Metagenomic characterization of the viral community of the South Scotia Ridge

Qingwei Yang[a], Chen Gao[a], Yong Jiang[a,b,c]#, Min Wang[a,b,c]#, Xinhao Zhou[a], Hongbing Shao[a], Zheng Gong[a], Andrew McMinn[a,d]

[a]College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China
[b]Key Lab of Polar Oceanography and Global Ocean Change, Ocean University of China, Qingdao 266003, China
[c]Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao 266003, China
[d]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia

Running Head: The viral community of the South Scotia Ridge

#Address correspondence to Yong Jiang, yongjiang@ouc.edu.cn and Min Wang, mingwang@ouc.edu.cn.

## Abstract

Viruses are the most abundant biological entities in aquatic ecosystems and harbor an enormous genetic diversity. While their great influence on the marine ecosystems is widely acknowledged, current information about their diversity remains scarce. A viral metagenomic analysis of two surfaces and one bottom water sample was conducted from sites on the South Scotia Ridge (SSR) near the Antarctic Peninsula, during the austral summer 2016. The taxonomic composition and diversity of the viral communities were investigated and a functional assessment of the sequences was determined. Phylotypic analysis showed that most viruses belonging to the order *Caudovirales*, in particular, the family *Podoviridae* (41.92-48.7%), which is similar to the viral communities from the Pacific Ocean. Functional analysis revealed a relatively high frequency of phage-associated and metabolism genes. Phylogenetic analyses of phage TerL and Capsid_NCLDV (nucleocytoplasmic large DNA viruses) marker genes indicated that many of the sequences associated with *Caudovirales* and NCLDV were novel and distinct from known complete phage genomes. High *Phaeocystis globosa* virus virophage (Pgvv) signatures were found in SSR area and complete and partial Pgvv-like were obtained which may have an influence on host-virus interactions in the area during summer. Our study expands the existing knowledge of viral communities and their diversities from the Antarctic region and provides basic data for further exploring polar microbiomes.

**Importance** In this study, we used high-throughput sequencing and bioinformatics analysis to analyze the viral community structure and biodiversity of SSR in the open sea near the Antarctic Peninsula. The results showed that the SSR viromes are novel, oceanic-related viromes and a high proportion of sequence reads was classified as unknown. Among known virus counterparts, members of the order *Caudovirales* were most abundant which is consistent with viromes from the Pacific Ocean. In addition, phylogenetic analyses based on the viral marker genes (TerL and

MCP) illustrate the high diversity among *Caudovirales* and NCLDV. Combining deep sequencing and a random subsampling assembly approach, a new Pgvv-like group was also found in this region, which may a signification factor regulating virus-host interactions.

## Introduction

Viruses exist wherever other life is found, including in the deep ocean and polar areas. Arguably, viruses are by far the most numerous, genetically diverse, and pervasive biological entities in the aquatic ecosystem(1, 2).They are critical mortality agents of both eukaryotes and prokaryotes, affecting the abundance and diversity of microbial communities as well as global biogeochemical processes and energy fluxes by causing lysis of a large proportion of both autotrophic and heterotrophic prokaryotes, shunting nutrients between particulate and dissolved phases(3-8) and modifying the efficiency of the carbon pump(9). Furthermore, bacteria and protists as vehicles for viral reproduction, and their genetic diversity was shaped by virus-mediated horizontal gene transfer, allowing viral genes to spread far and wide(2, 10).

The ecology of prokaryotes and protists, especially Antarctic phytoplankton during summer(11-16), together with the major role of viruses in prokaryotes and eukaryotic phytoplankton mortality(17-20) have been well studied. However, due to the geographical location and the difficulty of culturing viral hosts, an understanding of virus diversity and viral community structures in coastal regions of Antarctica including Sub-Antarctic areas is still lacking. So far, there have been few studies, based on culture-independent methods such as metagenomics and single-cell genomics, were conducted to analyze the DNA and RNA viral communities in Antarctic environments including freshwater habitats(21-25), the Southern Ocean closed to Western Antarctic Peninsula(26), sediment as well as soil(27, 28). These studies all pointed toward a high viral biodiversity in these Antarctic ecosystems. However, despite the virome diversity information derived from these special habitats in Antarctic, there have been few studies from the open sea near the Antarctic Peninsula.

In this study, we conduct an analysis of three viromes from South Scotia Ridge (SSR) seawater samples including two from the surface and one from the bottom (water depth=500 m) in an area influenced by Antarctic Circumpolar Current flow (ACC)(29). The major steps of the analysis were: (i) the determination of the composition of these three viromes and the dominant viral species, (ii) a comparison of these viromes with those from different habitats, (iii) the potential functional analysis, (iv) a phylogenetic and/or genomic analysis of the major viral groups that were present in these viromes.

## Results

### Overview of SSR viromes

The metagenome of three seawater samples collected from the SSR in Antarctic Peninsula region, ranged in temperature from -0.04 to -0.57 °C and in salinity from 34.37‰ to 34.57‰ (Fig. S1, Table S1). After extraction a total of 129,710,606 paired-end 150 bp sequences, with 109,923,264(84.75%) reads passing the quality screening (Table 1), were obtained.

Best BLAST Hit (E-value < 1e$^{-3}$) affiliations of unassembled high-quality reads from the three data sets are consistent with viral metagenomes analyses published so far, as more than three-quarters of the reads (75.7–88.24%) did not show any significant sequence similarity to current NCBI nr data (Fig. 1a). According to the NCBI nr and viral RefSeq annotation, 3.31-10.87% and

2.68-6.61% of the reads were classified as viruses respectively (Fig. 1b). A comparison of the annotation results of NCBI nr and viral RefSeq, found that the virus sequences annotated with virus in the NCBI nr database were more abundant than those in viral RefSeq (Fig. 1c), indicating that there is a number of sequences belonging to an unidentified virus that viral RefSeq excluded, such as uncultured Mediterranean phage uvMED.

**Taxonomic Diversity Analysis**
The BLAST data results (against viral RefSeq) of the virome composition are visualized using the Krona tool (Fig. S2, Fig. S3 and Fig. S4) (30)and show that, as excepted, the majority of viral reads (93.69-95.16%) with significant hits belonged to double-stranded DNA (dsDNA) viruses with no RNA stage. These were largely comprised of members of the *Caudovirales* comprising the families *Podoviridae*, *Siphoviridae* and *Myoviridae*, with similarities to single-stranded DNA (ssDNA)viruses and RNA viruses were also observed (Table 2 and Table 3). *Podoviridae* sequences (41.92%-48.7%) were the most abundant in all three viromes followed by *Myoviridae* (22.92-29.46%) and *Siphoviridae* (11.92-14.08%). Viruses from the *Phycodnaviridae* (infecting algae) and *Mimiviridae* (infecting amoebas and algae) were more abundant in surface waters than bottom water (D39s: 3.57% and 0.16%; DA4s: 2.22% and 0.22%; DA4b: 1.32% and 0.10%, respectively). There was a significant proportion of virophages that prey on phycodnaviruses in surface water, approximately 2% in D39s (Table 2).The top ten most abundant viral species (Fig. S5) including *Puniceispirillum* phage HMO-2011(31) (*Podoviridae*, circular genome), a phage infecting a bacterium of the SAR116 clade, was the most abundant in the SSR virome (18.50-25.75%), nearly accounting for 25% in DA4s, and Pelagibacter phage HTVC008M (32) (*Myoviridae*, linear) aT4-like myovirus infecting a SAR11 bacteriophage was second most abundance (8.6-11.11%).

**Comparison with other Virome**
To compare viromes between present study and previously published data sets, twenty viromes from different habitats were selected from MetaVir (see Materials and Methods for details). The result showed that the three SSR viromes were most closely related to ocean surface samples from the previous studies, except for the samples from ETSP-OMZ and SAR (Fig. 2), from which statistically significant differences were measured ($p < 0.001$). At the ocean surface, virome taxonomic at the family level was dominated by the *Caudovirales* (*Myoviridae*, *Siphoviridae*, *Podoviridae*), which collectively contributed 43.74-92.03% of the genomes. Viromes within special habitats including deep-ocean surface sediments, ETSP-OMZs, Antarctic freshwater, soil and hypolith are dominated by *Ciroviridae* and *Microviridae* members of ssDNA viruses which contributed 25.45-88.45% (Fig. 3, Table 3). Less than 5% of these viromes sequences showed any similarity (E-value $< 1e^{-3}$) to the SSR viromes (Table 4).

