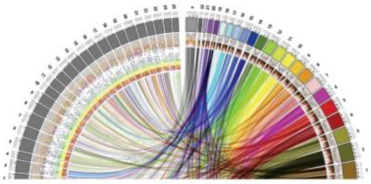

MSc IN BIOINFORMATICS

New Insights into Marine Viral Communities from Tropical and Subtropical Oceans

Author:

ESTER MARÍA LÓPEZ GARCÍA

Project supervisor:

FELIPE HERNANDES COUTINHO

Academic tutor:

JAIME MARTÍNEZ-URTAZA

Title: New Insights into Marine Viral Communities from Tropical and Subtropical Oceans

Author: Ester María López García

Date: July 2022

Approval of the project supervisor:

Signed,

Felipe Hernandes Coutinho

Author:

Signed,

Ester María López García

ABSTRACT

INTRODUCTION

Viruses, especially those that infect prokaryotes, are the most abundant and diverse biological entities in the oceans. Furthermore, viruses influence marine ecosystems through many different mechanisms (i.e., selective killing of their hosts, expression of metabolic genes during infection, acting as agents of genetic exchange, etc.). Significant advances have been made in the study of viral diversity from marine environments. Nevertheless, few studies have focused on the analysis of viral communities throughout depth gradients, and none have done so at the global scale. The aim of this study was to reveal novel marine viral diversity to provide insights on the influence of viruses over prokaryotic communities by means of mechanisms such as the listed above.

MATERIALS AND METHODS

Viral genomic sequences were identified within the assemblies from Malaspina vertical profiles metagenomes from the cellular fraction (0.22 µm). Next, protein sequences derived from the genomic sequences were annotated to identify metabolic genes of ecological relevance. Using state-of-the-art bioinformatics tools, the viral genomic sequences were classified taxonomically and linked to their putative hosts to estimate, considering the standardised relative abundance of viral genomes in the samples, the potential contributions of these viruses to biogeochemical cycles of global relevance. Furthermore, community ecology analyses were performed to assess beta- and alpha-diversity across samples from different water layers. Finally, associations among biological variables and environmental parameters were identified through correlation analyses.

RESULTS AND DISCUSSION

The Malaspina vertical profiles metagenomes yielded 101,219 viral genomic sequences, of which 299 represented complete genomes. Light availability was identified as the most important variable in driving differences in viral community composition. Furthermore, viral communities displayed clear shifts across gradients of temperature, salinity and nutrients concentration (NO_3, SiO_4 and PO_4). Taxonomic classification assigned most viral sequences to families of tailed viruses from the order *Caudovirales*, namely *Myoviridae* (46,466), *Podoviridae* (8,966) and *Siphoviridae* (7,648). Computational host prediction indicated that they infect abundant members of the marine microbiome, such as Cyanobacteria, Gammaproteobacteria, Alphaproteobacteria and Bacteroidetes.

CONCLUSION

These results provide new insights about the diversity and ecology of marine viruses throughout depth gradients and bring us closer to understanding their roles in biogeochemical cycles of global relevance.

Key words: Virus; Marine Ecology; Vertical Profiles; Host Prediction; Biogeochemical Cycles; Metagenomics.

TABLE OF CONTENTS

1. INTRODUCTION

Oceans cover more than 70% of the Earth's surface. They regulate the climate, provide a major portion of the protein consumed globally, and produce around half of the Earth's oxygen. Microorganisms drive the nutrient and energy cycles in the world's oceans, accounting for more than 90% of the sea's living biomass. Globally, marine viruses have been estimated to infect approximately 10^{23} microbes every second, removing 20%–40% of that biomass every day. Viruses, besides being agents of death, they are one of the world's largest reserves of unknown genetic variation (1).

Most ocean environmental variables are influenced directly or indirectly by depth and topography, including light penetration and photosynthesis, sedimentation, current movements and stratification, and hence temperature and oxygen gradients. As a result, these characteristics are likely to impact species distribution patterns and ocean production (2). The ocean is generally divided into five layers: Epipelagic zone (0 – 200 m deep), Mesopelagic zone (200 – 1,000 m deep), Bathypelagic zone (1,000 – 4,000 m deep), Abyssopelagic zone (4,000 – 6,000 m deep) and Hadalpelagic zone (>6,000 m deep). The vast ocean, distant from the coast, is referred as *pelagic*. Prefix “epi-” means "surface”; prefix “meso-”, "middle”; prefix “bathy-”, "deep”; prefix “abysso-”, "without bottom”; and prefix “hadal-”, “relating to the deepest region”. Furthermore, the *thermocline* is the transition zone between the epipelagic and mesopelagic zones. Some of the terms explained above are depicted in the following representation (Fig. 1.1).

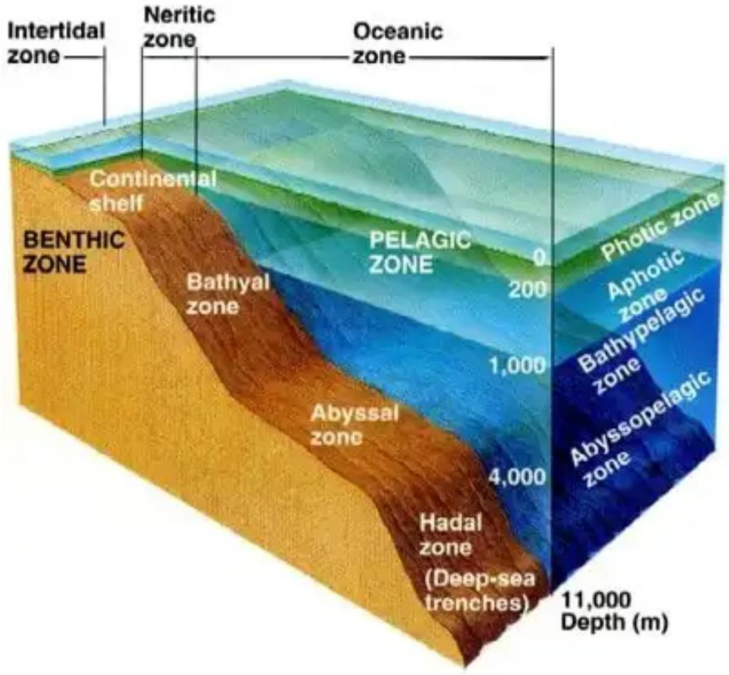

Figure 1.1. Tridimensional representation of marine water layers. The Photic zone encompasses the Epipelagic layer, while the Aphotic zone, the four water layers left (Mesopelagic, Bathypelagic Abyssalpelagic and Hadalpelagic). (Retrieved from (3)).

By international agreement, the largest saline water mass in the planet was divided into five oceans, which from smallest to largest are: the Arctic, Antarctic, Indian, Atlantic and Pacific. According to latitude (the angular distance of a place north or south of the Earth's equator), Indian, Atlantic and Pacific oceans can also be referred to as Tropical and Subtropical oceans. The Sun's heat drives the worldwide circulation of the Earth's seas and atmosphere. Much of that crucial solar energy first hits the tropics, where the Sun is practically directly above all year. Tropical ocean water temperatures often reach 20°C and remain rather stable throughout the year. (4, 5).

1.1. MARINE VIRAL ECOLOGY

The ecology of marine microorganisms is a complex scientific subject since it combines numerous disciplines such as oceanography, biogeochemistry, microbiology (including protistology and virology), physiology, evolution, and genomics (6).

The birth of microbial oceanography, at least in Central Europe, dates to the 19th century, when Bernard Fischer published his *Die Bakterien des Meeres* from 1894 (7). In regards to viruses, in the early 20th century, Russian oceanographers, such as B.L. Issatchenko (8), were among the first to study marine microbes, including protists and bacteriophages. The first genome of an isolated marine virus was published in 2000 (Rohwer et al. 2000). From then, progress in the understanding of marine viruses and their effects has been rapid and has been summarized in several comprehensive reviews (1, 9-16).

Despite these advances, many challenges remain in the study of viral communities (i.e., the documentation of diversity, host range, infection dynamics of marine viruses, as well as the subsequent effects of infection on both host cell metabolism and oceanic biogeochemistry).

1.1.1. Modes of infection

It has been shown that viruses exceed prokaryotes by an average ratio of 10:1 (17) and are far more prevalent than phytoplankton, zooplankton, or higher trophic level species. Viruses are thought to infect all marine organisms, ranging from small phytoplankton that play important roles in global carbon cycling to more commercially valuable organisms like invertebrates, fish, and whales (18). However, given that bacteria are by far the most prevalent cellular entities, the majority of the viruses are bacteriophages (i.e., viruses that infect bacteria).

Some viruses can undergo two mechanisms of replication, the lytic cycle and the lysogenic cycle (Fig. 1.2). In the first one, lytic viruses introduce their genetic material into the microbial cell and alter its metabolism to make new virions after effectively being in contact with their hosts' surface receptors. This type of infection cause nutrients to be released into the water column, making them available to competing microbes, a process known as viral shunt (explained in more detail in Section 1.1.2).

Regarding the lysogenic mechanism, temperate viruses can go latent in their host cells through a process known as lysogeny. In every infection, such viruses choose between the lytic and lysogenic cycles, in other words, whether to multiply and lyse their host or lysogenise and keep the host alive (19). The following diagram depicts the interplay between lytic and lysogenic cycles.

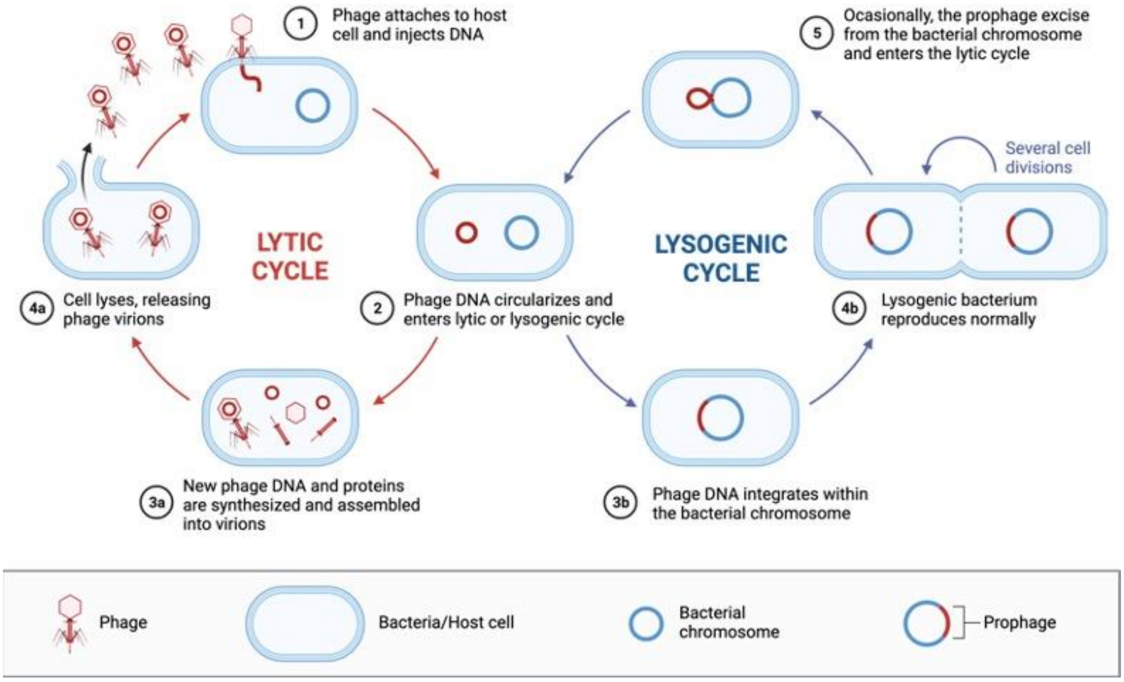
1 Phage attaches to host cell and injects DNA
LYTIC CYCLE
4a Cell lyses, releasing phage virions
2 Phage DNA circularizes and enters lytic or lysogenic cycle
3a New phage DNA and proteins are synthesized and assembled into virions
5 Ocasionally, the prophage excise from the bacterial chromosome and enters the lytic cycle
Several cell divisions
LYSOGENIC CYCLE
4b Lysogenic bacterium reproduces normally
3b Phage DNA integrates within the bacterial chromosome
Phage
Bacteria/Host cell
Bacterial chromosome
Prophage

Figure 1.2. The two main lifestyles of marine phages: lytic versus lysogenic. In the first step of both processes (1), phage injects its DNA into the cell. Secondly, phage DNA is circularised into the host cell (2). After that, in the lytic process, (3a) viral genetic material is replicated and proteins are synthesised, resulting in (4a) cell lysis and the release of offspring phage particles. With respect to temperate phages, they lysogenise their hosts, (3b) either integrating into the bacterial chromosome or remaining as an extrachromosomal element, (4b) where they divide as prophages with the bacterial cells until an environmental or (5) cellular trigger entails them to enter the lytic cycle. Although not depicted here, certain phages can cause a persistent infection in which phage particles are generated but the host bacterial cell does not die. (Reprinted from "Lytic and Lysogenic Cycle", by BioRender, April 2021, retrieved from (20)).

1.1.2. Roles of viruses in the marine food web

Despite in the previous section it has been claimed that the mechanism of mortality of bacteria is lysis due to phage lytic infections, there is another mechanism by which bacteria dye: grazing by protists. These two mechanisms together are thought to be responsible for approximately half of bacterial death. However, each process can also be studied separately to understand the operation of the complex network which flourish in the ocean.

Photosynthesis is an important process taking place in marine ecosystems. This chemical reaction is carried out by both bacterial autotrophs and eukaryotic phytoplankton, both of which fix carbon through this mechanism. After this fixation, grazing transports carbon up the food chain from bacteria to protozoa to zooplankton to larger animals. All trophic levels contribute Particulate Organic Matter (POM) pool, which sinks to the deep ocean, feeding into the biological carbon pump. Concomitantly, the microbial loop takes place. The relevance of this process is related to the two types of microbial mortality stated above: predation by unicellular eukaryotic grazers and viral lysis. When phage lyse bacteria, Dissolved Organic Matter (DOM) and inorganic nutrients from their cytoplasm are released in the water column. This readily available DOM is remineralised by prokaryotes within the microbial loop. This process, known as the viral shunt acts as the ocean's recycling mechanism of organic matter, since it reduces the transfer or organic matter and energy to the higher tropic levels, which otherwise would sink to the deep ocean. Thus, the viral shunt guarantees that primary productivity remains constant. (Fig. 1.3).

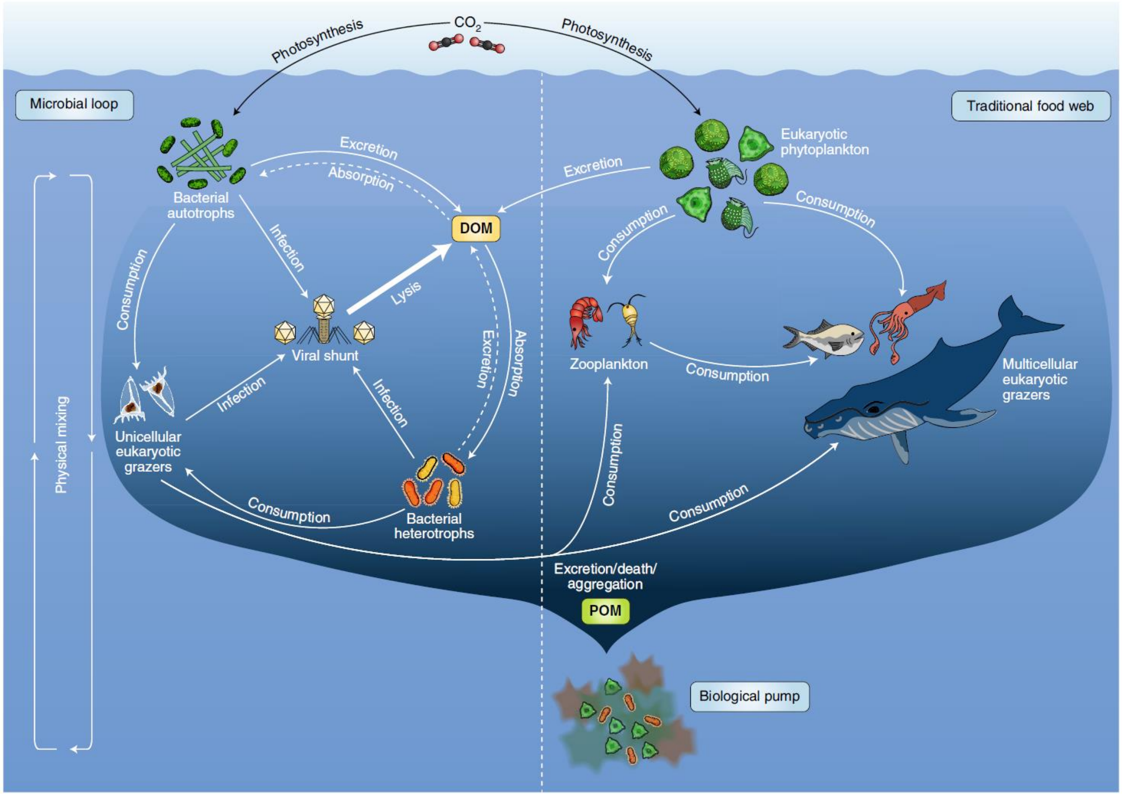

Figure 1.3. The marine food web. Both bacterial autotrophs and eukaryotic phytoplankton do photosynthesis in the waters, producing fixed carbon. DOM, the ocean's greatest carbon storage, is exclusively available for absorption by bacteria. Bacterial mortality can be classified into two types: predation by unicellular eukaryotic grazers and phage lysis. Regarding the first one (**right**, traditional food web), grazing transports carbon up the food chain from bacteria to protozoa to zooplankton to fish and bigger creatures, with all trophic levels contributing to the biological pump via POM sinking. With respect to the second one (**left**, microbial loop), when phages lyse bacteria, carbon and nutrients move via the viral shunt and are remineralized by bacteria within the microbial loop, allowing the viral shunt to serve as the ocean's recycling mechanism. (Figure retrieved from Breitbart et al., 2018 (15)).

Conversely, viruses have also been shown to contribute to the biological pump by producing adherent lysates that cluster and sink, effectively transporting organic carbon from the surface to the deep ocean and increasing the biological pump's efficiency (21, 22).

1.2. PROKARYOTE-VIRUS INTERACTIONS

For more than twenty years, the Kill-the-Winner (KtW) theory (23), which quantitatively defines the steady-state coexistence of many distinct phages and bacterial hosts in a particular niche, has led research in marine microbial and viral ecology. It is based in the Lotka-Volterra predator-prey dynamics. As stated by this theory, as a certain host becomes active, the number of phages capable of infecting that bacterial host increases. Phage infection causes the vulnerable host population to drop, which in turn promotes the phage population to fall since no more hosts are available. Consequently, as the initial dominant host population is no longer in the high command, a new host occupies the place. This one has the capability of being resistant to the phage which infected the previous host; therefore, this empty niche is taken up by a new emerging phage, and the cycle repeats (15).

Recent evidence has led an expansion of the KtW theory: the Piggyback-the-Winner theory (24). This theory is based on evidence that fewer viruses per host are observed when there are high levels of prokaryotes abundance, then host abundance is a primary driving force behind the shift from lytic to lysogenic infection (25, 26). This theory is more focused in the causes that lead to lytic-lysogenic infection switch and claims that, at high host abundance, there is an increased lysogenic infection. However, the conflict between KtW and PtW arises when the last is seen as a microbial strategy that confers lysogeny an advantage in the ecology of virus-host interaction.

At the present, the scientific community has not reached consensus regarding which hypothesis is the most suitable. In fact, the coexistence of KtW and PtW dynamics was discovered in an experimental environment (27), demonstrating the close relationship between microbial dynamics, diversity, and succession of the lysis-lysogeny switch. This shows that the KtW and PtW theories maybe are not mutually exclusive, but rather work together to explain viral control of the ocean's microbial ecology.

1.2.1. Auxiliary metabolic genes

Prokaryotic metabolic processes are rigorously regulated to provide a perfect balance of food intake, energy production, and biosynthetic activity, allowing for efficient synthesis and assembly of cellular components in precise proportions. Furthermore, with the aim of living and reproducing in dynamic natural conditions, prokaryotic organisms must be capable of rapidly changing their metabolism in reaction to variations in food availability and other environmental conditions (28).

Prokaryotic viruses must hijack the cellular machinery of their hosts to reproduce by creating viral particles. When the host cell is infected, it is termed a "virocell", a biological entity with a metabolism distinct from that of uninfected host cells (29). In the attempt of viruses to control the cellular metabolism, some of them encode and express Auxiliary Metabolic Genes (AMGs). These genes were discovered in cultivated viral isolates and can affect multiple aspects of the hosts molecular machinery, such as photosynthesis, phosphate scavenging and nitrogen metabolism, among others. These AMGs allow viruses to re-direct host metabolism towards pathways that promote viral proliferation (Fig. 1.4).

