the identical site. One bacteriophage was isolated using a host strain, IMCC19250, belong to the freshwater Methylotroph, LD28 group, and was named as P19250A. Another phage, P26218, was isolated using a strain belonging to the genus *Rhodoferax,* and the phage was name after its host, IMCC26218. Lastly, two phages infecting an identical host, *Curvibacter* sp. IMCC26059, were independently isolated and each were named as P26059A and P26059B. Bacteriophage P19250A and P26059A had icosahedral shaped heads with long contractile tails, indicating that those phages belong to the family *Siphoviridae*, while P26218 and P26059B had short tails, which classified these two phages into the family *Podoviridae*. For further analysis on the phages that were isolated, whole genome sequencing for all bacteriophages were performed through Illumina MiSeq sequencing platform.

The complete genome of the phage P19250A was 38,562 bp in length with 58 protein coding genes. Through competitive binning analysis performed in this study, P19250A, the first phage of the LD28 clade, appeared to be abundant in diverse freshwater environments, and suggested that the P19250A genome can lead to more appropriate interpretation of previously unidentified virome sequences, as demonstrated in marine environments by studies on phages infecting the SAR 11 and SAR116 clades (Kang *et al*., 2013; Zhao *et al*., 2013).

Lytic bacteriophage P26218 is the first virus identified that infects the genus *Rhodoferax*. The complete genome of the phage P26218 was 36,315 bp in length with 44 protein coding genes. All predicted ORFs from this phage genome were protein-coding, with 3 specifically coding for DNA replication, 7 for DNA metabolism, and 5 for packaging and structural proteins. The group of ORFs with

similar function was postulated to originate from different groups of viral families (*Podoviridae*, *Siphoviridae*, and *Myoviridae*), which was indicative of the mosaic property of the P26218 genome.

Bacteriophages P26059A and P26059B were two independent phages sharing an identical host, *Curvibacter* sp. IMCC26059. The P26059A genome was 84,008 bp in length and it carried total of 124 protein coding genes. The complete genome of P26059B was 41,471 bp in length with 46 predicted coding genes. When competitive binning analysis was performed against virome data prepared from both Lake Soyang and Lake Michigan, they were both detected at low frequencies, yet, they were consistently detected in virome data that were analyzed. Especially P26059A, which appeared more frequently than P26059B, had seasonal preference in summer (Lake Soyang-'15 Sept.). Also, P26059A had a synteny contig within a virome '15 Sept. data that showed high similarity (99% identity), implying for the presence of another bacteriophage infecting *Curvibacter* sp. IMCC26059. Many of the aspects, including physical characteristics, genome features, and ecological abundance were different between P26059A and P26059B, although they shared the identical host. This implies that sole information on phages, either physical characteristics or genomes, is not sufficient for their classification and categorization, but both are in need to correctly understand bacteriophage characteristics.

Recently, diverse viral metagenome studies were performed in attempts to mine for the novel bacteriophage genomes that could control bacterial population in diverse environments including aquatic, sediments, and clinical samples and possibly carry AMGs that influence their hosts in various ways. However, due to lack of conserved marker genes among phages and precedingly identified phage genomes, interpretation of viral metagenome data is highly limited. Thus, along with numerous viral metagenome studies that became available to the public, phage isolation and identification must be accompanied for better interpretation and classification of phage sequences retrieved from immense virome data.

CHAPTER 4.

Distribution of Bacteriophage Population and Antibiotic Resistance Genes Carried by Bacteriophage Genomes in an Urban River

ABSTRACT

With development of next generation sequencing technologies and establishment of viral metagenome preparation methods, increased number of viral metagenome studies have been done in diverse environments. However, compared to studies performed at diverse environments, those performed in river systems are significantly lacking. Therefore, in this study, 6 sites from Han River, one of the major river system in South Korea, were selected for viral metagenome study to observe viral population distribution and their changes along the river flow. Throughout the river, the taxonomic composition of viral assemblages remained constant with minor shifts between different viral strains, indicating that viral population is stably maintained in a single water system regardless of the distances between the sampling sites. Furthermore, since the Han River flows through the Seoul metropolitan area and is highly influenced by anthropological activities, antibiotic resistance genes (ARGs) carried by bacteriophage contigs were further studied among bacterial metabolic genes that are encoded by viral reads and diverse ARGs were detected throughout samples. To verify that these ARGs are truly carried by bacteriophage genomes, viral metagenome reads were assembled into contigs, then ARGs were searched within the contigs that were predicted to be viral origin. As a result, total of 19 contigs were found to be carrying ARGs and among them, 7 contigs were found to be carrying beta-lactamase genes. When beta-lactamase genes were further analyzed, all of them were found to have active sites, implying for functional ARGs that are carried by bacteriophage genomes. Viral metagenome study done in an urban river body revealed that bacteriophage community is relatively well maintained throughout the river flow. Also, environmental phages appeared to be functioning as reservoirs of bacterial protein genes, especially ARGs, calling for the need of interest in bacteriophage-carried ARGs.

1. INTRODUCTION

Viruses are the most abundant biological entities on The Earth, with the number of virus-like particles (VLPs) being estimated to be approximately 4.80×10^{31} (Cobián Güemes *et al*., 2016). The bacteriophage (phages) that infects bacterial cells represent the largest proportion of VLPs, and are present as 10-times more abundant than their hosts, bacteria (Ignacio-Espinoza *et al*., 2013). Recently, with development and standardization of highly efficient viral metagenome (virome) preparation methods (John *et al*., 2011; Thurber *et al*., 2009), high diversity and distribution of bacteriophages are being re-illuminated through large scale ocean virome studies. In 2009 to 2011, *Tara* Ocean expedition was set out to collect for marine biological samples including those for viral metagenome (Karsenti *et al*., 2011). Also during the same period, Pacific Ocean Virome (POV) expedition was also set out (Hurwitz and Sullivan, 2013) to examine microbial and viral community changes across the oceans. The global-wide virome expeditions provided deeper understanding of environmental viral community structures with large amount of predicted viral sequences.

Along with taxonomic annotations of metagenome reads for study of viral population distribution in environments, the virome analysis also provided information on the distribution and ecological roles of predicted viral functional genes. Bacteriophage genomes were previously known to be carrying metabolic genes that are indirectly related to phage reproduction by adjusting host metabolism, which are called auxiliary metabolic genes (AMG). The AMGs within bacteriophage particles are often subjected for horizontal gene transfer (HGT) from a bacterial cell to another, enriching bacterial genetic diversity through phage infection. Also, AMGs can be expressed within the host bacterial cells upon viral infection, participating in host metabolism. The most well-known AMGs within bacteriophage

genomes are those related to photosynthesis, *psbA*, *psbD*, *psaA*, PTOX, *petE*, *petF*, *hli*, and *et cetera*. (Hevroni *et al*., 2015; Ledermann *et al*., 2016; Mann *et al*., 2003; Millard *et al*., 2004; Sharon *et al*., 2009; Sullivan *et al*., 2006). Photosystem related AMGs are mainly found in phages that infect cyanobacteria and upon infection, these genes are expressed to assist the host cell photosystem to enhance cell metabolism, which will lead to improvement of replication efficiency of phage nucleic acids. Likewise, bacteriophages are known to carry diverse supplementary metabolic genes such as those involved in carbon, phosphate, nitrogen, and sulfur metabolisms (Breitbart, 2012; Hurwitz and U'Ren, 2016).

Other than AMGs that assist host cell metabolism, bacteriophages also carry accessory genes that act as a defensive system for their hosts. Stress response gene, *mazG*, that regulates programmed cell death under starvation stress, has been found within a cyanophage (Bryan *et al*., 2008). Also, some bacteriophages were revealed to be carrying antibiotic resistance genes (ARGs) that could defend their host bacteria from antibiotic attacks during infection (Lekunberri *et al*., 2017; Mazaheri Nezhad Fard *et al*., 2011; Modi *et al*., 2013). The ARGs found in bacteriophage genomes are considered more significant due to their high potential of HGT to different bacterial cells (Brown-Jaque *et al*., 2015) and safe carriage by phage capsids, which are less sensitive to environmental changes compared to bacterial cell membrane. Also, bacteriophage with ARGs were found not only in clinical environments such as animal system or fecal samples (Colomer-Lluch *et al*., 2011a), but also found in river waters (Colomer-Lluch *et al*., 2011b), implicating wide spread of ARGs in diverse systems by bacteriophages. As one of the major input of ARGs into the natural environment, the WWTP effluents were previously reported to be containing high copy number of ARGs, providing input of those genes into natural water systems (Colomer-Lluch *et al*., 2011b). The Han River system encompasses four urban WWTP effluent discharging sites, which are suspected to

be causing anthropologic influence on the river system. Therefore, as Han River flows from pristine upstream to downstream, ARGs were expected to increase in number, especially those encapsulated in phage capsids to be safely carried and conserved.

Unlike oceans, freshwater systems have highly diverse and independent characteristics per their locations. Also, as freshwater systems are located inland, they are more accessible to people and at the same time, influenced by them. Lotic freshwater systems have varying microbial community at different locations depending on environmental parameters such as water velocity, however key players of the microbial community structure are known to be stably maintained in freshwater environments (Staley *et al*., 2013). Compared to bacteria, bacteriophage population within running water bodies has been understudied (Cai *et al*., 2016; Rastrojo and Alcamí, 2016). The Han River is one of the most important river system located in South Korea that flows through the Seoul city, the capital of Korea. The Han River runs across the South Korea, experiencing numerous changes of surroundings. The river system encompasses 5 lakes that are well conserved to be used as water reservoirs. As the river flows toward the Yellow Sea, it flows through the Seoul metropolitan area, receiving wastewater treatment plant (WWTP) effluents. Therefore, as the river flows towards the downstream, viral community is expected to be changing.

In this research, 6 sites were selected from Han River, from upstream to downstream for survey of viral community distribution along the running water system. Along with taxonomic observation, protein coding genes encoded by viral metagenome reads were studied for estimation of their ecological roles. Then, considering that the selected river flows through densely populated cosmopolitan area with WWTP effluent discharging sites, ARGs carried by putative bacteriophage contigs obtained from virome reads were searched and analyzed.

2. MATERIALS AND METHODS

2.1. Sampling of surface water of Han River

On May 26th and 27th, 2016, approximately 20 L of surface water samples were collected from 6 selected sites on Han River system, located in South Korea (Fig. 4-1). All the samples were collected from the center of the river width to avoid possible bias that could be caused from the river banks. From upstream to downstream, the samples were named as H1 to H6. Sample from the site H1 was collected from the Guman bridge in Hwacheon county of Gangwon province. The Guman bridge is located on the North Han River, which is connecting two lakes, Lake Paro and Lake Hwacheon, the most northern freshwater lakes of South Korea. The site H2 is located in Gapyeong county of Gyeonggi province, surrounded by water recreational sites. Surface water sample of H3 site was collected from the Paldang bridge, which is located downstream of the Paldang dam, located in Hanam city of Gyeonggi province. The Han River continues to flow through the Seoul city and the sites H4 and H5 were selected within the Seoul city. The surface samples for sites H4 and H5 of Han River were collected from Hannam bridge and Haengju bridge, respectively. The two bridges are about 21 km apart from each other and both are located in the downstream of the urban WWTPs' effluent sites. The site H6 is located at the most downstream of the Han River and just outside the boundary of Seoul city. The surface water sample of H6 was collected from middle of the Ilsan bridge that connects Gimpo city and Ilsan city of Gyeonggi province. The water sample of H6 site was considered as brackish water with slight salinity of 0.15 PSU. Collected water samples were brought to the lab at 4˚C. The environmental parameters for each site were retrieved from the Water Information System operated by the Ministry of Environment of South Korea (http://water.nier.go.kr) (Table 4-1).

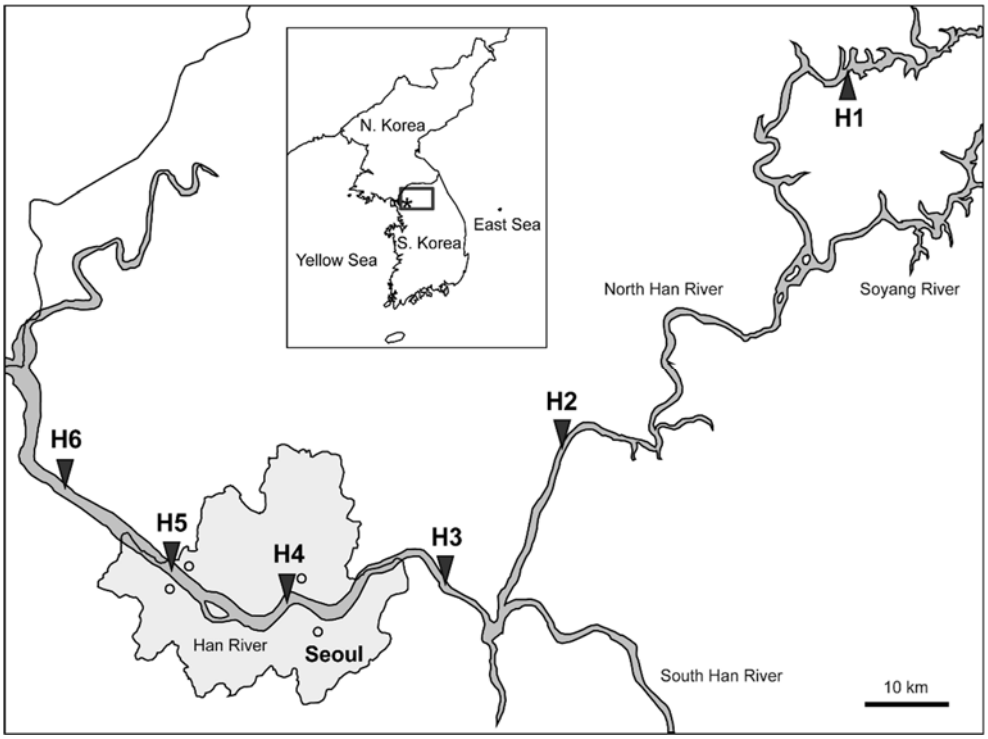

Figure 4-1. A map displaying sampling sites across the Han River body. The triangles indicate the sampling sites and black circles are representing waste water treatment plants located in Seoul. All sampling was performed in the middle of the river width.

Table 4-1. Environmental parameters of each sampling sites of the Han River.

Site	Longitude	Latitude	Sampling date[a]	Analysis date[b]	Temp. (°C)	pH	DO (mg/L)	BOD (mg/L)	COD (mg/L)	TN (mg/L)	TP (mg/L)	TOC (mg/L)
H1	38.093036	127.757051	16.05.27.	16.05.24.	14.20	8.0	12.40	0.50	2.00	1.30	0.02	1.80
H2	37.679676	127.381837	16.05.27.	16.05.16.	16.50	6.9	9.90	1.40	3.30	1.91	0.01	1.70
H3	37.545975	127.237466	16.05.26.	16.05.26.	20.00	7.9	9.60	1.10	2.90	1.94	0.04	2.90
H4	37.527282	127.012903	16.05.26	16.05.18.	20.80	7.8	9.40	1.10	4.10	2.49	0.04	2.10
H5	37.595556	126.817072	16.05.26	16.05.26.	21.00	7.3	6.40	2.80	5.90	5.31	0.12	4.90
H6	37.651395	126.717054	16.05.26	16.05.18.	23.80	7.6	9.90	3.90	6.90	4.48	0.35	3.10

[a] The date that the viral metagenome samples were collected

[b] The date that environmental parameters were measured

2.2. Sequencing of viral metagenome of Han River

Immediately after transporting to the lab, the collected water samples were filtered through a GF/A glass microfiber filter (Whatman, Maidstone, UK) to remove large sized particles and debris. Approximately 10 L of water samples were then filtered through a 0.2-μm Supor® PES Membrane filter (Pall Corporation, New York, USA) to remove particles larger than 0.2-μm in diameter, which included most of prokaryotic cells. As described in chapter 2, the filtered samples were treated with 0.01 g of $FeCl_3 \cdot 6H_2O$ per 10 L of sample and vigorously shaken to flocculate viral particles. After incubation of 1 hr in room temperature, the flocculated samples were filtered through a 0.8-μm Isopore polycarbonate filter (Merck Millipore, Darmstadt, Germany) to collect and concentrate viral particles (John *et al*., 2011). The polycarbonate filters with concentrate of 10 L portion of lake waters were treated with 0.1 M EDTA-0.2 M $MgCl_2$-0.2 M ascorbate buffer for an overnight on a rocking incubator in 4°C to dissolve flocculated viral particles and chelate iron molecules. DNase I and RNase A (Sigma-Aldrich, St. Louis, MO, USA) with final concentrations of 10 U/ml and 1 U/ml, respectively, were added to the dissolved solution for removal of external nucleic acids and the mixtures were incubated in 30°C for an hour. After incubation, the enzymes were deactivated by addition of 100 mM of EDTA and EGTA (Hurwitz *et al*., 2013). The viral particles were further concentrated and purified using CsCl step-gradient ultracentrifugation (Thurber *et al*., 2009). After 4 hrs of centrifugation at 24,000 rpm at 4˚C (Beckman Coulter L-90K ultracentrifuge), viral particles with densities between 1.5 and 1.35 g/cm^2 were collected with a syringe, which were specifically targeted for double stranded DNA phages. To remove any remaining CsCl in the sample, buffer exchange with SM buffer was performed using 50K Amicon centrifugal device (Merck Millipore). For final sterilization of the samples, 0.2-μm pore size Acrodisc® Syringe filter (Pall Corporation) was used to filter the collected samples and only the viral particles were

remaining in the sample. The viral DNA was extracted from the prepared samples using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). DNA samples were used for TruSeq library construction for Illumina MiSeq sequencing, which was performed at the ChunLab Inc. The overall process of the viral metagenome sample preparation is shown in Figure 4-2. After viral metagenome sequences were obtained, bacterial 16S SSU rRNA sequences were searched using MeTaxa program (Bengtsson-Palme *et al*., 2015) and confirmed that metagenome samples were free of bacterial contamination.