**Contigs and function analysis**
As the contigs assembled by the random subsampling approach could still contain redundant sequences derived from the same (or closely related) populations contigs derived from the same population were merged into clusters with 90% global average nucleotide identity by cd-hit-est, resulting in 145,023(D39s), 135,910(DA4s) and 234,648(DA4b) non-redundant genome fragments (>500 bp) respectively (Table 1). Of these 43.32%(D39s), 35.07%(DA4s) and

55.49%(DA4b) quality-filtered reads were assigned to nr contigs.

The putative functions of the annotated ORFs from the nr contigs dataset were predicted using MG-RAST, which assigns sequences to metabolic categories based on their Best BLAST Hit against the SEED database (E-value < 1e$^{-5}$). Using the subsystems approach, nearly 25% (17.54-26.46%) of the annotated proteins fell into the 'Phage, Prophage, Transposable elements or Plasmids' (Fig. 4). Phage structural, integration/excision and DNA metabolism-related proteins were most commonly identified and 10-11.96% of them classified into 'Clustering-based subsystems' with phage endolysin commonly found in this category. The other SEED functional annotation categories showed the metabolism of amino acids, carbohydrates, cofactors, vitamins, proteins, RNA, DNA, and nucleosides/nucleotides were the dominant annotations. In these categories, many proteins could be phage-related (or possible cellular origin) such as DNA polymerases and helicase. These hits were also found in the Pfam and COGs databases (see IMG system), with 'replication, recombination and repair' being the most common protein categories identified.

### Phylogenetic tree analysis

**Terminase phylogeny**. An ML phylogenetic analysis of the phage large terminase subunits identified in this work is shown in figure 5. The topology of the phylogenetic tree clearly shows that the majority of the SSR viromes' TerL amino acid sequences were widely distributed among the *Myo-*, *Sipho-* and *Podoviridae*, most of which were phylogenetically distant to known complete phage genomes (deposited in the NCBI database), and six new groups (in blue) did not cluster with any known species. This is supported by high bootstrap values, highlighting but important uncharacterized diversity for *Caudovirales* in SSR.

**Capsid_NCLDV phylogeny**. ML phylogenetic tree, based on the MCP including a group of putative MCP of Pgvv-like infected *Phaeocystis globosa* virus (Pgv), is shown here (Fig. 6). The MCP tree shows that several sequences from SSR viromes are closely related to known NCLDV, mainly those belonging to Phycodnaviruses, and these can be classified into Prasinovirus, Pgv, and Pgvv. The three clades differed from Phycodnaviruses and Mimivirus MCPs, forming three distinct groups with the well-supported clades; one of the clades, marked as Group3, was only found in the surface ocean of DA4 station. The Pgv group, included five new Pgvv-like MCPs, was distantly related to the Pgv group and had a higher relative abundance at the surface than at the bottom.

### Novel Pgvv group

From the MCP phylogenetic tree one distinct group of virophage was defined for which there is one known related virophage genome. However, it was still very different from the known Pgvv. An alignment of Pgvv-like group genomes is shown in figure 7. The Pgvv-like group genomes appear to have a relatively high (37.36-38.17%) GC content expect for the GC content of Pgvv-like 04 genome (GC, 35.85%) was similar to Pgvv (GC, 35.8%). All virophages share four homologous proteins or domains: 1) packaging ATPase (ATPase), 2) lipase, 3) major capsid protein (MCP), 4) minor capsid protein (mCP). In addition, Pgvv-like 02 also contain the OLV11-like tyrosine recombinase (Yrec) gene which is only distantly related to the OLV11-like family(33). Three genes with functional annotation (in yellow), absent in the Pgvv genome, were carried by Pgvv-like sequences, including putative primase-helicase and DNA methyltransferase

genes in Pgvv-like 02 and recombination endonuclease VII gene in Pgvv-like 04. These further indicate that these viruses maybe belong to a new Pgvv-like group.

## Discussion

Marine viral communities are still largely undescribed and many basic issues, such as their global ocean distribution and their actual genetic and species richness, remain unanswered(2, 9, 34). With the advent of metagenomic methods, associated with high-depth sequencing and meta-analyses of bioinformatics, an increasing number of studies have been conducted(35, 36). So far, only a few of these have focused on viral communities from the Antarctic region and most of these are from special habitats, such as freshwater lakes, hydrothermal vents and soils. For example, high diversity viral communities have been described from a lake (21), soils(28), and hydrothermal vents sites(37). As regards the sea water viromes, the important role of temperate viruses (that is, those capable of both lysogeny and lytic replication) in western Antarctic Peninsula dsDNA viral communities has recently been revealed(26) and differences in viral community composition between the subtropical Indian and Southern Oceans has been identified(38).

The number of reads identified as either bacteria or eukaryote was similar to that reported from viral metagenomes of other environments(34, 39), indicating a certain degree of bacterial and eukaryotic contamination of the metagenomes. In addition, the high proportion of no annotation sequences and relatively low number of rRNA and tRNA genes (< 1%) matching sequences (Table S2) indicates bacterial and eukaryotic contamination, previously reported to occur with TFF-based concentration methods(40). Because bacterial genes can be packaged into generalized transduced phage particles(41, 42), the bacterial-like sequences might have come from excised prophages mistakenly annotated as bacterial and/or from genes of bacterial origins that were transferred to their phages(39).

BLAST searches showed that >75% of the sequences before assembly did not have homologs in current sequence databases. This result is consistent with the results of the previously published viral metagenomic projects(38, 43-45) and such a high proportion of unknown sequences re-emphasizes that most of the biodiversity in the viral world is still undiscovered. The SSR viromes were largely dominated by *Caudovirales*, including Myoviruses, Siphoviruses, and Podoviruses, which are the dominant viral type recovered during metagenomic analyses of marine environments(2, 43). In the three SSR viromes investigated here, the largest number of reads (>40%) were related to Podoviruses and ca.13% of reads were of Siphoviruses (viruses which infect photosynthetic bacteria such as *Prochlorococcus* and *Synechococcus*) (in bold, Table S3). Consistent with previous investigation (31, 46), *Puniceispirillum* phage HMO-2011, that infects "*Candidatus Puniceispirillum marinum*" strain IMCC1322 of the SAR116 clade, and the *Pelagibacter* phage group (HTVC008M, HTVC010P, HTVC011P and HTVC019P), infects SAR11 populations were widespread and most abundant in SSR. Both of SAR11 and SAR 116 clades play important roles in oceanic Dimethyl sulfide (DMS) production and biogeochemical sulfur cycles, especially via bacteria-mediated dimethylsulfoniopropionate (DMSP) degradation(47, 48). Interestingly, Pgv is the tenth most abundant viral species in SSR region (2.53% in D93s), infecting the temperate algal species *Phaeocystis globosa*(49). In the Antarctic, however, one of the most abundant *Phaeocystis* species is *P. antarctica* (50) and the *P. antarctica* virus has still not been isolated and identified, which may indicate the high genome similarity between *P. antarctica* virus and Pgv.

Despite being in a cold marine environment with an average temperature below 0 °C, the SSR viral community had a structure similar to those found in the Pacific Ocean. However, there were still significant differences in nucleic acid levels and it is likely that the genotype of many viruses changed to allowing them to infect psychrophiles and thus evolve into new viral groups. The previously studied viromes from deep-ocean surface sediments, ETSP-OMZs, Antarctic freshwater, soil and hypolith, in which ssDNA viruses played dominated roles, were clearly different from those of the SSR. However, all of those viromes (except those from the deep-ocean surface sediment) were amplified using multiple displacement amplification (MDA) with phi29 polymerase, so the genomes of the ssDNA viruses could have been selectively amplified (51, 52), which may have led to an overestimation of the role of ssDNA viruses. Although existence bias from MDA in these studies and the prevalence of *Caudovirales* sequences has been observed in most marine viromes, previously published research on global morphological analysis of marine viruses, conducted by the Tara Oceans Expedition, showed that non-tailed viruses (largely ssDNA and RNA) numerically dominate the upper oceans (53)and small, non-tailed viruses were undoubtedly underestimated in SSR region also.