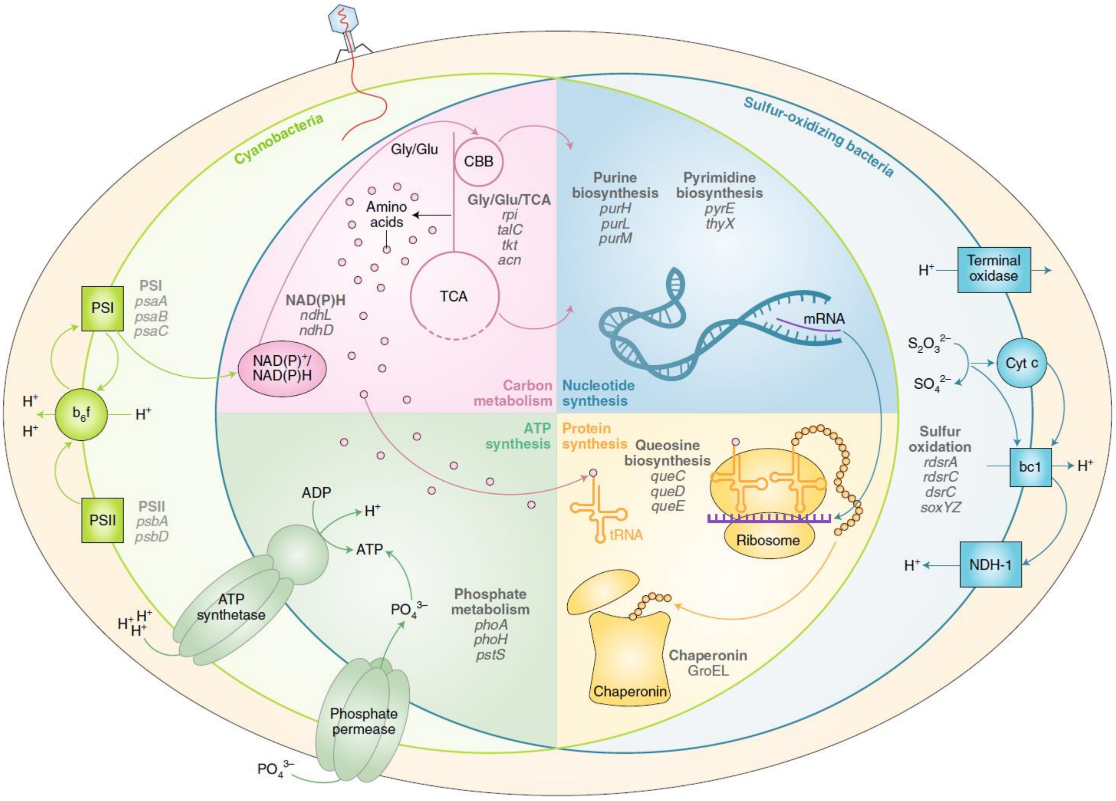

Figure 1.4. Examples of AMGs found among viruses of cyanobacterial (**left**) and sulfur-oxidizing (**right**) bacteria. Carbon metabolism, nucleotide metabolism, protein synthesis, ATP synthesis, photosynthesis and sulfur oxidation are depicted. The pathways in the central circle represent functions shared by both cells. Genes are written in italics. Gly, glycolysisn; Glu, gluconeogenesis; CBB, Calvin–Benson–Bassham cycle; TCA, tricarboxylic acid cycle; PSI/II, photosystem I/II; b6f, cytochrome b6f complex; Cyt c, cytochrome c; bc1, cytochrome bc1 complex; NDH-1, type I NAD(P)H dehydrogenase. (Figure retrieved from Breitbart et al., 2018 (15)).

The expression of AMGs contained in viral genomes, which actually are host-derived genes, contributes to a more targeted attack of these cells and ensures the maintenance of critical host cell processes that would otherwise be downregulated in response to a lytic infection mode (30). Phage-mediated redirection of host metabolism highlights the nutritional demand that production of phage progeny places on infected cells (31).

1.3. STUDYING VIRAL COMMUNITIES

The study of viral communities has recently begun to take off. Viruses were disregarded by microbial ecologists for decades due to a lack of adequate methods for quantifying and classifying them. Since the discovery that viruses are the most prevalent living entities in the oceans (32), microbiologists have begun to describe their ecological roles.

Another point worth noting is that, very often, technologies still had not been created or even they have been developed in other scientific fields, usually the medical one, and until they were not imported into the oceanography area, advances could not take place. One example of the last could be that of the flow cytometry, originally developed to enumerate human cells, and incorporated by Yentsch and Olson in 1983 (33), whose work allowed the discover of the cyanobacterium *Prochlorococcus.*

1.3.1. Diversity

Until the recent introduction of molecular biology tools, most microbial diversity was inaccessible. Even with the latter, however, there are considerable gaps in what these strategies can deliver. These blind spots have designated "microbial dark matter" (34). There are at least two reasons why it still there is a high uncertainty about microbes (prokaryotes, eukaryotes and viruses):

i) In some situations, universal probes and primers may not hybridise with the rRNA from all organisms which harbour it. Hence, in typical surveys, these microorganisms would go undiscovered. However, in case of viruses, they directly do not have rRNA nor any other universal taxonomic marker, hindering the advance in the viral field. The use of metagenomics, in both cases, is one answer to this challenge. Since primers are not utilised, all the nucleic acids of all microorganisms should be accessible for sequencing.

ii) Another way for microorganisms to go unnoticed is if they are so infrequent that they do not emerge in surveys. In case of viruses, the problem is not their number, since they exceed by several orders of magnitude the bacteria abundance (17), but their biomass. Due to their small size, sampling collections usually miss them during the process. One brute force way to solving this challenge is to get more sequences, as demonstrated by Crespo and colleagues in 2016 (35), who calculated that doubling the sequencing effort by a factor of four (up to 2 million rRNA tag sequences) would allow identification of 90% of the OTUs in the samples. A second technique, to improve specially virus detection would be make metagenome enrichments in viral fractions (i.e., by means of shotgun metagenomics applied specifically to the encapsidated fraction of viral DNA and/or RNA from a sample, throughout filtration, precipitation, and DNase/RNase treatment) as shown in Roux et al. 2021 (36).

1.3.2. Metagenomics

The study of viruses requires an understanding of how they interact with their hosts. Nonetheless, the restrictions already imposed in some prokaryotes by culturing are exacerbated when viruses are considered (37). Hence, it was only with the development of molecular tools that their variety could be accurately assessed. Culture-independent techniques have greatly improved the scientific community awareness of the variety of marine bacteria, archaea, microeukaryotes, and viruses, changing our view of Earth's evolution. Nowadays, the use of metagenomics to explore marine viruses is quickly expanding (38), resulting in massive volumes of data regarding the diversity and dynamics of viral communities. Furthermore, it has been demonstrated that their effect over host communities is far greater than previously thought.

However, one major challenge that has hampered the implementation of metagenomics is the extraordinarily low nucleic acid content of viruses compared to bacteria. To obtain enough nucleic acid for metagenomic sequencing while minimising contamination from bacteria and extracellular nucleic acids, three main steps have been used researchers consisting of (i) concentrating viruses from environmental samples, (ii) performing an accurate purification of concentrated viruses to reduce contamination, and (iii) either direct clone or amplify extracted viral nucleic acids before sequencing. (39, 40).

1.3.3. Computational advances

Despite metagenomics has been proposed in the previous section (Section 1.3.2) as one of the best ways of studying marine viruses, another problem arises besides the low nucleic acid content of viruses. It concerns that the vast majority of metagenomic sequences have no significant similarity to sequences in genomic databases (41). Thence, in order to tackle this problem, new bioinformatic strategies for extracting useful information from metagenomic data and comparing samples from diverse locations and investigations have been developed.

Some examples of solutions include making comparisons of raw metagenomic reads from different samples to assess unique or overrepresented sequences (42), using protein clusters to organise viral sequences based on Open Reading Frame (ORF) sequences similarity found among samples (41) and quantify host-associated viral diversity using viral-tagging metagenomics (38) which allowed the studied populations to be defined and detected within metagenomic sequences.

Furthermore, setting aside metagenomics issues, mathematical models, such as recently developed machine-learning approaches, can discover relationships between viruses, and other microorganisms, and environmental variables that are not clearly visible using simpler approaches such as correlations or exploratory analyses (32).

1.4. MALASPINA CIRCUMNAVIGATION EXPEDITION

Researchers are always challenged to sample the seas at suitable temporal and geographical scales. The deep ocean, which is likely the greatest ecosystem on Earth, has received far less attention than the upper layer (43). Because of the presence of many chemoautotrophic archaea and bacteria, the large role of particle-attached prokaryotes and the paradoxical lack of correspondence between measured carbon inputs and its use by bacteria, the bathypelagic ocean is particularly interesting (44).

With the aim of expanding the knowledge of the deeper ocean layer as well as of marine viral and microbial communities, the Spanish-led Malaspina expedition (45, 46) was accomplished. This expedition, carried out on the 200th anniversary of the death of Alessandro Malaspina (the navy commander and scientist who led the first Spanish circumnavigation with scientific purposes), departed from the same harbour of Cadiz to round the globe in order to expand the understanding of the ocean. The image below shows a comparison between the route explored in the XVIII century and the contemporary (Fig. 1.5).

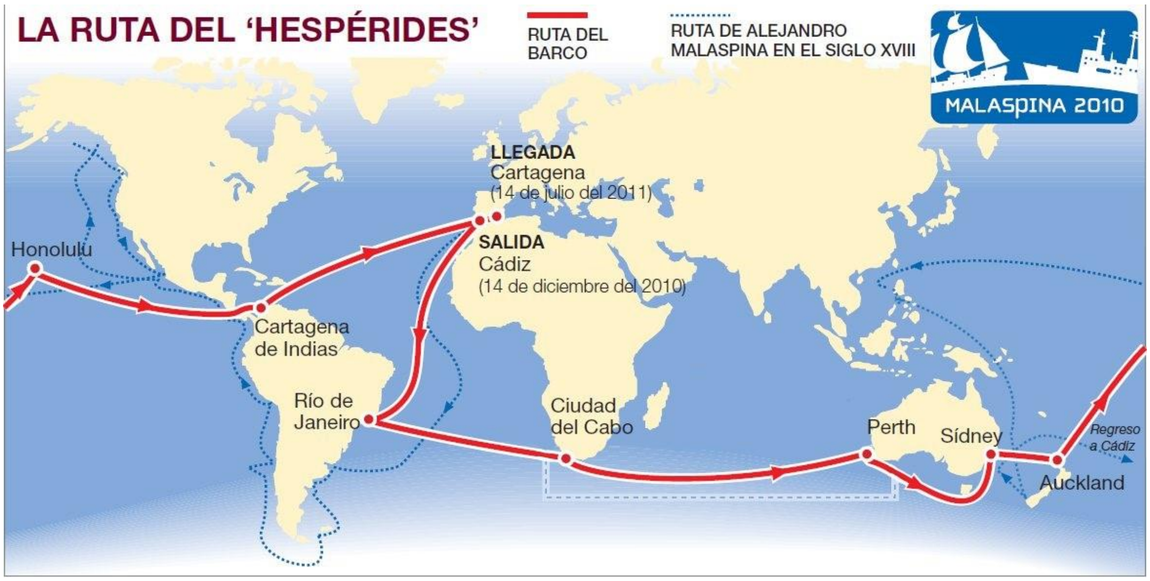

Figure 1.5. Comparison of routes explored by Alessandro Malaspina in the 18th century and by the Malaspina Circumnavigation Expedition in 2010 by the Hespérides ship. (Retrieved from (46)).

Funded by the Ministry of Science and Innovation, through Consolider-Ingenio 2010 (project CSD2008-00077) and the support of many other institutions from different parts of the globe, the total cost of the project was estimated at around 17 million euros. Coordinated by Carlos M. Duarte, research professor at the CSIC, about 400 researchers, more than fifty technicians, a hundred troops of the Spanish Navy sailors and 20 civilians were involved in the project. The ships used for the expedition were Hespérides and Sarmiento de Gamboa, which travelled 42,000 nautical miles. More than 300 sampling stations were established at sea, down to depths of 5,000 metres, which resulted in information (data, images, etc.) requiring over 5,500 GB of disk storage and more than 70,000 samples of air, water and plankton.

2. OBJECTIVES

The main objective of the current master thesis is expanding knowledge on the genetic diversity of marine viruses associated to the prokaryotic fraction of metagenomes retrieved from different water layers at tropical and subtropical latitudes.

- **OBJECTIVE 1 (OBJ. 1)**
  Uncover novel viral genomic diversity from marine ecosystems covering broad latitude and depth gradients.

- **OBJECTIVE 2 (OBJ. 2)**
  Describe viral community composition at the global scale and across depth gradients and perform macrodiversity studies of viral communities to analyse the change in diversity within (α-diversity) and between (β-diversity).

- **OBJECTIVE 3 (OBJ. 3)**
  Determine how viral community composition is associated with environmental parameters, regarding taxonomic, targeted host and AMG diversity.

3. METHODOLOGY

The diagram shown below (Fig. 3.1) represents the workflow followed to conduct all the analyses performed during the study. In addition, all the code created is stored in the following GitHub repository: https://github.com/Skogstokigg/master-thesis.

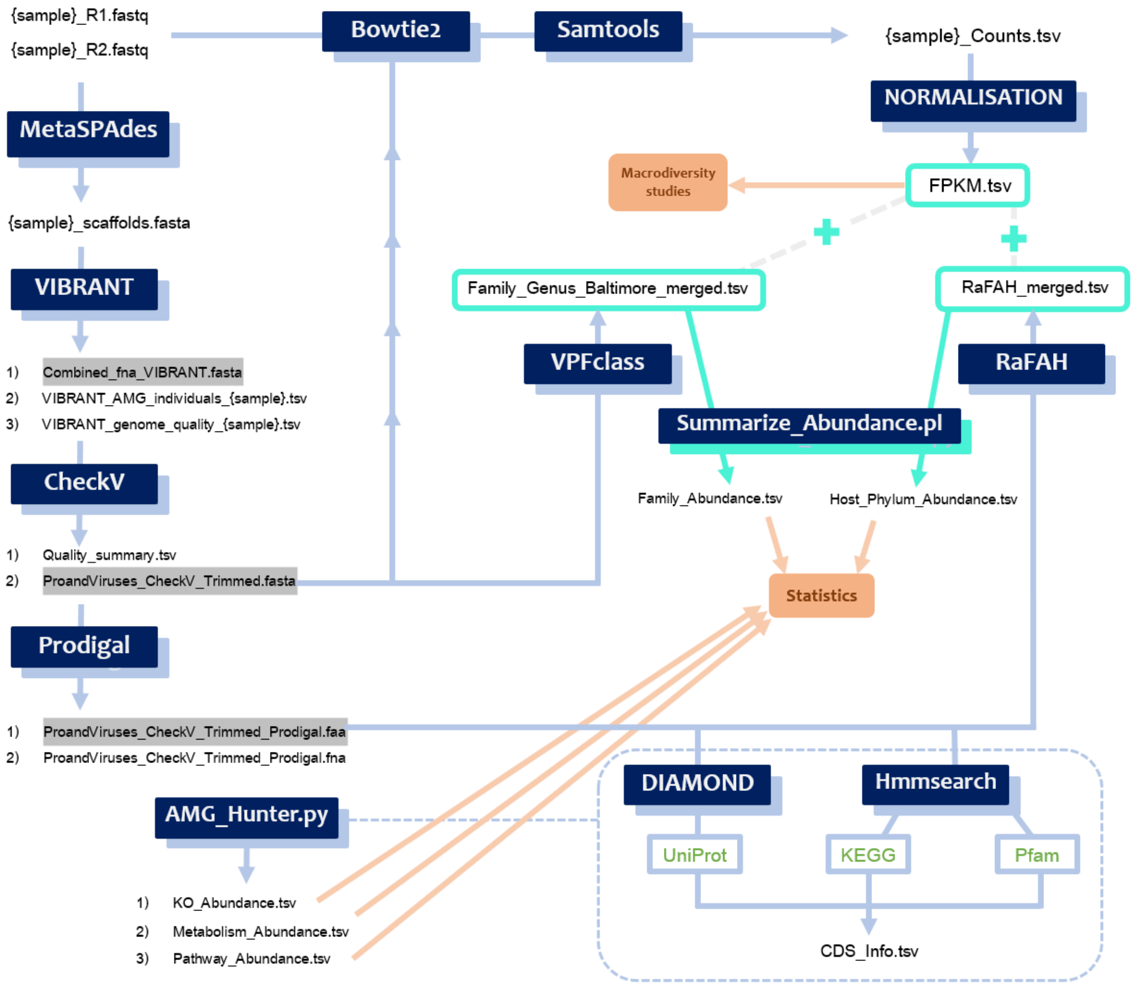

Figure 3.1. Diagram summarising all steps of methodology to perform the analysis.

3.1. DATASET

The dataset used for the current study was retrieved from the Malaspina Circumnavigation Expedition performed during 2010 and 2011 (45,46). From this expedition, more than 2,000 samples of microorganisms from different depths in the Atlantic, Indian and Pacific Ocean were collected and many datasets were generated. Nonetheless, to perform the current analyses, 76 samples from 11 vertical profiles (i.e., depth gradients) were used.

Vertical profiles were distributed throughout the world's oceans at tropical and subtropical latitudes (Fig. 3.2). From each vertical profile obtained at a single station, samples were gathered at different depths in the water column by means of Niskin bottles attached to a rosette coupled with a CTD profiler, which measured conductivity, temperature, fluorescence, oxygen and turbidity, along the water column. About 12 L of seawater were sequentially pre-filtered through a 200µm nylon mesh to remove large plankton, and then sequentially filtered, using a peristaltic pump, through a 20 µm nylon mesh, followed by a 3 µm and 0.2 µm polycarbonate filters of 47 mm diameter (Isopore, Millipore, Burlington, MA, USA). Filtration time was performed for approximately 15 minutes. After filtration, filters were flash-frozen in liquid N_2 and stored at -80ºC until downstream analyses.

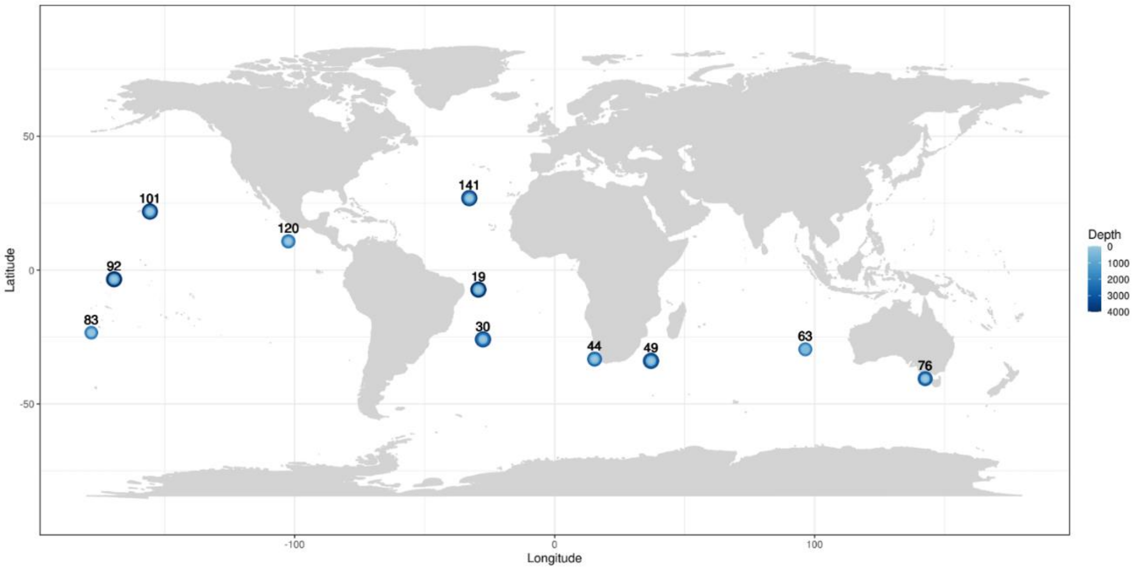

Figure 3.2. World map showing the geographical location of the 11 stations from which vertical profiles sampling collection was performed during the Malaspina Expedition 2010. Stations are distributed mainly throughout tropical and subtropical latitudes. From each station, several samples from different depths were retrieved. The colour and size of the points represents the sampling depth, which ranges from surface (light blue, 0 m depth) to bathypelagic layer (dark blue, 4000 m depth).

3.2. ENVIRONMENTAL PARAMETERS

Besides environmental parameters retrieved with the CTD, many were obtained by means of the analysis of water collected using Niskin bottles. For inorganic nutrients (NO_3^-, NO_2^-, PO_4^{3-}, SiO_2), samples were measured spectrophotometrically using an Alliance Evolution II autoanalyzer (47). In specific samples, where the previous method failed or was not applied, the nutrient concentration was estimated using the World Ocean Database (48). Moreover, spatial features were also considered, such as: Longhurst Provinces (49), Ocean (Atlantic, Indian, Pacific), Ocean Subdivision (Indian, North Atlantic, North Pacific, Pacific, South Atlantic, South Australian Bight, South Pacific), Depth (m), Latitude and Longitude.

3.3. METAGENOMIC ANALYSIS

All samples were sequenced using the *HiSeq 2000 Illumina* platform (2x101 bp) at Centre Nacional d'Análisi Genomica (CNAG) in Barcelona (50). Since paired-ends strategy was applied, two FASTQ files (R1 and R2) were created for each sample. Sequencing yielded 21.84 ± 0.57 Gbp (average ± standard deviation), which summed up total of 1.66 Tb of sequencing, with an average of 108.13 ± 2.83 million read-pairs generated per sample.

Before the assembly, adaptors and Phix174 reads were removed using *Fastx_clipper* from *FASTX-Toolkit v0.0.14* (51). Furthermore, metagenomic raw reads were trimmed for Phred quality scores of ≥33, length ≥45 bp and adapter length ≥10 bp after having performed the quality control of sequences with *FastQC v0.11.7* (52).