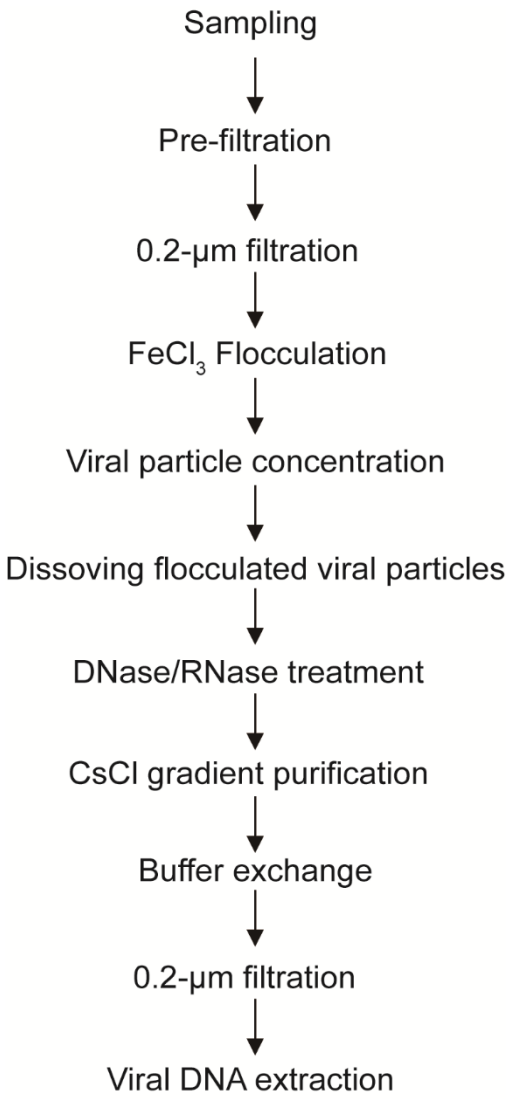

Figure 4-2. Flow chart of viral metagenome sample processing steps.

2.3. Quality trimming of sequencing data, assembly of virome reads, and analysis of similarity between viromes

Before analysis, the sequencing reads were mapped to the phiX174 genome to remove putative contamination by sequencing control reads, using CLC Genomics Workbench (Qiagen). Then, the sequences were trimmed using Trimmomatic based on quality score and length (Bolger *et al*., 2014). Using the trimmed reads from each virome, contigs were assembled using SPAdes version 3.8.2 (Bankevich *et al*., 2012) with metaspades.py assembler option. The assembled contigs were further screened and those shorter than 10 kb were removed and those longer than 10 kb were used for further analysis.

All the virome contigs obtained from SPAdes assembler were determined if they are viral origin or not, using VirSorter algorithm (Roux *et al*., 2015) that is available in iPlant Discovery Environment (http://de.cyverse.org/de/). The VirSorter algorithm detects probable virus and prophage genomes based on BLAST search against viral genome specific database. Only 0.39% to 0.93% of the viral metagenome contigs were identified as either viral or prophage origin.

The dissimilarity/distance between virome samples were analyzed using MASH algorithm (Ondov *et al*., 2016), which estimates distance between genome/metagenome reads based on Jaccard index calculated using reduced representation of k-mer profile of sequence data (sketch). Based on the dissimilarity/distances matrix calculated by MASH, Non-metric multidimensional scaling (NMDS) plot and principal coordinate analysis (PCoA) plot were constructed to ordinate virome samples using the Vegan package and hclust in R (Oksanen *et al*., 2007).

2.4. Phylogenetic and functional annotation of virome reads using metagenome analysis pipelines

The virome samples were analyzed using two online pipelines, MG-RAST and IMG/M ER server. Raw virome reads were uploaded unto the MG-RAST server for their taxonomic and functional gene annotation (http://metagenomics.anl.gov) (Glass *et al*., 2010). The raw sequences were uploaded because the MG-RAST operated its own quality control and pre-processing pipelines. The MG-RAST provided taxonomic and functional annotations on each shotgun sequencing reads based on multiple databases, maximizing the number of annotated genes. The assembled contigs were uploaded unto the IMG/M ER webserver (Markowitz *et al*., 2012) for their taxonomic and functional annotations. The IMG/M ER server does not provide quality-control service and assembling algorithm, thereby only the contigs that were quality controlled and assembled were uploaded. Also, the IMG/M ER server provided functional annotation for each protein coding genes within assembled contigs, thereby allowing prediction of functional genes present in putative bacteriophage genomes. The assembled viral metagenome data are available on IMG/M ER webserver (Gp0175588, Gp0175592, Gp0175603, Gp0175595, Gp0175596, and Gp0175601).

2.5. Antibiotic resistance gene search and sequence analysis

Among the AMGs carried by bacteriophages, ARGs were specifically screened from Han River virome. Assembled virome contigs that were classified as virus or prophages were used as query to screen for AMGs that are carried by putative phage genomes. The query sequences were analyzed by BLAST against two ARG-specific databases. The Comprehensive Antibiotic Resistance gene Database (CARD, downloaded on Feb. 2016) (Jia *et al*., 2017) was used to search for ARGs in general using local BLAST (Lavigne *et al*., 2008). For more reliable detection of

ARGs, only the BLAST results with e-values lower than 0.001, bitscores larger than 40, and percent identity higher than 80% were accepted. After finding that most of the ARGs that were detected were related to beta-lactamase genes, more specific database, Resfams AR (Antibiotic Resistance, downloaded on Aug. 2016) database (Gibson *et al*., 2015), were used. To screen for ARGs against the Resfams database, hmmscan was used (Söding, 2005). From the results, only those with e-values lower than 0.001 and scores higher than 40 were accepted. Putative ARG sequences obtained from viral metagenome contigs were aligned with representative ARG sequences obtained from the NCBI nr database using ClustalW embedded in MEGA 6 (Tamura *et al*., 2013). Then, the alignment of sequences were visualized using Jalview version 2 (Waterhouse *et al*., 2009).

3. RESULTS

3.1. Analysis on viral metagenome reads obtained from Han River

On May 2016, surface water samples from 6 selected sites on Han River were collected for viral metagenome analysis. After flocculation and concentration of viral particles using $FeCl_3$ from 10 L of collected samples, sequencing was performed using Illumina MiSeq platform. Thus, 4.3 million to 8.3 million reads were obtained from each sample and after removing low-quality reads, 3.6 million to 6.6 million reads were retrieved as a result (Table 4-2). The biodiversity indices indicated high diversity of virome samples (Table 4-3). Using quality controlled virome reads, dissimilarity/distance between the viromes was calculated based on k-mer profiles and distance matrix was constructed. To observe relationship between virome samples, a dendrogram was constructed based on the distance matrix. Although the samples had low dissimilarity (0.03-0.08), 6 virome samples were able to be grouped into 2 groups. One group consisted of samples collected from H1, H2, and H5 sites while the other consisted of those collected from H3, H4, and H6 sites (Fig. 4-3). Because these groupings were rather inconsistent with locations along river flow or distances between sampling sites, PCoA and NMDS plots were constructed using the same distance matrix, and the impact of environmental parameters were analyzed by fitting the parameters onto the plot ('envfit' function of R). However, the *p*-values for correlation between environmental factors and ordination coordinates of virome samples were too high (p=0.45-0.98 for PCoA and p=0.59-0.99 for NMDS), indicating that influence of environmental parameters on similarity/dissimilarity of viral metagenomes were insignificant (Fig. 4-4 and Fig. 4-5). Prior to taxonomic and functional analysis of viral metagenome data, contamination by bacterioplankton cells were determined using MeTaxa program. Within metagenome reads, only 0.00007% to 0.00025% of the reads appeared to

Table 4-2. Viral metagenome data statistics after each quality control step

Site	Raw sequence	ΦX 174 adaptor sequence removed		Quality trimmed		Assembled contigs	Assembled contigs (> 10kb)
	No.	No	% surviving	No	% surviving	No.	No.
H1	5,823,653	5,703,934	98	4,739,263	83	93,140	610
H2	5,132,489	5,042,179	98	4,253,428	84	113,682	976
H3	4,366,147	4,314,519	99	3,649,892	85	275,322	985
H4	7,159,288	7,067,388	99	6,006,517	85	280,209	1,420
H5	8,338,536	8,186,499	98	6,630,128	81	248,253	787
H6	5,820,537	5,753,723	99	4,818,906	84	328,339	760

Table 4-3. Shannon-Wiener and Simpson's index of viral metagenomes prepared from the Han River

Sample	H'	^2D
H1	4.5656	0.9774
H2	4.6923	0.9819
H3	4.9026	0.9856
H4	4.7096	0.9811
H5	4.8502	0.9844
H6	4.8869	0.9863

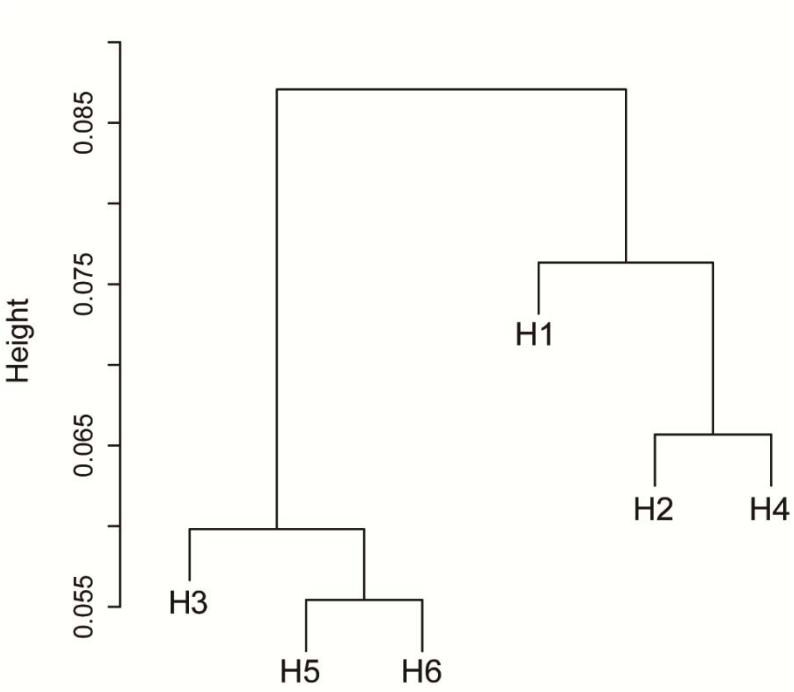

Figure 4-3. Dendrogram showing the clustering pattern of the Han River viral metagenomes. The distance matrix between viromes was calculated using the MASH algorithm, and then used for clustering based on UPGMA.

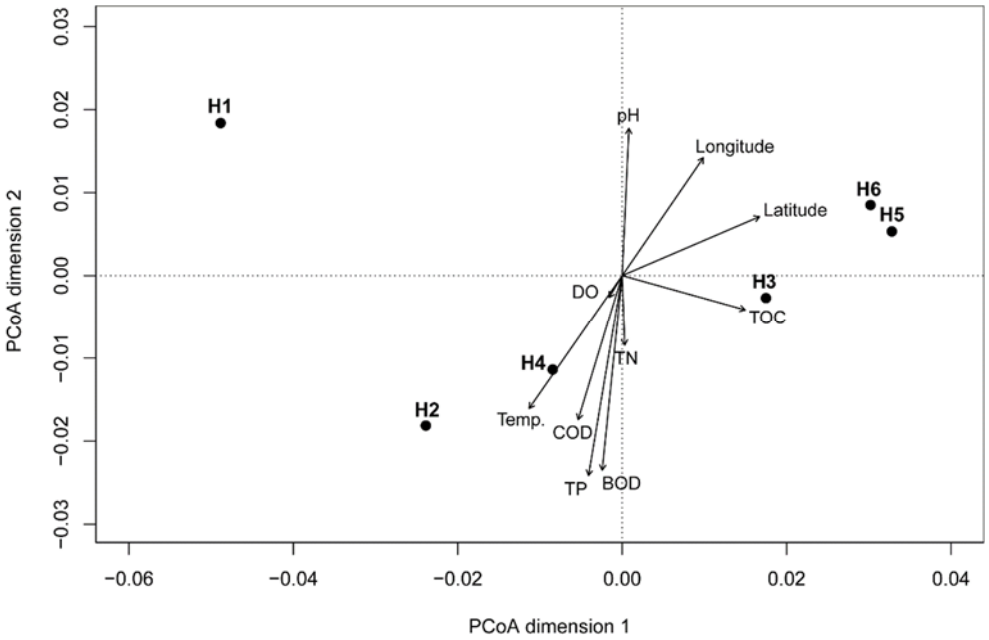

Figure 4-4. Principal coordinate analysis (PCoA) plot of six virome samples obtained from the Han River body. The distance was calculated based on raw virome reads using MASH algorithm. Environmental vectors were added to the PCoA plot and they are depicted in arrows. However, no significant correlation between environmental vectors and virome samples were observed (p=0.4474-0.9806).

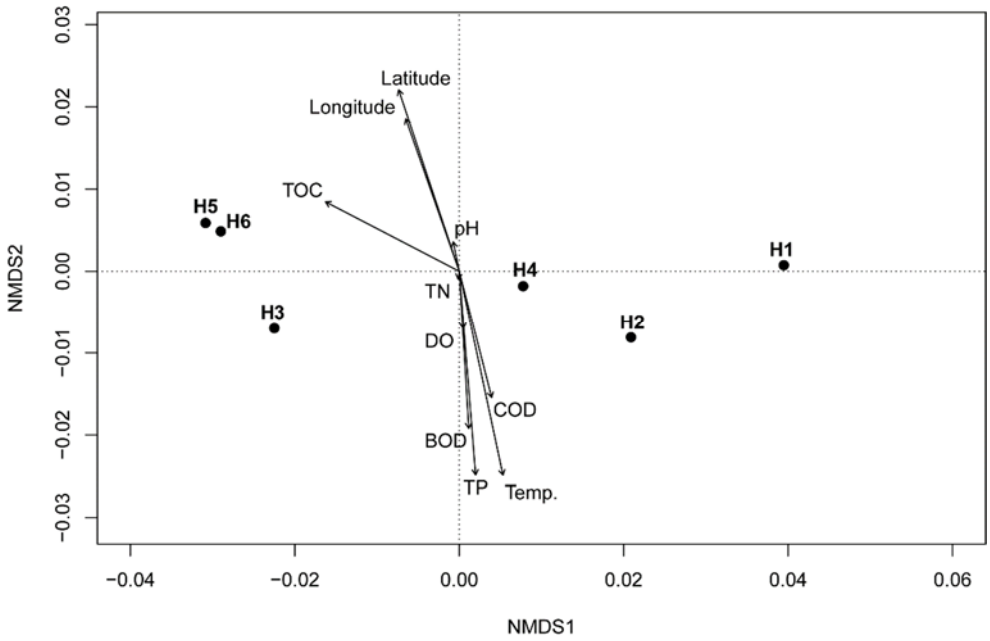

Figure 4-5. Non-metric multidimensional scaling (NMDS) plot of six virome samples obtained from the Han River body. The distance was calculated based on raw virome reads using MASH algorithm. Environmental vectors were added to the NMDS plot and they are depicted in arrows. No significant correlation between environmental vectors and virome samples were observed (p=0.6500-0.9986).

have similarity with bacterial 16S rRNA sequences, indicating that no bacterial contamination was present in Han River virome samples (Table 4-4). After quality control of metagenome data, the virome reads were uploaded onto the MG-RAST webserver for taxonomic and functional analysis. Prior to the analysis, the MG-RAST calculated average length of the submitted virome reads and it was shown that virome read lengths were evenly distributed (Table 4-5). Therefore, it was assumed that no bias was caused by sequence length and further analysis was performed.