The deep sequencing method, combined with a random subsampling assembly approach, made it possible to obtain a nearly complete viral genome and create phylogenetic analyses on marker genes. Analysis of the major viral groups found in the SSR viromes showed a very broad diversity and many previously unknow virotypes. The terminase gene which is responsible for DNA recognition and initiation of DNA packaging, is an essential component of all head-tail phages (*Caudovirales*), encoding the molecular movements that translocate DNA into empty capsids(54). There is a large diversity of terminases, which can be used to resolve different Caudoviruses groups(55). The NCLDV comprises a monophyletic group of viruses infecting both animals and a diverse range of unicellular eukaryotes, including the *Phycodna-*, *Mimi-*, *Asco-, Asfar-*, *Irido-* and *Poxviridae* families. The MCP of NCLDV (capsid_NCLDV), a redox protein that encodes complex DNA replication and transcription systems and involved in the formation of disulfide bond in virion membrane proteins, is relatively conservative among NCLDVs evolution(56-58). Using phylogenetic trees based on these two viral marker genes (TerL and MCP), illustrated the high diversity among *Caudovirales* and NCLDV. A high proportion of TerL sequences were distributed both far from the reference and also far from each other, highlighting both the richness of *Caudovirales* in the SSR communities and also the absence of closely-related reference sequences. In addition, some SSR virome sequences appear to form a new clade (Group 6) related to the T4 viruses, one of the best described *Caudovirales* families.

The topology of the MCP tree and genomic comparisons strongly suggest that the five putative virophage genomes are more closely related to the Pgvv than to other NCDLV families, including the Pgvv host. The Pgvv-like group also has a high relative abundance. The Lotka-Volterra simulation demonstrated that virophages promote secondary production through the microbial loop by reducing overall mortality of the algal cell after a bloom and increasing the frequency of blooms during the summer(18). According to the above model, it can be inferred that the Pgvv-like group plays a previously unrecognized role in regulating virus-host interactions in SSR area during summer.

## Methods and materials
### Sample Collection and Sequencing

Seawater samples, including two surface waters and one bottom water (Table S1), were collected during the austral summer (December 2016) from two sites (D39 close to the edge of the Powell Basin and DA4 near to Clarence and Elephant Islands, Fig. S1) on the southern flank of SSR situated between South America and Antarctica Peninsula. Seawater temperature and salinity were recorded with a CTD profiler (SBE9/11 plus V5.2, Sea-Bird Inc., USA). Water for biological and chemical analysis was collected with Niskin bottles attached to the CTD profiler and was prefiltered with a 20 μm mesh to remove large particles.

The virome samples were processed immediately according to Sun et al(59). Briefly, the samples were sequentially filtered through 3 μm and 0.22 μm pore size filters to remove any microorganisms, and then a Two-step tangential flow filtration (TFF) with 50-kDa cartridge (Millipore, MA, USA) was used to concentrate the viruses to a final volume of ca. 50 ml and stored at -80 °C. The samples were further concentrated by Polyethylene glycol (PEG-8000) precipitation (10% w/v) and incubated at 4 °C overnight. The concentrate was then centrifuged at 8000 g for 80 min at 4 ºC and suspended in 200 μl SM buffer. Finally, DNA was extracted using the phenol/chloroform/isoamyl method and precipitated with ethanol without random amplification. High-throughput sequencing was performed by Novogene (Beijing, China) using Illumina Hiseq X ten (Paired End sequencing, 2×150 bp).

**Virome Composition Analysis**

A series of quality-screenings were undertaken to further remove low quality reads; these followed the quality control protocols of the sequencing company. First, adaptor sequences were trimmed by cutadapt (v 1.14). Subsequently, sequences that had a Phred score of at least 20 with a minimum length of 100 bp and had a Phred score of at least 30 with a minimum length of 88 bp were removed from libraries.

In order to avoid chimeras, SSR virome sequences were analyzed without assembly and queried by Diamond (60)against the NCBI non-redundant(nr) protein database(ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) and RefSeq complete viral genomes protein references (viral RefSeq) database(ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/), setting a maximum E-value of $1e^{-3}$. Taxonomic identification was assigned based on best similarities and the relative taxonomic was normalized against complete viral genome length.

**Virome Composition Analysis**

Twenty previously published viromes taxonomic compositions, with the same maximum E-value based on reads number were selected from MetaVir to compare with this study(61).These were obtained from a variety of habitats, including six temperate freshwater lakes(Lake Bourget and Lake Pavin(62), Antarctic lakes (21), Lough Neagh(63), Tilapia Channel(64)), nine seawater sites from eastern tropical South Pacific Oxygen minimum zones (ETSP-OMZ)(65),the Indian Ocean(66),the high salinity Jiulong River Estuary (67), Dunk Island, Fitzroy Island, LA26S and M1CS of POV(44), the Arctic Ocean and Sargasso Sea (SAR)(68)), three deep-sea surface sediments (Arctic Ocean, Black Sea and Mediterranean sea(69), soil and hypolith(70)) (Table 3). The relative taxonomic composition was normalized as described above. The similarity search algorithm BLAST was performed on the three SSR viromes against the twenty viromes obtained in MetaVir. The distance matrix of taxonomic composition with relative abundance was used in a nMDS analysis to plot viromes (the metaMDS function with a Bray-Curtis dissimilarity index of

the package VEGAN from R software(71)) and a PERMANOVA test (P-test) was also performed.

## Metagenomic assembly and function analysis

SSR virome assemblies were performed via a random subsampling approach as previously described(72), designed to obtain as long as possible contigs by reducing the microdiversity within the samples(73, 74). Briefly, the assembling strategy was based on random selection of a subset of the reads: 1% (75×), 5% (50×), 10% (50×), 25% (25×), 75% (25×), 100% (1×) from each sample and then assembling these subsets individually with IDBA_UD (v 1.1.2)(75) using the default parameters. Combing contigs derived from all the assemblies of the same samples and removing those < 500 bp. To this end, contigs were clustered at 90% global average nucleotide identity with cd-hit-est (v 4.7, options: -c 0.9 -n 8)(76). The relative abundance of each non-redundancy(nr) contigs was determined based on the mapping of the quality-filtered reads to the contigs, computed with bowtie2(v 2.3.3.1)(77) and SAMtools (78), using the default parameters (the total length of reads mapping to the contig divided by the contig length). Then the nr contigs with a relative abundance were uploaded to the IMG system (https://img.jgi.doe.gov/)(79), and analyzed with the standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4)(80). Finally, the IMG genomes are 3300028548, 3300028550 and 3300028925 respectively, were obtained. The functional content was further characterized using MG-RAST (81)(with MG-RAST accession number 4808192.3, 4808195.3 and 4808193.3 respectively), an online metagenome annotation service (http://metagenomics.anl.gov/), which compared data to the SEED Subsystems database using a maximum E-value of $1e^{-5}$, a minimum identity of 60%, and a minimum alignment length of 15.

## Phylogenetic trees

Two dsDNA markers: the phage terminase large-subunit domains (TerL) was present in phages of the order *Caudovirales* [Terminase_6, PF03237], and the major capsid protein (MCP) gene was present in large eukaryotic DNA virus [Capsid_NCLDV, PF04451]. Both of them were used to construct the phylogenetic trees and the TerL sequences were dereplicated at the 97% nucleotide level using cd-hit(76). These markers from the SSR virome genes were screened by the DOE-JGI Metagenome Annotation Pipeline and compared to the viral RefSeq database using BLASTP (E-value < $1e^{-5}$) to recruit relevant reference sequences. All sequences were aligned at the amino acid level using MUSCLE(82) (using default parameters) and both maximum likelihood(ML) trees (MCP and TerL) with 1000 bootstraps were constructed with the program FastTree (v2.1.10) (83)using a JTT+CAT model and an estimation of the gamma parameter. Finally, visualized and displayed using iTOL (Interactive Tree of Life)(84).

## Genomic comparison

The Pgvv-like genomes were annotated with RAST and predicted open reading frames (ORFs) were searched against the NCBI reference viral protein(taxid:10239) with online BLASTP(85). The partial functional annotations of the Pgvv reference sequence was obtained from Yutin et al(33). Visualization of genomes map comparisons was generated with EasyFig(86).

## Conclusion

Analysis of the SSR viromes has showed that there are novel, oceanic-related viromes. A high

proportion of sequence reads was classified as unknown, with only 3.31-10.87% having known virus counterparts, among these members of the order *Caudovirales* were most abundant. This pattern is consistent with previously described viromes from the Pacific Ocean as well as from a range of different biomes. The diversity of the *Caudovirales* and NCLDV in the SSR viromes is high, suggesting that in gelid environments viral diversity is high. However, the abundance and diversity of ssDNA and RNA viruses need further research. The strong signatures of Pgvv were found in the SSR, which may indicate that the virophage play an important role in regulating virus-host interaction.

## Acknowledgments

## Reference

1. Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399:541-548.

2. Suttle CA. 2005. Viruses in the sea. Nature 437:356-361.

3. Proctor LM, Fuhrman JA. 1990. Viral Mortality of Marine-Bacteria and Cyanobacteria. Nature 343:60-62.

4. Suttle CA. 1994. The Significance of Viruses to Mortality in Aquatic Microbial Communities. Microbial Ecology 28:237-243.

5. Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M, Kimmance SA, Middelboe M, Nagasaki K, Paul JH, Schroeder DC, Suttle CA, Vaque D, Wommack KE. 2008. Global-scale processes with a nanoscale drive: the role of marine viruses. Isme Journal 2:575-578.

6. Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. Nature 459:207-212.

7. Weitz JS, Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 Biol Rep 4:17.

8. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nature Reviews Microbiology 12:519-528.

9. Suttle CA. 2007. Marine viruses - major players in the global ecosystem. Nature Reviews Microbiology 5:801-812.

10. Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: Bacterial photosynthesis genes in a virus. Nature 424:741-741.

11. Kuparinen J, Bjornsen PK. 1992. Spatial-Distribution of Bacterioplankton Production across the Weddell-Scotia Confluence during Early Austral Summer 1988-1989. Polar Biology 12:197-204.

12. Archer SD, Leakey RJG, Burkill PH, Sleigh MA. 1996. Microbial dynamics in coastal waters

of East Antarctica: Herbivory by heterotrophic dinoflagellates. Marine Ecology Progress Series 139:239-255.

13. Vaque D, Calderon-Paz JI, Guixa-Boixereu N, Pedros-Alio C. 2002. Spatial distribution of microbial biomass and activity (bacterivory and bacterial production) in the northern Weddell Sea during the austral summer (January 1994). Aquatic Microbial Ecology 29:107-121.

14. Christaki U, Obernosterer I, Van Wambeke F, Veldhuis M, Garcia N, Catala P. 2008. Microbial food web structure in a naturally iron-fertilized area in the Southern Ocean (Kerguelen Plateau). Deep-Sea Research Part Ii-Topical Studies in Oceanography 55:706-719.

15. Montes-Hugo M, Doney SC, Ducklow HW, Fraser W, Martinson D, Stammerjohn SE, Schofield O. 2009. Recent Changes in Phytoplankton Communities Associated with Rapid Regional Climate Change Along the Western Antarctic Peninsula. Science 323:1470-1473.

16. Moschonas G, Gowen RJ, Paterson RF, Mitchell E, Stewart BM, McNeill S, Glibert PM, Davidson K. 2017. Nitrogen dynamics and phytoplankton community structure: the role of organic nutrients. Biogeochemistry 134:125-145.

17. Guixa-Boixereu N, Vaque D, Gasol JM, Sanchez-Camara J, Pedros-Alio C. 2002. Viral distribution and activity in Antarctic waters. Deep-Sea Research Part Ii-Topical Studies in Oceanography 49:827-845.

18. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA, Cavicchioli R. 2011. Virophage control of antarctic algal host-virus dynamics. Proceedings of the National Academy of Sciences of the United States of America 108:6163-6168.

19. Malits A, Christaki U, Obernosterer I, Weinbauer MG. 2014. Enhanced viral production and virus-mediated mortality of bacterioplankton in a natural iron-fertilized bloom event above the Kerguelen Plateau. Biogeosciences 11:6841-6853.

20. Vaque D, Boras JA, Torrent-Llagostera F, Agusti S, Arrieta JM, Lara E, Castillo YM, Duarte CM, Sala MM. 2017. Viruses and Protists Induced-mortality of Prokaryotes around the Antarctic Peninsula during the Austral Summer. Frontiers in Microbiology 8.

21. Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A, Alcami A. 2009. High Diversity of the Viral Community from an Antarctic Lake. Science 326:858-861.

22. Zawar-Reza P, Arguello-Astorga GR, Kraberger S, Julian L, Stainton D, Broady PA, Varsani A. 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). Infection Genetics and Evolution 26:132-138.

23. de Carcer DA, Lopez-Bueno A, Pearce DA, Alcami A. 2015. Biodiversity and distribution of polar freshwater DNA viruses. Science Advances 1.

24. de Carcer DA, Lopez-Bueno A, Alonso-Lobo JM, Quesada A, Alcami A. 2016. Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the Antarctic Peninsula. Fems Microbiology Ecology 92.

25. Cavicchioli R, Erdmann S. 2015. The discovery of Antarctic RNA viruses: a new game changer. Molecular Ecology 24:4809-4811.

26. Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. 2016. Seasonal time

bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. Isme Journal 10:437-449.

27.     Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M, Breitbart M. 2014. Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. Isme Journal 8:2093-2103.

28.     Adriaenssens EM, Kramer R, Van Goethem MW, Makhalanyane TP, Hogg I, Cowan DA. 2017. Environmental drivers of viral community composition in Antarctic soils identified by viromics. Microbiome 5.

29.     Howe JA, Livermore RA, Maldonado A. 1998. Mudwave activity and current-controlled sedimentation in Powell Basin, northern Weddell Sea, Antarctica. Marine Geology 149:229-241.

30.     Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. Bmc Bioinformatics 12.

31.     Kang I, Oh HM, Kang D, Cho JC. 2013. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. Proceedings of the National Academy of Sciences of the United States of America 110:12343-12348.

32.     Eggleston EM, Hewson I. 2016. Abundance of Two Pelagibacter ubique Bacteriophage Genotypes along a Latitudinal Transect in the North and South Atlantic Oceans. Frontiers in Microbiology 7.

33.     Yutin N, Shevchenko S, Kapitonov V, Krupovic M, Koonin EV. 2015. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. Bmc Biology 13.

34.     Edwards RA, Rohwer F. 2005. Viral metagenomics. Nature Reviews Microbiology 3:504-510.

35.     Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. Current Opinion in Virology 2:63-77.

36.     Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nature Reviews Microbiology 13:147-159.

37.     Millard AD, Pearce D, Zwirglmaier K. 2016. Biogeography of bacteriophages at four hydrothermal vent sites in the Antarctic based on g23 sequence diversity. Fems Microbiology Letters 363.

38.     Flaviani F, Schroeder DC, Lebret K, Balestreri C, Highfield AC, Schroeder JL, Thorpe SE, Moore K, Pasckiewicz K, Pfaff MC, Rybicki EP. 2018. Distinct Oceanic Microbiomes From Viruses to Protists Located Near the Antarctic Circumpolar Current. Frontiers in Microbiology 9.

39.     Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C. 2013. Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. Isme Journal 7:359-369.

40.     Hurwitz BL, Deng L, Poulos BT, Sullivan MB. 2013. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environmental Microbiology 15:1428-1440.

41.     Beumer A, Robinson JB. 2005. A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. Applied and Environmental Microbiology 71:8301-8304.

42.     Ghosh D, Roy K, Williamson KE, White DC, Wommack KE, Sublette KL, Radosevich M.

2008. Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzN genes in viral-community DNA. Applied and Environmental Microbiology 74:495-502.

43. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99:14250-14255.

44. Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. Plos One 8.

45. Roux S, Enault F, Ravet V, Colombet J, Bettarel Y, Auguet JC, Bouvier T, Lucas-Staat S, Vellet A, Prangishvili D, Forterre P, Debroas D, Sime-Ngando T. 2016. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environmental Microbiology 18:889-903.

46. Zhao YL, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean. Nature 494:357-360.

47. Howard EC, Sun SL, Biers EJ, Moran MA. 2008. Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. Environmental Microbiology 10:2397-2410.

48. Choi DH, Park KT, An SM, Lee K, Cho JC, Lee JH, Kim D, Jeon D, Noh JH. 2015. Pyrosequencing Revealed SAR116 Clade as Dominant dddP-Containing Bacteria in Oligotrophic NW Pacific Ocean. Plos One 10.

49. Baudoux AC, Brussaard CPD. 2005. Characterization of different viruses infecting the marine harmful algal bloom species Phaeocystis globosa. Virology 341:80-90.

50. DiTullio GR, Grebmeier JM, Arrigo KR, Lizotte MP, Robinson DH, Leventer A, Barry JB, VanWoert ML, Dunbar RB. 2000. Rapid and early export of Phaeocystis antarctica blooms in the Ross Sea, Antarctica. Nature 404:595-598.

51. Lasken RS, Stockwell TB. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol 7:19.

52. Kim KH, Bae JW. 2011. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. Applied and Environmental Microbiology 77:7663-7668.

53. Brum JR, Schenck RO, Sullivan MB. 2013. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. Isme Journal 7:1738-1751.

54. Maniloff J, Ackermann HW. 1998. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. Archives of Virology 143:2051-2063.

55. Mizuno CM, Rodriguez-Valera F, Garcia-Heredia I, Martin-Cuadrado AB, Ghai R. 2013. Reconstruction of Novel Cyanobacterial Siphovirus Genomes from Mediterranean Metagenomic Fosmids. Applied and Environmental Microbiology 79:688-695.