All samples of the Malaspina vertical profiles dataset (45) were individually assembled using the meta*SPAdes v3.15.4* (53) mode of SPAdes (54). 500 Gb of memory and 48 threads were used. All other parameters were set to default.

To assess the outcome of the scaffolds obtained from the assemblies, both N50 and N90 were calculated for each sample and in total, before and after having filtered out sequences with less than 1,000 bp length (Appendix: Table A2, Table A3). In addition, the number of gaps in each sample, their average length by scaffold and the standard deviation were also calculated (Appendix: Table A4).

3.4. VIRAL GENOMES IDENTIFICATION

VIBRANT v1.2.1 (55) was used to automate recovery and annotation of bacterial and archaeal viruses. VIBRANT identifies scaffolds putatively derived from viral genomes based on neural network models that rely on protein annotation signatures. VIBRANT is capable of identifying both complete and fragmented viral genomes, including proviruses integrated into larger scaffolds.

3.5. GENOMIC SEQUENCE QUALITY CONTROL

CheckV v0.8.1 (56) was applied in order to assess the quality (i.e., completeness and host contamination) of the viral scaffolds identified in the previous step. Provirus regions potentially contaminated with host regions were excised from scaffolds prior to subsequent analyses.

3.6. RELATIVE ABUNDANCE CALCULATION

The process to calculate the relative abundance of each single scaffold in all samples consisted in first creating an indexed DB with the post-QC viral scaffolds using *Bowtie2 v2.4.3* (57). Paired-end reads from each one of the 76 metagenomes were queried against the aforementioned database using the local alignment (`--sensitive-local`) option. Output SAM formatted files were converted to BAM formatted files, sorted, and indexed using *Samtools v1.8* (58). SAM files were converted into BAM, sorted and indexed. Finally, `idxstats` was used to obtain the number of paired reads mapped to each scaffold in each metagenome.

3.7. STANDARDISATION

An important aspect of working with metagenomics is to apply proper standardisation procedures to the absolute counts, to avoid possible biases raised due to different samples size or variable scaffolds length, among others. Thus, the relative abundances of the viral scaffolds were calculated as Fragments Per Kilobase of scaffold per Million reads mapped (FPKM), which is represented by the equation below:

$$FPKM = \frac{Number\ of\ fragments\ mapped\ to\ scaffold \cdot 10^3 \cdot 10^6}{Total\ number\ of\ reads\ mapped\ to\ sample \cdot Scaffold\ length\ in\ bp} \qquad \text{(Eq.3.1)}$$

where 10^3 normalises for scaffold length and 10^6 for sequencing depth factor.

3.8. MACRODIVERSITY ANALYSES

Macrodiversity is the measure of population diversity within a community. While some diversity measurements rely strictly on the presence or absence of populations, many rely on the relative abundances of populations within communities (i.e., Bray-Curtis distances, Shannon's H index, etc.) (59). Given that it has been shown that metrics that rely on relative abundances are more robust for metagenomic data since they are less susceptible to uneven sampling of rare taxa (60), for the current work these ones were selected.

3.8.1. Analysis of β-diversity

All macrodiversity studies were conducted in R using package *Vegan v2.5-7* (61). This analysis was used to measure differences among samples according to the Bray-Curtis dissimilarity (62). First of all, *vegdist* function was used to generate a dissimilarity matrix showing distances between samples pairs. Next, a non-metric multidimensional scaling (NMDS) analysis was made to reduce the dimensionality of the data and facilitate visualization using function *metaMDS*. Accordingly, two new dimensions, which gathered all the variance, were generated and represented in a bidimensional plot using package G*gplot2* (63).

On the grounds that any of the two dimensions gave information about which variable was separating samples according to their genomic information, a correlogram was created, using package *GGally* (64), which correlated the 1st and 2nd dimension of the NMDS with sample metadata (i.e., depth, temperature, nutrients concentration, [O_2], etc.).

An Analysis of Similarities (ANOSIM) was also applied to statistically determine whether there was a significant difference between two or more groups of sampling units though *anosim* function within *Vegan v2.5-7* package. The distance metric was “bray” and input data consisted of both metadata and scaffolds relative abundances expressed as FPKM.

3.8.2. Analysis of α-diversity

The second approach to study macrodiversity consisted in analysing α-diversity. This analysis was used to measure the diversity by sample. In doing so, two factors: richness and evenness, were taken into account. The measure of the number of different kinds of organisms present in a particular community is defined as richness, while evenness compares the uniformity of the population size of each of the species present (65).

The Shannon’s index (66), sometimes called the Shannon-Wiener index, considers both species richness and evenness. The *diversity* function of *Vegan v2.5-7* allowed to calculate the Shannon index by specifying “shannon” before calling the function.

3.9. GENE CALLING

The PROkaryotic DYnamic programming Gene-finding Algorithm or *Prodigal v2.6.3* (67) in metagenomic mode was applied to predict coding DNA sequences (CDS) within each scaffold.

3.10. CLASSIFICATION OF VIRAL SCAFFOLDS

VPF-Class (68) was used to perform the taxonomic classification of viral scaffolds from Malaspina profiles. This is a tool that automates the taxonomic classification of viral contigs/scaffolds and host prediction of viral contigs/scaffolds based on the assignment of their proteins to a set of classified Viral Protein Families (VPFs).

Despite this software could perform both the taxonomical classification of viruses and the host prediction, only the first function was implemented with the Malaspina sequences. The reason is that for the host prediction a more complete program was used (see Section 3.12). Regarding the steps followed by *VPF-Class*, they consisted in parsing a FASTA file for obtaining proteins of each virus with *Prodigal v2.6.X* (67) and, then, performing a *Hmmsearch v3.2+* (69) against the given HMMS file generated previously with the VPFs to obtain a classification.

For each scaffold, the best classification was made according to the membership ratio, which is a metric of the score associated to a single protein from a scaffold, after having been classified in a taxonomy, divided by the total score of all proteins from the given scaffold. Furthermore, a `--chunk-size` of 1000 was applied, to select this number of sequences to be processed at once, and the default threads were used.

3.11. HOST PREDICTION

To perform the host prediction, a recently developed method was applied in this work, known as Random Forest Assignment of Hosts (*RaFAH v0.3)* (70), which is a classifier program written in *R* and *Perl* that combines the precision of manual curation, the recall of alignment-free approaches, and the speed and flexibility of machine learning. *RaFAH* uses random forests technique to classify protein content of viral sequences and to predict putative virus-host associations.

The default cut-off of 0.14 was applied, which referred to the minimum score to consider a host prediction as valid. All predictions which equalled or surpassed that value at phylum level had at least a 95% of accuracy in the prediction. 48 threads were used, instead of the default.

3.12. FUNCTIONAL ANNOTATION OF VIRAL CODING DNA SEQUENCES

To perform a complete functional annotation, many databases and repositories must be queried to extract different information. The current annotation utilised data from three repositories: UniProt (71, 72), to obtain the proteins taxonomical classification; KEGG (73, 74), to retrieve metabolic pathways information; and Pfam (75) to gather all domains associated to each protein.

The first program used for the functional annotation was *DIAMOND v2.0.7* (76, 77). The arguments used with the command were the following: `blastp`, to align amino acid query sequences against a protein reference database, which in this case was UniRef100 (78); `--matrix` referred to the score matrix for protein alignment (default was BLOSUM62 but, in this case, BLOSUM45 was used to find similarities between more divergent sequences); `--more-sensitive` to enable the more sensitive mode; `--query` made reference to the FASTA files containing the protein sequences derived from the viral scaffolds; `--evalue` to select the maximum e-value to report alignments (1E-05); `--max-`

`target-seqs` to establish the maximum number of target sequences to report alignments for (100 sequences).

The second program used was *Hmmsearch v3.3* (69) which made faster and more sensitive searches of subject protein HMMs against the query proteins, unlike *Diamond* (76, 77), whose searches consisted in comparing query protein sequences against subject protein sequences. The DBs employed to make the searches were KEGG (73, 74) (`/mnt/lustre/scratch/elopez/KOfam/All_KOs.hmm`) and Pfam (75) (`/mnt/lustre/repos/bio/databases/public/pfam/pfam_release_34.0/Pfam-A.hmm`). Argument `--noali` was used to avoid writing the alignment in the output file.

Besides the two programs explained above, to estimate the relative abundances of KOs, metabolisms and pathways in which there were AMGs implicated, *AMG_Hunter* was run. Using as input a file containing the protein sequences derived from the viral scaffolds, *AMG_Hunter* applied *Diamond* and *Hmmsearch* to perform the analyses, taking into account the relative abundances of the scaffolds in which the KO/metabolism/pathway were encoded. In this case, since functional annotation with *Diamond* and *Hmmsearch* had already been performed, this step was skipped with command `–parse_only True`.

3.13. HOST COMMUNITY COMPOSITION BASED ON METAGENOMIC READS

Taxonomic composition of the cellular organisms in the metagenomes (i.e., the host community which is infected by the viruses) was assessed through mTAGs analysis. In order to extract mTAGs, trimmed pair-end reads (Section 3.3) were merged to increase the overall read length using *PEAR v0.9.6* (79) with options: `-b 33` for the base PHRED quality score (33), `-p 0.01` to specify the p-value of the assembly (0.01), `-g 2` to use the acceptance probability for small overlap sizes (2) and `-v 5` to specify the minimum overlap size (5). Then, an input FASTA file with merged pairs and unmerged forward pairs was built and split in smaller files using *FASTA splitter v0.2.6* (80) (using `-n 6` to have the file divided in 6 parts) for *Cdbfasta v1.00* (81) to work properly (it only indexes 4GB). *Cdbfasta* indexed the large multi-FASTA files for quick retrieval of any sequences (82).

Next, *Hmmsearch v3.3* (69) was run to align the indexed sequences against a HMM database containing models of genes from the large subunit (LSU) rRNA and small subunit (SSU) rRNA for bacteria, archaea and eukarya (`-i <test.merged.fna> -o <test.merged.rRNA> -m <ssu,lsu> -k <bac,arc,euk>`). From this alignment, the best hit was retrieved using a script (`parse_rna_hmm3_output.pl <test.merged.rRNA>`) developed in *Perl v5.28* (83) which accepted as input the output file of *Hmmsearch v3.3*. After that, the sequences of the parsed mTAGs were converted into FASTA format (`extract_rrna_seqs.pl <test.merged.rRNA.parsed> 1 90`, where 90 referred to the minimum length of the tag to be extracted) and merged into a single file by sample.

The following step consisted in mapping all the 76 mTAG files against SILVA132 (84, 85) at 99% identity. This DB, located in the cluster (`/mnt/lustre/repos/bio/databases/public/SILVA/SILVA_132_SSURef_Nr99_tax_silva_trunc.final.accession-only.fasta`) provided a manually curated taxonomy for all three domains mapped, based on representative phylogenetic trees for the SSU and LSU rRNA genes. Then, the abundance tables of each rank based on SILVA132 taxonomy per sample were generated using R and, eventually, all samples outputs were merged into a single file by taxonomic level (`<output>/Merged_{taxonomy}.tsv`).

3.14. STATISTICAL ANALYSIS

3.14.1. Redundancy analysis

Redundancy analysis (RDA) is a direct gradient analysis technique which summarises linear relationships between components of response variables that are redundant with a set of explanatory variables (86). One way of studying the data distribution according to metadata information, with the aim of acquiring a general overview of samples behaviour by layer, was performing an RDA.

Variables selected for RDA analysis were Temperature (ºC), Salinity (Practical Salinity Units, PSU), NO_3, PO4, and SiO4 concentrations (µmol/L), Oxygen (mL/L) and Layer. The reasons to choose them were lack of dependency, the more complete dataset (fewer missing values) and their relevance in marine ecology.

All analyses were performed in R using function *rda*.

3.14.2. Mann-Whitney U test

Once the relative abundance of scaffolds was obtained and grouped according to predicted hosts, taxonomic affiliation and functional gene content, statistical analyses were carried out to determine how the relative abundances of these variables changed across samples or groups of samples, and whether these differences were significant. To that end, three statistical analyses were applied: Shapiro-Wilk (87) as normality test, Fligner-Killeen (90) to analyse homoscedasticity and Mann-Whitney U test (91) to compare the differences between two independent samples when the sample distributions are not normally distributed.

Shapiro-Wilk (87) was used as test for normality. This test is a more appropriate method for small sample sizes (<50 samples) although it can also be handling on larger sample sizes. For the Shapiro-Wilk normality test, null hypothesis (H_0) states that data are taken from normal distributed population (88, 89), while rejecting the null hypothesis (accept H_1) is synonym of having a non-normal distribution of data. To perform the analysis, the `shapiro.test` R function was applied.

After having discarded the gaussian distribution of data, another analysis is needed to evaluate homoscedasticity. In order to be able to apply the Mann-Whitney U test, it is advisable that variance be equal in the two groups. Hence, the Fligner-Killeen test (90) was used. This is a non-parametric test for homogeneity of group variances based on ranks. Since it is based on medians comparison, it is the best option when data are non-normally distributed. The null hypothesis (H_0) states that the two populations variances are equal, whereas the alternative hypothesis (H_1) affirms that there are differences. To perform the analysis, the `fligner.test` R function was applied.

The Mann-Whitney U test (91) is used to compare the differences between two independent samples when the sample distributions are not normally distributed and the sample sizes are small ($n < 30$). It is considered to be the nonparametric equivalent to the two-sample independent t-test.

Using `wilcox.test` function from R repository, this test was applied for all pairs of groups: epipelagic – mesopelagic, epipelagic – bathypelagic, mesopelagic – bathypelagic.

4. RESULTS

The following sections gather all results generated after having performed the correspondent analyses of the 76 metagenomics samples of the vertical profiles, as well as of its associated metadata, retrieved from the Malaspina Circumnavigation Expedition. Information of the sample code, location and sampling depth are located in Appendix (Table A1).

4.1. METADATA ANALYSIS

From the 76 samples obtained from vertical profiles, a total of 23 samples came from the epipelagic layer (from surface to 200 m), 28 from mesopelagic (200 – 1000 m) and 25 from bathypelagic (1000 – 4000 m).

An RDA was performed to assess how the different physical and chemical parameters varied across samples. The results of this analysis are shown in an RDA-biplot (Fig. 4.1), where samples are represented by their code and the arrows represent variables influencing the samples distribution. Temperature and salinity were higher in epipelagic samples, followed by mesopelagic and bathypelagic ones. Generally, mesopelagic and bathypelagic samples contained higher amounts of inorganic nutrients (NO_3, SiO_4 and PO_4) than epipelagic samples. Finally, oxygen concentrations varied among samples, regardless of the depth zone from which they were obtained.

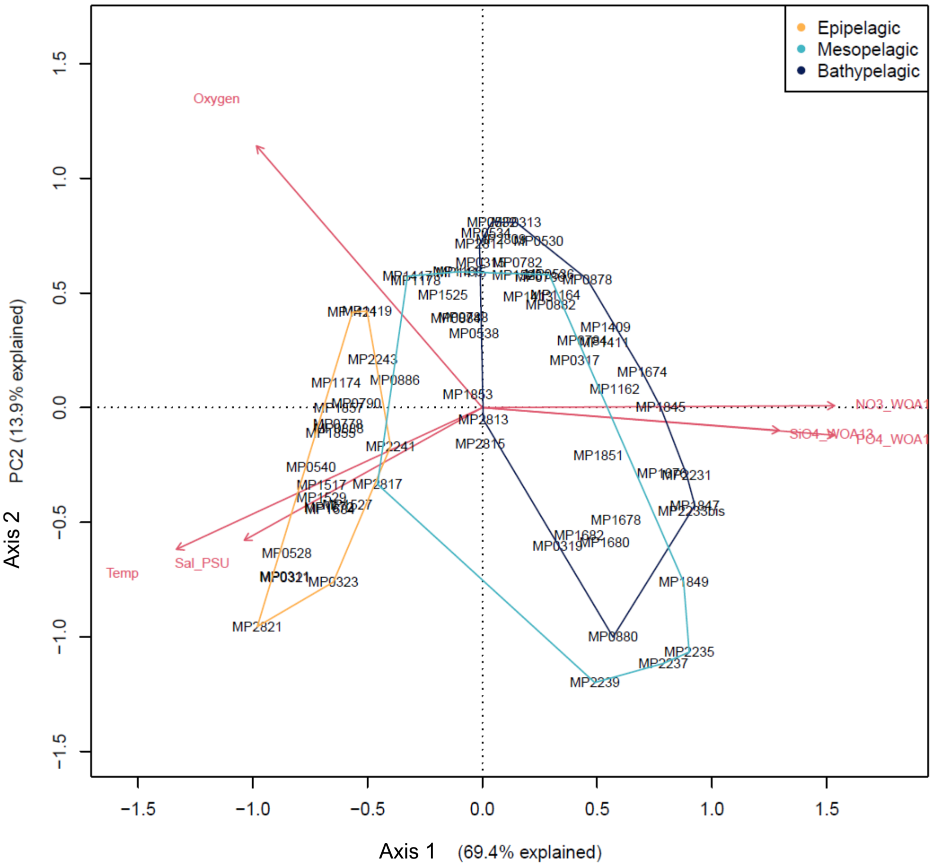

Figure 4.1. Redundancy analysis biplot with the environmental data of the 76 samples. Axis 1 explained 69.4% of variance, while the second one (Axis 2) explained the 13.9%. Points represent objects (in this case, samples are presented by their ID code). Samples were clustered according to the physical and chemical parameters. Then, their sampling depth (Epipelagic – yellow; Mesopelagic – Light blue; Bathypelagic – Dark blue) was drawn around each group *a posteriori* to help visualisation. Red vectors represent the original variables (oxygen (µmol/kg), temperature (ºC), salinity (PSU), [NO_2] and [SiO_4]) used to build the axis. Temperature and salinity were higher in epipelagic samples, followed by mesopelagic and bathypelagic ones. Generally, mesopelagic and bathypelagic samples contained higher amounts of inorganic nutrients (NO_2 and SiO_4) than epipelagic samples, despite there was more data dispersion. There were almost no differences between samples when taking oxygen into account.

4.2. RAW SEQUENCES QUALITY CONTROL

The analysis of all sequences gave as a result 23,616,830 ± 5,526,122 (average ± standard deviation) total sequences per sample; any of them were flagged as poor quality; with an average length of 101 bp and a mean GC content of 41.40 ± 4.52 %.

4.3. ASSEMBLY QUALITY CONTROL

The assembly step generated a total of 25,914,084 scaffolds. The longest scaffold was 2,458,067 bp long and the minimum was established at 1,000 bp length. All scaffolds shorter than the minimum were removed, leading to a total of 7,931,351 scaffolds, which added up to ~ 21 Gbp (20,967,998,656 bp).

Before filtering, the sample with the largest N50 (4,321 bp) was MP0780, as well as the one with the largest N90 (684 bp). After filtering, the sample with the largest N50 (7,678 bp) was MP0782, whereas the one with the largest N90 (1,420 bp) was MP0780. These two samples came from the South Atlantic Ocean (Table A1) Values of the rest of samples are shown in Appendix (Table A2 and Table A3).

Another way of assessing the assembly outcome was calculating the number of gaps, their average length, the maximum length and the minimum length, for both each sample and all samples together. A total of 1,846,493 gaps were generated in total in all samples, being sample MP0790, from the South Atlantic Ocean, the one with the greatest number of gaps (53,021). The maximum gap size was 60 bps and, the minimum, 2 bps. On average, the size was 13.57 ± 12.47 bp. Information of each sample is displayed in Appendix (Table A3).

4.4. VIRAL SCAFFOLDS RETRIEVAL

VIBRANT identified a total of 101,221 scaffolds putatively derived from viral genomes. 98,320 came from full viruses, while 2,901 corresponded to proviruses. The histogram below represents the length distribution of viral scaffolds (Fig. 4.2).

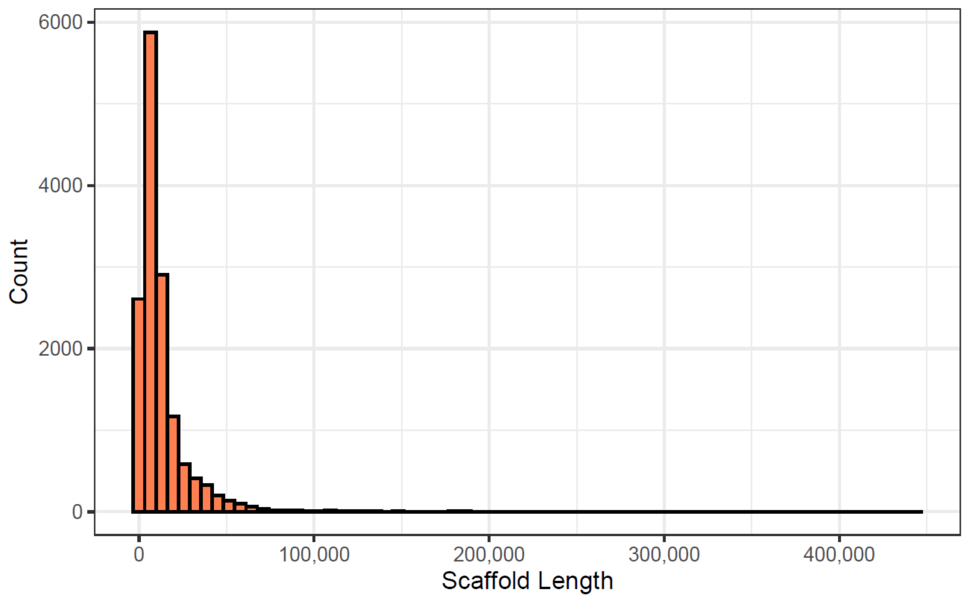

Figure 4.2. Histogram of the scaffolds length distribution for viral reads in the samples. The most part of scaffolds had lengths lower than 100,000 bp, as expected.