The quality controlled viral metagenome reads were assembled using SPAdes program with metagenome option and 93,140 to 328,299 contigs were constructed from each sample (Table 4-2). Among those, only the contigs with 10 kb or longer in length were selected for further analysis. Among the assembled contigs, those that are predicted to be viral or prophage origin were screened using VirSorter which identifies viral or prophage ORFs within given sequences. Per VirSorter, less than 1% of the assembled contigs were found to be viral or prophage origin (Table 4-6).

Table 4-4. Percent of 16S rRNA bacterial SSU sequences in the Han River viral metagenome data

Site	Total bp in virome	Total 16S rRNA bp in virome	% of 16S rRNA seq. in virome
H1	1,456,458,749	3,664	0.00025%
H2	1,288,843,440	926	0.00007%
H3	1,902,716,563	5,668	0.00030%
H4	3,147,663,020	9,480	0.00030%
H5	2,000,621,402	1,881	0.00009%
H6	2,509,404,691	3,290	0.00013%

Table 4-5. Average read lengths of viral metagenome collected from the Han River

Site	**Mean seq. length (before QC)**	**Mean seq. length (after QC)**
H1	298 ± 17	210 ± 75
H2	298 ± 16	211 ± 76
H3	297 ± 19	219 ± 72
H4	298 ± 17	222 ± 71
H5	298 ± 18	203 ± 76
H6	298 ± 18	218 ± 72

Table 4-6. Number of viral metagenome contigs that were identified as virus or prophage r

Site	Viruses		Prophages	
	Contigs	**> 10kb**	**Contigs**	**> 10kb**
H1	616	158	110	73
H2	961	220	101	49
H3	954	105	118	75
H4	1,388	291	155	97
H5	1,316	142	134	82
H6	1,174	145	123	68

3.2. Taxonomic and functional annotation of Han River virome reads

3.2.1. Viral taxonomic distribution in Han River

The raw viral metagenome reads were first mapped to phiX174 genome and mapped reads were removed to avoid bias caused by sequencing control reads. Then the virome reads that were unmapped to phiX174 genome were recollected and they were uploaded unto the MG-RAST webserver. The MG-RAST annotation pipeline included removal of low quality reads, thereby no prior trimming process was necessary. After the metagenome sequences were uploaded, the MG-RAST pipeline annotated each read to a non-redundant M5nr database (Wilke *et al*., 2012). Although bacterial cells and any possible external nucleic acid was removed during the viral metagenome sample preparation, more than 70% of the annotated reads were classified as bacteria (Fig. 4-6a). However, this is a common phenomenon in viral metagenome analyses due to limitation in complete removal of bacterial cells during sample preparation, existence of prophage regions in many bacterial genomes, and dearth of viral gene database. Compared to number of bacterial genomes that have been sequenced, those of viral genomes are much less. Accordingly, most of the novel viral genes retrieved from viral metagenome samples are often annotated as bacterial or unknown not being able to be classified with a known query. Within the reads that were annotated as viruses, which comprised of approximately 13 to 24% of annotated reads in all samples, proportion of predicted viral family was analyzed. Proportion of viral reads that were assigned to the *Podoviridae* and *Siphoviridae* family increased as the river flowed to the downstream while that of unclassified viruses decreased (Fig. 4-6b). However, the viral families encompass broad range of viruses regardless of phage hosts that no definitive conclusion on viral population distribution was able to be made. When viral species that were annotated from viral metagenome reads were inspected, *Acanthocystis turfaceae Chlorella* virus 1, *Bordetella* phages, and *Prochlorococcus* phages appeared to be most abundantly

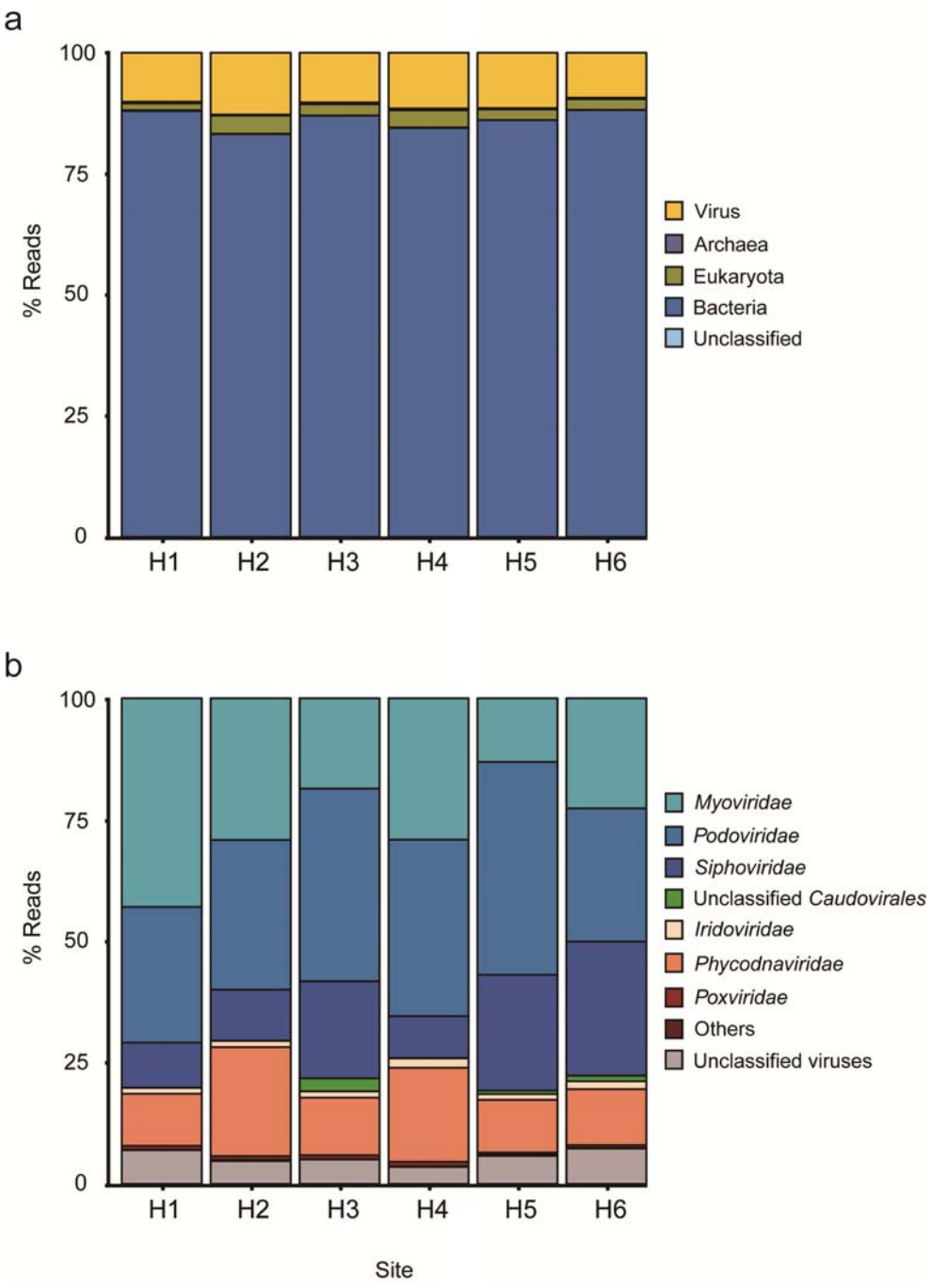

Figure 4-6. Taxonomic annotation of the Han River virome samples by metagenome analysis server. (a) Proportion of different domains that were annotated from Han River virome raw reads are depicted. (b) Proportion of different families of virus are shown.

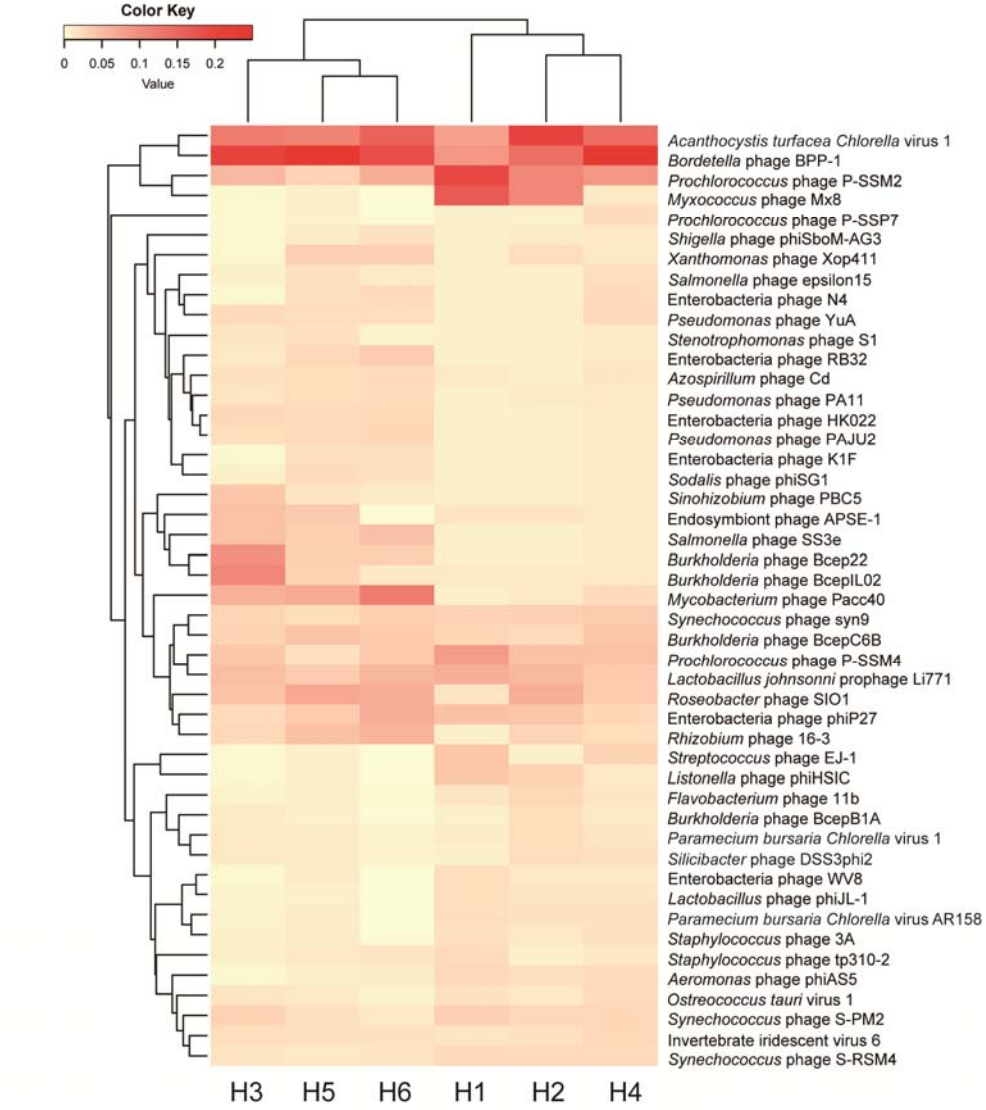

Figure 4-7. Heatmap showing the taxonomic composition of the viromes obtained from the Han River body. The taxonomic assignment of virome reads was performed by MG-RAST server and only viral species that had relative abundance of more than 1% in at least 1 sample were shown here.

assigned throughout all 6 virome samples while *Myxococococcus* phage Mx8 appeared with high abundance only in H1 and H2 samples, the most upstream samples (Fig. 4-7). The host of *Bordetella* phage BPP-1, *Bordetella bronchiseptica,* is an animal pathogen that could cause bronchitis and both the host and the bacteriophage are known to be able to persist well in natural environment (Coote, 2001). On the contrary, *Prochlorococcus* phage P-SSM2 is known to be present in marine environments, infecting *Prochlorococcus marinus* str. NATL1A. Origin of annotated viral species of all Han River virome samples were highly diverse (Table 4-7 to 12). This may indicate presence of diverse bacteriophages in Han River, but at the same time, it may indicate that some of viral metagenome reads have been inadequately assigned to viruses from different origins due to shortness in number of bacteriophages and viruses isolated from freshwater. Other than minor shifts among phage populations, no specific pattern was observed in viral population along the river flow. Thereby, it could be concluded that major components of the viral population remain consistent across the river.

3.2.2. Functional protein distribution in Han River

The MG-RAST server predicted protein coding genes of viral metagenome reads and functionally annotated them based on diverse protein databases such as the SEED subsystems, UniProt, COG (Conserved Ortholog Groups), NOG (Non-supervised Ortholog Groups), and KEGG. Based on the functional protein annotation from the SEED subsystem, which provides the most detailed annotation, functional protein distribution was observed among Han River virome samples. On the contrary to the taxonomic annotation, 40 to 66% of the annotated reads were identified and grouped into phages and prophage-related genes (Fig. 4-8, Table 4-13). Among the phage-related genes that were found, r1t-like Streptococcal phage proteins were most frequently annotated, followed by proteins related to phage packaging machinery and replication (Table 4-14). High frequency of phage-related

Table 4-7. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H1 sample

H1		
Species name	**%**	**Origin of isolation**
Prochlorococcus phage P-SSM2	18.13	Marine
Myxococcus phage Mx8	15.48	Soil
Bordetella phage BPP-1	7.76	Animal lung
Prochlorococcus phage P-SSM4	7.31	Marine
Acanthocystis turfacea Chlorella virus 1	6.36	Freshwater
Lactobacillus johnsonii prophage Lj771	5.16	Gut
Enterobacteria phage phiP27	3.66	Gut
Streptococcus phage EJ-1	3.16	-
Listonella phage phiHSIC	3.16	Marine
Synechococcus phage S-PM2	2.29	Marine
Synechococcus phage syn9	2.11	Marine
Burkholderia phage BcepC6B	1.86	Plant root
Aeromonas phage phiAS5	1.73	River
Staphylococcus phage tp310-2	1.49	Soil (animal farm)
Synechococcus phage S-RSM4	1.33	Marine
Enterobacteria phage WV8	1.25	River
Lactobacillus phage phiJL-1	1.23	Vegetable fermentation
Paramecium bursaria Chlorella virus AR158	1.10	Freshwater
Staphylococcus phage 3A	1.01	-
Mycobacterium phage Chah	0.86	-

Table 4-8. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H2 sample

H2		
Species name	**%**	**Origin of isolation**
Acanthocystis turfacea Chlorella virus 1	17.96	Freshwater
Bordetella phage BPP-1	11.86	Animal lung
Prochlorococcus phage P-SSM2	9.01	Marine
Myxococcus phage Mx8	8.82	Soil
Roseobacter phage SIO1	4.96	Marine
Lactobacillus johnsonii prophage Lj771	4.08	Gut
Prochlorococcus phage P-SSM4	3.47	Marine
Enterobacteria phage phiP27	2.91	Gut
Synechococcus phage syn9	2.30	Marine
Listonella phage phiHSIC	2.00	Marine
Rhizobium phage 16-3	1.85	Plant root
Synechococcus phage S-PM2	1.58	Marine
Flavobacterium phage 11b	1.45	Sea-ice
Burkholderia phage BcepC6B	1.35	Plant root
Synechococcus phage S-RSM4	1.31	Marine
Silicibacter phage DSS3phi2	1.09	Marine
Burkholderia phage BcepB1A	1.08	Plant root
Paramecium bursaria Chlorella virus 1	1.07	Freshwater
Xanthomonas phage Xop411	1.04	Plant
Aeromonas phage phiAS5	0.96	River