56. Tidona CA, Schnitzler P, Kehm R, Darai G. 1998. Is the major capsid protein of iridoviruses a suitable target for the study of viral evolution? Virus Genes 16:59-66.

57. Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. Journal of Virology 75:11720-11734.

58. Larsen JB, Larsen A, Bratbak G, Sandaa RA. 2008. Phylogenetic analysis of members of

the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. Applied and Environmental Microbiology 74:3048-3057.

59.  Sun G, Xiao J, Wang H, Gong C, Pan Y, Yan S, Wang Y. 2014. Efficient purification and concentration of viruses from a large body of high turbidity seawater. MethodsX 1:197-206.

60.  Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60.

61.  Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27:3074-3075.

62.  Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. Plos One 7.

63.  Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, Watson C, Arkhipova K, Lavigne R, Kulakov LA. 2016. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. Plos One 11.

64.  Rodriguez-Brito B, Li LL, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pasic L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F. 2010. Viral and microbial community dynamics in four aquatic environments. Isme Journal 4:739-751.

65.  Cassman N, Prieto-Davo A, Walsh K, Silva GGZ, Angly F, Akhter S, Barott K, Busch J, McDole T, Haggerty JM, Willner D, Alarcon G, Ulloa O, DeLong EF, Dutilh BE, Rohwer F, Dinsdale EA. 2012. Oxygen minimum zones harbour novel viral communities with low diversity. Environmental Microbiology 14:3043-3065.

66.  Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, McCrow JP, Tovchigrechko A, Yooseph S, Venter JC. 2012. Metagenomic Exploration of Viruses throughout the Indian Ocean. Plos One 7.

67.  Cai LL, Zhang R, He Y, Feng XY, Jiao NZ. 2016. Metagenomic Analysis of Virioplankton of the Subtropical Jiulong River Estuary, China. Viruses-Basel 8.

68.  Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. 2006. The marine viromes of four oceanic regions. Plos Biology 4:2121-2131.

69.  Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, Tamburini C, Weinbauer M. 2008. Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454:1084-U27.

70.  Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D. 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl Environ Microbiol 80:6888-97.

71.  Dixon P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14:927-930.

72.  Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics shed light

on viral strategies throughout the oceans. Nature Communications 8.

73. Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nature Reviews Genetics 14:157-167.

74. de Crcer DA, Angly FE, Alcami A. 2014. Evaluation of viral genome assembly and diversity estimation in deep metagenomes. Bmc Genomics 15.

75. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-1428.

76. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-9.

77. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-U54.

78. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

79. Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang JH, Andersen E, Huntemann M, Varghese N, Hadjithomas M, Tennessen K, Nielsen T, Ivanova NN, Kyrpides NC. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Research 45:D507-D516.

80. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, Palaniappan K, Szeto E, Pillay M, Chen IMA, Pati A, Nielsen T, Markowitz VM, Kyrpides NC. 2016. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). Standards in Genomic Sciences 11.

81. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. Bmc Bioinformatics 9.

82. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

83. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. Plos One 5.

84. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research 44:W242-W245.

85. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

86. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27:1009-1010.

**FIGURES**
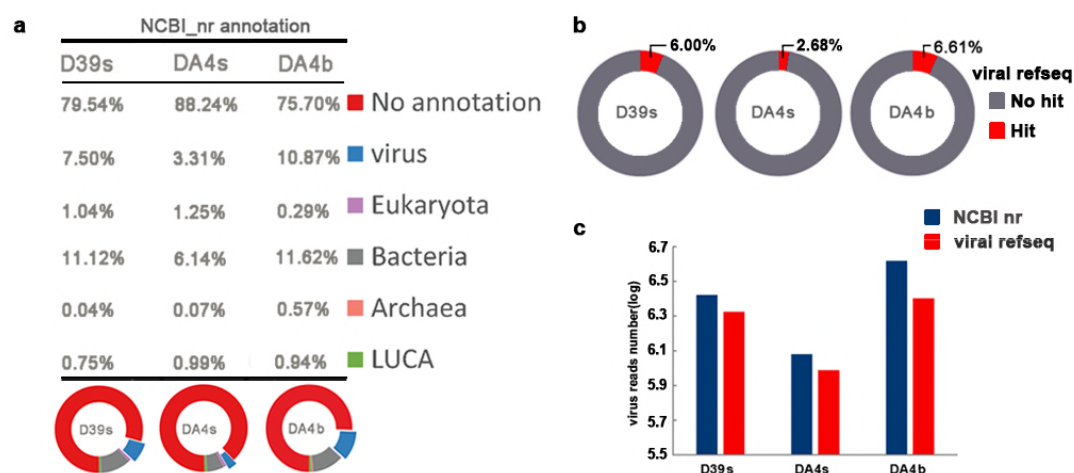


**Fig. 1** Taxonomic assignment of metagenomic reads (a) percentage of the sequence reads classified by the taxonomic grouping based on BLASTX similarity search with NCBI nr database (E-value$<1e^{-3}$). Sequences with no hits with E-value $> 1e^{-3}$ were regarded as unidentified reads ("no annotation" category in the table and red in the pie graphs). LUCA (green) denotes reads that could not unambiguously be assigned to a domain of life. (b) taxonomic assignment of metagenomic reads based on BLASTX similarity search with viral RefSeq database (E-value$<1e^{-3}$). (c) compare the annotation results of NCBI nr and viral RefSeq.
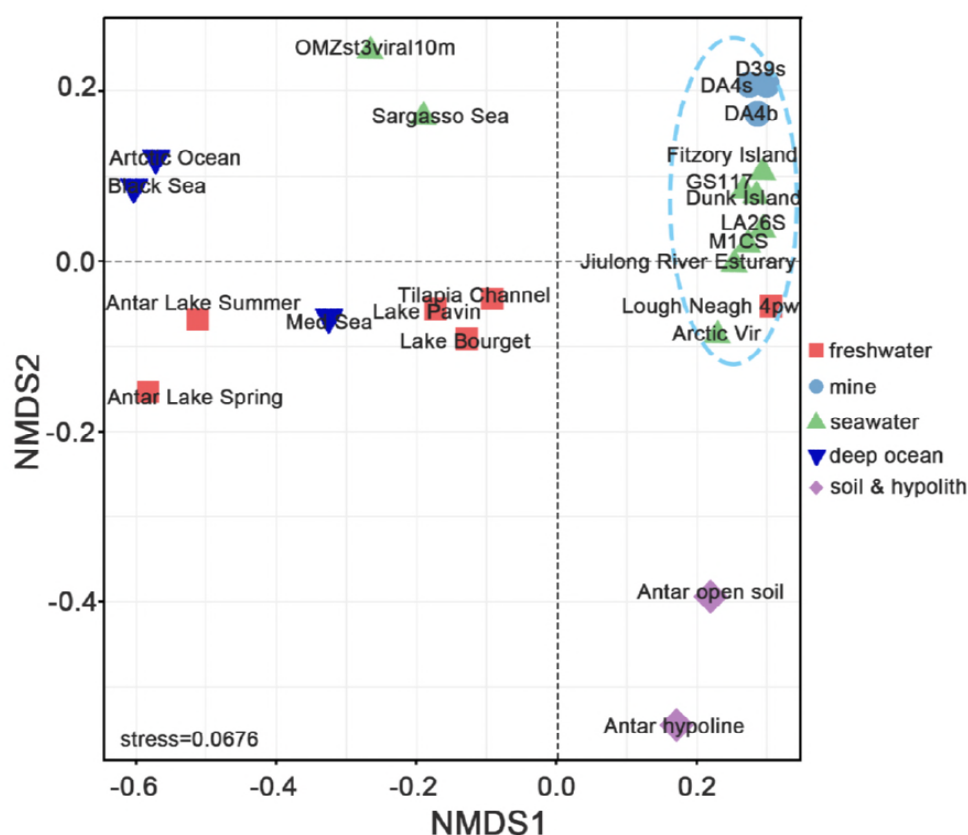