To assess quality of viral scaffolds, *CheckV* was used. Sequences were classified to one of five different quality tiers (complete, high quality, medium quality, low quality, undetermined) based on their estimated completeness (100% complete, >90% complete, 50-90% complete, <50% complete, undetermined), as shown in Fig. 4.3.

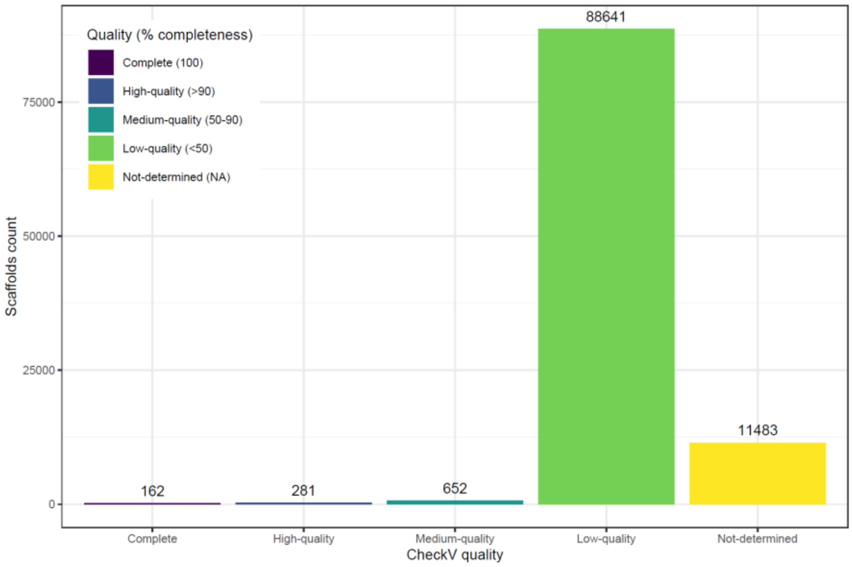

Figure 4.3. Histogram of CheckV quality classification for viral scaffolds. 162 scaffolds were Complete; 261 scaffolds, of High-quality; 652 scaffolds, of Medium-quality; 88,641 scaffolds, of Low-quality; and 11,483 scaffolds, Not-determined.

From what has been shown above, completeness of viral scaffolds was mainly below 50%. The average of completeness was 4.43 ± 10.08 %. Regarding contamination, the average in all viral scaffolds was 1.03 ± 6.94 %.

4.5. MACRODIVERSITY

4.5.1. Analysis of β-diversity

Relative abundance of viral scaffolds expressed as FPKM was used as input for Bray-Curtis dissimilarity analysis. Results were gathered in a 76 x 76 square matrix, which was used as input. To perform the NMDS analysis, set to two dimensions were set, which were represented in a scatterplot (A, Fig. 4.4). The first dimension separated photic (i.e., epipelagic) from aphotic (i.e., mesopelagic and bathypelagic) samples. Furthermore, both the Pearson (Fig. A1, Appendix) and Spearman (Fig. A2, Appendix) correlations of NMDS1 with different variables, displayed a significantly strong positive correlation with Temperature (P: 0.891***; S: 0.893***) and a significantly strong negative correlation with Depth (P: -0.646***; S: -0.892***), being "***" a significance of p-value < 0.001. Eventually, after performing the ANOSIM test, the null hypothesis (H_0), which stated that there were no differences between the means of two or more groups of ranked dissimilarities, was rejected (P-value < 0.001; R statistic: 0.5744). An R value close to 1.0 would suggest dissimilarity between groups, while an R value close to 0, an even distribution of high and low ranks within and between groups. The R statistic was closer to 1, therefore there was significant dissimilarity between groups.

The 76 samples were distributed across NMDS1 according to their water layer (A, Fig. 4.4). This meant that the main factor influencing samples differences was the sampling depth. In order to discover new factors influencing in samples separation by layer, the same analyses were performed for samples coming from each layer: calculation of Bray-Curtis dissimilarity, NMDS analysis and correlation of the first dimension of each analysis with environmental variables. The results of the NMDS analyses were plotted together with the variable with the strongest correlation influencing in samples separation in each layer (B-D, Fig.4.4).

In all plots, stress values were lower than 0.2, therefore representations of reduced dimensions are reliable. With respect to the NMDS analyses by layer, variable with the strongest correlation influencing in samples separation by the first dimension in the epipelagic layer was depth (m); in the mesopelagic layer, NO_3 (µmol/L); and in the bathypelagic layer, SiO_4 (µmol/L).

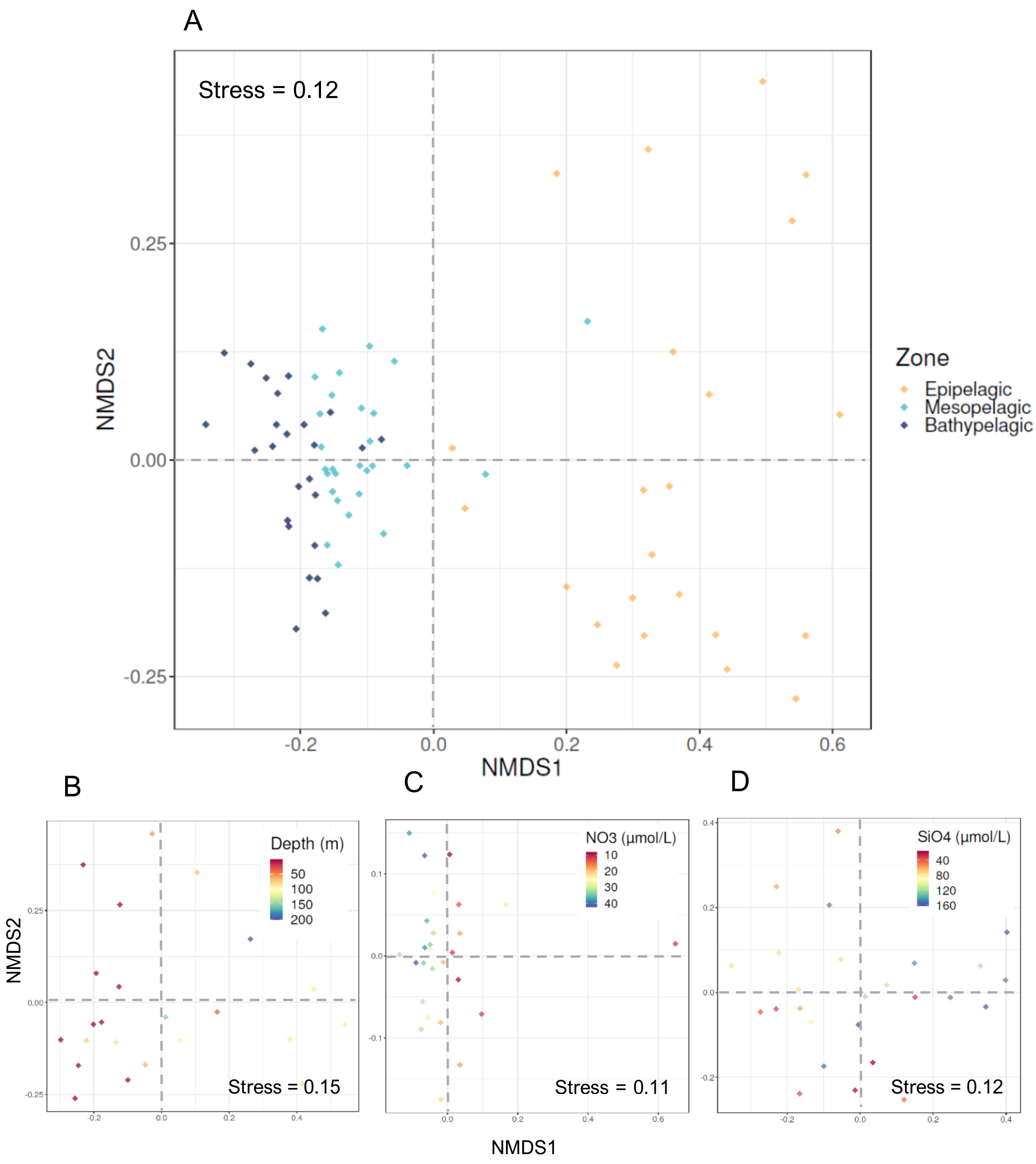

Figure 4.4. Non-metric multidimensional scaling plots. For all graphics (**A-D**), points represent samples. Samples that are more similar to one another are ordinated closer together. Stress values in all plots are < 0.2, therefore representations of reduced dimensions are reliable. The first graphic, (**A**), was generated from the 76 Malaspina vertical profiles samples. Samples were clustered according to their water layer (Epipelagic – in yellow, Mesopelagic – in light blue and Bathypelagic – in dark blue). In addition, there were marked differences among samples located in the photic zone (Epipelagic, on the right) and the aphotic zone (Mesopelagic and Bathypelagic, on the left). The second NMDS plot, (**B**), represented the 23 samples retrieved from the Epipelagic layer. The variable with the strongest correlation influencing in samples separation by the first dimension in Epipelagic layer was Depth (m). The third plot, (**C**), represented the 28 samples retrieved from the Mesopelagic layer. The variable with the strongest correlation influencing in samples separation by the first dimension in Mesopelagic layer was NO_3 (µmol/L). The fourth plot, (**D**), represented the 25 samples retrieved from the Bathypelagic layer. The variable with the strongest correlation influencing in samples separation by the first dimension in Bathypelagic layer was SiO_4 (µmol/L).

4.5.2. Analysis of α-diversity

The Shannon's Index was obtained for each one of the 76 samples. Higher viral diversity was observed among samples from upper layers. Diversity tended to decrease until ~ 700 m deep, after which it stabilised (Fig. 4.5).

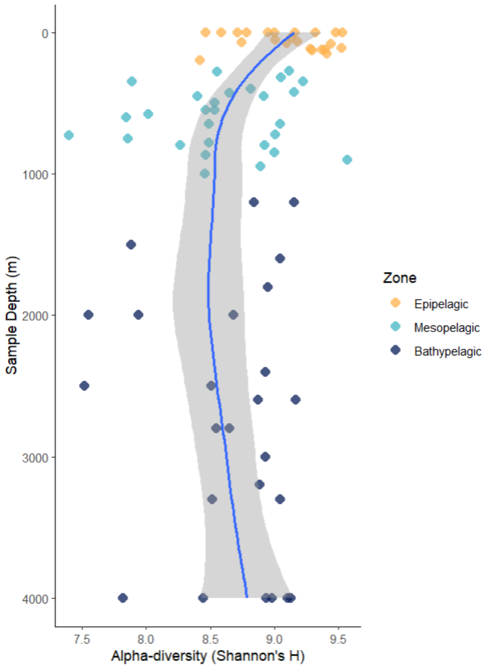

Figure 4.5. LOESS smooth plot showing the depth distribution of macropopulation diversity (Shannon's H). The line represents the LOESS best fit, while the lighter band corresponds to the 95% confidence window of the fit. Points represent samples and colours, the samples water layer (Epipelagic – yellow, Mesopelagic – light blue, Bathypelagic – dark blue). From up to down, diversity tended to decrease until ~ 700 m deep, from which it stabilised.

4.6. TAXONOMIC CLASSIFICATION OF VIRAL SEQUENCES

After applying *VPF-Class* for virus classification, only the hit with the best membership ratio by scaffold was retrieved, with the aim of making results easier to handle.

According to the Baltimore classification, scaffolds were classified as dsDNA (84,572), ssDNA (182), RT (2) and dsRNA (0). With respect to the family classification, scaffolds were labelled as Myoviridae (46,466), Podoviridae (8,966) or Siphoviridae (7,648), among others. Finally, in relation to genus classification, scaffolds were labelled as T4virus (32,854), M12virus (7,551) or Lambdavirus (4,246), among others.

Afterwards, the relative abundances of taxonomic groups in each sample were calculated based on the standardised relative abundances of scaffolds (calculations in Section 3.6 and 3.7) and their taxonomic classification, both at family and genus levels

(Fig. 4.6 and Fig. A4, respectively). With respect to the relative abundances analysis by water layer, the viral classification at family level was: *Myoviridae* the most abundant in the three water layers, followed by *Microviridae*, *Siphoviridae* (in Bathypelagic) and *Podoviridae*. None of these taxa showed relevant changes across all water layers, except for *Siphoviridae*, which was more abundant in Bathypelagic compared to Epipelagic and Mesopelagic layers.

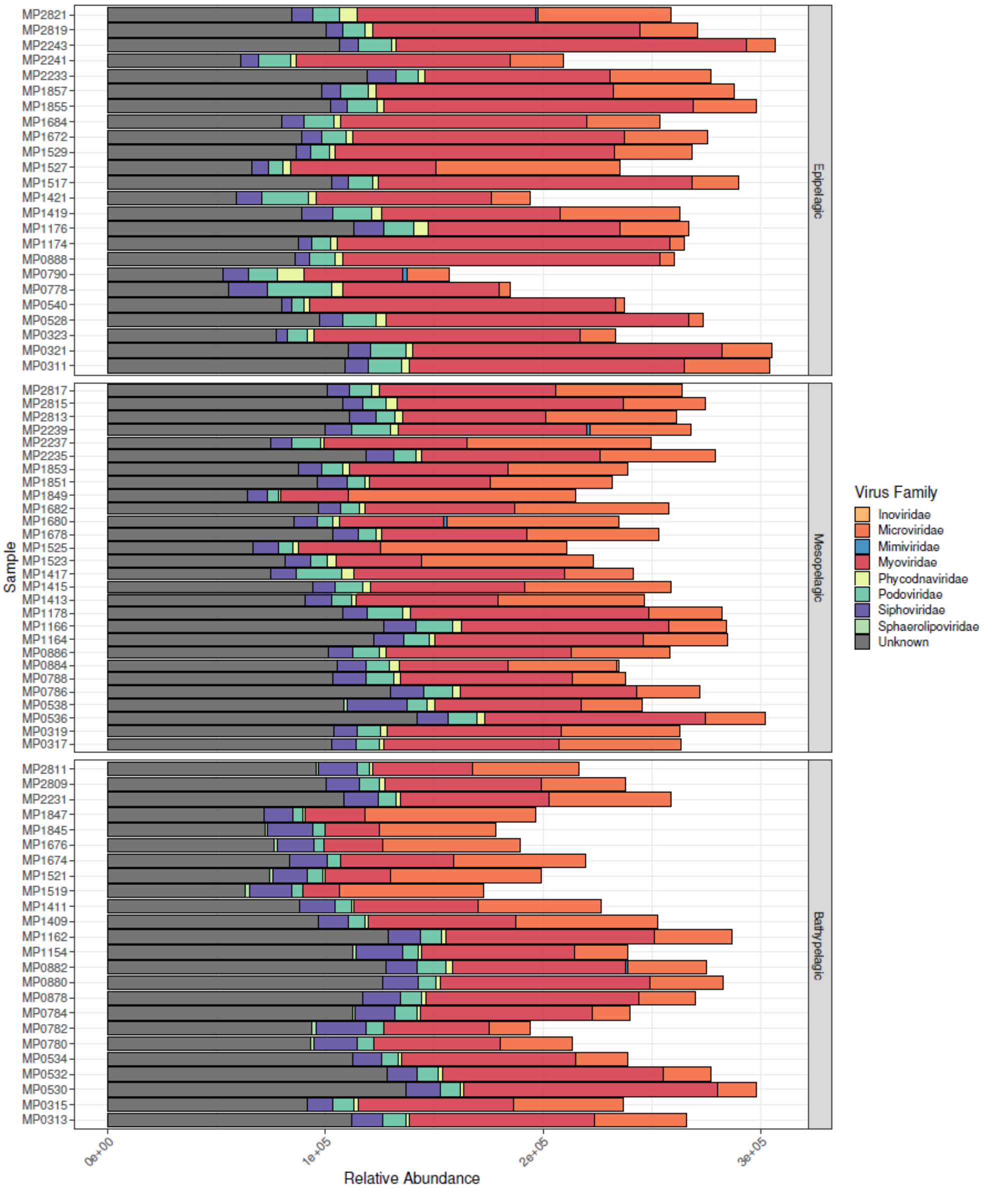

Figure 4.6. Relative abundance of viruses according to family along all the water column. The y-axis represents samples and the x-axis, the reads per kilobase per million mapped reads (FPKM). In the epipelagic layer, Myoviridae was the most abundant, followed by Microviridae. In mesopelagic and bathypelagic layers, both Myoviridae and Microviridae were the most abundant, being the latter slightly more abundant than in the more superficial layer. Another point is that Phycodnaviridae is slightly more abundant in epipelagic and mesopelagic than in bathypelagic layer. In contrast, Siphoviridae abundance is greater in the deepest layer than in the two others.

4.7. VIRAL HOSTS

After applying *RaFAH v0.3* for host prediction, the 101,221 scaffolds were classified as *Pseudomonas* (72,630), *Candidatus Pelagibacter* (6,686), *Nostoc* (1,416) and *Synechococcus* (1,058), among others.

The standardised relative abundances of scaffolds (calculations in Section 3.6 and 3.7) was merged with the number of scaffolds classified in each host taxonomic category at phylum level (Fig. 4.7).

With respect to the relative abundances analysis by water layer, it was shown that, according to viral host prediction classification, phylum *Proteobacteria* (specially class *Gammaproteobacteria* and *Alphaproteobacteria*) was notably the most abundant in all water column, followed by *Bacteroidetes* and *Cyanobacteria*. *Proteobacteria* was predominant in all water layers and *Bacteroidetes* did not show abundance variations across depth. Nevertheless, *Cyanobacteria* were more abundant in the Epipelagic layer compared to deeper water layers.

Since the results of *RaFAH* provided a computational prediction of bacterial and archaeal hosts for each sample, the bacterial abundance within the metagenomes was also retrieved in order to compare the taxonomic composition of the relative abundances of viruses against that of their host. The results are shown in Fig. 4.8. Regarding the two figures representing relative abundances (Fig. 4.7 and Fig. 4.8), there is a predominance of phylum *Proteobacteria* in both. *Bacteroidetes* remains constant in abundance in both cases. Finally, in the Epipelagic layer, there is a higher abundance of *Cyanobacteria* in comparison to the other layers. These results show how predicted viral hosts match properly with the bacterial abundances from the same samples.

4.8. FUNCTIONAL DIVERSITY

For the functional annotation, *DIAMOND v2.0.7* was utilised to obtain information of the UniProt database, while the information of KEGG and Pfam repositories was extracted with *Hmmsearch v3.3*.

748,072 CDSs were identified across all scaffolds. 1,955 KOs were assigned which were included in 241 pathways which, in turn, were embedded in 26 different metabolisms. 652,174 / 748,072 CDS were assigned to KOs.

The relative abundances of KEGG metabolic pathways in the metagenomes was estimated based on the relative abundances of the viral scaffolds and on the functional annotation of the protein encoding genes identified in them (Fig. 4.9). Results suggested there was no predominance of a specific metabolic activity in any water layer, except for “replication and repair”, which was highly represented. Moreover, there was a higher relative abundance of metabolic genes among samples from the Epipelagic layer.

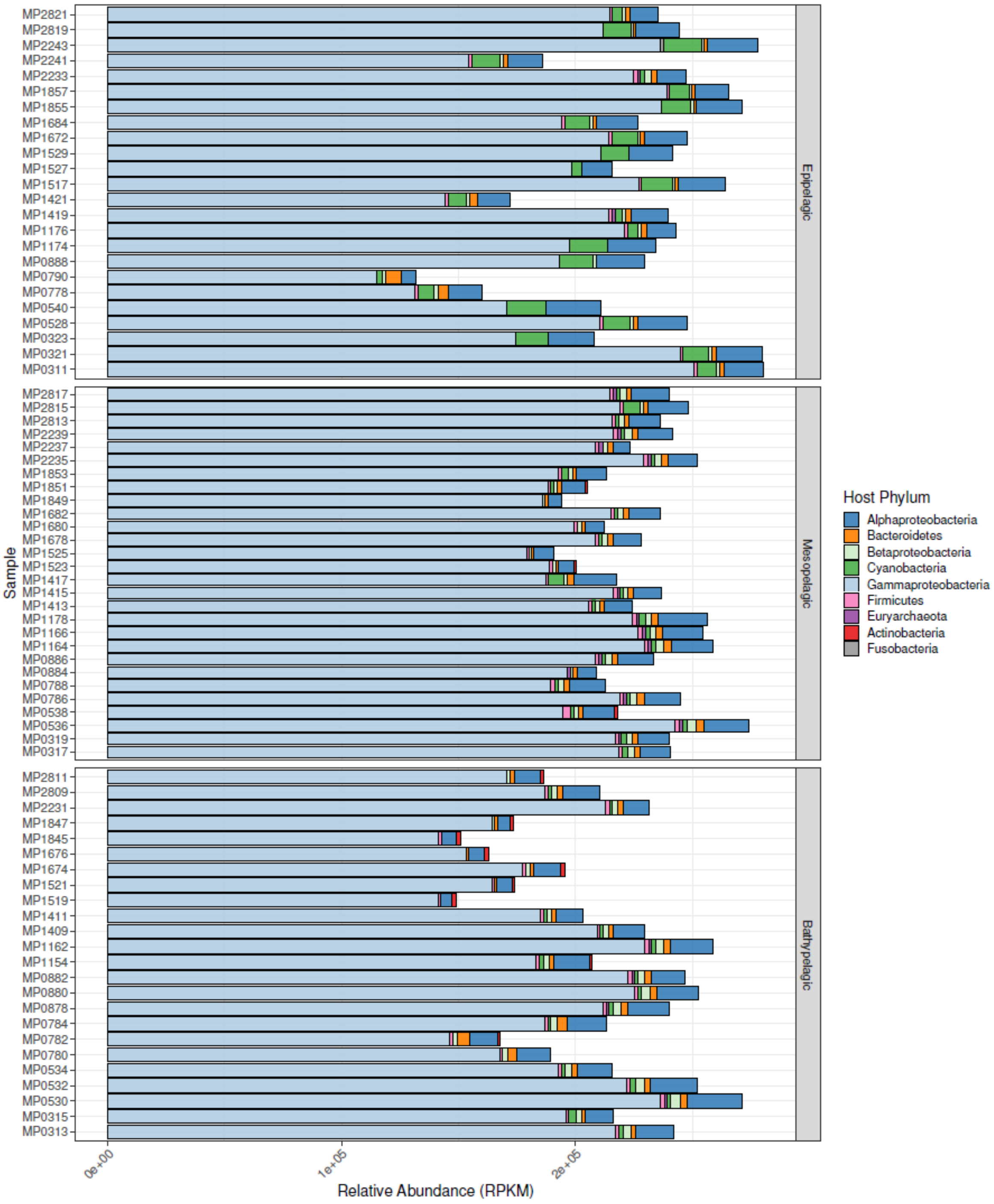

Figure 4.7. Relative abundance of viruses according to the predicted hosts phylum along all the water column. The y-axis represents samples and the x-axis, the reads per kilobase per million mapped reads (FPKM). There is a predominance of class *Gammaproteobacteria* in all water layers. Abundance of *Cyanobacteria* was greater in the Epipelagic layer than in the other layers.