Table 4-9. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H3 sample

H3		
Species name	**%**	**Origin of isolation**
Bordetella phage BPP-1	14.23	Animal lung
Acanthocystis turfacea Chlorella virus 1	8.06	Freshwater
Burkholderia phage BcepIL02	7.23	Plant root
Burkholderia phage Bcep22	6.26	Plant root
Mycobacterium phage Pacc40	3.96	-
Prochlorococcus phage P-SSM2	3.40	Marine
Lactobacillus johnsonii prophage Lj771	3.03	Gut
Endosymbiont phage APSE-1	3.00	Aphid
Salmonella phage SS3e	2.86	Sewage
Roseobacter phage SIO1	2.71	Marine
Sinorhizobium phage PBC5	2.62	Plant root
Prochlorococcus phage P-SSM4	2.57	Marine
Burkholderia phage BcepC6B	1.98	Plant root
Synechococcus phage S-PM2	1.83	Marine
Synechococcus phage syn9	1.74	Marine
Enterobacteria phage HK022	1.55	Gut
Rhizobium phage 16-3	1.51	Plant root
Pseudomonas phage YuA	1.33	Freshwater
Enterobacteria phage phiP27	1.32	Gut
Invertebrate iridescent virus 6	1.23	*Drosophila*

Table 4-10. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H4 sample

H4		
Species name	**%**	**Origin of isolation**
Bordetella phage BPP-1	24.91	Animal lung
Acanthocystis turfacea Chlorella virus 1	14.49	Freshwater
Prochlorococcus phage P-SSM2	8.49	Marine
Prochlorococcus phage P-SSM4	3.82	Marine
Burkholderia phage BcepC6B	3.14	Plant root
Lactobacillus johnsonii prophage Lj771	3.00	Gut
Synechococcus phage syn9	2.69	Marine
Roseobacter phage SIO1	2.62	Marine
Streptococcus phage EJ-1	2.11	-
Synechococcus phage S-PM2	1.61	Marine
Enterobacteria phage phiP27	1.38	Gut
Invertebrate iridescent virus 6	1.38	*Drosophila*
Pseudomonas phage YuA	1.37	Freshwater
Synechococcus phage S-RSM4	1.34	Marine
Ostreococcus tauri virus 1	1.29	Marine
Enterobacteria phage N4	1.28	Gut
Mycobacterium phage Pacc40	1.24	-
Aeromonas phage phiAS5	1.24	River
Prochlorococcus phage P-SSP7	1.15	Marine
Salmonella phage epsilon15	1.00	-

Table 4-11. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H5 sample

H5		
Species name	**%**	**Origin of isolation**
Bordetella phage BPP-1	21.61	Animal lung
Acanthocystis turfacea Chlorella virus 1	9.52	Freshwater
Roseobacter phage SIO1	5.60	Marine
Mycobacterium phage Pacc40	5.51	-
Rhizobium phage 16-3	3.37	Plant root
Burkholderia phage BcepC6B	3.08	Plant root
Endosymbiont phage APSE-1	2.61	Aphid
Enterobacteria phage phiP27	2.55	Gut
Lactobacillus johnsonii prophage Lj771	2.55	Gut
Burkholderia phage Bcep22	2.37	Plant root
Xanthomonas phage Xop411	2.29	Plant
Salmonella phage SS3e	2.27	Sewage
Prochlorococcus phage P-SSM2	2.07	Marine
Burkholderia phage BcepIL02	2.04	Plant root
Enterobacteria phage RB32	1.61	Gut
Pseudomonas phage PAJU2	1.34	Sputum/River
Enterobacteria phage HK022	1.34	Gut
Sodalis phage phiSG1	1.26	Tsetse fly
Synechococcus phage syn9	1.22	Marine
Azospirillum phage Cd	1.22	Plant

Table 4-12. List of 20 viruses that were most frequently assigned within the viral metagenome reads of the H6 sample

H6		
Name	**%**	**Origin of isolation**
Bordetella phage BPP-1	11.24	Animal lung
Acanthocystis turfacea Chlorella virus 1	9.53	Freshwater
Mycobacterium phage Pacc40	7.47	-
Enterobacteria phage phiP27	4.02	Gut
Prochlorococcus phage P-SSM2	3.87	Marine
Roseobacter phage SIO1	3.86	Marine
Rhizobium phage 16-3	3.68	Plant root
Lactobacillus johnsonii prophage Lj771	3.39	Gut
Salmonella phage SS3e	3.05	Sewage
Burkholderia phage BcepC6B	2.42	Plant root
Prochlorococcus phage P-SSM4	2.30	Marine
Enterobacteria phage RB32	2.17	Gut
Synechococcus phage syn9	2.08	Marine
Xanthomonas phage Xop411	2.05	Plant
Burkholderia phage Bcep22	2.03	Plant root
Enterobacteria phage HK022	1.64	Gut
Pseudomonas phage PAJU2	1.59	Marine
Pseudomonas phage PA11	1.53	Marine
Enterobacteria phage N4	1.45	Gut
Azospirillum phage Cd	1.42	Plant

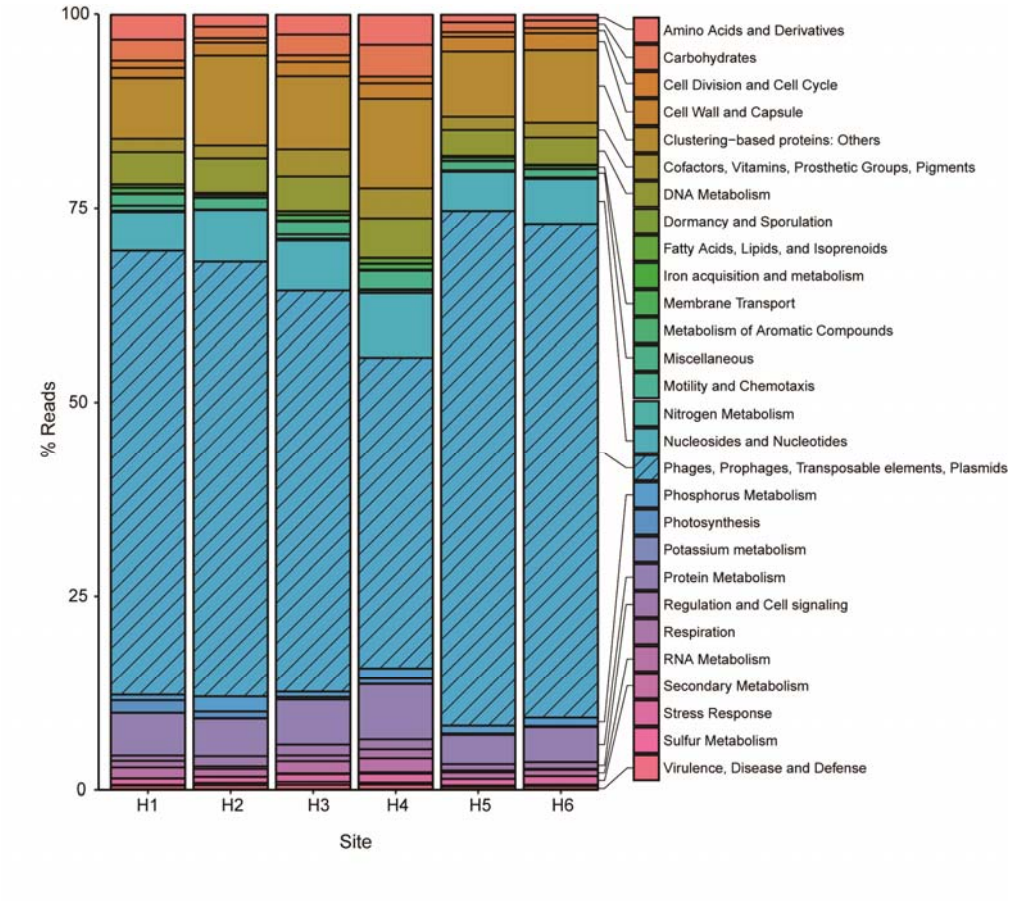

Figure 4-8. Functional gene annotation of the Han River virome samples. The functional gene annotation was performed by MG-RAST annotation server. The annotation was done based on the SEED subsystem database and annotation matches with e-value threshold of 1.00E-5 were taken into consideration for data analysis.

Table 4-13. Proportion of each Han River virome reads that were assigned to function annotation categories

Function category	H1 (%)	H2 (%)	H3 (%)	H4 (%)	H5 (%)	H6 (%)
Amino Acids and Derivatives	3.21	1.54	2.56	3.88	0.97	0.74
Carbohydrates	2.72	1.48	2.68	4.11	1.29	0.98
Cell Division and Cell Cycle	0.95	0.58	0.86	0.88	0.61	0.66
Cell Wall and Capsule	1.30	1.67	1.84	2.00	1.89	2.19
Clustering-based subsystems	7.78	11.53	9.39	11.49	8.42	9.31
Cofactors, Vitamins, Prosthetic Groups, Pigments	1.72	1.67	3.50	3.91	1.64	1.91
DNA Metabolism	4.16	4.47	4.50	5.14	3.36	3.51
Dormancy and Sporulation	0.03	0.03	0.03	0.04	0.02	0.02
Fatty Acids, Lipids, and Isoprenoids	0.41	0.17	0.44	0.69	0.15	0.07
Iron acquisition and metabolism	0.02	0.03	0.03	0.04	0.03	0.04
Membrane Transport	0.71	0.34	0.70	0.75	0.42	0.39
Metabolism of Aromatic Compounds	0.09	0.09	0.13	0.14	0.05	0.06
Miscellaneous	1.51	1.50	1.72	2.41	1.20	1.05
Motility and Chemotaxis	0.64	0.03	0.57	0.14	0.06	0.05
Nitrogen Metabolism	0.20	0.05	0.25	0.32	0.12	0.17
Nucleosides and Nucleotides	5.03	6.72	6.43	8.29	5.10	5.85
Phages, Prophages, Transposable elements, Plasmids	57.32	56.13	51.79	40.15	66.49	63.77
Phosphorus Metabolism	0.71	1.94	0.68	1.30	1.01	1.08
Photosynthesis	1.61	0.86	0.31	0.70	0.15	0.13
Potassium metabolism	0.02	0.04	0.02	0.04	0.03	0.01
Protein Metabolism	5.51	4.86	5.83	7.14	3.79	4.50
RNA Metabolism	1.39	0.99	1.57	1.80	0.84	0.80
Regulation and Cell signaling	0.67	1.32	1.39	1.28	0.81	0.86
Respiration	0.85	0.34	0.75	1.17	0.24	0.14
Secondary Metabolism	0.04	0.03	0.11	0.19	0.02	0.01
Stress Response	0.81	0.76	0.99	1.14	0.82	1.10
Sulfur Metabolism	0.08	0.30	0.37	0.26	0.16	0.19
Virulence, Disease and Defense	0.48	0.52	0.57	0.60	0.33	0.39

Table 4-14. Proportion of Han River virome reads that were assigned to viral functional proteins in public protein database.

Function annotation	**H1 (%)[a]**	**H2 (%)**	**H3 (%)**	**H4 (%)**	**H5 (%)**	**H6 (%)**
r1t-like Streptococcal phage protein	24.41	24.25	23.18	20.48	28.33	20.12
Phage packaging machinery	18.62	17.46	13.56	9.03	20.49	22.89
Phage replication	5.33	5.08	6.25	5.77	8.92	10.73
Gene Transfer Agent	3.12	1.59	0.77	0.48	0.86	1.18
Staphylococcal pathogenicity islands, SaPI	1.32	1.15	0.90	1.01	0.58	0.95
Phage tail fiber protein	1.55	1.97	0.52	0.40	0.79	0.66
Phage capsid protein	10.01	1.50	2.50	1.00	2.59	2.56
Phage entry and exit-related protein	0.84	1.21	2.06	0.66	1.19	0.95
Phage *nin* genes	0.46	0.22	0.08	0.12	0.23	0.23
Phage integration and excision proteins	0.25	1.18	1.17	0.85	1.20	1.67
Others	0.41	0.51	0.79	0.35	1.18	1.84
Total % reads assigned to viral functional proteins	57.32	56.13	51.79	40.15	66.49	63.77

[a] (Number of viral metagenome reads that were assigned to the functional protein) / (Total number of viral metagenome reads that were assigned to a functional protein) x 100

genes obtained by functional annotation of viral metagenome reads, while only 13 to 24% of the identical virome reads were taxonomically annotated to belong to viral species, may derive two inferences; most of the protein coding genes that were predicted to be coding for phage-related genes were carried by bacterial cells, indicative of high proportion of prophages, and that Han River viromes mostly consist of unknown bacteriophages that carry essential bacteriophage protein genes but were inappropriately assigned to bacterial species putatively due to sequence uniqueness.

Bacteriophages are known to be carrying AMGs that are not essential but helpful for phage reproduction by adjusting host metabolism upon infection. Various AMGs have been identified from bacteriophage genomes and they encode proteins involved in functions such as carbon utilization, ammonia assimilation, sulfur oxidation, nitrogen regulation, and photosynthesis. Within Han River virome samples, diverse AMGs have also been predicted from virome reads (Table 4-15). Although they were found in low frequencies, many AMGs found within the Han River virome consisted of enzymes participating for carbohydrate metabolism, which are known to assist host cell metabolism during phage infection. Also, AMGs related to photosynthesis and respiration were found as reported before. Along with metabolic genes, virome reads were also annotated as defensive proteins against toxic substances such as cobalt, zinc, and cadmium. Interestingly, diverse ARGs were also found within virome reads (Table 4-16). Considering that gene transfer agents, such as plasmids, integrons, and transposons were found at relatively high proportion among virome reads (Table 4-14), and diverse ARGs were also found, it may indicate that bacteriophage community in Han River are acting as couriers and transporters of diverse bacterial genes including ARG.