**Fig. 2** Comparison viromes between SSR area and other environmental viromes depending on

taxonomic composition. Twenty environmental viromes were available on MetaVir2, obtained from different habitats including freshwater, seawater, deep-sea surface sediments, soil and hypoline.
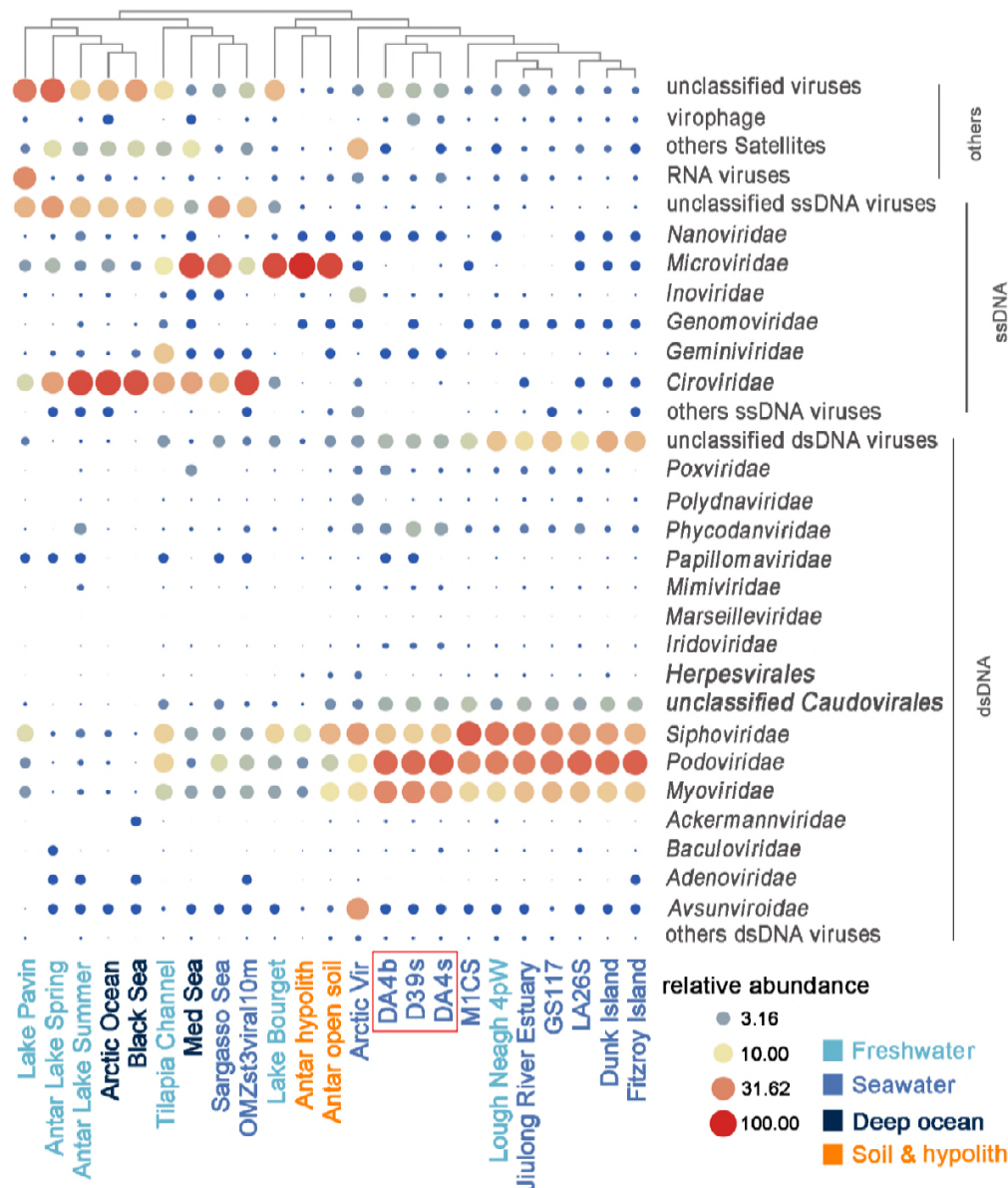


**Fig. 3** The relative abundance of viral sequences (normalized with genome length) largely at the family level in each different habitat virome. Points size indicate the value of relative abundance.
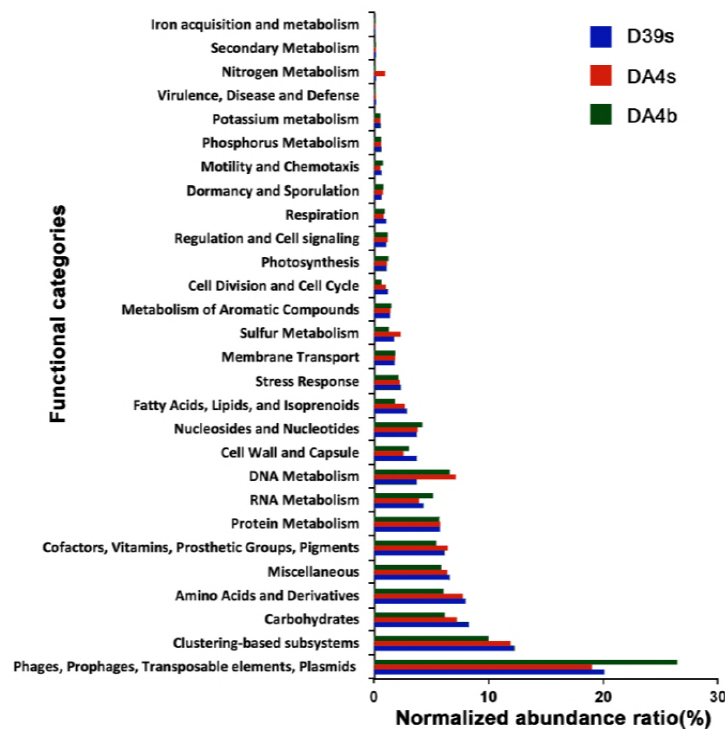
**Fig. 4** Composition of predicted functional genes of the SSR contigs. The CDSs were compared with the SEED database using subsystems in MG-RAST. The metabolic categorization is based on the sequences Best BLAST Hits in the SEED database curated subsystems (E-value <10⁻⁵).
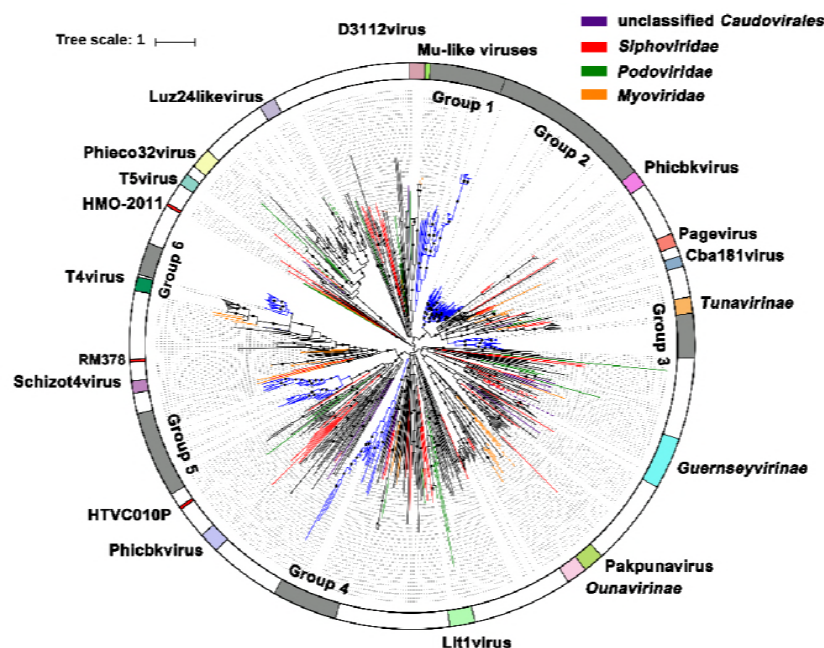