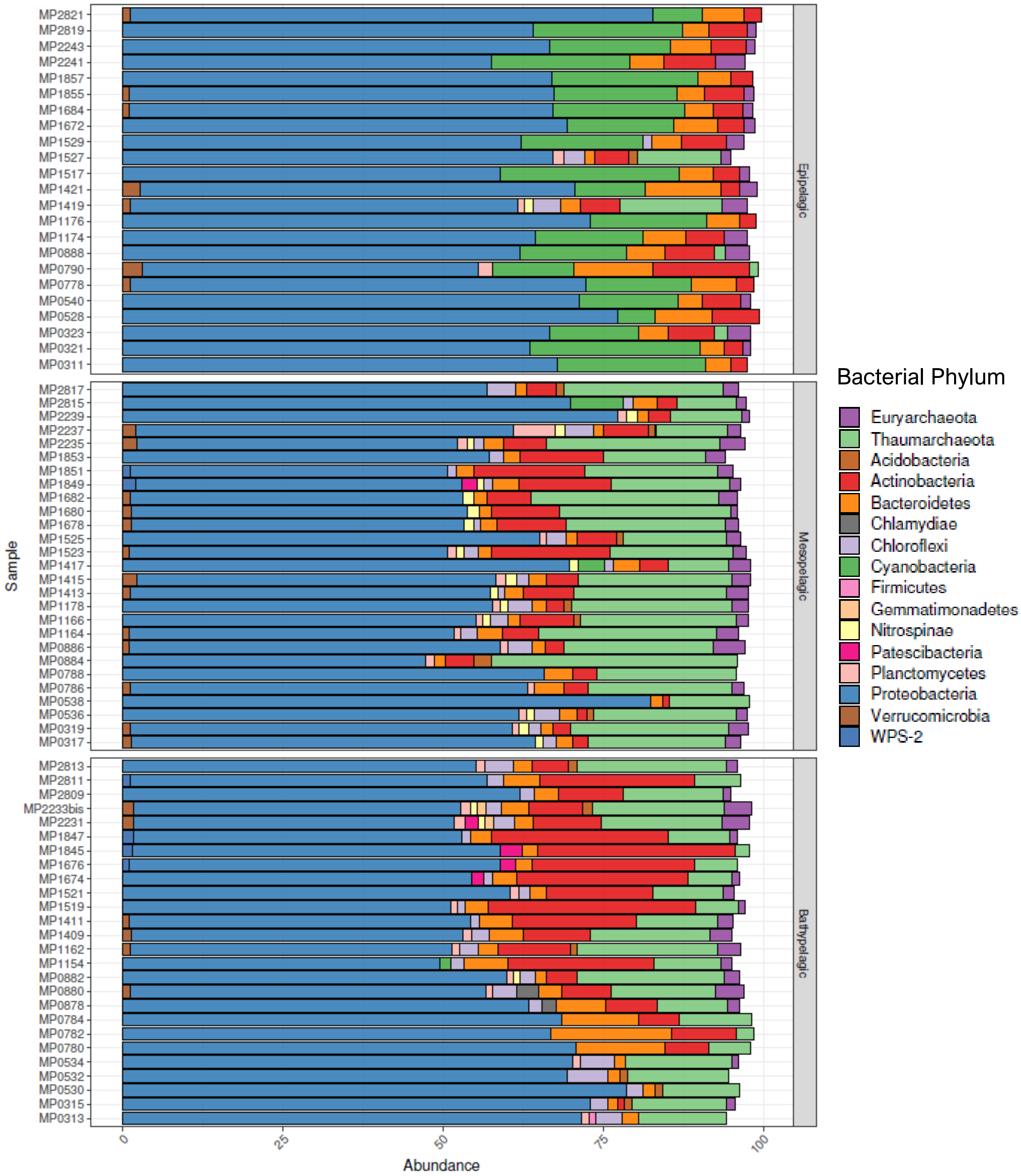

Figure 4.8. Relative abundance of bacteria along all the water column. The y-axis represents samples and the x-axis, the reads per kilobase per million mapped reads (FPKM). There is a predominance of phylum proteobacteria. In the epipelagic layer, there is a high abundance of Cyanobacteria in comparison to the other layers. Bacteroidetes are more abundant in epipelagic and bathypelagic. Finally, phylum Actinobacteria is highly abundant in bathypelagic compared to the other layers.

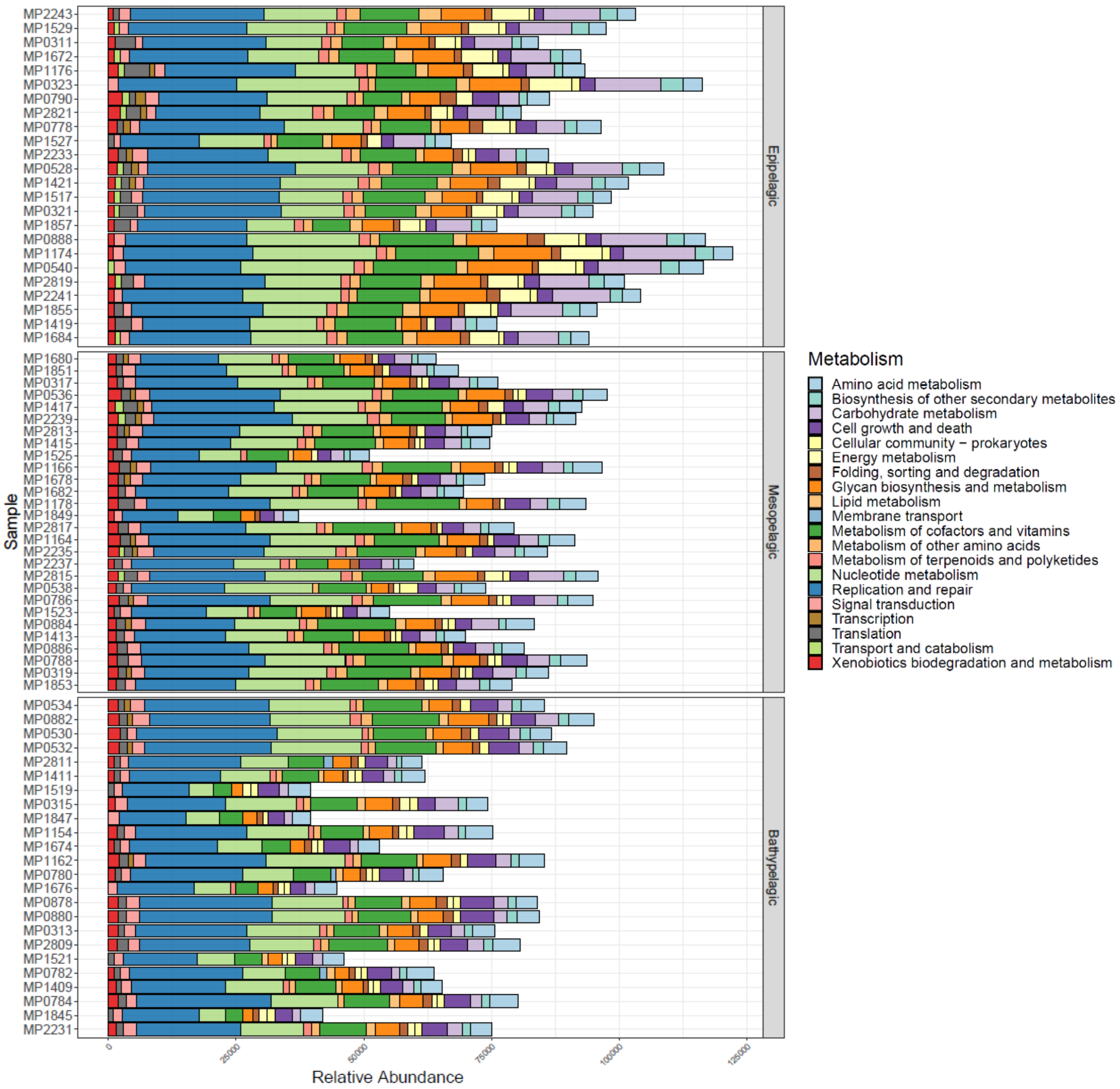

Figure 4.9. Relative abundance of KEGG metabolic pathways along all the water column. The y-axis represents samples and the x-axis, the reads per kilobase per million mapped reads (FPKM). There is no predominance of a specific metabolic activity in any water layer, except for “replication and repair”, which is highly represented. Moreover, there is a higher abundance of metabolisms in general in the epipelagic layer.

4.9. SIGNIFICANT DIFFERENCES AMONG WATER LAYERS

The normality test rejected the null hypothesis (p-value < 0.05), indicating that the data of scaffolds showing relative abundance of viral taxonomy, host taxonomy and functional abundances were non-normally distributed. Regarding the homoscedasticity test, 4,213 groups out of 6,825 rejected the null hypothesis (p-value < 0.05), which means that most part of data (61.7%) were not homoscedastic.

Once rejected both normality and homoscedasticity, Mann-Whitney test was applied to study whether differences in variables abundances among the three water layers were significant.

The results were the following: 5,025 out of 6,825 (73.6%) pairwise comparisons showed significant differences rejecting the null hypothesis (p-value < 0.05). After having adjusted the p-value with the Bonferroni correction, 2,225 out of 6,825 (32.6%) pairwise comparisons showed significant differences rejecting the null hypothesis (Adjusted p-value < 0.05). The following table shows the Mann-Whitney test results for each type of variable (Table 4.1).

Table 4.1. Results of the Mann-Whitney test for groups comparison.

File	Rejected H_0/Total (p-value < 0.05)	%	Rejected H_0/Total (Adjusted p-value < 0.05)	%
KO	4,354/**5,862**	74.3	1,920/**5,862**	32.8
Pathway	502/**720**	69.7	232/**720**	32.2
Viral family	83/**117**	70.9	33/**117**	28.2
Metabolism	50/**75**	66.7	25/**75**	33.3
Host phylum	32/**51**	62.7	15/**51**	29.4

More specifically, the host phylum which showed the more significant difference was *Cyanobacteria* when comparing Epipelagic and Bathypelagic populations, followed by *Cyanobacteria* when comparing Epipelagic and Mesopelagic ones. In both cases, this phylum was more abundant in the surface layer. *Crenarchaeota* was the third and fifth more significant when comparing Epipelagic and Bathypelagic, and Epipelagic and Mesopelagic population abundances, respectively. In this case, the Epipelagic layer showed the fewer abundance for both comparisons.

Regarding metabolic pathways, the most significant changes in abundances where in the “energetic metabolism”, both when comparing Epipelagic and Bathypelagic as when comparing Epipelagic and Bathypelagic. In both cases, the most abundant was in the Epipelagic layer. The following most significant changes were in the “metabolism of carbohydrates” and “lipids metabolism”, when comparing Epipelagic-Bathypelagic and Epipelagic-Mesopelagic layers. In both cases, the more abundant occurred in Epipelagic. Another important result was that “cell motility” was also significantly different in Epipelagic and Bathypelagic, being more abundant in Bathypelagic.

Next, with respect to the viral family, the most significant change corresponded to *Sphaerolipoviridae*, for the three comparisons (Epipelagic – Bathypelagic, Epipelagic – Mesopelagic and Bathypelagic – Mesopelagic). Its abundance increased with depth. Another relevant result was the significant change in abundance between Epipelagic and Mesopelagic, and Epipelagic and Bathypelagic of *Lavidaviridae* viral family, being more abundant in the deepest layers.

5. DISCUSSION

Multiple analyses were performed to increase the body of knowledge of marine viral communities thriving at the three water layers of the ocean, specially from Atlantic, Pacific and Indian oceans, at tropical and subtropical latitudes. Since 3 specific objectives were proposed for the current study (Section 2), an assessment of their completeness will be performed during this section throughout the discussion of results.

The first objective of the project (OBJ. 1) was uncovering novel viral genomic diversity from marine ecosystems encompassing broad latitude and depth gradients. To that aim, the study of samples distribution according to environmental variables having into account metadata information was first performed. Results showed how samples separated according to the physical and chemical parameters, as expected. However, there were differences in how distant samples in the RDA plot (Fig. 4.1) were depending on the water layer. While samples from the Epipelagic layer showed more similar environmental conditions, Mesopelagic and Bathypelagic samples were more spread over the plot. It can be explained by the fact that these two layers encompass a greater length (200 – 1000 m and 1000 – 4000 m, respectively) in comparison to the shallowest layer (only 200 m depth), hence more variability in data was obtained.

Regarding variables influencing samples separation, the most influential were both salinity and temperature, on the grounds that samples were distributed over the arrows of the RDA-biplot representing variables, being clearly separated by sampling depth. The correlation between depth and temperature has been widely documented (92, 93, 94). In the ocean, solar energy is reflected in the upper surface or rapidly absorbed with depth, meaning that the deeper descended into the ocean, the less sunlight there is. This results in less warming of the water. It is a fact that temperature profiles vary at different latitudes but given that, in this study, samples came from tropical and subtropical latitudes, profiles were practically identical during all seasons of the year. Concerning salinity, some studies (95, 96) demonstrate how the decrease in temperature with depth is directly correlated to lower Practical Salinity Units (PSU) levels in many ocean water columns, as shown in the current analysis. With respect to inorganic nutrients concentrations variables (NO_3, SiO_4 and PO_4), Mesopelagic and Bathypelagic samples showed greater concentrations, whereas in the Epipelagic layer these levels were lower. Most nutrients are removed from the euphotic zone and transferred to the deeper ocean as dead organisms sink to the ocean floor. In the deeper layers, organic matter is remineralized, what it means that nutrients are brought back into solution.

Once the environmental data had been analysed, in order to complete OBJ. 1, the genomic information was included in subsequent analyses, which leads concurrently to the proposal of the second objective of the project (OBJ. 2). This objective was describing the viral community composition at the global scale and across depth gradients and performing macrodiversity studies of these viral communities. To that aim, different analyses were conducted: (i) beta-macrodiversity and alpha-macrodiversity analyses, as well as (ii) a taxonomical classification.

With regard to the analysis of β-diversity, the Bray-Curtis dissimilarity was calculated. As a first proxy, viral scaffolds abundances from all samples were included in the analysis. Samples separation was influenced by sampling depth and temperature as indicated by strong correlation between depth and NMDS1. This result was akin to the analysis of samples distribution according to metadata information explained above, where temperature was the main variable influencing in samples distribution by water layer. However, with this result, a new variable is being included: genetic information. Viral communities' differences according to the Bray-Curtis dissimilarity of taxonomical groups, were mainly influenced by temperature and, consequently by depth. It supports what other studies have obtained from analysing β-diversity in bacterial communities (97-100): temperature, appears to be one of the major components contributing to the vertical β-diversity.

Another noteworthy point related to what it can be inferred from Fig. 4.4 (A) is that differences among Epipelagic layer samples according to Bray-Curtis dissimilarity were greater than those from Mesopelagic and Bathypelagic, since the dispersion of the points (samples) was greater in the more superficial samples compared to the more deep ones. Studies carried out with bacterial metagenomes have shown the opposite tendence (101-103). Consequently, the same analysis was made using all scaffolds obtained from the 76 samples (Fig. A3, Appendix), without removing prokaryotic scaffolds, and results were similar to those from the analysis made with viral scaffolds: more differences in community composition in Epipelagic than in the deeper layers. One possible explanation to these outcomes is that, despite latitudes from which samples were retrieved corresponded to tropical and subtropical oceans, longitude spanned all the globe. It is relevant to highlight at this point that analyses of vertical profiles at such large scale had not been carried out to date, therefore this result deepens the current knowledge of viral communities' diversity covering broad latitude and depth gradients (as proposed by OBJ. 1). Regarding the origin of the greater differences among Epipelagic samples, this could mean variations due to environmental conditions differences, especially in superficial waters.

Following with the analysis of β-diversity, the result achieved after separating samples by water layer (B-D, Fig. 4.4) showed differences among samples in which variables which correlated the most with the first NMDS dimension were depth, NO_3 and SiO_4 for each one of the three water layers, respectively, from top to bottom. This result supports results from the above metadata analysis, in which the deepest layers contained the largest concentrations of nutrients. Moreover, it was already known that at the photic zone the variables driving the viral communities were temperature/depth/light, but any study had reported that, at the aphotic zone, inorganic nutrients appear to be one of the major components contributing to the driving differences in viral community composition.

To complete the macrodiversity analysis, another study was performed, this time to evaluate the α-diversity or mean diversity of species in each sample. Result depicted in Fig. 4.5, representing the Shannon's index, shows how diversity tended to decrease until ~ 700 m deep, from which it stabilised. This tendency was also observed by Luo and colleagues (104) when studying vertical profiles in the North Pacific Subtropical Gyre, from where they revealed a peak in virioplankton diversity at the base of the euphotic zone. According to the study, this peak in diversity could reflect both habitat variability and transitions in microbial metabolic diversity. However, in the current analysis, there is not a noticeable peak, therefore the first hypothesis of habitat variability is the more coherent. In addition, this variability could also explain the large dispersion showed by the NMDS plot analysing β-diversity from Fig. 4.4 (A) and discussed before.

The other analysis to reach OBJ. 2, in which it was intended to describe the viral community composition at the global scale and across depth gradients, consisted in performing a taxonomic classification of viruses to deepen knowledge of which communities ruled each water layer. With respect to the relative abundances analysis by water layer, it was shown that *Myoviridae* was the most abundant in the three water layers, followed by *Microviridae*, *Siphoviridae* (in Bathypelagic) and *Podoviridae*. This result is in agreement with previous evidence that postulated these three families as the most abundant in oceanic ecosystems (105-107) although recently a new non-tailed family of viruses has also been discovered to be highly abundant (108). *Myoviridae* has a long contractile tail; *Siphoviridae*, a long noncontractile tail; and *Podoviridae*, a short noncontractile tail. All have an icosahedral head with a portal vertex connected to a neck structure followed by the tail. Regarding *Microviruses*, although they do not come from the same order and in spite of being tail-less, they follow an entry pathway similar to that of tailed phages during infection (109). In this mechanism, viruses adsorb reversibly to the cells, akin to the contact of tailed phage tail fibres with the host cell. After this adsorption, which allows the host bacteria to be detected, comes irreversible adsorption. The virus binds to lipopolysaccharide, causing conformational changes in the spike proteins and the release of DNA from the viral capsid to the host cytosol. The clear success of *Caudovirales* is thought to correspond to the fact that their long genome (they are defined as jumbo phages for having genomes sizes >200 kbp) confers them unusually complex functional capabilities, such as encoding entire transcriptional apparat (110) or sophisticated anti-CRISP defense mechanisms (111, 112). Regarding *Microviridae*, despite they show a mechanism of infection similar to *Caudovirales*, any hypothesis has arisen to justify their high abundance, since their genome is not as large as that of *Caudovirales*. In this regard, it is noticeable that another research (113) centred in the study of ubiquitous DNA viruses in agricultural soils also found order *Caudovirales* together with family *Microviridae* viruses as the most abundant taxa in those ecosystems.

Besides the taxonomical classification, a Mann-Whitney test was applied to assess the significance of changes in viral taxa relative abundance among water layers. The most significant change corresponded to *Sphaerolipoviridae*, for the three comparisons (Epipelagic – Bathypelagic, Epipelagic – Mesopelagic and Bathypelagic – Mesopelagic). Its abundance increased with depth. Members of this viral family include both archaeal viruses and bacteriophages (114, 115), and are conformed by a tailless icosahedral capsid with internal membrane (116). It is interesting that viruses from this family are commonly related to deep sea thermophilic bacteria and halophilic archaea (116). The present study supports the fact that there are significantly more viruses from family *Sphaerolipoviridae* in the deep ocean compared to upper layers. Furthermore, it has been reported that the abundance of archaea increases with depth (114), hence the higher abundance of these viruses in Mesopelagic and Bathypelagic. With respect to the second most significant change among layers, *Lavidaviridae* viral family was significantly more abundant also in the deep layers. As stated in bibliographical sources (117-120), *Lavidaviridae* viruses have the peculiarity of infecting other viruses and are therefore called virophages. The current analysis indicates by first time that the abundance of these viruses is significantly greater in the deep oceans.