Table 4-15. List of AMG products that were commonly found in bacteriophage genomes and proportion of viral metagenome reads that were assigned to each AMG

Functional groups	Protein functions (Accession no.)	H1[a]	H2	H3	H4	H5	H6
Carbohydrates	6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)	0.02	0.01	-	-	-	-
Carbohydrates	Ribose 5-phosphate isomerase B (EC 5.3.1.6)	0.01	0.01	0.01	0.04	-	0.01
Carbohydrates	Transaldolase (EC 2.2.1.2)	0.03	0.01	0.01	0.02	-	0.01
Carbohydrates	Ribulose-phosphate 3-epimerase (EC 5.1.3.1)	0.01	-	0.01	0.01	0.01	-
Carbohydrates	Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1)	0.09	0.03	0.09	0.09	0.02	0.01
Carbohydrates	Transketolase (EC 2.2.1.1)	0.07	0.18	0.23	0.20	0.25	0.23
Carbohydrates	Pyruvate dehydrogenase E1 component (EC 1.2.4.1)	0.11	0.07	0.11	0.14	0.03	0.01
Photosynthesis	Photosystem II protein D1 (PsbA)	1.58	0.83	0.29	0.66	0.14	0.12
Photosynthesis	Photosystem II protein D2 (PsbD)	0.02	0.01	0.02	0.02	-	0.01
Phosphorous metabolism	Phosphate starvation-inducible protein PhoH	0.52	2.36	0.44	1.00	1.02	1.18
Respiration	NADH-ubiquinone oxidoreductase chains (EC 1.6.5.3)	0.41	0.12	0.32	0.51	0.08	0.04

Table 4-15. (continued)

Functional groups	Protein functions	H1[a]	H2	H3	H4	H5	H6
RNA Metabolism	Queuosine Biosynthesis QueC, E, D	0.02	0.01	0.09	0.01	0.06	0.05
Virulence, Disease and Defense	Cobalt-zinc-cadmium resistance protein	-	-	0.02	0.01	0.01	-
Virulence, Disease and Defense	Staphylococcal pathogenicity islands, SaPI	2.98	2.19	2.12	2.19	1.40	2.01

[a] (Number of viral metagenome reads assigned to a AMG) / (Total number of viral metagenome reads that were assigned to a protein function) x 100

Table 4-16. Number of viral metagenome reads that were assigned to ARG-related genes

Antibiotic resistance genes (ARG)	**H1**[a]	**H2**	**H3**	**H4**	**H5**	**H6**
ABC-type multidrug transport system	14	5	16	27	24	14
Beta-lactamase	3	-	-	1	5	1
Metallo-beta-lactamase family protein, RNA-specific	1	1	4	13	6	3
Penicillin-binding protein	1	-	-	2	4	2
Acriflavin resistance protein	4	3	5	4	7	10
Vancomycin B-type resistance protein VanW	2	-	-	-	-	-
Methicillin resistance protein	11	5	8	19	2	2
Proportion of ARG-related viral metagenome reads[b]	0.06%	0.03%	0.06%	0.08%	0.05%	0.04%

[a] Number of viral metagenome reads that were assigned to a ARG

[b] (Number of viral metagenome reads that were assigned to ARG-related genes) / (Total number of viral metagenome reads that were assigned to a functional protein) * 100

3.3. Antibiotic resistance genes within viral metagenome and viral contigs

3.3.1. Search of ARG from general protein database

Although ARGs were found in viral metagenome reads, whether they are carried by true bacteriophage genomes and whether they are functional were not able to be judged. Therefore, assembled viral metagenome contigs were analyzed. To predict protein coding genes within the assembled contigs and annotate their functions, the virome contigs that are 10 kb in length or longer, were uploaded unto IMG/M ER webserver. From the IMG/M ER annotation, 53 contigs were found to be carrying ARGs. Those ARGs were collected and further analyzed by BLAST against NCBI nr database, and 33 ARGs in 33 contigs were eliminated for high BLAST e-values. Among the 15 contigs that survived (Table 4-17), bona fide bacteriophage contigs were searched by annotating all the ORFs found in the contig. To double-check, the 15 contigs were reannotated by the RAST server (Aziz *et al*., 2008). Among 15 contigs, only 3 contigs were carrying bacteriophage-related genes such as terminase and capsid gene that they were determined to be bona fide bacteriophage genomes with ARG (Fig. 4-9). All three contigs, H2-260, H4-1399, and H5-411 carried Beta-lactamase genes with different domains. The contig H2-260 that has been retrieved from site H2 carried a beta-lactamase 2 gene with three conserved active sites of H-X-H-X-D-H, H-D, and D (Fig. 4-10), implying for possible activity of the beta-lactamase. Besides those active site, beta-lactamase genes also share highly conserved active site, S-X-X-K. The contig H5-411 was shown to be carrying metallo-beta-lactamase gene but this ARG does not carry the S-X-X-K active site within the protein coding gene. Yet, interestingly, this metallo-beta-lactamase gene contained amino acid sequences of K-X-X-S, a reversed form of the conserved active site of the beta-lactamases.

Table 4-17. List of ARGs that were found within assembled viral metagenome contigs

Site	Contig	Length (bp)	Predicted function	NCBI BLAST match[a]	Query cover	e-value	% ident.	Accession
H1	746	12,382	Penicillin-binding protein-related factor A, recombinase	*Clostridiales* bacterium 52-15	75%	2.00E-06	30%	OKZ68299.1
H2	260	17,580	Beta-lactamase superfamily protein	*Capnocytophaga* sp. CM59	88%	2.00E-65	56%	EJF32237.1
H3	119	28,856	Dihydrofolate reductase, trimethoprim resistance protein	*Burkholderiales* bacterium JOSHI_001	80%	8.00E-23	79%	EHR70718.1
H3	388	15,525	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Thauera* sp. 28	92%	3.00E-35	37%	WP_002932368.1
H4	9	78,317	Dihydrofolate reductase, trimethoprim resistance protein	*Burkholderiales* bacterium JOSHI_001	80%	8.00E-23	79%	EHR70718.1
H4	45	54,878	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Thauera* sp. 28	92%	3.00E-35	37%	WP_002932368.1
H4	77	40,802	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Aeromonas* sp. RU39B	90%	2.00E-43	40%	WP_076574305.1
H4	1399	10,102	Beta-lactamase superfamily, PASTA domain	*Butyrivibrio* sp. XPD2006	67%	5.00E-06	43%	WP_022764105.1
H5	122	33,152	Dihydrofolate reductase, trimethoprim resistance protein	*Burkholderiales* bacterium JOSHI_001	80%	8.00E-23	79%	EHR70718.1
H5	139	31,091	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Thauera* sp. 28	89%	6.00E-35	37%	WP_002932368.1

Table 4-17. (continued)

Site	Contig	Length (bp)	Predicted function	NCBI BLAST match[a]	Query cover	e-value	% ident.	Accession
H5	145	30,808	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Aeromonas* sp. RU39B	90%	2.00E-43	40%	WP_076574305.1
H5	411	17,068	Metallo-beta-lactamase	*Flavobacterium spartansii*	96%	4.00E-15	64%	WP_070907730.1
H6	8	77,134	Dihydrofolate reductase, trimethoprim resistance protein	*Burkholderiales* bacterium JOSHI_001	80%	8.00E-23	79%	EHR70718.1
H6	105	31,811	D-alanyl-D-alanine endopeptidase, penicillin-binding protein 7	*Aeromonas* sp. RU39B	90%	2.00E-43	40%	WP_076574305.1
H6	586	12,048	Multidrug resistance efflux pump	*Marinobacter lutaoensis*	98%	1.00E-23	37%	WP_079724739.1

[a] The best BLAST match of the ARG found within the viral metagenome contigs

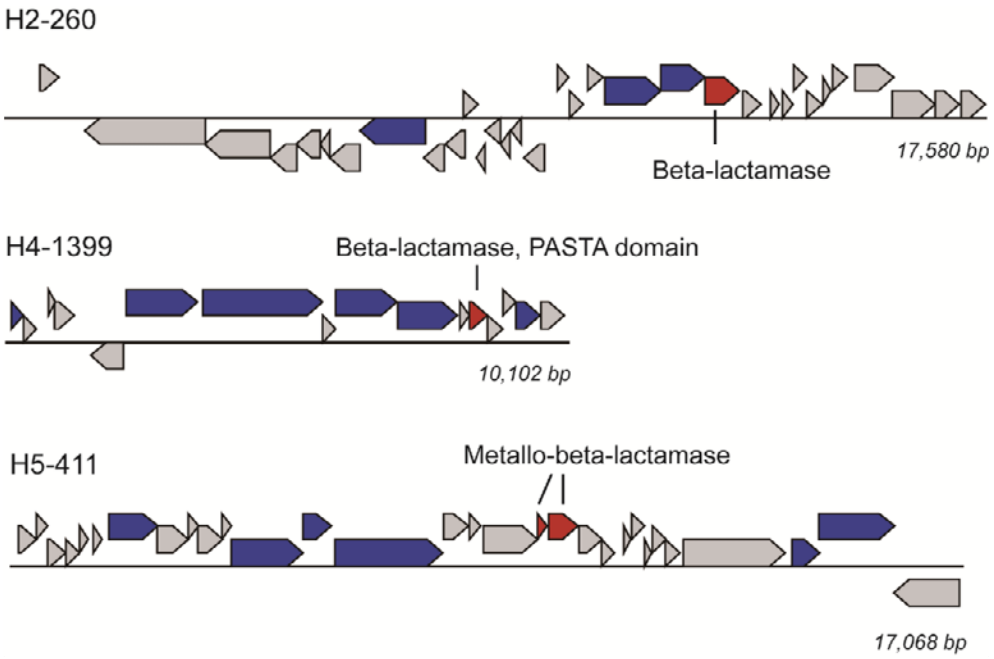

Figure 4-9. Genome map of three putative bacteriophage contigs retrieved from the Han River virome that carry ARGs. The ORFs that code for bacteriophage-related genes are shown in blue while ORFs coding for ARG are shown in red.

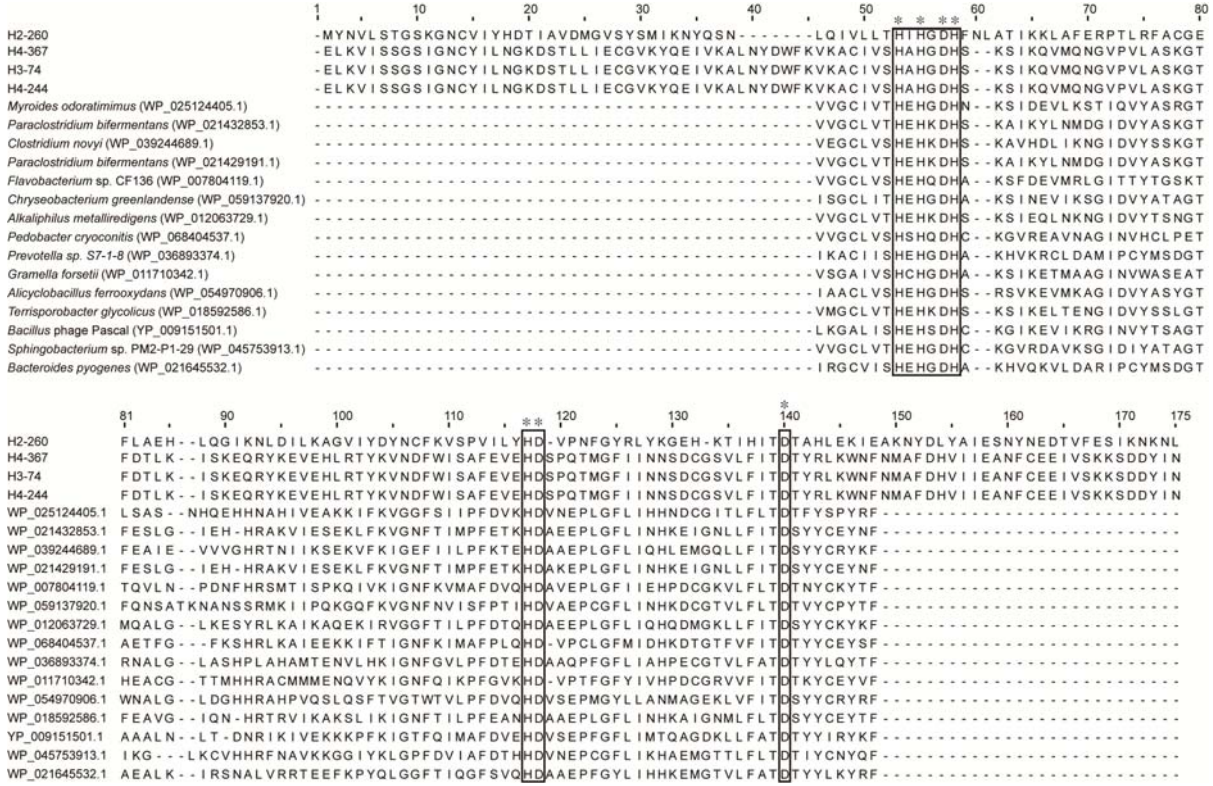

Figure 4-10. Sequence alignment of Lactamase B-2 (PF12706.2), a member of Metallo-beta-lactamase superfamily. Lactamase B-2 genes retrieved from assembled virome data and NCBI nr database were aligned using ClustalW embedded in MEGA 6 and sequence alignment was drawn by JalView. The beta-lactamase gene found in contig H2-260 was retrieved from IMG annotation server while the other genes were found from Resfams database. The colored regions indicate the conserved active sites of the Lactamase B-2 gene, HXHXDH, HD, and D.

The contig H4-1399 carried a PASTA domain (Penicillin-binding-protein (PBP) and Serine/Threonine kinase Associated protein), which are found at the C-termini of the Penicillin-binding-proteins. The PASTA domains are “mutational hotspots” that could provide large diversity to PBPs (Yeats *et al*., 2002). The PASTA domain found in the H4-1399 contig showed large variation from its reference sequences (Fig. 4-11), suggesting diversified form of the PASTA domain. Also, until recently, the PASTA domain has not been found in a viral genome before, thereby the contig H4-1399 is the first putative bacteriophage genome that was revealed to be carrying the PASTA domain.

3.3.2. ARG-specified databases

Inspired by the ARGs found in virome contigs, more specified and sensitive ARG search was performed using the CARD. For local BLAST search, assembled and virus-sorted virome contigs were used as a query and CARD, which contains 2,341 reference sequences was used as a database. Regardless of the sampling sites, large number of genes were detected to have similarity with ARGs. To refine the search results, threshold of e-value≤0.001, percent identity ≥ 80%, and bitscore ≥ 40 were applied. Unfortunately, no BLAST result satisfied all three thresholds given (Fig. 4-12). However, seeing that high number of genes had a significant match with ARG with at least one of the threshold criteria, the candidate ARGs present in the viral metagenome data were suspected to be diverged and CARD was too narrow database to be used for metagenome reads.

Based on the fact that all three ARGs detected from IMG database were related to beta-lactamase genes, beta-lactamase gene specific search was performed using Resfams database. The Resfams is a protein database that includes CARD and curated beta-lactamase protein sequences. The VirSorter program, which was used to sort virome contigs of viral origin, provided translated protein coding sequences

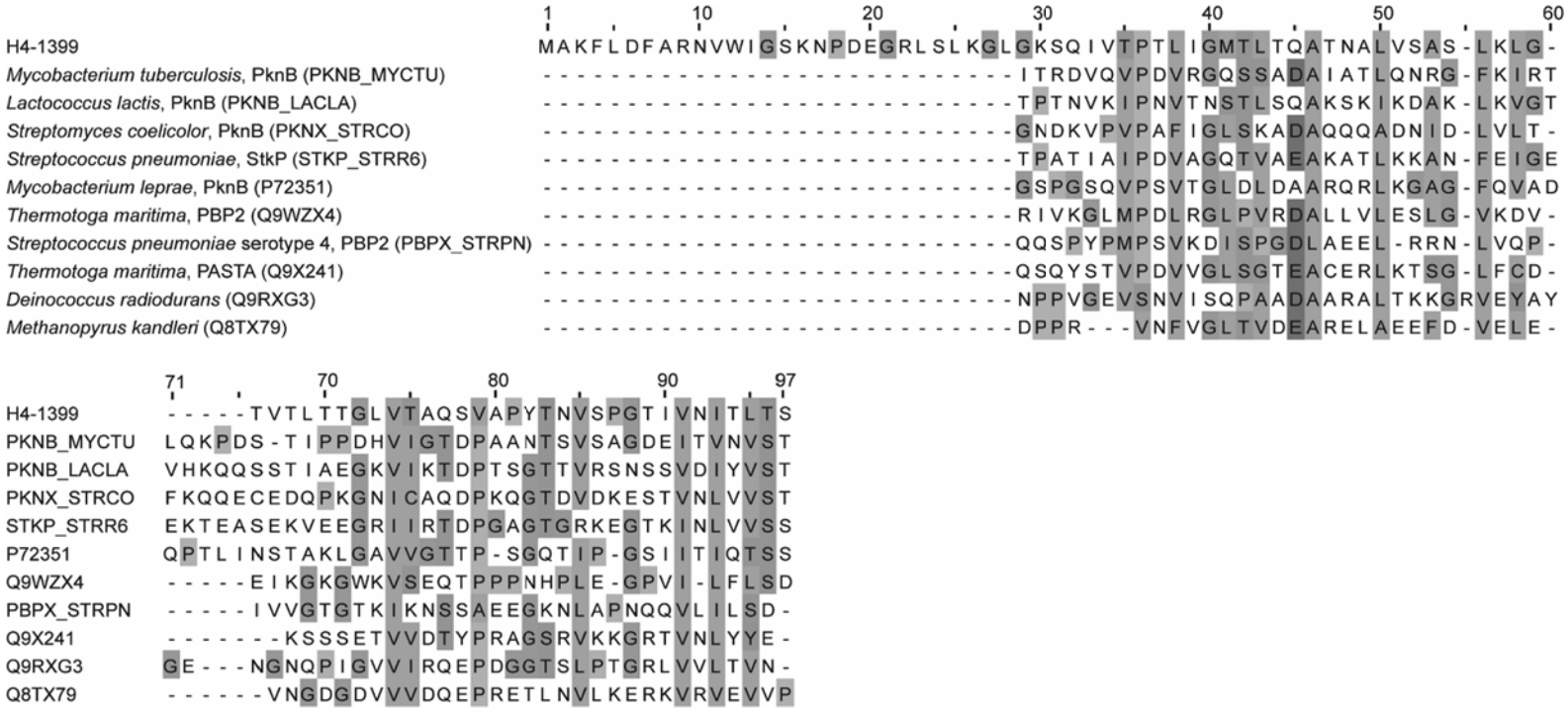

Figure 4-11. Sequence alignment of Beta-lactamase PASTA domains found in Serine/threonine kinase and Penicillin binding protein 2. The representative sequences were collected from UniProt database and they were aligned using ClustalW embedded in MEGA 6 and sequence alignment was drawn by JalView.