**Fig. 5 Terminase phylogeny.** A maximum-likelihood phylogenetic tree of *Caudovirales* terminase large-subunit domains (PF03237) is shown (1000 iterations, JTT+G model). Only bootstrap values of > 50% are indicated at the nodes of the tree and bootstrap scores greater than 90% are indicated with a black dot. Average branch length distance of leaves less than 0.4 were collapsed and shown as triangles. Reference sequences are marked (see color legend at the top) and new groups are highlighted in blue.
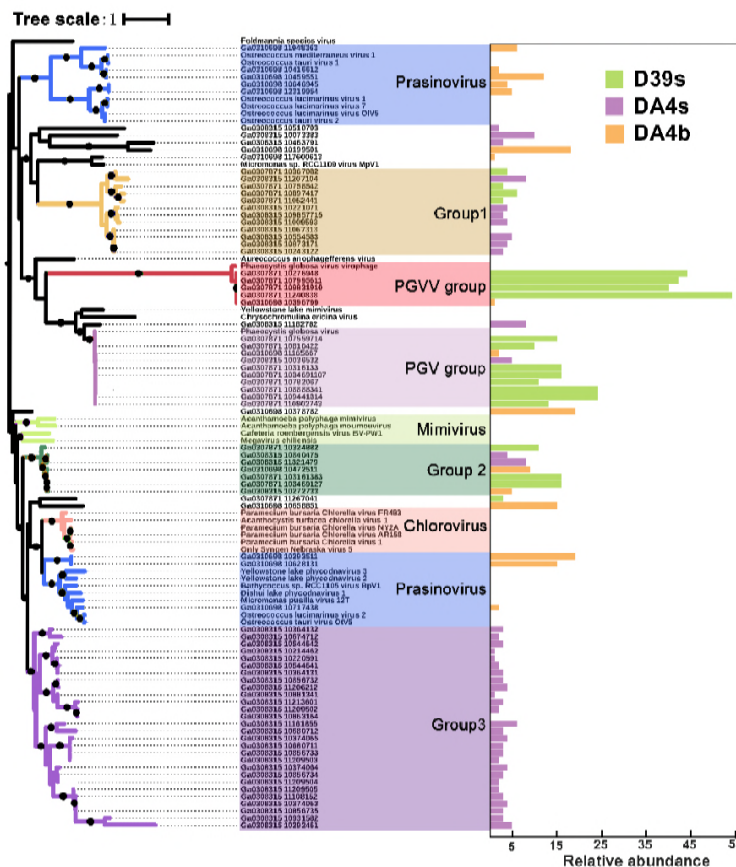
**Fig. 6 Capsid_NCLDV phylogeny**. A maximum-likelihood phylogenetic tree drawn from the capsid_NCLDV (PF04451) and six virophage putative major capsid (MCP) protein multiple alignment is shown (1000 iterations, JTT+G model). Bootstrap scores greater than 90% are marked with black dots. Each MCP is associated with an abundance profile (right) that displays the relative abundance of the contig across the three SSR viromes (based on normalized coverage).
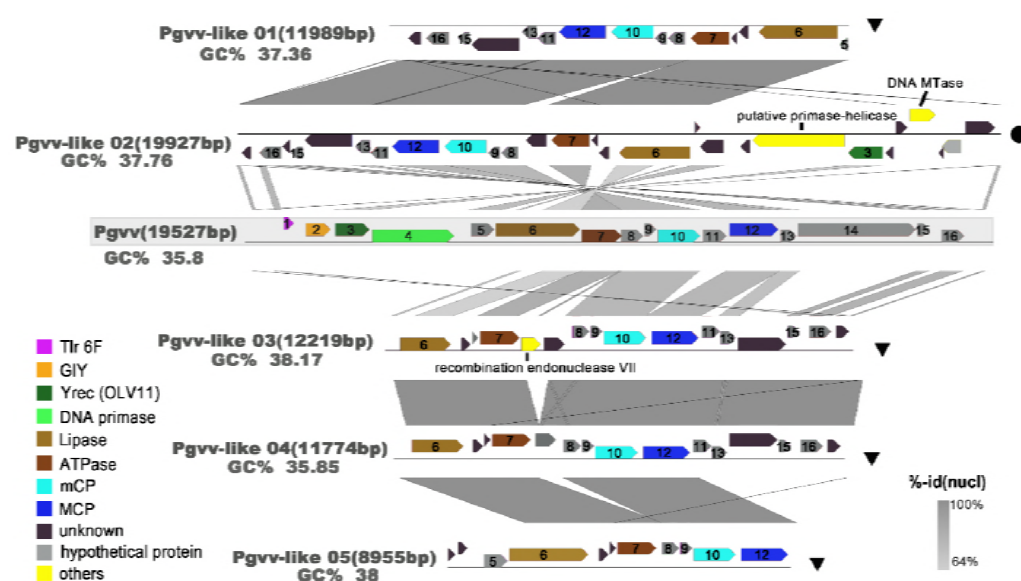
**Fig. 7** Overview of genomic synteny and similarities between Pgvv-like group. The complete Pgvv reference was covered by grey shadow. A color scale for percent identity (nucleic) is shown at the bottom right. The name, percent GC content (GC%), and length for each genome are also indicated. Genes are colored according to their functional affiliation. Tlr 6F, Toll-like receptor 6 family; GIY, GIY-YIG family nuclease; MCP, major capsid protein; mCP, minor capsid protein; Yrec (OLV11), OLV11-like tyrosine recombinase.

## TABLES

**Table 1.** The remaining number of quality-controlled reads and nonredundancy contigs

| Process | | SSR viromes | | |
| --- | --- | --- | --- | --- |
| | | D39s | DA4s | DA4b |
| Quality control | Raw reads | 42547324 | 43791908 | 43371374 |
| | Cut adaptor | 40670620(95.59%) | 41883752(95.64%) | 41353540(95.35%) |
| | Q20>20% | 35357306(83.10%) | 37567224(85.79%) | 38397082(88.53%) |
| | Q30>30% | 35334370(83.05%) | 36385270(83.09%) | 38203624(88.08%) |
| Assemble | All assembled contigs(>500bp) | 2418081 | 3699559 | 1693019 |
| | Non-redundancy contigs | 145023 | 135910 | 234648 |
| Mapping | Mapped reads | 15307526(43.32%) | 12760519(35.07%) | 21197289(55.49%) |

**Table 2.** Classification of reads from viromes hitting viral sequences

| Group | Order | Family | Relative abundance (%) | | |
| --- | --- | --- | --- | --- | --- |
| | | | D39s | DA4s | DA4b |
| dsDNA | *Caudovirales* | *Podoviridae* | 41.92 | 48.70 | 42.15 |
| | *Caudovirales* | *Myoviridae* | 28.34 | 22.92 | 29.46 |
| | *Caudovirales* | *Siphoviridae* | 11.92 | 13.38 | 14.08 |
| | *Caudovirales* | unclassified | 3.56 | 2.99 | 2.90 |
| | *Caudovirales* | *Phycodnaviridae* | 3.57 | 2.22 | 1.32 |
| | -- | *Mimiviridae* | 0.16 | 0.22 | 0.10 |
| | -- | *Poxviridae* | 0.14 | 0.22 | 1.13 |
| | -- | *Iridoviridae* | 0.37 | 0.44 | 0.30 |
| | unclassified | | 3.13 | 2.84 | 3.17 |
| ssDNA | -- | *Inoviridae* | 0.10 | 0.13 | 0.12 |
| | -- | *Microviridae* | 0.00 | 0.10 | 0.04 |
| | -- | *Circoviridae* | 0.01 | 0.11 | 0.01 |
| | -- | unclassified | 0.09 | 0.05 | 0.05 |
| **virophage** | -- | | 2.00 | 0.58 | 0.19 |
| others | *Ortervirales* | *Retroviridae* | 0.26 | 0.29 | 0.21 |
| | *Ortervirales* | *Caulimoviridae* | 0.07 | 0.76 | 0.04 |
| Unclassified phage/viruses | -- | | 3.71 | 3.01 | 4.07 |

**Table 3.** Distribution of sequences from the SSR viromes and twenty previously published viromes as determined by the indicated BLAST comparison to viral Refseq(E-value<1e-3)

| Virome | MetaVir ID. | Reads | Viral Hits | dsDNA viruses, no RNA stage | | | | | | | | | | | | | RNA | ssDNA | | | | | | Satellite | virophage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Ackermann | Baculo | Myo | Podo | Sipho | Herpes | Irido | Mimi | Papilloma | Phycodn | Polyom | Pox | others | | Circo | Gemini | Ino | Micro | Nano | others | | |
| Antar Lake Spring | 10 | 41322 | 8476 | 3 | 0 | 150 | 131 | 286 | 2 | 2 | 46 | 0 | 91 | 4 | 3 | 58 | 1 | 1736 | 29 | 9 | 631 | 10 | 1876 | 332 | 2 |
| Antar Lake Summer | 11 | 38475 | 8490 | 1 | 10 | 416 | 190 | 353 | 40 | 25 | 480 | 0 | 4237 | 11 | 26 | 170 | 4 | 1567 | 16 | 9 | 70 | 11 | 411 | 44 | 56 |
| Lake Pavin | 6 | 649290 | 162563 | 138 | 65 | 36625 | 18206 | 42746 | 246 | 259 | 2447 | 0 | 4998 | 80 | 191 | 10114 | 82 | 3780 | 52 | 217 | 2089 | 19 | 9430 | 313 | 1610 |
| Lake Bourget | 7 | 593084 | 222162 | 207 | 79 | 44714 | 32811 | 51952 | 309 | 468 | 2269 | 4 | 6719 | 118 | 265 | 16612 | 44 | 762 | 16 | 16 | 53771 | 62 | 805 | 80 | 880 |
| Tilapia Channel | 33 | 264844 | 24300 | 57 | 11 | 5210 | 7447 | 6180 | 15 | 74 | 301 | 0 | 342 | 34 | 14 | 1826 | 8 | 512 | 271 | 42 | 506 | 3 | 328 | 63 | 10 |
| Lough Neagh | 4925 | 2295055 | 565796 | 587 | 1154 | 104482 | 163643 | 193720 | 2223 | 1403 | 2600 | 51 | 11118 | 99 | 579 | 77234 | 62 | 1 | 7 | 146 | 31 | 0 | 385 | 0 | 269 |
| Jiulong River Estuary | 6305 | 498957 | 156678 | 1164 | 179 | 42292 | 43486 | 40181 | 361 | 209 | 2661 | 5 | 5731 | 84 | 236 | 18861 | 62 | 0 | 1 | 26 | 1 | 1 | 43 | 8 | 111 |
| GS117 | 1479 | 480375 | 186407 | 139 | 96 | 63122 | 55028 | 34963 | 276 | 224 | 1488 | 3 | 4300 | 63 | 143 | 25097 | 23 | 1 | 1 | 6 | 2 | 2 | 23 | 6 | 224 |
| M1CS | 1440 | 303519 | 99950 | 30 | 71 | 28472 | 33611 | 20883 | 196 | 272 | 1398 | 3 | 2980 | 58 | 81 | 10943 | 9 | 2 | 2 | 2 | 0 | 1 | 19 | 5 | 70 |
| Dunk Island | 1357 | 1165256 | 42053 | 138 | 20 | 11242 | 14440 | 7705 | 46 | 46 | 265 | 3 | 703 | 15 | 18 | 7332 | 2 | 0 | 1 | 1 | 0 | 0 | 3 | 3 | 37 |
| Fitzroy Island | 1358 | 82739 | 27019 | 10 | 14 | 7413 | 10783 | 3834 | 28 | 34 | 193 | 2 | 556 | 11 | 9 | 3994 | 3 | 0 | 1 | 1 | 0 | 0 | 5 | 0 | 34 |
| LA26S | 1396 | 165256 | 47012 | 11 | 74 | 15274 | 15050 | 7709 | 92 | 98 | 876 | 1 | 3047 | 69 | 93 | 4144 | 7 | 0 | 1 | 3 | 0 | 0 | 4 | 4 | 35 |
| D39s | -- | 17667185 | 2115364 | 2957 | 1219 | 980773 | 457752 | 160981 | 833 | 8387 | 19182 | 0 | 320111 | 208 | 2408 | 88748 | 626 | 6 | 0 | 159 | 2 | 0 | 6 | 6 | 9289 |
| DA4s | -- | 19101812 | 972948 | 1071 | 1725 | 410657 | 288733 | 91713 | 769 | 4537 | 14512 | 3 | 82102 | 232 | 2705 | 44752 | 344 | 27 | 3 | 110 | 48 | 0 | 1 | 1 | 1846 |
| DA4b | -- | 18192635 | 2523383 | 4589 | 1018 | 1334685 | 596892 | 245673 | 740 | 8775 | 19444 | 0 | 108271 | 84 | 3447 | 107902 | 657 | 6 | 2 | 252 | 53 | 0 | 5 | 5 | 1753 |
| Sargasso Sea | 12 | 397939 | 46952 | 23 | 7 | 12948 | 11573 | 3813 | 6 | 109 | 502 | 0 | 917 | 32 | 11 | 5280 | 14 | 1320 | 0 | 0 | 8355 | 1 | 1758 | 28 | 9 |
| Arctic Vir | 15 | 686209 | 8913 | 13 | 11 | 2513 | 1196 | 2244 | 47 | 33 | 643 | 1 | 861 | 183 | 69 | 769 | 21 | 4 | 1 | 78 | 0 | 0 | 2 | 24 | 3 |
| Arctic Ocean | 1158 | 79646 | 11271 | 3 | 11 | 459 | 244 | 559 | 21 | 26 | 219 | 1 | 494 | 53 | 49 | 237 | 9 | 5489 | 12 | 48 | 444 | 9 | 1321 | 210 | 0 |
| Black Sea | 1155 | 78436 | 12922 | 0 | 1 | 221 | 114 | 231 | 12 | 8 | 108 | 1 | 656 | 3 | 60 | 70 | 2 | 6768 | 106 | 5 | 265 | 12 | 1768 | 442 | 1 |
| Med Sea | 1161 | 65340 | 10542 | 5 | 10 | 2072 | 555 | 1731 | 78 | 121 | 321 | 1 | 518 | 10 | 126 | 371 | 6 | 507 | 0 | 3 | 3630 | 0 | 110 | 117 | 0 |
| OMZst3viral10m | 897 | 128441 | 19982 | 14 | 15 | 7341 | 3387 | 1911 | 16 | 48 | 463 | 0 | 1737 | 12 | 236 | 972 | 2 | 2277 | 0 | 4 | 714 | 4 | 523 | 40 | 65 |
| Antar hypolith | 2726 | 1057535 | 134697 | 150 | 170 | 13536 | 8392 | 47445 | 973 | 186 | 563 | 44 | 768 | 154 | 66 | 2337 | 68 | 8 | 2 | 153 | 58450 | 0 | 267 | 8 | 5 |
| Antar open soil | 2727 | 870687 | 147941 | 101 | 372 | 44977 | 16234 | 46475 | 860 | 553 | 1635 | 30 | 3383 | 35 | 478 | 9971 | 86 | 1 | 0 | 14 | 20944 | 0 | 39 | 2 | 7 |

**Table 4.** The table shows the percentage of reads in other published viromes obtained from MetaVir with a significant similarity (BLASTN, E-value<1e-3) to the SSR viromes

| Biome | Virome | MetaVir project ID. | No. of reads | South Scotia Sea | | |
|---|---|---|---|---|---|---|
| | | | | D39s | DA4s | DA4b |
| Antarctic seawater | D39s | - | 35334370 | 100% | 50.35% | 49.60% |
| Antarctic seawater | DA4s | - | 36385270 | 48.90% | 100% | 28.37% |
| Antarctic seawater | DA4b | - | 38203624 | 45.88% | 27.02% | 100% |
| Seawater | OMZst3viral10m | 897 | 128441 | 6.50% | 6.86% | 10.69% |
| Seawater | GS117 | 1479 | 480375 | 8.61% | 9.09% | 14.57% |
| Arctic seawater | Arctic Vir | 15 | 686209 | 1.45% | 1.33% | 2.10% |
| Seawater | Sargasso Sea | 12 | 397939 | 4.68% | 4.97% | 7.69% |
| POV seawater | Dunk Island | 1357 | 1165256 | 0.55% | 0.71% | 1.14% |
| POV seawater | Fitzroy Island | 1358 | 82739 | 7.44% | 9.11% | 14.04% |
| POV seawater | LA26S | 1396 | 165256 | 16.75% | 14.72% | 21.02% |
| POV seawater | M1CS | 1440 | 303519 | 14.77% | 15.59% | 23.12% |
| Deep Ocean | **Arctic Ocean** | 1158 | 79646 | 1.64% | 2.17% | 4.22% |
| Deep Ocean | **Black Sea** | 1155 | 78436 | 0.57% | 0.62% | 0.69% |
| Deep Ocean | **Med Sea** | 1161 | 65340 | 0.86% | 0.89% | 2.04% |
| Freshwater | Lake Bourget | 7 | 593084 | 0.94% | 0.87% | 1.66% |
| Freshwater | Lake Pavin | 6 | 649290 | 0.25% | 0.23% | 0.44% |
| Antarctic freshwater | **Antar Lake Spring** | 10 | 41322 | 0.07% | 0.06% | 0.16% |
| Antarctic freshwate | **Antar Lake Summer** | 11 | 38475 | 0.41% | 0.43% | 1.87% |
| Freshwater | Lough Neagh | 4925 | 2295055 | 0.31% | 0.30% | 0.52% |
| Freshwater | Jiulong River Estuary | 6305 | 498957 | 5.74% | 6.06% | 9.61% |
| Freshwater | **Tilapia Channel** | 33 | 264844 | 0.14% | 0.15% | 0.31% |
| Antarctic soil | **Antar open soil** | 2727 | 870687 | 0.43% | 0.42% | 0.59% |
| Antarctic hypolith | **Antar hypolith** | 2726 | 1057535 | 0.26% | 0.42% | 0.36% |