At this stage of discussion, it is worth mentioning that both OBJ. 1 and OBJ. 2 were achieved. The first (OBJ. 1), with the description of samples variability at environmental and genomic levels through the analysis of new viral genomes never studied before, especially the ones from Mesopelagic and Bathypelagic ocean layers. The most relevant discovery was that inorganic nutrients concentration (NO_3 and SiO_4) were the most important variable in driving differences in viral community composition at the aphotic zone. Then, with regard to the second objective (OBJ. 2), besides the macrodiversity studies, which led to the discovery of new variables influencing aphotic samples differences, the taxonomical classification of viruses allowed to describe viral community composition at the global scale and across depth gradients. Novel insights regarding *Siphoviridae, Lavidaviridae* and *Sphaerolipoviridae* were obtained. The first, in addition to be one of the most abundant families in the ocean, it was found to be especially abundant in Bathypelagic. With respect to *Lavidaviridae* and *Sphaerolipoviridae*, they were found to be significantly more abundant in Bathypelagic than in the other layers.

The last objective (OBJ. 3) consisted in determining how viral communities were associated with environmental parameters, regarding taxonomic, targeted host and AMG diversity. To face this challenge, the analyses carried out consisted in: (i) the taxonomical classification of viruses (this was already discussed above for the purpose of achieving OBJ. 2), (ii) the viruses classification according to their predicted hosts and (iii) the functional annotation and classification of genes in search of AMGs. Analyses (ii) and (iii) are discussed below.

To perform the host prediction of viral genomes in the samples, first, the relative abundances analysis by water layer was conducted. As a result, it was shown that phylum *Proteobacteria* was notably the most abundant in all water column. It was followed by *Bacteroidetes* and *Cyanobacteria*. According to many studies (121-123), phylum *Proteobacteria* is not only the most abundant bacterial phylum in the oceans but also is one of the most widespread around the globe. An important detail is that the program used to predict host, *RaFAH*, specified that when it was applied the default cut-off of 0.14 (the minimum score to consider a host prediction as valid), all predictions which equalled or surpassed that value would have at least a 95% of accuracy in the prediction at phylum level. However, since the abundance of *Proteobacteria* was so high, the class level was represented. Results showed a clear predominance of *Gammaproteobacteria* and *Alphaproteobacteria*, in comparison with *Betaproteobacteria*. These results also support the bibliographical sources consulted (122, 124). With respect to phylum *Bacteroidetes*, studies also support their great abundance in the ocean (121-123, 125). Eventually, regarding phylum *Cyanobacteria*, numerous studies have reached to the same result

(106, 122, 126), showing that this phylum is also one of the most abundant, specially near the surface.

Then, the Mann-Whitney test was used to assess the significance of changes in viral relative abundance showing predicted hosts information among water layers. The most significant and interesting change corresponded to the predicted host phylum *Cyanobacteria*, one of the phyla which precisely showed a greater abundance in Epipelagic layer both when predicting the host (Fig. 4.7) as when analysing the abundance of that phylum in the samples (Fig. 4.8). It was expected to be a significantly higher abundance of this taxon in Epipelagic, given that they are most investigated oxygenic photosynthetic bacteria. Studies have demonstrated that *Cyanobacteria* are the main primary producers in marine habitats, emitting O_2 and fixating CO_2 from the atmosphere, and therefore have a fundamental role on sustaining marine ecosystems (14, 127, 128).

With respect to the study of AMGs diversity, embedded into OBJ. 3, a functional profiling of samples and depths was performed, this way stablishing the functions driven by viruses depending on the water column and discovering roles that both bacteria and viruses developed at each depth. At first sight (Fig. 4.9), for KEGG metabolic pathways abundance, there was no predominance of a specific metabolic activity in any water layer, except for “replication and repair”, which was highly represented along all the water column, despite it is relevant to point out that genes included in this pathway are viral replication genes, not AMGs. Moreover, there was a higher abundance of metabolisms in general in the Epipelagic layer. With respect to the “replication and repair” pathway, many studies corroborate its high abundance (129-131) on the grounds that several of the genes associated with viral genome replication (e.g., DNA polymerases, helicases, etc.) belong to this metabolic pathway.

After that, the Mann-Whitney test was applied to assess the significance of changes in metabolic pathways relative abundances among water layers. The most significant change was an improved “Energy Metabolism” in the Epipelagic layer in comparison to the other two layers. This higher influence of viruses in the energetic metabolism of their host could be related with a higher efficiency and lower cost of synthesis of nucleotides and amino acids which would contribute to boost the viral infection. This is supported by other studies (131-133). The second most significant abundance difference was “Cell Motility”, when compared Epipelagic and Bathypelagic layers, being more abundant in Bathypelagic. Many bacteria are motile and propel themselves by rotating helical flagella driven by molecular motors (134), a process in which genes related had also been observed in viromes (135). This result coincides with what Hurwitz and colleagues (131) hypothesises of that such genes for “chemotaxis and motility” may boost viral hosts’ motility to improve nutrient acquisition in the deep sea.

In a nutshell, the description of viral communities at global level, emphasising in their hosts phyla, found that *Proteobateria* (specially *Gammaproteobacteria* and *Alphaproteobacteria*), *Bacteroidetes* and *Cyanobacteria* were the most abundant predicted hosts in the ocean, being *Cyanobacteria* significantly more abundant in

Epipelagic layer compared to the deep ocean. Then, regarding functional annotation, the "Replication and Repair" metabolic pathway abundance was widely spread over broad latitude and depth gradients. In addition, it was also found that both AMGs related to "Energy Metabolism" and "Cell Motility" pathways were significantly more abundant in Epipelagic and Bathypelagic layers, respectively. The previous discoveries confirm that the third objective (OBJ. 3) was properly achieved.

As final remark, it is important to mention that for all results obtained after performing the Mann-Whitney test of significance, it is assumed a 29.3% of probability of inflation of Type I Error Probability (TIEP), which consists in rejecting the null hypothesis when it is actually true. The reason is that part of data are not homoscedastic, so several studies show that this condition alters the significance of results (136, 137).

6. CONCLUSIONS

After having studied the viral communities from vertical profiles retrieved from the Malaspina Expedition at tropical and subtropical latitudes, making use of metagenomics, some conclusions can be elucidated:

- Variables influencing the most in samples distribution according to environmental metadata are temperature, salinity and inorganic nutrients concentrations (NO_3, SiO_4 and PO_4). Moreover, temperature appears to be one of the major components contributing to the vertical β-diversity in viral populations.

- Light availability is the most important variable in driving differences in viral community composition at the photic zone. In addition, this study reports by first time that inorganic nutrients concentration (NO_3 and SiO_4) is the most important variable in driving differences in viral community composition at the aphotic zone.

- The higher mean diversity of species by sample is observed in the Epipelagic layer. This diversity decreases with depth until ~700 m, from which it stabilises. Changes in surface environmental conditions along the different longitudes may contribute to this result.

- With respect to viral taxonomic classification, family *Myoviridae* was the most abundant in the three water layers, followed by *Microviridae*, *Siphoviridae* and *Podoviridae*. This result supports those from other studies, as well as contribute with new information related to *Siphoviridae*, which has been observed to be especially abundant in the deep ocean.

- Regarding novel viral genomic diversity, this analysis indicates by first time that *Lavidaviridae* viruses, characterised by infecting other viruses, are significantly more abundant in the deep oceans. Furthermore, in respect of family

Sphaerolipoviridae, commonly related to deep sea thermophilic bacteria and halophilic archaea, the present study unravels that there are significantly more viruses from this family in the deep ocean compared to upper layers.

- The description of viral communities at global level, emphasising in their hosts phyla, have found that *Proteobateria* (specially *Gammaproteobacteria* and *Alphaproteobacteria*), *Bacteroidetes* and *Cyanobacteria* are the most abundant predicted hosts in the ocean. Moreover, *Cyanobacteria* is significantly more abundant in Epipelagic layer compared to the deep ocean, where they act as the main primary producers in marine habitats.

- It has been shown that AMGs are mostly related to the “Replication and Repair” metabolic pathways along broad latitude and depth gradients. Besides, this study underlines new insights regarding “Energy Metabolism” and “Cell Motility” pathways to enhance viral infections effectiveness. The “Energy Metabolism” is more significantly exploited by viruses during infection in the shallow ocean (Epipelagic), while “Cell Motility” is significantly more relevant in the deep ocean (Bathypelagic).

BIBLIOGRAPHY

1. Suttle, C. A. (2007). Marine viruses—Major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(10), 801-812. https://doi.org/10.1038/nrmicro1750
2. Costello, M. J., Cheung, A., & De Hauwere, N. (2010). Surface Area and the Seabed Area, Volume, Depth, Slope, and Topographic Variation for the World's Seas, Oceans, and Countries. *Environmental Science & Technology*, *44*(23), 8821-8828. https://doi.org/10.1021/es1012752
3. *The Deep Sea ~ MarineBio Conservation Society*. (2018, junio 17). https://www.marinebio.org/oceans/deep-sea/
4. Turk, D., McPhaden, M. J., Busalacchi, A. J., & Lewis, M. R. (2001). Remotely sensed biological production in the equatorial Pacific. *Science (New York, N.Y.)*, *293*(5529), 471-474. https://doi.org/10.1126/science.1056449
5. *Ocean Circulation—2nd Edition*. (s. f.). Retrieved on july 4th, 2022, from https://www.elsevier.com/books/ocean-circulation/open-university/978-0-7506-5278-0
6. Gasol, J. M., & Kirchman, D. L. (2018). *Microbial Ecology of the Ocean: Third edition*. https://digital.csic.es/handle/10261/172945
7. Fischer, B. (1894). *Die Bakterien des Meeres nach den Untersuchungen der Plankton-Expedition: Unter gleichzeitiger Berücksichtigung einiger älterer und neuerer Untersuchungen*. Lipsius & Tischer.
8. Issatchenko, B. L. 1914. Investigations on the bacteria of the glacial Artic Ocean. Monograph, Petrograd, 300 pp
9. Wilhelm, S. W. & Suttle, C. A. Viruses and nutrient cycles in the sea. *Bioscience* **49**, 781–788 (1999).
10. Fuhrman, J. A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548.
11. Wommack, K. E. & Colwell, R. R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114.
12. Suttle, C. A. (2005) Viruses in the sea. *Nature* **437**, 356–361.
13. Weinbauer, M. G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181.
14. Coutinho, F. H., Silveira, C. B., Gregoracci, G. B., Thompson, C. C., Edwards, R. A., Brussaard, C. P. D., Dutilh, B. E., & Thompson, F. L. (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, *8*, 15955. https://doi.org/10.1038/ncomms15955
15. Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology*, *3*(7), 754-766. https://doi.org/10.1038/s41564-018-0166-y
16. Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., Weinbauer, M. G., Luo, T., Chen, F., & Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: Long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, *8*(8), 593-599. https://doi.org/10.1038/nrmicro2386
17. Wigington, C. H., Sonderegger, D., Brussaard, C. P. D., Buchan, A., Finke, J. F., Fuhrman, J. A., Lennon, J. T., Middelboe, M., Suttle, C. A., Stock, C., Wilson, W. H., Wommack, K. E., Wilhelm, S. W., & Weitz, J. S. (2016). Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology*, *1*(3), 15024. https://doi.org/10.1038/nmicrobiol.2015.24
18. Munn, C.B. (2006). Viruses as pathogens of marine organisms - from bacteria to whales. *J. Mar. Biol. Ass. U.K. 86(3)*: 453-467. https://dx.doi.org/10.1017/S002531540601335X
19. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., & Sullivan, M. B. (2017). Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *The ISME Journal*, *11*(7), 1511-1520. https://doi.org/10.1038/ismej.2017.16
20. *Lytic and Lysogenic Cycle*. (s. f.). Recuperado 2 de julio de 2022, de https://app.biorender.com/biorender-templates
21. Laber, C. P., Hunter, J. E., Carvalho, F., Collins, J. R., Hunter, E. J., Schieler, B. M., Boss, E., More, K., Frada, M., Thamatrakoln, K., Brown, C. M., Haramaty, L., Ossolinski, J., Fredricks, H., Nissimov, J. I., Vandzura, R., Sheyn, U., Lehahn, Y., Chant, R. J., … Bidle, K. D. (2018). Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology*, *3*(5), 537-547. https://doi.org/10.1038/s41564-018-0128-4
22. Sullivan, M. B., Weitz, J. S., & Wilhelm, S. (2017). Viral ecology comes of age. *Environmental Microbiology Reports*, *9*(1), 33-35. https://doi.org/10.1111/1758-2229.12504
23. Thinstad, Tf., & Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology*, *13*(1), 19-27. https://doi.org/10.3354/ame013019
24. Chen, X., Weinbauer, M. G., Jiao, N., & Zhang, R. (2021). Revisiting marine lytic and lysogenic virus-host interactions: Kill-the-Winner and Piggyback-the-Winner. *Science Bulletin*, *66*(9), 871-874. https://doi.org/10.1016/j.scib.2020.12.014
25. Silveira, C. B., & Rohwer, F. L. (2016). Piggyback-the-Winner in host-associated microbial communities. *Npj Biofilms and Microbiomes*, *2*(1), 1-5. https://doi.org/10.1038/npjbiofilms.2016.10
26. Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., … Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, *531*(7595), 466-470. https://doi.org/10.1038/nature17193
27. Chen, X., Ma, R., Yang, Y., Jiao, N., & Zhang, R. (2019). Viral Regulation on Bacterial Community Impacted by Lysis-Lysogeny Switch: A Microcosm Experiment in Eutrophic Coastal Waters. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.01763
28. Jacobson, T. B., Callaghan, M. M., & Amador-Noguez, D. (2021). Hostile Takeover: How Viruses Reprogram Prokaryotic Metabolism. *Annual Review of Microbiology*, *75*(1), 515-539. https://doi.org/10.1146/annurev-micro-060621-043448
29. Forterre, P. (2013). The virocell concept and environmental microbiology. *The ISME Journal*, *7*(2), 233-236. https://doi.org/10.1038/ismej.2012.110
30. Hurwitz, B. L., & U'Ren, J. M. (2016). Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology*, *31*, 161-168. https://doi.org/10.1016/j.mib.2016.04.002
31. Tuttle, M. J., & Buchan, A. (2020). Lysogeny in the oceans: Lessons from cultivated model systems and a reanalysis of its prevalence. *Environmental Microbiology*, *22*(12), 4919-4933. https://doi.org/10.1111/1462-2920.15233
32. Coutinho, F. H., Gregoracci, G. B., Walter, J. M., Thompson, C. C., & Thompson, F. L. (2018). Metagenomics Sheds Light on the Ecology of Marine Microbes and Their Viruses. *Trends in Microbiology*, *26*(11), 955-965. https://doi.org/10.1016/j.tim.2018.05.015
33. Sosik, H., Olson, R., & Armbrust, E. (1970). *Flow Cytometry in Phytoplankton Research* (pp. 171-185). https://doi.org/10.1007/978-90-481-9268-7_8
34. Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A., & Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences*, *104*(29), 11889-11894. https://doi.org/10.1073/pnas.0704662104
35. Crespo, B. G., Wallhead, P. J., Logares, R., & Pedrós-Alió, C. (2016). Probing the Rare Biosphere of the North-West Mediterranean Sea: An Experiment with High Sequencing Effort. *PLOS ONE*, *11*(7), e0159195. https://doi.org/10.1371/journal.pone.0159195
36. Roux, S., Matthijnssens, J., & Dutilh, B. E. (2021). Metagenomics in Virology. *Encyclopedia of Virology*, 133-140. https://doi.org/10.1016/B978-0-12-809633-8.20957-6
37. Payne, S. (2017). Methods to Study Viruses. *Viruses*, 37-52. https://doi.org/10.1016/B978-0-12-803109-4.00004-0
38. Willner, D., & Hugenholtz, P. (2013). From deep sequencing to viral tagging: Recent advances in viral metagenomics. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *35*(5), 436-442. https://doi.org/10.1002/bies.201200174
39. Brum, J. R., & Sullivan, M. B. (2015). Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nature Reviews Microbiology*, *13*(3), 147-159. https://doi.org/10.1038/nrmicro3404
40. Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: Read Length Matters. *Applied and Environmental Microbiology*, *74*(5), 1453-1463. https://doi.org/10.1128/AEM.02181-07
41. Hurwitz, B. L., & 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*(2), e57355. https://doi.org/10.1371/journal.pone.0057355

42. Hurwitz, B. L., Westveld, A. H., Brum, J. R., & Sullivan, M. B. (2014). Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(29), 10714-10719. https://doi.org/10.1073/pnas.1319778111
43. Arístegui, J., Gasol, J. M., Duarte, C. M., & Herndld, G. J. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, *54*(5), 1501-1529. https://doi.org/10.4319/lo.2009.54.5.1501
44. Herndl, G. J., & Reinthaler, T. (2013). Microbial control of the dark end of the biological pump. *Nature Geoscience*, *6*(9), 718-724. https://doi.org/10.1038/ngeo1921
45. Duarte, C. M. (2015). Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, *24*(1), 11-14. https://doi.org/10.1002/lob.10008
46. *Malaspina Expedition* (s. f.). Recuperado 2 de julio de 2022, de http://www.expedicionmalaspina.es/
47. Grasshoff, K., Ehrhardt, M. and Kremling, K. (1983) Methods of Seawater Analysis. 2nd Edition, Verlag Chemie Weinhein, New York, 419 p.
48. Boyer, T. P., Antonov, J. I., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Johnson, D. R., Locarnini, R. A., Mishonov, A. V., O'Brien, T. D., Paver, C. R., Reagan, J. R., Seidov, D., Smolyar, I. V., & Zweng, M. M. (2013). *World Ocean Database 2013.* [Report]. NOAA Printing Office. https://doi.org/10.25607/OBP-1454
49. Longhurst, A. R. (2007). *Ecological geography of the sea*. Academic Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=182190
50. *Malaspinomics: Sequencing the genome of the deep ocean | cnag.crg.eu*. (s. f.). Recuperado 2 de julio de 2022, de https://www.cnag.cat/news/malaspinomics-sequencing-genome-deep-ocean
51. Hannon. (2010)*. FASTX-Toolkit*. Recuperado 2 de julio de 2022, de http://hannonlab.cshl.edu/fastx_toolkit/
52. Andrews. (2010)*. Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. Recuperado 2 de julio de 2022, de https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
53. Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834. https://doi.org/10.1101/gr.213959.116
54. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. https://doi.org/10.1002/cpbi.102
55. Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1), 90. https://doi.org/10.1186/s40168-020-00867-0
56. Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578-585. https://doi.org/10.1038/s41587-020-00774-7
57. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923
58. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352
59. Gregory, A. C., Gerhardt, K., Zhong, Z.-P., Bolduc, B., Temperton, B., Konstantinidis, K. T., & Sullivan, M. B. (2022). MetaPop: A pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome*, *10*, 49. https://doi.org/10.1186/s40168-022-01231-0
60. Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., & Weitz, J. S. (2013). Robust estimation of microbial diversity in theory and in practice. *The ISME Journal*, *7*(6), 1092-1101. https://doi.org/10.1038/ismej.2013.10
61. *vegan: An R package for community ecologists*. (2022). [R]. vegandevs. https://github.com/vegandevs/vegan (Original work published 2012)
62. Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, *27*(4), 325-349. https://doi.org/10.2307/1942268
63. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
64. Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Toomet, O., Crowley, J., Hofmann, H., & Wickham, H. (2021). *GGally: Extension to «ggplot2»* (2.1.2) [Computer software]. https://CRAN.R-project.org/package=GGally
65. Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., Lee, J.-H., & Isaacson, H. B. K. and R. E. (2017). *Deciphering Diversity Indices for a Better Understanding of Microbial Communities*. *27*(12), 2089-2093. https://doi.org/10.4014/jmb.1709.09027
66. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
67. Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119
68. Pons, J. C., Paez-Espino, D., Riera, G., Ivanova, N., Kyrpides, N. C., & Llabrés, M. (2021). VPF-Class: Taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics*, *37*(13), 1805-1813. https://doi.org/10.1093/bioinformatics/btab026
69. Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*(1), 431. https://doi.org/10.1186/1471-2105-11-431
70. Coutinho, F. H., Zaragoza-Solas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, *2*(7). https://doi.org/10.1016/j.patter.2021.100274
71. *UniProt*. (s. f.). Recuperado 2 de julio de 2022, de https://www.uniprot.org/
72. The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158-D169. https://doi.org/10.1093/nar/gkw1099
73. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. (s. f.). Recuperado 2 de julio de 2022, de https://www.genome.jp/kegg/
74. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353-D361. https://doi.org/10.1093/nar/gkw1092
75. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412-D419. https://doi.org/10.1093/nar/gkaa913
76. Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176
77. Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366-368. https://doi.org/10.1038/s41592-021-01101-x
78. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282-1288. https://doi.org/10.1093/bioinformatics/btm098
79. Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614-620. https://doi.org/10.1093/bioinformatics/btt593
80. *FASTA Splitter*. (s. f.). Recuperado 2 de julio de 2022, de http://kirill-kryukov.com/study/tools/fasta-splitter/
81. Pertea, G. (2018). *Gpertea/cdbfasta* [C++]. https://github.com/gpertea/cdbfasta (Original work published 2017)
82. *Cdb tools for fasta files*. (s. f.). Recuperado 2 de julio de 2022, de https://vcru.wisc.edu/simonlab/bioinformatics/programs/cdbfasta/cdbfasta_usage.html
83. *The Perl Programming Language—Www.perl.org*. (s. f.). Recuperado 2 de julio de 2022, de https://www.perl.org/