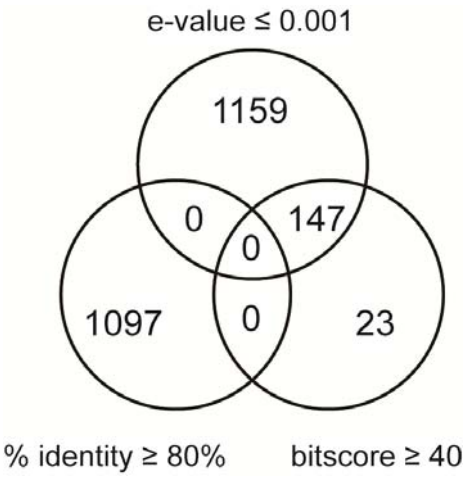

Figure 4-12. A Venn diagram displaying number of viral metagenome contigs that were found to be carrying ARG based on CARD. Virome contigs were analyzed with BLAST against the CARD to screen for ARG and contigs that satisfy each threshold levels, e-value, percent identity, and bitscore were counted. No contigs were found to be satisfying all the three thresholds.

for all the contigs submitted. Therefore, the translated contig sequences that were classified as phage-origin were used as a query and the hmmscan was performed. As a result, total of 4 contigs were revealed to contain beta-lactamase 2 genes with two different domains (Table 4-18). The beta-lactamase 2 genes within contigs H3-74, H4-367, and H4-244 had an identical active site as the H2-260 contig, which was identified earlier (Fig. 4-10). The contig H4-441 had a different domain from the others, and it carried an active site of S-X-X-K, the universal active site of beta-lactamase 2 genes (Fig. 4-13).

Using number of different databases to search for ARGs within virome samples, total of 7 bacteriophage contigs were discovered to be carrying ARG. Five of 7 contigs were retrieved from site H4, which was Hannam Bridge located in the Seoul city. The site is located downstream of an urban WWTP efflux site, thereby it could be suspected to be responsible for increase in number of viral contigs containing ARGs.

Table 4-18. List of ARGs that were found within assembled viral metagenome contigs

Site	Contig no.	Contig length (bp)	ARG	e-value	Score	Pfam domain
H3	74	34,930	Beta-lactamase-2	6.40E-20	65.10	PF12706.2
H4	244	29,404	Beta-lactamase-2	2.40E-20	65.10	PF12706.2
H4	367	26,437	Beta-lactamase-2	2.40E-20	65.10	PF12706.2
H4	441	20,541	Beta-lactamase-2	6.20E-14	45.50	PF13354.1

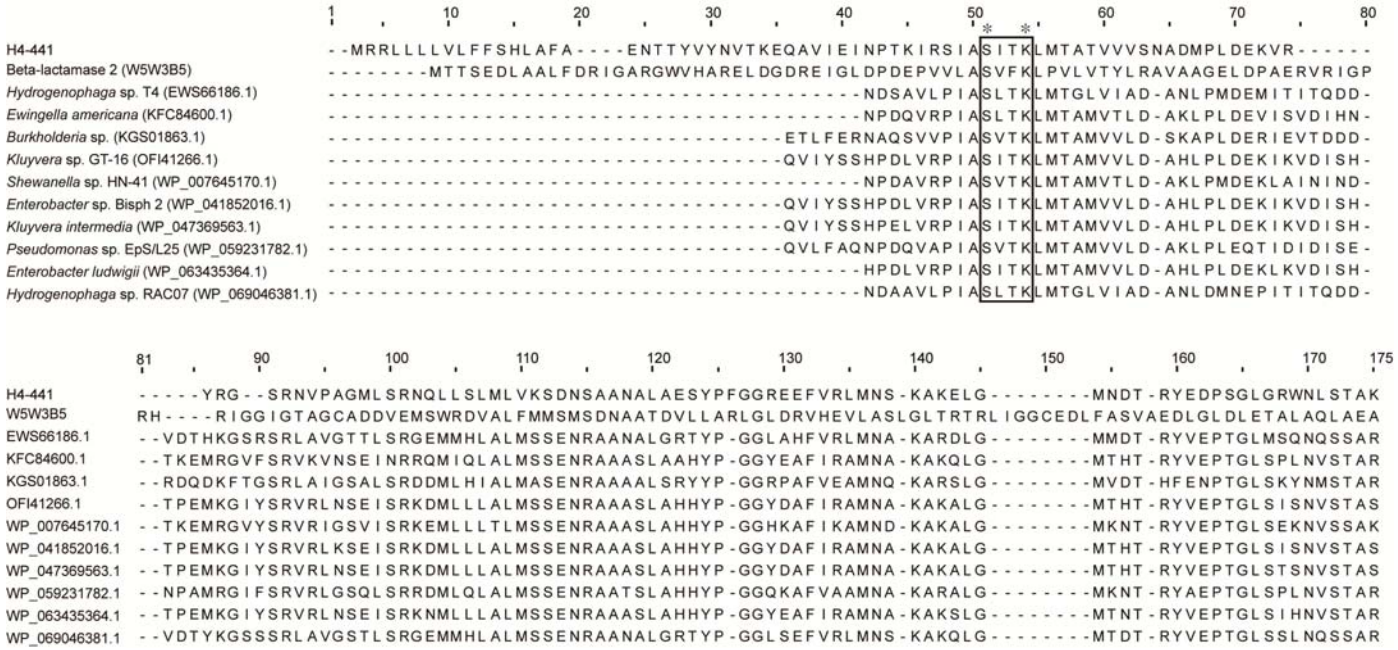

Figure 4-13. Sequence alignment of Beta-lactamase, group 2 genes (PF13354.1). Six Beta-lactamase 2 genes retrieved from assembled virome data and representative genes of Beta-lactamase 2 were aligned using ClustalW embedded in MEGA 6 and alignment was drawn using JalView. The colored region indicates the conserved active site of the Beta-lactamase genes, SXXK.

4. DISCUSSION

Bacteriophages, especially those found in natural environments play important ecological roles through controlling bacterial population, interference of geochemical cycling, and transfer of genetic materials from one cell to another. Therefore, in here, bacteriophage distribution in urban river waters were studied using viral metagenomics, a culture-independent method. The Han River, selected study site, flows across the northern area of South Korea and through the Seoul city. From the most upstream sampling site to the most downstream one, which flows for approximately 180 km, the river experiences diverse changes of environments; conserved water reservoir lakes, recreational sites, and urban river that receives WWTP effluents. Therefore, along with the water flow, the bacteriophage population composition was expected to change. However, the overall distribution of viral populations did not show much variation. The most frequently assigned bacteriophages in sampling sites H1 and H2 were *Thalassomonas* phage BA3 and *Prochlorococcus* phage P-SSM2, which are marine bacteriophages. Taxonomic assignment of viromes reads to marine bacteriophages in inland freshwater river indicates either narrowness of current bacteriophage database that hampers analysis of bacteriophage community of freshwater systems or that the presence of close relatives of marine bacteriophages thriving in river systems, yet, to be taxonomically unidentified. In sampling sites H3 to H6, the most abundant bacteriophages were *Bordetella* phage BPP-1, which infects human pathogen bacteria, *Bordetella bronchiseptica*. Since site H3 to H6 are located close or within the metropolitan city, increased abundance of *Bordetella* phage BPP-1 may imply the intervention of human influence to the water system.

Compared to the taxonomic annotation of the Han River virome samples, which annotated approximately 70% of virome reads as bacterial origin, the

functional annotation revealed that more than 40 % of the reads were annotated into bacteriophage-related functional group, which mostly consisted of phage structural genes. In protein sequence databases, numerous phage-related sequences are taxonomically assigned as bacterial origin, largely due to carriage of prophages in bacterial genomes. Consequently, imbalance between taxonomic and functional annotation may occur as seen in Han river virome. Besides well-known bacteriophage related genes, diverse metabolic genes were also detected, which were predicted to be AMGs of the bacteriophages. Along with the AMGs that are known to assist the host metabolism such as carbohydrate, nitrogen, phosphorous, and sulfur metabolism auxiliary defensive genes were also found that may contribute in defensive mechanism against heavy metals through expression of resistance genes and heavy metal efflux pumps. Also, genes to defend against antibiotics were also found. Once the bacterial cells are infected by phages that carry defensive genes, until the cell lysis due to bacteriophage replication and bursting, the cells will survive against toxic materials, which will provide enough time for active phage replication.

Among defensive AMGs detected from the Han River virome, ARGs were investigated again at the assembled contigs level. Also, to ensure that ARGs are carried by bacteriophages, not prokaryotic cells, the assembled contigs were checked twice, before and after the detection of ARGs within the contig. Through utilization of three databases, two of which were ARG-specified databases, total of 7 bona fide bacteriophage contigs were found to be carrying beta-lactamase genes. These ARGs had highly conserved active sites, anticipating that these genes would be functional. When each of the ORFs of those contigs were analyzed using protein BLAST to predict their hosts, no decisive conclusion was able to be drawn, for most of the ORFs had best BLAST matches of diverse organisms, including environmental virome reads. Although virome contigs that carry true ARGs were found at very low frequency, presence of active and diverse ARGs within environmental bacteriophage

contigs suggest their role as reservoirs of diverse ARGs within environments. Novel ARGs with various protein folds and variated sequences often arise from different environmental bacteria. Hence, the fact that environmental bacteriophage contigs are carrying active ARGs imply the possibility that those genes could be transferred to the next host bacteria and lead to arise of pathogenic bacterial strains with new ARGs (Vaz-Moreira *et al*., 2014; Wright, 2010). Bacteriophages that were isolated from animal system and clinical wastes were previously reported to be carrying ARGs (Colomer-Lluch *et al*., 2011a; Colomer-Lluch *et al*., 2011b; Modi *et al*., 2013). However, there has been debate whether ARGs present in bacteriophage genomes are functional (Enault *et al*., 2016). In this study, the ARGs found in freshwater virome contigs were shown to contain conserved active sites, suggesting their viable activity.

Although samples for virome analysis were taken from different sites along a lotic freshwater system, viral population and functional annotation did not vary significantly. Minor shifts in dominating bacteriophage were observed between 6 virome samples but overall population composition remained consistent. The functional composition of the virome reads also remained consistent throughout the river flow, indicating that viral population is not influenced by distances or different sites, within an identical water system. Within 6 viromes, 7 of ARGs with conserved active sites were found in bacteriophage contigs. This result provides evidence for their role as reservoirs of bacterial genes, and at the same time, their role as mediator for ARGs in the environment.

CHAPTER 5.

Conclusions

Despite their high abundance and ecological roles as microbial population controllers and reservoirs of bacterial genes, environmental bacteriophages were underappreciated. However, recently, with development of comprehensive methods to efficiently concentrate viral particles from aquatic environments and easier access to metagenome technologies, environmental bacteriophages have been re-illuminated with their extensive variability of genetic diversity and wide distribution. However, interpretation of viral metagenome data was still limited due to lack of isolated and sequenced individual bacteriophages from environment, leaving most of the viral metagenome reads as unknown. Therefore, from Lake Soyang, an oligotrophic freshwater reservoir, viral metagenome study was accompanied with culturing bacteriophage particles and sequencing of dominantly found bacteriophages in the freshwater environment. Hence, in this study, viral metagenome and bacteriophage isolation and culturing were performed simultaneously to better understand the bacteriophage population dynamics in freshwater environments.

From Lake Soyang, a large oligotrophic lake, seasonal distribution of bacteriophages was observed using viral metagenome. However, no significant seasonal variation was observed among bacteriophages, except for cyanophages, which bloomed in summer season and gradually decreased as winter approached. From Lake Soyang virome project, only the seasonal variability of cyanophages was observable because they are the most studied and identified environmental bacteriophage. About 90% of the analyzed viral metagenome reads appeared to be unidentified by the existing databases, restricting analysis of metagenome data. Since database-based classification of virome reads were limiting, protein sequence-based analysis was performed using metagenome-assembled contigs. As a result, approximately 2,000 contigs of Lake Soyang were found to be unreported sequences. Furthermore, identification of phage taxonomy and putative hosts of these contigs

were performed through manual curation of each ORFs carried by contigs. Among those, 976 contigs were predicted to have a host within the phylum *Proteobacteria*, 315 of them to have a host within the phylum *Actinobacteria,* and 59 of them were predicted to infect bacteria belonging to the phylum *Bacteroidetes*. Yet, prediction on specific bacterial host strain of the virome contigs were not able to be made. Each virome contigs were constructed with ORFs that are predicted to be accumulated from diverse bacterial host through HGT over long history of infection and co-evolution. Therefore, conclusion on single putative bacterial host was hard to be made. Genomic studies of bacteriophages and their metagenome data are also in need of more accurate annotation of their sequences, which could only be achieved by actual culturing and sequencing of bacteriophage isolates.

Therefore, from the identical site, Lake Soyang, novel bacteriophages were screened and isolated using bacterial strains isolated from the same lake. Using three bacteria strains that each belonging to LD28 group, *Curvibacter* species, and *Rhodoferax* species, four distinctive bacteriophages were isolated. Using the pure culture of bacteriophages obtained, whole genome sequencing was performed. As a result, it was revealed that the bacteriophage P19250A, which infects a strain belonging to the LD28 group, was the most abundantly found bacteriophage in Lake Soyang, specifically in winter seasons. Also, through binning analysis, four bacteriophage genomes obtained from Lake Soyang were found to be detected in various viral metagenome samples prepared from freshwater lakes of different countries. Especially in Lough Neagh, located in United Kingdom, the genome of P19250A appeared to be the most abundantly found bacteriophage genome, providing more refined interpretation of Lough Neagh virome study. Also, in bacteriophage evolution perspectives, the fact that the phage P19250A genome was found in diverse lakes of different countries show that phages appear to be infecting hosts of the same taxonomic clades may have developed independently and provide

evidence for convergent evolution of bacteriophages.

Along with bacterial strain belonging to LD28 clade, two representative bacterial strains, IMCC26218 and IMCC26059 belonging to *Rhodoferax* sp. and *Curvibacter* sp., respectively, were used to screen for novel bacteriophages from Lake Soyang. P26218, which infects *Rhodoferax* sp., IMCC26218 appeared to be a member of the family *Podoviridae*. P26059A and P26059B, which infect *Curvibacter* sp., belonged to different bacteriophage families, the family *Siphoviridae* and *Podoviridae*, respectively.

Using the same method as in Lake Soyang, viral population changes were observed through viral metagenome in running freshwater water, the Han River, one of the major river bodies of South Korea, six sampling sites were selected along the river flow to observe viral population distribution along the river flow. However, viral population distribution along the river flow was relatively consistent, displaying stably maintained viral community over 180 km of the rive length. From the Han River viral metagenome, various ARGs carried by bacteriophage contigs were discovered. These ARGs had highly conserved active sites indicating their possibility to be expressed within bacterial cells upon phage infection. Such event can lead to transfer of ARGs to diverse bacterial stains, leading to establishment of antibiotic resistance strains. Since viability of bacteriophage-carried ARGs have been in debate by many researchers, the virome-origin ARGs with highly conserved functional domains will provide solid proof that bacteriophages are the reservoirs and carriers of ARGs that could benefit the infected hosts.

This study has observed bacteriophage population, specifically dsDNA bacteriophages within freshwater environments, including both lentic and lotic waters, obtaining large number of unreported virome contigs. Also, from viral metagenome study performed in the Han River, an urban river, bacteriophage

genomes appeared to be carrying ARGs with novel sequences with conserved active sites, indicating that bacteriophages are the reservoirs of ARGs that occur in natural environments. Along with viral metagenome studies, this study isolated four novel bacteriophages from Lake Soyang. Among them, P19250A, an LD28 clade phage, appeared to be the most abundant bacteriophage in many lakes, including Lake Soyang, especially during winter seasons, and contributed to better understanding of freshwater bacteriophage ecology.

REFERENCES

Abedon, S.T. (2015). Bacteriophage secondary infection. *Virol Sin* 30, 3-10.

Ackermann, H.-W., and Heldal, M. (2010). Basic electron microscopy of aquatic viruses. In *Manual of Aquatic Viral Ecology,* Wilhelm, S.W., Weinbauer, M.G., and Suttle, C.A. ed. (Waco, USA: American Society of Limnology and Oceanography), pp. 182-192.

Adriaenssens, E.M., and Cowan, D.A. (2014). Using signature genes as tools to assess environmental viral ecology and diversity. *Appl Environ Microbiol* 80, 4470-4480.

Adriaenssens, E.M., Van Zyl, L., De Maayer, P., Rubagotti, E., Rybicki, E., Tuffin, M., and Cowan, D.A. (2015). Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ Microbiol* 17, 480-495.

Adriaenssens, E.M., van Zyl, L.J., Cowan, D.A., and Trindade, M.I. (2016). Metaviromics of Namib desert salt pans: a novel lineage of haloarchaeal salterproviruses and a rich source of ssDNA viruses. *Viruses* 8, 14.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., and *et al*. (2006). The marine viromes of four oceanic regions. *PLoS Biol* 4, e368.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 32, D115-D119.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-

cell sequencing. *J Comput Biol* 19, 455-477.

Beck, D.A., McTaggart, T.L., Setboonsarng, U., Vorobev, A., Kalyuzhnaya, M.G., Ivanova, N., Goodwin, L., Woyke, T., Lidstrom, M.E., and Chistoserdova, L. (2014). The expanded diversity of *Methylophilaceae* from Lake Washington through cultivation and genomic sequencing of novel ecotypes. *PLoS One* 9, e102458.

Bellas, C.M., and Anesio, A.M. (2013). High diversity and potential origins of T4-type bacteriophages on the surface of Arctic glaciers. *Extremophiles* 17, 861-870.

Bengtsson-Palme, J., Hartmann, M., Eriksson, K.M., Pal, C., Thorell, K., Larsson, D.G.J., and Nilsson, R.H. (2015). Metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* 15, 1403-1414.

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B.L., and Sullivan, M.B. (2016). iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J* 11, 7-14.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114-2120.

Breitbart, M. (2012). Marine viruses: truth or dare. *Annu Rev Mar Sci* 4, 425-448.

Breitbart, M., Miyake, J.H., and Rohwer, F. (2004a). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* 236, 249-256.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *PNAS* 99, 14250-14255.

Breitbart, M., Wegley, L., Leeds, S., Schoenfeld, T., and Rohwer, F. (2004b). Phage Community Dynamics in Hot Springs. *Appl Environ Microbiol* 70, 1633-1640.

Brown-Jaque, M., Calero-Cáceres, W., and Muniesa, M. (2015). Transfer of antibiotic-resistance genes via phage-related mobile elements. *Plasmid* 79, 1-7.

Bruder, K., Malki, K., Cooper, A., Sible, E., Shapiro, J.W., Watkins, S.C., and Putonti, C. (2016). Freshwater metaviromics and bacteriophages: A current assessment of the state of the art in relation to bioinformatic challenges. *Evol Bioinform Online* 12, 25.

Brum, J.R., Hurwitz, B.L., Schofield, O., Ducklow, H.W., and Sullivan, M.B. (2015a). Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME* J 10, 437-449.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., and *et al*. (2015b). Patterns and ecological drivers of ocean viral communities. *Science* 348, 1261498.

Brum, J.R., Schenck, R.O., and Sullivan, M.B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* 7, 1738-1751.

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

Bryan, M.J., Burroughs, N.J., Spence, E.M., Clokie, M.R., Mann, N.H., and Bryan, S.J. (2008). Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. *PLoS One* 3, e2048.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12, 59-60.

Cai, L., Zhang, R., He, Y., Feng, X., and Jiao, N. (2016). Metagenomic analysis of virioplankton of the subtropical Jiulong River estuary, China. *Viruses* 8, 35.

Carini, P., Steindler, L., Beszteri, S., and Giovannoni, S.J. (2013). Nutrient requirements for growth of the extreme oligotroph '*Candidatus*

Pelagibacter ubique' HTCC1062 on a defined medium. *ISME J* 7, 592-602.

Cheng, H., Shen, N., Pei, J., and Grishin, N.V. (2004). Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease. *Prot Sci* 13, 2260-2269.

Chistoserdova, L. (2011). Methylotrophy in a lake: From metagenomics to single-organism physiology. *Appl Environ Microbiol* 77, 4705-4711.

Chistoserdova, L. (2015). Methylotrophs in natural habitats: current insights through metagenomics. *Appl Environ Microbiol* 99, 5763-5779.

Chistoserdova, L., and Lidstrom, M.E. (2013). Aerobic Methylotrophic Prokaryotes. In *The Prokaryotes*, Rosenberg, E., DeLong, E.F., Lory, S., Stackerbrandt, E., Thompson, F., ed. (New York:Springer), pp. 267-285.

Cho, J.-C., and Giovannoni, S.J. (2004). Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. *Appl Environ Microbiol* 70, 432-440.

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

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as winners in the game of life. *Annu Rev Virol* 3, 197-214.

Colomer-Lluch, M., Imamovic, L., Jofre, J., and Muniesa, M. (2011a). Bacteriophages carrying antibiotic resistance genes in fecal waste from cattle, pigs, and poultry. *Antimicrob Agents Chemother* 55, 4908-4911.

Colomer-Lluch, M., Jofre, J., and Muniesa, M. (2011b). Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS One* 6, e17549.

Colson, P., Yutin, N., Shabalina, S.A., Robert, C., Fournous, G., La Scola, B., Raoult, D., and Koonin, E.V. (2011). Viruses with more than 1,000 genes: Mamavirus, a new *Acanthamoeba polyphagamimivirus* strain, and reannotation of Mimivirus genes. *Genome Biol Evol* 3, 737-742.

Coote, J.G. (2001). Environmental sensing mechanisms in *Bordetella*. *Adv Microb Physiol* 44, 141-181.

Cottrell, M.T., Waidner, L.A., Yu, L., and Kirchman, D.L. (2005). Bacterial diversity of metagenomic and PCR libraries from the Delaware River. *Environ Microbiol* 7, 1883-1895.

Culley, A.I. (2013). Insight into the unknown marine virus majority. *PNAS* 110, 12166-12167.

Daims, H., Lebedeva, E.V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., Jehmlich, N., Palatinszky, M., Vierheilig, J., Bulaev, A., *et al.* (2015). Complete nitrification by *Nitrospira* bacteria. *Nature* 528, 504-509.

de Cárcer, D.A., López-Bueno, A., Pearce, D.A., and Alcamí, A. (2015). Biodiversity and distribution of polar freshwater DNA viruses. *Sci Adv* 1, e1400127.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27, 4636-4641.

DeLong, E.F., and Beja, O. (2010). The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. *PLoS Biol* 8, e1000359.

Dougan, D.A., Reid, B.G., Horwich, A.L., and Bukau, B. (2002). ClpS, a substrate modulator of the ClpAP machine. *Mol Cell* 9, 673-683.

Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. *Nature Rev Microbiol* 3, 504-510.

Eiler, A., Zaremba-Niedzwiedzka, K., Martínez-García, M., McMahon, K.D., Stepanauskas, R., Andersson, S.G., and Bertilsson, S. (2014). Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environ Microbiol* 16, 2682-2698.

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M.B., and Petit, M.-A. (2016). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analysis. *ISME J* 11, 237-247.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J. and *et al*. (2013). Pfam: the protein families database. *Nucleic Acids Res* 42, D222-D230.

Gaillot, O., Pellegrini, E., Bregenholt, S., Nair, S., and Berche, P. (2000). The ClpP serine protease is essential for the intracellular parasitism and virulence of *Listeria monocytogenes*. *Mol Microbiol* 35, 1286-1294.

Ghai, R., Mehrshad, M., Megumi Mizuno, C., and Rodriguez-Valera, F. (2017). Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *ISME J* 11, 304-308

Ghylin, T.W., Garcia, S.L., Moya, F., Oyserman, B.O., Schwientek, P., Forest, K.T., Mutschler, J., Dwulit-Smith, J., Chan, L.-K., Martinez-Garcia, M., *et al.* (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI actinobacteria lineage. *ISME J* 8, 2503-2516.

Gibson, M.K., Forsberg, K.J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 9, 207-216.

Gifford, S.M., Sharma, S., Booth, M., and Moran, M.A. (2013). Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* 7, 281-298.

Glass, E.M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* pdb.prot5368.

Glöckner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000). Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl Environ Microbiol* 66, 5053-5065.

Goldsmith, D.B., Crosti, G., Dwivedi, B., McDaniel, L.D., Varsani, A., Suttle, C.A.,

Weinbauer, M.G., Sandaa, R.-A., and Breitbart, M. (2011). Development of phoH as a novel signature gene for assessing marine phage diversity. *Appl Environ Microbiol* 77, 7730-7739.

Grabow, W. (2004). Bacteriophages: update on application as models for viruses in water. *Water Sa* 27, 251-268.

Green, J.C., Rahman, F., Saxton, M.A., and Williamson, K.E. (2015). Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA. *Aquat Microb Ecol* 75, 117-128.

Green, M.R., and Sambrook, J. (2012). Molecular cloning: a laboratory manual (New York: Cold Spring Harbor Laboratory Press).

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J.R., *et al.* (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465-470.

Hahn, M.W., Jezberová, J., Koll, U., Saueressig-Beck, T., and Schmidt, J. (2016). Complete ecological isolation and cryptic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. *ISME J* 10, 1642-1655.

Hahn, M.W., Koll, U., Jezberová, J., and Camacho, A. (2015). Global phylogeography of pelagic *Polynucleobacter* bacteria: Restricted geographic distribution of subgroups, isolation by distance and influence of climate. *Environ Microbiol* 17, 829-840.

Halsey, K.H., Carter, A.E., and Giovannoni, S.J. (2012). Synergistic metabolism of a broad range of C1 compounds in the marine methylotrophic bacterium HTCC2181. *Environ Microbiol* 14, 630-640.

Hanson, R.S. (1998). Ecology of methylotrophic bacteria. In *Techniques in microbial ecology*, Burlage, R.S., Atlas, R., Stahl, D., Geesey, G., and Sayler, G. ed. (New York: Oxford University Press), pp. 137-161.

Hendrix, R.W. (2010). Recoding in Bacteriophages. In *Recoding: Expansion of*

decoding rules enriches gene expression, Atkins, J.F., Gesteland, R.F. ed. (New York: Springer), pp. 249-258.

Hevroni, G., Enav, H., Rohwer, F., and Beja, O. (2015). Diversity of viral photosystem-I psaA genes. *ISME J* 9, 1892-1898.

Hewson, I., and Fuhrman, J.A. (2008). Chapter 25 - Viruses, Bacteria, and the Microbial Loop. In *Nitrogen in the Marine Environment* (2nd Edition), Capone, D.G., Bronk, D.A., Mulholland, M.R., and Carpenter, E.J. eds. (San Diego: Academic Press), pp. 1097-1134.

Hingamp, P., Grimsley, N., Acinas, S.G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmento, H., Villar, E., and Lima-Mendez, G. (2013). Exploring nucleo-cytoplasmic large DNA viruses in *Tara* Oceans microbial metagenomes. *ISME J* 7, 1678-1695.

Hiraishi, A., Hoshino, Y., and Satoh, T. (1991). *Rhodoferax fermentans* gen. nov., sp. nov., a phototrophic purple nonsulfur bacterium previously referred to as the "*Rhodocyclus gelatinosus*-like" group. *Archiv Microbiol* 155, 330-336.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol* 3, reviews0003.0001-reviews0003.0008.

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, 1428-1440.

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.

Hurwitz, B.L., Brum, J.R., and Sullivan, M.B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core'and 'flexible'Pacific Ocean Virome. *ISME J* 9, 472-484.

Hurwitz, B.L., and U'Ren, J.M. (2016). Viral metabolic reprogramming in marine

ecosystems. *Curr Opin Micribiol* 31, 161-168.

Ignacio-Espinoza, J.C., Solonenko, S.A., and Sullivan, M.B. (2013). The global virome: not as big as we thought? *Curr Opin Virol* 3, 566-571.

Jezbera, J., Jezberova, J., Kasalicky, V., Simek, K., and Hahn, M.W. (2013). Patterns of *Limnohabitans* microdiversity across a large set of freshwater habitats as revealed by Reverse Line Blot Hybridization. *PLoS One* 8, e58527.

Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N., *et al*. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 45, D566-D573.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K., Kern, S., Brum, J., Polz, M.F., Boyle, E.A., and Sullivan, M.B. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* 3, 195-202.

Kaden, R., Sproer, C., Beyer, D., and Krolla-Sidenstein, P. (2014). *Rhodoferax saidenbachensis* sp. nov., a psychrotolerant, very slowly growing bacterium within the family *Comamonadaceae*, proposal of appropriate taxonomic position of *Albidiferax ferrireducens* strain T118T in the genus *Rhodoferax* and emended description of the genus *Rhodoferax*. *Int J Syst Evol Microbiol* 64, 1186-1193.

Kang, I., Oh, H.M., Kang, D., and Cho, J.C. (2013). Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *PNAS* 110, 12343-12348.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., and Claverie, J.-M. (2011). A holistic approach to marine eco-systems biology. *PLoS Biol* 9, e1001177.

Kinch, L.N., Ginalski, K., Rychlewski, L., and Grishin, N.V. (2005). Identification of novel restriction endonuclease-like fold families among hypothetical

proteins. *Nucleic Acids Res* 33, 3598-3605.

King, A.M., Adams, M.J., and Lefkowitz, E.J. (2012). *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*, Vol 9 (Amsterdam: Elsevier).

King, G., and Murray, N.E. (1995). Restriction alleviation and modification enhancement by the Rac prophage of *Escherichia coli* K-12. *Mol Microbiol* 16, 769-777.

Knowles, B., Silveira, C., Bailey, B., Barott, K., Cantu, V., Cobián-Güemes, A., Coutinho, F., Dinsdale, E., Felts, B., and Furby, K. (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466-470.

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.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., and Koonin, E. *et al.* (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100-104.

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.

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.W., and Kropinski, A.M. (2008). Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. *Res Microbiol* 159, 406-414.

Ledermann, B., Béjà, O., and Frankenberg-Dinkel, N. (2016). New biosynthetic pathway for pink pigments from uncultured oceanic viruses. *Environ Microbiol* 18, 4337-4347.

Lekunberri, I., Subirats, J., Borrego, C.M., and Balcázar, J.L. (2017). Exploring the contribution of bacteriophages to antibiotic resistance. *Environ Pollut* 220, Part B, 981-984.

Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., and Alcami,

A. (2009). High diversity of the viral community from an Antarctic lake. *Science* 326, 858-861.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25, 955-964.

Lu, M.J., and Henning, U. (1989). The immunity (imm) gene of *Escherichia coli* bacteriophage T4. *J Virol* 63, 3472-3478.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res* 26, 1107-1115.

Madigan, M.T., Jung, D.O., Woese, C.R., and Achenbach, L.A. (2000). *Rhodoferax antarcticus* sp. nov., a moderately psychrophilic purple nonsulfur bacterium isolated from an Antarctic microbial mat. *Arch Microbiol* 173, 269-277.

Mak, A.N.-S., Lambert, A.R., and Stoddard, B.L. (2010). Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kI. *Structure* 18, 1321-1331.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* 424, 741-741.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., and Gonzales, N.R. *et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39, D225-D229.

Markowitz, V.M., Chen, I.M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., *et al*. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* 40, D115-D122.

Maurice, C., Bouvier, C., Wit, R., and Bouvier, T. (2013). Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. *Environ Microbiol* 15,

2463-2475.

Mazaheri Nezhad Fard, R., Barton, M., and Heuzenroeder, M. (2011). Bacteriophage-mediated transduction of antibiotic resistance in enterococci. *Lett Appl Microbiol* 52, 559-564.

Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004). Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *PNAS* 101, 11007-11012.

Minchin, P.R., O'Hara, R., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2015). Vegan: Community Ecology Package. R package version 2.3-0.

Modi, S.R., Lee, H.H., Spina, C.S., and Collins, J.J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499, 219-222.

Mohiuddin, M., and Schellhorn, H.E. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* 6, 960.

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* 75, 14-49.

Niño-García, J.P., Ruiz-González, C., and del Giorgio, P.A. (2016). Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *ISME J* 10, 1755-1766.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., and Suggests, M. (2007). The vegan package. Community ecology package. Version 2.0-7 10, 631-637.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). MASH: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 17, 132.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., *et al.* (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42, D206-D214.