84. *Silva*. (s. f.). Recuperado 2 de julio de 2022, de https://www.arb-silva.de/
85. Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2014). The SILVA and «All-species Living Tree Project (LTP)» taxonomic frameworks. *Nucleic Acids Research*, *42*(Database issue), D643-648. https://doi.org/10.1093/nar/gkt1209
86. Buttigieg, P. L., & Ramette, A. (2014). A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, *90*(3), 543-550. https://doi.org/10.1111/1574-6941.12437
87. Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591-611. https://doi.org/10.2307/2333709
88. Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, *22*(1), 67-72. https://doi.org/10.4103/aca.ACA_157_18
89. Flores M., P., & Ocaña, J. (2018). Heteroscedasticity irrelevance when testing means difference. *SORT : statistics and operations research transactions*, *42*(1), 0059-0072. https://doi.org/10.2436/20.8080.02.69
90. Fligner, M., & Killeen, T. (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of The American Statistical Association - J AMER STATIST ASSN*, *71*, 210-213. https://doi.org/10.1080/01621459.1976.10481517
91. Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50-60. https://doi.org/10.1214/aoms/1177730491
92. *EWOCE*. (s. f.). Recuperado 2 de julio de 2022, de https://www.ewoce.org/index.html
93. Forrest, J., Marcucci, E., & Scott, P. (2007). *Geothermal gradients and subsurface temperatures in northern Gulf of Mexico*. *55*.
94. Kopylov, A. I., Zabotkina, E. A., Romanenko, A. V., Kosolapov, D. B., & Sazhin, A. F. (2020). Viruses in the water column and the sediment of the eastern part of the Laptev Sea. *Estuarine, Coastal and Shelf Science*, *242*, 106836. https://doi.org/10.1016/j.ecss.2020.106836
95. *Temperature–Salinity Structure of the North Atlantic Circulation and Associated Heat and Freshwater Transports in: Journal of Climate Volume 29 Issue 21 (2016)*. (s. f.). Recuperado 2 de julio de 2022, de https://journals.ametsoc.org/view/journals/clim/29/21/jcli-d-15-0798.1.xml
96. Cutter, G. A., & Measures, C. I. (1999). The 1996 IOC contaminant baseline survey in the Atlantic Ocean from 33°S to 10°N: Introduction, sampling protocols, and hydrographic data. *Deep Sea Research Part II: Topical Studies in Oceanography*, *46*(5), 867-884.
97. *Frontiers | Vertical Beta-Diversity of Bacterial Communities Depending on Water Stratification*. (s. f.). Recuperado 2 de julio de 2022, de https://www.frontiersin.org/articles/10.3389/fmicb.2020.00449/full#F3
98. Cram, J. A., Chow, C.-E. T., Sachdeva, R., Needham, D. M., Parada, A. E., Steele, J. A., & Fuhrman, J. A. (2015). Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *The ISME Journal*, *9*(3), 563-580. https://doi.org/10.1038/ismej.2014.153
99. Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359. https://doi.org/10.1126/science.1261359
100. Lawes, J. C., Neilan, B. A., Brown, M. V., Clark, G. F., & Johnston, E. L. (2016). Elevated nutrients change bacterial community composition and connectivity: High throughput sequencing of young marine biofilms. *Biofouling*, *32*(1), 57-69. https://doi.org/10.1080/08927014.2015.1126581
101. Walsh, E. A., Kirkpatrick, J. B., Rutherford, S. D., Smith, D. C., Sogin, M., & D'Hondt, S. (2016). Bacterial diversity and community composition from seasurface to subseafloor. *The ISME Journal*, *10*(4), 979-989. https://doi.org/10.1038/ismej.2015.175
102. DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W., & Karl, D. M. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science (New York, N.Y.)*, *311*(5760), 496-503. https://doi.org/10.1126/science.1120250
103. Brown, M. V., Philip, G. K., Bunge, J. A., Smith, M. C., Bissett, A., Lauro, F. M., Fuhrman, J. A., & Donachie, S. P. (2009). Microbial community structure in the North Pacific ocean. *The ISME Journal*, *3*(12), 1374-1386. https://doi.org/10.1038/ismej.2009.86
104. Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R., & DeLong, E. F. (2020). Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *The ISME Journal*, *14*(5), 1304-1315. https://doi.org/10.1038/s41396-020-0604-8
105. Iwasaki, T., Yamashita, E., Nakagawa, A., Enomoto, A., Tomihara, M., & Takeda, S. (2018). Three-dimensional structures of bacteriophage neck subunits are shared in Podoviridae, Siphoviridae and Myoviridae. *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, *23*(7), 528-536. https://doi.org/10.1111/gtc.12594
106. Sullivan, M. B., Waterbury, J. B., & Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature*, *424*(6952), 1047-1051. https://doi.org/10.1038/nature01929
107. López-Pérez, M., Haro-Moreno, J. M., de la Torre, J. R., & Rodriguez-Valera, F. (2019). Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled from metagenomes. *Environmental Microbiology*, *21*(6), 1980-1988. https://doi.org/10.1111/1462-2920.14462
108. *A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria | Nature*. (s. f.). Recuperado 2 de julio de 2022, de https://www.nature.com/articles/nature25474
109. Poranen, M. M., & Domanska, A. (2008). Virus Entry to Bacterial Cells. En B. W. J. Mahy & M. H. V. Van Regenmortel (Eds.), *Encyclopedia of Virology (Third Edition)* (pp. 365-370). Academic Press. https://doi.org/10.1016/B978-012374410-4.00746-9
110. Ceyssens, P.-J., Minakhin, L., Van den Bossche, A., Yakunina, M., Klimuk, E., Blasdel, B., De Smet, J., Noben, J.-P., Bläsi, U., Severinov, K., & Lavigne, R. (2014). Development of Giant Bacteriophage φKZ Is Independent of the Host Transcription Apparatus. *Journal of Virology*, *88*(18), 10501-10510. https://doi.org/10.1128/JVI.01347-14
111. Malone, L. M., Warring, S. L., Jackson, S. A., Warnecke, C., Gardner, P. P., Gumy, L. F., & Fineran, P. C. (2020). A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nature Microbiology*, *5*(1), 48-55. https://doi.org/10.1038/s41564-019-0612-5
112. Mendoza, S. D., Nieweglowska, E. S., Govindarajan, S., Leon, L. M., Berry, J. D., Tiwari, A., Chaikeeratisak, V., Pogliano, J., Agard, D. A., & Bondy-Denomy, J. (2020). A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature*, *577*(7789), 244-248. https://doi.org/10.1038/s41586-019-1786-y
113. Han, L.-L., Yu, D.-T., Zhang, L.-M., Shen, J.-P., & He, J.-Z. (2017). Genetic and functional diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. *Scientific Reports*, *7*, 45142. https://doi.org/10.1038/srep45142
114. Rambo, I. M., Langwig, M. V., Leão, P., De Anda, V., & Baker, B. J. (2022). Genomes of six viruses that infect Asgard archaea from deep-sea sediments. *Nature Microbiology*, *7*(7), 953-961. https://doi.org/10.1038/s41564-022-01150-8
115. Pawlowski, A., Rissanen, I., Bamford, J. K. H., Krupovic, M., & Jalasvuori, M. (2014). Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. *Archives of Virology*, *159*(6), 1541-1554. https://doi.org/10.1007/s00705-013-1970-6
116. Demina, T. A., Pietilä, M. K., Svirskaitė, J., Ravantti, J. J., Atanasova, N. S., Bamford, D. H., & Oksanen, H. M. (2017). HCIV-1 and Other Tailless Icosahedral Internal Membrane-Containing Viruses of the Family Sphaerolipoviridae. *Viruses*, *9*(2), 32. https://doi.org/10.3390/v9020032
117. Duponchel, S., & Fischer, M. G. (2019). Viva lavidaviruses! Five features of virophages that parasitize giant DNA viruses. *PLoS Pathogens*, *15*(3), e1007592. https://doi.org/10.1371/journal.ppat.1007592
118. Wilhelm, S. W., Coy, S. R., Gann, E. R., Moniruzzaman, M., & Stough, J. M. A. (2016). Standing on the Shoulders of Giant

Viruses: Five Lessons Learned about Large Viruses Infecting Small Eukaryotes and the Opportunities They Create. *PLOS Pathogens*, *12*(8), e1005752. https://doi.org/10.1371/journal.ppat.1005752

119. La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., & Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, *455*(7209), 100-104. https://doi.org/10.1038/nature07218
120. Fischer, M. G. (2021). The Virophage Family Lavidaviridae. *Current Issues in Molecular Biology*, *40*, 1-24. https://doi.org/10.21775/cimb.040.001
121. Zhou, Z., Tran, P. Q., Kieft, K., & Anantharaman, K. (2020). Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation. *The ISME Journal*, *14*(8), 2060-2077. https://doi.org/10.1038/s41396-020-0669-4
122. Wang, J., Kan, J., Borecki, L., Zhang, X., Wang, D., & Sun, J. (2016). A snapshot on spatial and vertical distribution of bacterial communities in the eastern Indian Ocean. *Acta Oceanologica Sinica*, *35*, 85-93. https://doi.org/10.1007/s13131-016-0871-4
123. Kong, J., Liu, X., Wang, L., Huang, H., Ou, D., Guo, J., Laws, E. A., & Huang, B. (2021). Patterns of Relative and Quantitative Abundances of Marine Bacteria in Surface Waters of the Subtropical Northwest Pacific Ocean Estimated With High-Throughput Quantification Sequencing. *Frontiers in Microbiology*, *11*. https://www.frontiersin.org/articles/10.3389/fmicb.2020.599614
124. Buijs, Y., Bech, P. K., Vazquez-Albacete, D., Bentzon-Tilia, M., Sonnenschein, E. C., Gram, L., & Zhang, S.-D. (2019). Marine Proteobacteria as a source of natural products: Advances in molecular tools and strategies. *Natural Product Reports*, *36*(9), 1333-1350. https://doi.org/10.1039/C9NP00020H
125. Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Acinas, S. G., González, J. M., & Pedrós-Alió, C. (2013). Ecology of marine Bacteroidetes: A comparative genomics approach. *The ISME Journal*, *7*(5), 1026-1037. https://doi.org/10.1038/ismej.2012.169
126. Partensky, F., Hess, W. R., & Vaulot, D. (1999). Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance. *Microbiology and Molecular Biology Reviews*, *63*(1), 106-127. https://doi.org/10.1128/MMBR.63.1.106-127.1999
127. Schuurmans, R. M., Alphen, P. van, Schuurmans, J. M., Matthijs, H. C. P., & Hellingwerf, K. J. (2015). Comparison of the Photosynthetic Yield of Cyanobacteria and Green Algae: Different Methods Give Different Answers. *PLOS ONE*, *10*(9), e0139061. https://doi.org/10.1371/journal.pone.0139061
128. Bižić, M., Klintzsch, T., Ionescu, D., Hindiyeh, M. Y., Günthel, M., Muro-Pastor, A. M., Eckert, W., Urich, T., Keppler, F., & Grossart, H.-P. (2020). Aquatic and terrestrial cyanobacteria produce methane. *Science Advances*, *6*(3), eaax5343. https://doi.org/10.1126/sciadv.aax5343
129. Mara, P., Vik, D., Pachiadaki, M. G., Suter, E. A., Poulos, B., Taylor, G. T., Sullivan, M. B., & Edgcomb, V. P. (2020). Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *The ISME Journal*, *14*(12), 3079-3092. https://doi.org/10.1038/s41396-020-00739-3
130. Shapiro, J. A. (1979). Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings of the National Academy of Sciences*, *76*(4), 1933-1937. https://doi.org/10.1073/pnas.76.4.1933
131. Hurwitz, B. L., Brum, J. R., & Sullivan, M. B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *The ISME Journal*, *9*(2), 472-484. https://doi.org/10.1038/ismej.2014.143
132. Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., Hess, M., Sullivan, M. B., Walsh, D. A., Roux, S., & Anantharaman, K. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-23698-5
133. Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., & Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences*, *108*(39), E757-E764. https://doi.org/10.1073/pnas.1102164108
134. Stocker, R., & Seymour, J. R. (2012). Ecology and Physics of Bacterial Chemotaxis in the Ocean. *Microbiology and Molecular Biology Reviews : MMBR*, *76*(4), 792-812. https://doi.org/10.1128/MMBR.00029-12
135. Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., … Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. *Nature*, *452*(7187), 629-632. https://doi.org/10.1038/nature06810
136. Flores, P, Ocaña, J. (2018) "Heteroscedasticity irrelevance when testing means difference". *SORT-Statistics and Operations Research Transactions*, [online], , Vol. 1, Num. 1, pp. 59-72, https://raco.cat/index.php/SORT/article/view/338210 [View: 4-07-2022].
137. Overall, J.E., R.S. Atlas and J.M. Gibson (1995). Tests that are robust against variance heterogeneity in k × 2 designs with unequal cell frequencies. Psychological Reports, 76, 1011–1017

APPENDICES

Table A1. Information of depth and location from Malaspina Expedition metagenomic samples.

Sample code	Station	Ocean	Depth
MP0311	19	South Atlantic	Epipelagic
MP0313			Bathypelagic
MP0315			Bathypelagic
MP0317			Mesopelagic
MP0319			Mesopelagic
MP0321			Epipelagic
MP0323			Epipelagic
MP0528	30	South Atlantic	Epipelagic
MP0530			Bathypelagic
MP0532			Bathypelagic
MP0534			Bathypelagic
MP0536			Mesopelagic
MP0538			Mesopelagic
MP0540			Epipelagic
MP0778	44	South Atlantic	Epipelagic
MP0780			Bathypelagic
MP0782			Bathypelagic
MP0784			Bathypelagic
MP0786			Mesopelagic
MP0788			Mesopelagic
MP0790			Epipelagic
MP0878	49	Indian	Bathypelagic
MP0880			Bathypelagic
MP0882			Bathypelagic
MP0884			Mesopelagic
MP0886			Mesopelagic
MP0888			Epipelagic
MP1154	63	Indian	Bathypelagic
MP1162			Bathypelagic
MP1164			Mesopelagic
MP1166			Mesopelagic
MP1174			Epipelagic
MP1176			Epipelagic
MP1178			Mesopelagic
MP1409	76	Indian (Great Australian Bight)	Bathypelagic
MP1411			Bathypelagic
MP1413			Mesopelagic
MP1415			Mesopelagic
MP1417			Mesopelagic
MP1419			Epipelagic
MP1421			Epipelagic
MP1517	83	Western Pacific	Epipelagic

MP1519			Bathypelagic
MP1521			Bathypelagic
MP1523			Mesopelagic
MP1525			Mesopelagic
MP1527			Epipelagic
MP1529			Epipelagic
MP1672	92	Western Pacific	Epipelagic
MP1674			Bathypelagic
MP1676			Bathypelagic
MP1678			Mesopelagic
MP1680			Mesopelagic
MP1682			Mesopelagic
MP1684			Epipelagic
MP1845	101	North Pacific	Bathypelagic
MP1847			Bathypelagic
MP1849			Mesopelagic
MP1851			Mesopelagic
MP1853			Mesopelagic
MP1855			Epipelagic
MP1857			Epipelagic
MP2231	120	Eastern Pacific	Bathypelagic
MP2233			Bathypelagic
MP2235			Mesopelagic
MP2237			Mesopelagic
MP2239			Mesopelagic
MP2241			Epipelagic
MP2243			Epipelagic
MP2809	141	North Atlantic	Bathypelagic
MP2811			Bathypelagic
MP2813			Bathypelagic
MP2815			Mesopelagic
MP2817			Mesopelagic
MP2819			Epipelagic
MP2821			Epipelagic

Table A2. Number of contigs, number of bases, maximum contig length, N50 and N90 of multi-fasta files obtained from metagenomes before having filtered sequences greater than 1000 bp.

File	Scaffolds	Bases	Max	N50	N90
MP0311.fasta	381437	411681402	169994	1111	564
MP0313.fasta	297837	416220625	449642	1818	592
MP0315.fasta	300562	381960611	226773	1480	583
MP0317.fasta	275537	330074471	806207	1344	573
MP0319.fasta	314347	363491885	102976	1262	571
MP0321.fasta	406066	429998605	189365	1078	561
MP0323.fasta	488430	525472018	189298	1116	566

MP0528.fasta	370674	472929182	255067	1462	594
MP0530.fasta	303792	453674918	1592702	2035	610
MP0532.fasta	332603	458628264	766578	1737	596
MP0534.fasta	316539	453514377	583047	1891	603
MP0536.fasta	381998	468076497	613280	1368	581
MP0538.fasta	199493	284204181	1208588	1797	598
MP0540.fasta	422834	503623910	364958	1314	580
MP0778.fasta	430240	590292540	303241	1681	604
MP0780.fasta	161058	328423991	962570	4321	684
MP0782.fasta	150257	275230603	792149	3528	650
MP0784.fasta	205992	340390142	1885427	2549	631
MP0786.fasta	273291	364409468	881066	1598	592
MP0788.fasta	284214	385676095	967705	1664	592
MP0790.fasta	346028	585215644	607124	2542	655
MP0878.fasta	271150	439742452	1082667	2427	630
MP0880.fasta	280847	432177344	1056618	2165	621
MP0882.fasta	295606	390593166	848140	1584	598
MP0884.fasta	349193	444115265	558882	1454	588
MP0886.fasta	356661	426223713	645563	1312	575
MP0888.fasta	365043	421957612	157496	1260	573
MP1154.fasta	305091	492700341	1175840	2384	629
MP1162.fasta	313967	465440663	573606	2010	610
MP1164.fasta	311572	410376636	1056421	1578	591
MP1166.fasta	353891	444551028	572396	1462	587
MP1174.fasta	395734	445378458	155175	1206	573
MP1176.fasta	366445	446781446	445741	1373	588
MP1178.fasta	417822	500654579	161394	1321	581
MP1409.fasta	319902	475718351	897297	2096	603
MP1411.fasta	311411	459901826	933877	1976	605
MP1413.fasta	300892	363782817	827084	1333	576
MP1415.fasta	333120	369327670	173342	1161	566
MP1417.fasta	463489	520331295	191491	1190	567
MP1419.fasta	449337	525342473	572431	1266	571
MP1421.fasta	389798	478033751	185072	1362	575
MP1517.fasta	398761	459017024	252678	1224	574
MP1519.fasta	277266	429181591	740565	2184	620
MP1521.fasta	308784	450380531	725129	1979	599
MP1523.fasta	299527	356537138	572481	1306	574
MP1525.fasta	441001	481183980	730354	1121	562
MP1527.fasta	628310	687946749	507528	1128	565
MP1529.fasta	431886	482103574	1053974	1162	562
MP1672.fasta	366972	452626388	503252	1389	577
MP1674.fasta	280408	476952959	1818184	2824	629
MP1676.fasta	310743	482917898	723362	2198	620
MP1678.fasta	287930	366030856	799118	1482	578
MP1680.fasta	279533	359481732	1169690	1476	578

MP1682.fasta	311415	364984137	897428	1274	573
MP1684.fasta	470601	522565082	269979	1162	568
MP1845.fasta	249041	440312957	826662	3079	644
MP1847.fasta	307422	479157036	2458067	2160	618
MP1849.fasta	338382	476652742	897432	1774	601
MP1851.fasta	291070	398355286	1189902	1659	589
MP1853.fasta	388994	454664331	807958	1252	568
MP1855.fasta	423431	451287203	1231286	1084	559
MP1857.fasta	422585	487524189	1786432	1233	570
MP2231.fasta	328583	442106047	455488	1622	588
MP2233.fasta	341695	423300930	1082733	1382	577
MP2235.fasta	276504	337058072	272529	1394	578
MP2237.fasta	366839	453632259	315077	1402	578
MP2239.fasta	228512	290433883	253245	1479	583
MP2241.fasta	389139	466617158	249726	1325	574
MP2243.fasta	377456	443235090	211269	1268	575
MP2809.fasta	351141	422352456	991898	1285	566
MP2811.fasta	318604	486211511	810075	2141	607
MP2813.fasta	330929	385727554	840347	1267	569
MP2815.fasta	326179	324949926	157261	994	552
MP2817.fasta	386296	386252187	150636	988	551
MP2819.fasta	423810	438803419	231819	1057	560
MP2821.fasta	360134	401892948	166357	1172	569

Table A3. Number of scaffolds, number of bases, maximum contig length, N50 and N90 of multi-fasta files obtained from metagenomes after having filtered sequences greater than 1000 bp.