Pan, D., Watson, R., Wang, D., Tan, Z.H., Snow, D.D., and Weber, K.A. (2014). Correlation between viral production and carbon mineralization under nitrate-reducing conditions in aquifer sediment. *ISME J* 8, 1691-1703.

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, 273-284.

Patel, A., Noble, R.T., Steele, J.A., Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc* 2, 269-276.

Payet, J.P., and Suttle, C.A. (2013). To kill or not to kill: the balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol Oceanogr* 58, 465-474.

Pennell, S., Déclais, A.-C., Li, J., Haire, L.F., Berg, W., Saldanha, J.W., Taylor, I.A., Rouse, J., Lilley, D.M., and Smerdon, S.J. (2014). FAN1 activity on asymmetric repair intermediates is mediated by an atypical monomeric virus-type replication-repair nuclease domain. *Cell Rep* 8, 84-93.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41, D590-D596.

Rappe, M.S., Kemp, P.F., and Giovannoni, S.J. (1997). Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. *Limnol Oceanogr* 42, 811-826.

Rasmussen, M., Jacobsson, M., and Björck, L. (2003). Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. *J Biol Chem* 278, 32313-32316.

Rastrojo, A., and Alcamí, A. (2016). Aquatic viral metagenomics: Lights and shadows. *Virus Res*.

Raymond, P.A., Hartmann, J., Lauerwald, R., Sobek, S., McDonald, C., Hoover, M., Butman, D., Striegl, R., Mayorga, E., and Humborg, C., *et al.* (2013). Global carbon dioxide emissions from inland waters. *Nature* 503, 355-359.

Reavy, B., Swanson, M.M., Cock, P.J., Dawson, L., Freitag, T.E., Singh, B.K., Torrance, L., Mushegian, A.R., and Taliansky, M. (2015). Distinct circular single-stranded DNA viruses exist in different soil types. *Appl Environ Microbiol* 81, 3934-3945.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., *et al*. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J* 4, 739-751.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., *et al*. (2016a). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689-693.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B., Coleman, M.L., Breitbart, M., and Sullivan, M.B. (2016b). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4, e2777.

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

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., and Debroas, D. (2012). Assessing the diversity and

specificity of two freshwater viral communities through metagenomics. *PLoS One* 7, e33641.

Salcher, M.M. (2014). Same same but different: ecological niche partitioning of planktonic freshwater prokaryotes. *J Limnol* 73, 74-87.

Salcher, M.M., Neuenschwander, S.M., Posch, T., and Pernthaler, J. (2015). The ecology of pelagic freshwater methylotrophs assessed by a high-resolution monitoring and isolation campaign. *ISME J* 9, 2442-2453.

Salcher, M.M., Pernthaler, J., Frater, N., and Posch, T. (2011). Vertical and longitudinal distribution patterns of different bacterioplankton populations in a canyon-shaped, deep prealpine lake. *Limnol Oceanogr* 56, 2027-2039.

Sanyal, S.J., Yang, T.-C., and Catalano, C.E. (2014). Integration host factor assembly at the cohesive end site of the bacteriophage lambda genome: Implications for viral DNA packaging and bacterial gene regulation. *Biochemistry* 53, 7459-7470.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., and *et al.*, (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75, 7537-7541.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R.Y., Partensky, F., Koonin, E.V., Wolf, Y.I., *et al.* (2009). Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461, 258-262.

Sime-Ngando, T. (2014). Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Front Microbiol* 5, 355.

Skvortsov, T., de Leeuwe, C., Quinn, J.P., McGrath, J.W., Allen, C.C., McElarney, Y., Watson, C., Arkhipova, K., Lavigne, R., and Kulakov, L.A. (2016). Metagenomic characterisation of the viral community of Lough Neagh, the

largest freshwater lake in Ireland. *PLoS One* 11, e0150361.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics* 21, 951-960.

Sowell, S.M., Abraham, P.E., Shah, M., Verberkmoes, N.C., Smith, D.P., Barofsky, D.F., and Giovannoni, S.J. (2011). Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J* 5, 856-865.

Srinivasiah, S., Lovett, J., Polson, S., Bhavsar, J., Ghosh, D., Roy, K., Fuhrmann, J.J., Radosevich, M., and Wommack, K.E. (2013). Direct assessment of viral diversity in soils by random PCR amplification of polymorphic DNA. *Appl Environ Microbiol* 79, 5450-5457.

Staley, C., Unno, T., Gould, T., Jarvis, B., Phillips, J., Cotner, J., and Sadowsky, M. (2013). Application of Illumina next-generation sequencing to characterize the bacterial community of the Upper Mississippi River. *J Appl Microbiol* 115, 1147-1158.

Stingl, U., Cho, J.-C., Foo, W., Vergin, K., Lanoil, B., and Giovannoni, S. (2008). Dilution-to-extinction culturing of psychrotolerant planktonic bacteria from permanently ice-covered lakes in the McMurdo Dry Valleys, Antarctica. *Microbiol Ecol* 55, 395-405.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* 4, e234.

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

Swanson, M.M., Reavy, B., Makarova, K.S., Cock, P.J., Hopkins, D.W., Torrance, L., Koonin, E.V., and Taliansky, M. (2012). Novel bacteriophages containing a genome of another bacteriophage within their genomes. *PLoS One* 7, e40683.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6:

molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30, 2725-2729.

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.

Thompson, J.D., Gibson, T., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, 2.3. 1-2.3. 22.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4, 470-483.

Tranvik, L.J., Downing, J.A., Cotner, J.B., Loiselle, S.A., Striegl, R.G., Ballatore, T.J., Dillon, P., Finlay, K., Fortino, K., Knoll, L.B., *et al.* (2009). Lakes and reservoirs as regulators of carbon cycling and climate. *Limnol Oceanogr* 54, 2298-2314.

Tseng, C.H., Chiang, P.W., Shiah, F.K., Chen, Y.L., Liou, J.R., Hsu, T.C., Maheswararajah, S., Saeed, I., Halgamuge, S., and Tang, S.L. (2013). Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* 7, 2374-2386.

Vartoukian, S.R., Palmer, R.M., and Wade, W.G. (2010). Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol Lett* 309, 1-7.

Vaz-Moreira, I., Nunes, O.C., and Manaia, C.M. (2014). Bacterial diversity and antibiotic resistance in water habitats: searching the links with the human microbiome. *FEMS Microbiol Rev* 38, 761-778.

Warnecke, F., Amann, R., and Pernthaler, J. (2004). Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. *Environ Microbiol* 6, 242-253.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis

workbench. *Bioinformatics* 25, 1189-1191.

Watkins, S.C., Kuehnle, N., Ruggeri, C.A., Malki, K., Bruder, K., Elayyan, J., Damisch, K., Vahora, N., O'Malley, P., and Ruggles-Sage, B. (2015). Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar Freshwater Res* 67, 1700-1708.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E.M., Kyrpides, N., Mavrommatis, K., and Meyer, F. (2012). The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 13, 141.

Willems, A. (2014). The family comamonadaceae. In *The Prokaryotes* (New York: Springer), pp. 777-851.

Wommack, K.E., Nasko, D.J., Chopyk, J., and Sakowski, E.G. (2015). Counts and sequences, observations that continue to change our understanding of viruses in nature. *J Microbiol* 53, 181-192.

Wommack, K.E., Williamson, S.J., Sundbergh, A., Helton, R.R., Glazer, B.T., Portune, K., and Cary, S.C. (2004). An instrument for collecting discrete large-volume water samples suitable for ecological studies of microorganisms. *Deep-sea Res PT I* 51, 1781-1792.

Wright, G.D. (2010). Antibiotic resistance in the environment: a link to the clinic? *Curr Opin Microbiol* 13, 589-594.

Yang, S.-J., Kang, I., and Cho, J.-C. (2016). Expansion of Cultured Bacterial Diversity by Large-Scale Dilution-to-Extinction Culturing from a Single Seawater Sample. *Microb Ecol* 71, 29-43.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R., and Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Rev Microbiol* 12, 635-645.

Yeats, C., Finn, R.D., and Bateman, A. (2002). The PASTA domain: a β-lactam-

binding domain. *Trends Biochem Sci* 27, 438-440.

Yoshida, M., Yoshida-Takashima, Y., Nunoura, T., and Takai, K. (2015). Genomic characterization of a temperate phage of the psychrotolerant deep-sea bacterium Aurantimonas sp. *Extremophiles* 19, 49-58.

Yu, P., Mathieu, J., Li, M., Dai, Z., and Alvarez, P.J. (2016). Isolation of Polyvalent Bacteriophages by Sequential Multiple-Host Approaches. *Appl Environ Microbiol* 82, 808-815.

Zawar-Reza, P., Arguello-Astorga, G.R., Kraberger, S., Julian, L., Stainton, D., Broady, P.A., and Varsani, A. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect Genet Evol* 26, 132-138.

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

Zwart, G., Crump, B.C., Kamst-van Agterveld, M.P., Hagen, F., and Han, S.-K. (2002). Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat Microb Ecol* 28, 141-155.

국문초록

바이러스는 지구상에서 가장 많은 수로 존재하고 있으며 그 중, 박테리오파지 (파지)는 세균을 감염시킬 수 있는 바이러스로써 바이러스 중 가장 많은 비율을 차지하고 있다. 그들의 숙주인 세균은 지금까지 알려진 거의 모든 환경에서 발견됨에 따라 박테리오파지 역시 해양, 토양, 온천, 극지방, 사막 등 다양한 환경에서 존재함이 확인되었다. 그러나 이러한 생존 능력과 개체의 풍부함에도 불구하고 그들의 숙주인 세균 배양의 한계로 인해 지금까지의 환경 박테리오파지의 연구는 매우 제한적이었다. 최근, 이러한 제한을 극복하기 위해 바이러스 메타게놈 방법이 제시되었다. 바이러스 메타게놈은 박테리오파지를 배양하지 않고 그들의 유전자 정보에만 기반하여 박테리오파지 개체군을 연구할 수 있게 하였다. 따라서 바이러스 메타게놈을 사용하여, 대규모 해양 바이러스 프로젝트가 수행되었다. 그러나 내륙 담수 환경이 다양한 미생물 군락의 저장고로서의 역할을 하고 있음에도 불구하고 담수 환경에서의 박테리오파지 연구는 아직까지 소수에 불과하다. 따라서 본 연구에서는 담수 환경에서의 박테리오파지 개체군 분포와 세균 유전자 운반체로서의 생태계 내 역할의 이해를 위해 바이러스 메타게놈 시퀀싱과 박테리오파지의 배양 방법을 함께 수행하여 담수 환경 내의 박테리오파지 연구를 진행하였다. 우선, 계절 변동성에 따른 지표 호수내의 박테리오파지 개체 수 분포를 연구하기 위해 국내 담수 호수인 소양호에서 계절별 시료를 채취하여 바이러스의 메타게놈 표본을 준비하였다. 여섯 개의 시료 간 염기서열 유사성을 비교 하였을 때, 계절에 따른 명확한 변동성은 관찰되지 않았지만, 시간 경과에 따른 바이러스 메타게놈 유전자의 점진적 변화가 관찰

되었다. 또한 바이러스 메타게놈 데이터에서 얻어진 바이러스 contig 와 NCBI 에 등록된 바이러스 유전자 서열들과 함께 단백질 서열 유사성을 기반하여 분류하였을 때, 693개의 단백질 서열 그룹이 생성되었고, 그 중, 211개의 그룹이 소양호에서 얻은 contig 들로만 이루어져 있음을 알 수 있었다. 이 contig 들은 기존의 데이터베이스를 통해 유사한 박테리오파지, 또는 바이러스를 확인할 수 없는 소양호 특이적 박테리오파지로써, 기존의 BLAST 방법을 통해 계통학적 분류 분석을 수행할 수 없었다. 따라서, 211개 그룹에 속해있는 contig 의 open reading frame (ORF)를 각각 분석하여, 박테리오파지 분류 기준의 하나인, 그들의 숙주 세균을 예측하고자 하였다. 그 결과, 976 개의 contig 가 속해 있는 23개의 그룹은 *Proteobacteria* 문에 속하는 세균을 숙주로 가지고 있을 것으로 예측되었고, 315개의 contig 가 속해 있는 1개의 그룹은 *Actinobacteria* 문에 속하는 세균을 숙주로 가지고 있을 것으로 예측되었다. 하지만, 해당 박테리오파지 contig 의 좀 더 정확한 계통학적 분류는 이루어지지 않았다. 이는, 공공 유전체 데이터베이스 내에 담수 박테리오파지 유전체의 부족으로 인해 발생하는 현상으로써 바이러스 메타게놈의 결과가 충분히 해석되지 못하고 있음을 나타냈다. 따라서 이러한 현상을 해소하고 담수 박테리오파지 군집 이해에 기여하기 위해, 소양호에서 분리 된 배양 균주들을 이용하여 그들의 박테리오파지를 선별 배양 하였다. 그 결과, 소양호에서 총 4개의 새로운 박테리오파지가 분리되었다. 그 중 하나인 P19250A 박테리오파지는 *Methylophilaceae* 계통의 균주를 감염시키며 겨울철 소양호에서 가장 높은 빈도를 보이며 나타나는 것을 확인하였다. 이는 겨울철에 만연하는 P19250A 의 숙주 균주인 LD28 clade 의 계절성을 함께 보이는 것으로 나타났다. 또한, 소양호와 해외 담수 호수에서 얻은 바이러스

메타게놈 데이터를 대상으로 binning 분석을 진행 하였을 때, 소양호의 5개 샘플을 포함한 해외의 담수 호수에서 P19250A가 가장 높은 빈도로 나타나는 것을 확인하였다. 담수 호수 내 우점하는 세균 그룹의 하나인, *Comamonadaceae*과 내 두개의 세균을 선정하여 박테리오파지 분리 실험에 사용하였다. 그 결과, *Rhodoferax*속에 속하는 세균을 감염하는 박테리오파지인 P26218을 분리하는데 성공했으며, *Curvibacter*속의 세균인 IMCC26059를 감염하는 박테리오파지 P26059A 와 P26059B 를 분리하였다. 분리 후, 이들 박테리오파지들은 전체 유전체 시퀀싱을 통해 유전체를 얻었으며 이는 같은 지역에서 준비한 바이러스 메타게놈과 비교하여 이들의 유전체 풍부도를 확인할 수 있었다. 그러므로 신종 박테리오파지의 분리와 그들의 유전체 분석은 담수 바이러스 메타게놈 분석에 필요한 필수적인 자원임을 보여주었다. 내륙 담수 환경 중 하나인 강에서의 박테리오파지 분포를 확인하기 위해, 한국의 가장 북단에 위치한 한강에서 6개의 지점을 선정 후 표층수 시료를 채취하여 바이러스 메타게놈 연구를 진행하였다. 여섯 개의 시료의 전체 유전자 서열 유사성을 비교하였을 때, 모든 시료는 모두 낮은 비 유사성을 보였다. 또한 바이러스의 계통학적 분류를 통한 분석 결과, 바이러스 속의 분포는 강의 흐름에 따른 유의한 변화를 가지지 않았다. 이는 약 180 km 에 걸쳐 흐르는 한강은 안정적인 바이러스 및 박테리오파지 개체군 분포를 가지고 있음을 시사하였다. 또한 바이러스 메타게놈에서 얻은 바이러스 contig 내에서 박테리오파지에 의하 운반되는 보조 대사 유전자의 하나인 항생제 내성 유전자의 분포를 관찰하였다. 그 결과, 총 15개의 바이러스 contig는 활성 항생제 내성 유전자를 보유하고 있는 것으로 나타났으며, 이는 파지 감염 시, 이들 유전자가 다음 숙주 세균 세포로 옮겨져 항생제 내성 균주의

발생으로 이어질 수 있음을 시사하였다. 따라서 내륙 담수에서 서식하는 박테리오파지가 세균의 유전자 이동에 기여하는 중요한 매개체임을 확인하였다. 본 연구는 이와 같이, 생태계에서 가장 작은 개체인 박테리오파지의 신종 발견과 메타게놈 분석을 통해 국내 담수 환경에 서식하는 박테리오파지의 분포도와 생태학적 역할을 확인하였으며, 나아가 해외 담수 호수의 박테리오파지 군집 분석에도 기여하였다.

주요어 : 박테리오파지, dsDNA 바이러스, 호수, 담수 지표수, 바이러스 메타게놈, 신종 박테리오파지, 전체 게놈 시퀀싱, 항생제 내성 유전자