File	Scaffolds	Bases	Max	N50	N90
MP0311.fasta	102973	225403065	169994	2280	1130
MP0313.fasta	93830	279649555	449642	4199	1230
MP0315.fasta	94788	243868057	226773	3033	1183
MP0317.fasta	82331	201140057	806207	2791	1171
MP0319.fasta	93233	215584584	102976	2531	1155
MP0321.fasta	105425	229544756	189365	2263	1129
MP0323.fasta	136616	289398149	189298	2172	1127
MP0528.fasta	124690	305944176	255067	2738	1170
MP0530.fasta	105401	319965626	1592702	4092	1248
MP0532.fasta	110131	308894451	766578	3590	1216
MP0534.fasta	109123	313801334	583047	3765	1232
MP0536.fasta	117735	290072137	613280	2792	1165
MP0538.fasta	62631	192028708	1208588	4375	1219
MP0540.fasta	131621	307193240	364958	2549	1155
MP0778.fasta	150300	400125954	303241	3197	1198
MP0780.fasta	65004	263413091	962570	7432	1420
MP0782.fasta	54564	210345106	792149	7678	1344

MP0784.fasta	74781	251539784	1885427	5092	1285
MP0786.fasta	88293	239752265	881066	3326	1194
MP0788.fasta	90334	255222204	967705	3637	1206
MP0790.fasta	141084	445005490	607124	4388	1283
MP0878.fasta	100028	323660387	1082667	4790	1276
MP0880.fasta	102259	311112701	1056618	4136	1253
MP0882.fasta	103580	260476918	848140	2888	1187
MP0884.fasta	111506	283580409	558882	2905	1175
MP0886.fasta	105787	257918323	645563	2703	1161
MP0888.fasta	109756	250601820	157496	2470	1151
MP1154.fasta	111559	361527466	1175840	4649	1273
MP1162.fasta	109842	327631371	573606	4024	1239
MP1164.fasta	102879	269906688	1056421	3107	1190
MP1166.fasta	118166	285579701	572396	2704	1178
MP1174.fasta	118626	258881381	155175	2309	1141
MP1176.fasta	121037	280356514	445741	2523	1159
MP1178.fasta	130494	306836322	161394	2580	1156
MP1409.fasta	105278	331791902	897297	4575	1255
MP1411.fasta	104388	320492301	933877	4229	1242
MP1413.fasta	88701	221540694	827084	2798	1165
MP1415.fasta	92620	208534975	173342	2368	1139
MP1417.fasta	130289	297661159	191491	2468	1146
MP1419.fasta	128098	310629520	572431	2713	1159
MP1421.fasta	110708	291359829	185072	3202	1176
MP1517.fasta	115898	268677721	252678	2516	1145
MP1519.fasta	99957	309294820	740565	4225	1257
MP1521.fasta	97220	308452715	725129	4746	1246
MP1523.fasta	89461	215765680	572481	2665	1159
MP1525.fasta	115353	264233519	730354	2392	1137
MP1527.fasta	168694	380376819	507528	2374	1134
MP1529.fasta	112498	270074134	1053974	2675	1150
MP1672.fasta	108323	279608085	503252	3041	1177
MP1674.fasta	96903	353160301	1818184	6760	1306
MP1676.fasta	107004	344620272	723362	4977	1248
MP1678.fasta	85888	231242635	799118	3270	1189
MP1680.fasta	79139	225806221	1169690	3794	1191
MP1682.fasta	92301	218078027	897428	2567	1153
MP1684.fasta	131882	295494681	269979	2355	1138
MP1845.fasta	88965	331562008	826662	7083	1312
MP1847.fasta	106843	343480052	2458067	4747	1246
MP1849.fasta	112722	324350387	897432	3629	1217
MP1851.fasta	88634	262563589	1189902	3916	1209
MP1853.fasta	106000	266408748	807958	2841	1160
MP1855.fasta	108370	241760786	1231286	2352	1132
MP1857.fasta	118470	284087662	1786432	2683	1154
MP2231.fasta	100453	289154413	455488	3833	1204

MP2233.fasta	98473	260489789	1082733	3165	1178
MP2235.fasta	87114	210329945	272529	2697	1172
MP2237.fasta	109439	281438544	315077	3004	1179
MP2239.fasta	70914	184737841	253245	3085	1186
MP2241.fasta	114432	282821459	249726	2809	1165
MP2243.fasta	110475	263920588	211269	2623	1153
MP2809.fasta	89234	248701955	991898	3468	1172
MP2811.fasta	99917	338816888	810075	5858	1250
MP2813.fasta	94852	228394564	840347	2648	1157
MP2815.fasta	79579	161542898	157261	2044	1114
MP2817.fasta	90453	190896672	150636	2141	1116
MP2819.fasta	113043	231038341	231819	2066	1119
MP2821.fasta	101959	228643727	166357	2377	1139

Table A4. Number of gaps, average length and standard deviation. For all samples, the maximum length of gap was 60 bps and the minimum, 2 bps.

File	Number_gaps	Average	SD
MP1672.fasta	8303	21,17	19,08
MP2243.fasta	6310	23,29	20,01
MP1413.fasta	6805	19,58	18,34
MP0317.fasta	5509	23,62	20,17
MP1853.fasta	8511	15,41	15,26
MP1680.fasta	6260	21,95	19,52
MP2233.fasta	7235	21,58	19,32
MP1417.fasta	7998	17,04	16,65
MP2821.fasta	5998	23,15	19,98
MP1421.fasta	8489	23,22	19,92
MP0315.fasta	6864	20,39	18,73
MP1519.fasta	9570	16,87	16,48
MP1517.fasta	6932	18,99	17,96
MP1682.fasta	7288	18,23	17,48
MP1527.fasta	12597	12,89	12,34
MP1174.fasta	26524	12,38	10,37
MP0321.fasta	5087	19,77	18,47
MP2241.fasta	8466	18,55	17,78
MP1176.fasta	28295	13,24	11,41
MP2813.fasta	6635	17,98	17,32
MP1678.fasta	7624	21,2	19,11
MP1162.fasta	33002	13,45	11,95
MP1525.fasta	8729	16,09	15,79
MP0780.fasta	16060	14,8	13,72
MP0882.fasta	31897	13,66	11,75
MP0784.fasta	18647	16,02	14,85
MP1674.fasta	9385	20,23	18,69
MP0888.fasta	19853	12,03	10,24

MP1419.fasta	9187	15,95	15,75
MP2811.fasta	8974	18,61	17,77
MP2809.fasta	7246	18,83	17,82
MP0319.fasta	6596	19,16	18,11
MP2817.fasta	5287	17,05	16,73
MP0534.fasta	23816	14,52	13,66
MP1523.fasta	6908	16,89	16,61
MP0530.fasta	19187	14,18	13,36
MP2237.fasta	11287	19,3	18,16
MP0878.fasta	28729	14,95	13,73
MP0778.fasta	38419	13,65	12,23
MP0884.fasta	25269	11,79	9,5
MP2239.fasta	5290	21,41	19,26
MP1847.fasta	10491	14,19	13,96
MP2235.fasta	6420	21,73	19,38
MP1409.fasta	9733	19,38	18,23
MP2819.fasta	5892	15,68	15,58
MP1166.fasta	40483	13,27	11,23
MP0532.fasta	21889	14,53	13,68
MP0790.fasta	53021	15,15	13,69
MP0311.fasta	5311	21,48	19,3
MP1529.fasta	7545	15,49	15,3
MP0536.fasta	29565	13,3	11,66
MP0788.fasta	19085	14,32	13,07
MP1855.fasta	5504	13,88	13,77
MP0786.fasta	21795	15,64	14,26
MP0886.fasta	22599	13,09	12,01
MP1411.fasta	9051	19,35	18,23
MP1684.fasta	9296	15,68	15,45
MP1415.fasta	7055	17,55	17,04
MP0528.fasta	25849	14,2	13,43
MP1857.fasta	7596	16,28	16
MP1676.fasta	11458	17,34	16,8
MP1521.fasta	9099	17,86	17,19
MP1154.fasta	18027	15,64	15,05
MP1178.fasta	37739	12,88	10,85
MP2231.fasta	9557	23,71	20,12
MP1845.fasta	9666	12,76	12,07
MP0782.fasta	14672	17,91	16,54
MP1849.fasta	12254	15,14	14,85
MP1851.fasta	7202	17,94	17,27
MP0313.fasta	9146	22,04	19,47
MP0323.fasta	10364	13,34	12,98
MP2815.fasta	4810	15,77	15,68
MP1164.fasta	23920	13,46	11,99
MP0880.fasta	22627	14,47	13,35

MP0540.fasta	27329	12,72	10,89
MP0538.fasta	16234	14,59	13,18

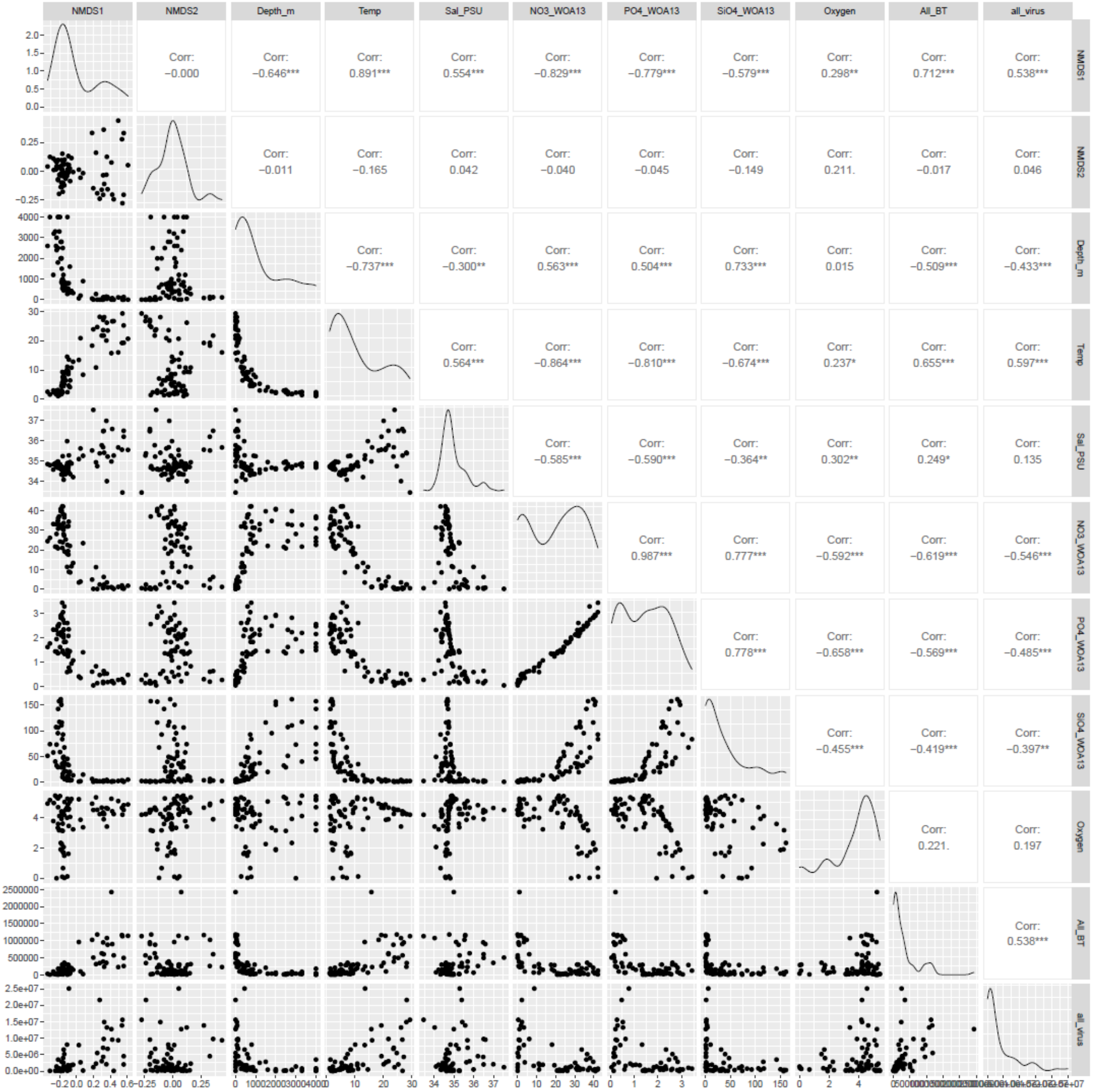

Figure A1. Correlogram of NMDS dimensions with metadata using Pearson's correlation. Level of significance: "·" p-value < 0.1, "*" p-value < 0.05, "**" p-value < 0.01, "***" p-value < 0.001.

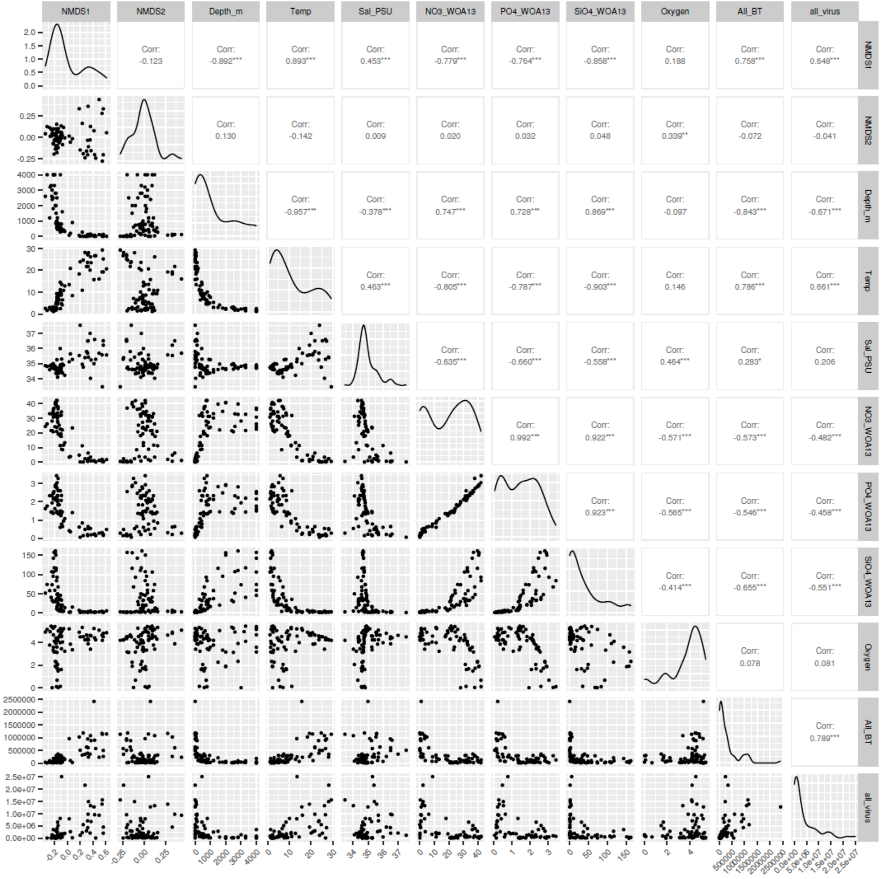

Figure A2. Correlogram of NMDS dimensions with metadata using Spearman's correlation. Level of significance: "·" p-value < 0.1, "*" p-value < 0.05, "**" p-value < 0.01, "***" p-value < 0.001.

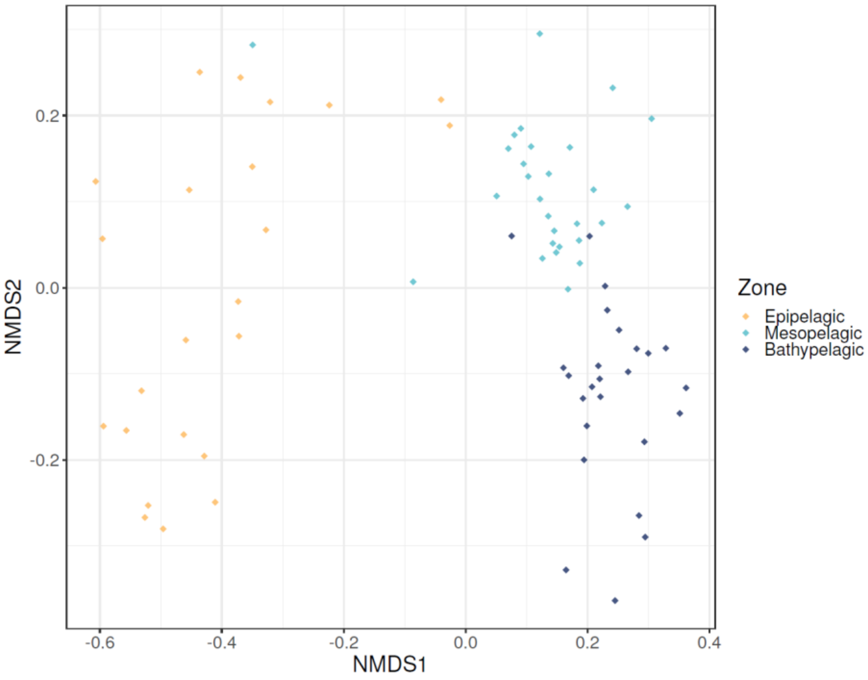

Figure A3. Non-metric multidimensional scaling plots. Points represent samples. Samples that are more similar to one another are ordinated closer together.

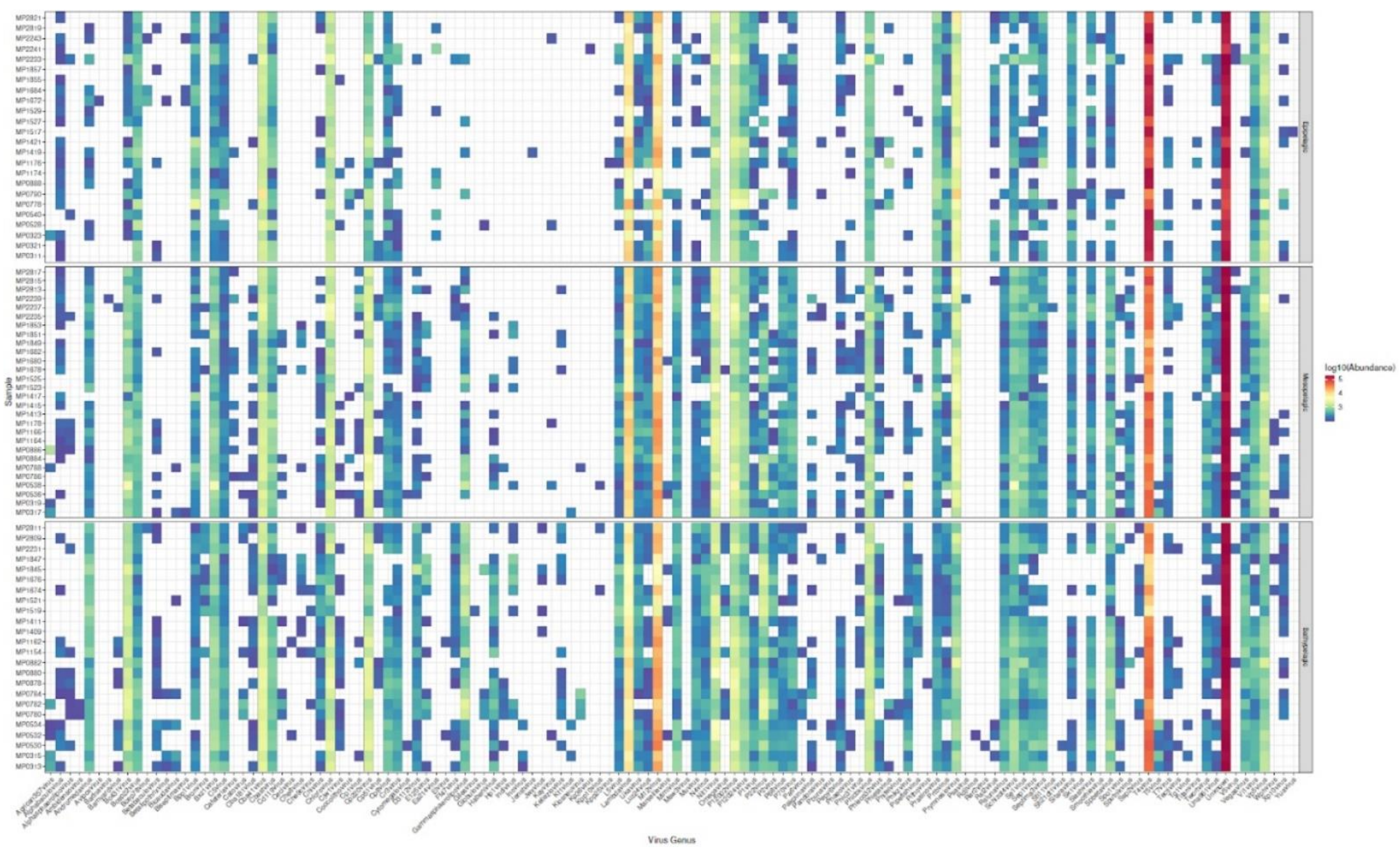

Figure A4. Relative abundance of viruses according to genus along all the water column. The y-axis represents samples and the x-axis, the reads per kilobase per million mapped reads (FPKM). T4virus and M12virus were the most abundant genus across all the water layers. Moreover, T4virus showed a higher abundance n Epipelagic, while M12virus was more abundant in Bathypelagic.

ABBREVIATIONS

ANOSIM	*Analysis of Similarities*
AMG	*Auxiliary Metabolic Gene*
bp	*Base Pair*
CTD	*Conductivity, Temperature, and Depth*
CDS	*Codifying DNA Sequence*
CNAG	*Centre Nacional d'Análisi Genomica*
DB	*Data Base*
DNA	*Desoxyribonucleic Acid*
DOM	*Dissolved Organic Matter*
FC	*Fold Change*
FPKM	*Fragments Per Kilobase of scaffold per Million reads mapped*
Gbp	*Giga base pair*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
KtW	*Kill-the-Winner*
LSU	*Large Subunit*
mTAG	*Taxonomic assignation of metagenomic reads*
POM	*Particulate Organic Matter*
Prodigal	*PROkaryotic DYnamic programming Gene-finding Algorithm*
PtW	*Piggyback-the-Winner*
RaFAH	*Random Forest Assignment of Hosts*
RDA	*Redundancy Analysis*
SAM	*Sequence Alignment Map format*
SSU	*Small Subunit*
TIEP	*Type I Error Probability*
VPF	*Viral Protein Families*

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS