Bernardo José Costa Bica

Cracking the code: exploring the virome of the marine sponge *Spongia officinalis*

[2022]

Mestrado em Biologia Molecular e Microbiana

[2022]

Universidade do Algarve

Faculdade de Ciências e Tecnologias

Cracking the code: exploring the virome of the marine sponge *Spongia officinalis*

Bernardo José Costa Bica

Dissertação orientada pelo Professor Doutor Rodrigo Costa e

coorientada pela Professora Doutora Filomena Fonseca

Mestrado em Biologia Molecular e Microbiana

[2022]

Cracking the code: exploring the virome of the marine sponge *Spongia officinalis*

Declaração de Autoria do Trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Bernardo José Costa Bica

Direitos de cópia ou Copyright
©Copyright: (Bernardo José Costa Bica).
A Universidade do Algarve tem o direito, perpétuo e sem limites geográficos, de arquivar e publicitar este trabalho através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, de o divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Acknowledgments

A word of appreciation to all my family, friends, and my supervisors.

A sincere thank you to Professor Rodrigo Costa for teaching me, more than knowledge, true wisdom and leadership.

Thank you Cymon Cox for installing and helping me run DRAM, a crucial step of this work.

Many thanks to Francisco Coroado for pushing me and providing me with the enthusiasm to unlock the potential of R. You are a carrier of a pure hearth and tireless in your friendship.

Dedication

I dedicate this work to my mother, brother and to you Margarida.

Thank you, mother, for holding my hand always, from my very first steps to where I'm standing. The greatest joy of doing things right is trying to make you proud.

I give my biggest hopes to you brother, my best friend in life.

Even through the future is scary in all its uncertainty, I'm not afraid if I hold you, and you hold me. In your love I find comfort and strength, armed with these we can move forward without fear. Because of you I believe in luck. I am forever grateful, Margarida.

And finally, to all professors, doctors and nurses of this country.

"Certain mystes aver that the real world has been constructed by the human mind, since our ways are governed by the artificial categories into which we place essentially undifferentiated things, things weaker than our words for them. [...] We believe we invent symbols. The truth is that they invent us; we are their creatures, shaped by their hard, defining edges."

Gene Wolfe, The Book of the New Sun#2

Resumo

Os vírus estabelecem um vasto conjunto de interações com o mundo vivo e químico no oceano, com grande impacto nos ciclos biogeoquímicos. Estas entidades biológicas podem estabelecer complexas interações com as comunidades microbianas, com estratégias para manipular genoma do hospedeiro no sentido de reconduzir os recursos energéticos para a formação de mais vírus, ou ser o próprio vírus a contribuir com um reportório genético que poderá auxiliar, por exemplo, na produção de energia ou obtenção de um nutriente.

As esponjas marinhas alojam complexas comunidades microbiológicas com um grande interesse ecológico, evolutivo e metabólico. Os vírus, apesar de demonstrarem grande potencial na mediação da resposta imunitária da esponja e na estruturação da comunidade bacteriana simbionte, são ainda pouco estudados neste contexto. Este trabalho explora as comunidades virais da esponja marinha *Spongia officinalis*, a partir de dados metagenómicos obtidos por sequenciação em Hiseq Illumina. Após o passo de verificação da qualidade das sequências, estas foram assembladas e feita a previsão dos *contigs* virais, que foram anotados taxonomicamente e funcionalmente com o recurso a ferramentas bioinformáticas. Os dados obtidos foram manipulados e visualizados através de software específico ou do RStudio. As operações efetuadas ao longo do trabalho foram adaptadas para um pacote do R e integradas numa aplicação nomeada de MetaViral, de forma a tornar o trabalho mais reprodutível e acessível.

As abundâncias relativas de diferentes níveis taxonómicos e os índices de diversidade ecológica calculados revelaram algumas diferenças entre os viromas presentes na *S. officinalis* e na água do mar. As amostras da esponja apresentaram uma diversidade maior para a generalidade dos índices e ainda abundâncias relativas comparáveis para a grande maioria dos grupos taxonómicos. Foram verificadas mais diferenças para vírus pertencentes à família *Siphoviridae*, *Myoviridae* e *Podoviridae*. Este contraste foi confirmado por análise estatística multivariada, onde se pode observar o agrupamento de amostras pertencentes ao mesmo bioma. Foram criadas redes neuronais e identificados os grupos virais únicos para cada bioma, assim como atribuição de um domínio.

O perfil funcional das comunidades revelou uma menor carga genética viral para as amostras da *S. officinalis*. Verificou-se a presença de mais previsões funcionais relacionadas com integrases e *ABC transporters* nas amostras de esponja, o que pode estar relacionado com uma preferência pelo ciclo viral lisogénico. A ausência de correspondências com proteínas relacionas com compostos de ferro levanta questões sobre a obtenção deste micronutriente por parte dos simbiontes bacterianos e de como esta carência funcional pode justificar vias alternativas de disponibilidade e obtenção do ferro onde os vírus associados ao microbioma simbionte da

esponja podem ter um papel facilitador, de remineralização do micronutriente, ou oportunista como estratégia de entrada na parede celular por incorporação de compostos de ferro na cauda viral.

Duas enzimas, *dTDP-4-dehydrorhamnose 3,5-epimerase* e a provável *UDP-N acetylglucosamine pyrophosphorylase,* com correspondências apenas encontradas nas amostras de esponja, sugerem a existência de uma via metabólica alternativa relacionada com a biossíntese da rhamnose. Ao nível das polimerases, duas foram exclusivamente encontradas em amostras de vírus associados à esponja, *sigma-70 factor region 4* and *RNA polymerase sigma-G factor*. Estas polimerases podem estar envolvidas em sistemas de reconhecimento do genoma bacteriano mais específicos, revelando uma especiação que não está presente nas partículas virais livres.

Estas diferenças realçam o potencial papel que os vírus podem ter enquanto membros do viroma da *S. officinalis*, podendo estar envolvidos em diversos processos metabólicos e de troca genética com as comunidades simbiontes bacterianas.

Palavras-chave: *Spongia ofiicinalis*, esponja, marinha, fago, vírus, viroma, simbiose, bactéria, metagenómica, bioinformática.

Abstract

Viruses establish a vast array of interactions with the living and chemical world in the ocean, with great impact on biogeochemical cycles. These biological entities can establish complex interactions with microbial communities, with strategies to manipulate the host genome in order to redirect energy resources towards the formation of more viruses, or the virus itself can contribute a genetic repertoire that may assist, for example, in energy production or nutrient acquisition.

Marine sponges host complex microbiological communities with great ecological, evolutionary, and metabolic interest. Despite showing great potential in mediating the sponge immune response and structuring the symbiotic bacterial community, viruses are still poorly studied in this context. This study explores the viral communities of the marine sponge *Spongia officinalis*, using metagenomic data obtained through Illumina Hiseq sequencing. After quality checking, the sequences were assembled, and viral contigs were predicted, taxonomically and functionally annotated with the aid of bioinformatics tools. The obtained data were manipulated and visualized using specific software or RStudio. The operations performed throughout the study were adapted to an R package and integrated into an application named MetaViral, in order to make the work more reproducible and accessible.

Relative abundances of different taxonomic levels and calculated ecological diversity indices revealed some differences between the viromes present in *S. officinalis* and seawater. Sponge samples presented higher diversity for most indices and comparable relative abundances for the vast majority of taxonomic groups. More differences were found for viruses belonging to the families Siphoviridae, Myoviridae, and Podoviridae. This contrast was confirmed by multivariate statistical analysis, where samples belonging to the same biome clustered together. Networks were created and unique viral groups were identified for each biome, as well as domain attribution.

The functional profile of the communities revealed a lower viral genetic load for the *S. officinalis* samples. More functional predictions related to integrases and ABC transporters were found in sponge samples, which may be related to a preference for the lysogenic viral life cycle. The absence of correspondences with proteins related to iron compounds raises questions about the acquisition of this micronutrient by the bacterial symbionts and how this functional deficiency can justify alternative pathways of availability and acquisition of iron where symbiotic viruses may have a facilitating role in ensuring their own replication.

Two enzymes, dTDP-4-dehydrorhamnose 3,5-epimerase and the probable UDP-N-acetylglucosamine pyrophosphorylase, with matches only found in sponge samples, suggest the existence of an alternative metabolic pathway related to rhamnose biosynthesis. At the polymerase level, two were exclusively found in virus samples associated with the sponge, sigma-70 factor region 4 and RNA polymerase sigma-G factor. These polymerases may be involved in more specific recognition systems of the bacterial genome, revealing a speciation that is not present in free viral particles. These differences highlight the potential role that viruses can play as members of the S. officinalis virome, potentially being involved in various metabolic processes and genetic exchange with symbiotic bacterial communities.

Keywords: *Spongia ofiicinalis*, sponge, phage, virus, virome, symbiosis, bacteria, metagenomics, bioinformatics.

TABLE OF CONTENTS

PAGE

INDEX OF FIGURES

PAGE

CHAPTER 1. INTRODUCTION

1.1 How viruses shape marine life

When the word virus comes to thought it is most likely associated with disease, and for good reason. Human infecting viruses have since long shaped our history, a history that teaches us the value of life, and to value health and science. All this brings a ”bad” reputation to viruses that greatly overshines their potential in the development of novel therapeutical procedures (Mietzsch & Agbandje-McKenna, 2017), including the great promise they bear as antibiotic replacements (phage therapy) (Loc-Carrillo & Abedon, 2011; Skurnik & Strauch, 2006).

The ocean is an excellent example of how impactful is the role in which viruses is partake. Infecting all types of organisms, regardless of size or taxonomy group, viruses greatly contribute to the regulation of population growth in the marine environment. The ecology of marine viruses is an emerging field of study that witnessed an exponential interest since the last two decades. Although incredible small in size, with an average of ~50 nm (Bar-On & Milo, 2019) and in many cases overlapping what is considered dissolved organic matter (DOM), viruses are the absolute dominant marine organisms in terms of density while only representing ~1% of the total biomass, while protists and bacteria encompass approximately two-thirds in biomass (Bar-On & Milo, 2019). It has been estimated that 20-40% of all prokaryotes in surface waters are daily removed through viral lysis (Suttle, 2007), with proportions comparable to the mortality resulted by grazing (R. Zhang et al., 2020). Viral infection of microbial cells often will culminate in death by lysis, this allows the release of DOM and POM (particulate organic matter) that can be easily assimilated by microorganisms. The carbon released by viral shunt sinks at a much slower rate than fecal pellets from zooplankton grazing, and the nutrients serve as fuel for its re-assimilation by the phytoplankton, preventing this way the sequestration of carbon in the deep ocean (Suttle, 2007) and consequently, the increase in the concentration of carbon dioxide in the atmosphere (Jiao & Zheng, 2011). The viral shunt also contributes with the release of dissolved organic nitrogen and dissolved organic phosphorus (DOP), as well as new viruses that can retain up to 5% of the total DOP pool in surface waters (Jover et al., 2014).

Not all phages commit to cell lysis upon infection, the viral life cycle can in some cases include a step where integration of the viral genome into the host genome occurs or is kept as an extrachromosomal element. In this “stealthy” state phages are called prophages and will propagate to cell daughters upon division. When certain conditions are met, the prophages enter the lytic state, replicating and resulting in cell death, this is the lysogenic cycle that temperate viruses can undergo.

There is also the case where viral infection does not result in the disruption of the host cells, this is true for both lytic and temperate phages, named chronic and chronic temperate phages, respectively (Hobbs&Abedon, 2016).

Temperate viruses, while integrated into the host genome, avoid facing some of the challenges that free phage particles undergo, such as UV inactivation, proteolytic digestion and grazing (Paul, 2008). This implies that it is in the best interest of temperate viruses that the infected cell thrives. To this end there are some advantages to the host for having viral genes integrated to its chromosome such as granting immunity to homologous viruses and potentiate the genetic variability. The genetic material of the virus may contain auxiliary metabolic genes (AMGs) that expand the metabolic repertoire of the infected cell providing a source for rapid adaptation and genetic variability (Perez Sepulveda et al., 2016). The AMGs may have originated from an infected cell or acquired, via lateral gene transfer, to a globally connected viral gene pool (Bryan et al., 2008).

Studies on cyanophages that infect mainly *Prochlorococcus* and *Synechococcus*, hint that the most prevalent AMGs associated with photosynthesis are selected via vertical inheritance (Crummett et al., 2016). These can induce the host cell to focus on deoxynucleoside triphosphate (dNTP) biosynthesis by inhibiting the Calvin cycle and channeling energy molecules to the Pentose Phosphate Pathway, fueling the dNTP biosynthesis through NADH and ribose 5-phosphate, ultimately providing genetic material to the offspring phages. (Thompson et al., 2011). Whether through the expression of Calvin cycle inhibitor genes or conversion of glucose (glucose-6-phosphate) to glycogen (Hurwitz & U'Ren, 2016), phages can this way simulate a starvation state for the host and redirect the metabolic flux to dNTP biosynthesis, favoring phage replication.

Regardless of the efforts in accessing and elucidating the global dynamics in which marine viruses partake, there are several factors that add to the complexity of this task. Abiotic factors such as seasonal variation and degree of stratification of the water column greatly influence the viral communities' densities (Finke, Hunt, Winter, Carmack, & Suttle, 2017), along with, temperature, salinity, organic and inorganic particles, UV radiation and nutrient stoichiometry, among other parameters (Mojica & Brussaard, 2014). Biotic factors on the other hand are much more difficult to predict with the host's possible phenotypes adding to the equation. The lack of knowledge regarding virus-host dynamics in the ocean and the inability to predict responses to shifts in the populations and the environment are just a few of the setbacks of this field of study, especially from an ecological standpoint.

Even the way in which the abundance of bacteria influences phage numbers is not as simple and linear as once was thought (Wigington et al., 2016), as contact rates and interactions between host and phage do not seem to follow a constant pattern and are unique to each dynamic microniche (Breitbart, Bonnain, Malki, & Sawaya, 2018). The proposed models that apply to

marine microbial systems dynamics are (i) the Kill-the-winner model, where the competition specialist (not necessarily the most abundant population) is targeted by phages, enabling this way more diversity among bacterial communities (Winter et al., 2010) and constantly resetting the "race" for the nutrient uptake (Marantos et al., 2022); and (ii) The Red Queen coevolutionary hypothesis, that applies to viruses at a intrapopulation and species level, where fluctuations in the virus genotype occur as a response to the host's defense mechanisms, such as the CRISPR and restriction enzymes (Ignacio-Espinoza et al., 2020). A conjunction of both models probably occurs, where dominance of a certain viral species is a consequence of active hosts that will suffer a rapid decline due to viral infection. The gap that is left is occupied by bacteria that developed resistance to infection and so are able to prosper, until the phages "catch up", and new viral species or new variants of the previously dominant viral species, stop again the dominant bacterial progression (Breitbart et al., 2018).

1.2 Accessing marine viruses

Accessing marine viruses is diving into, probably, the most diverse and unknow group, taxonomically, morphologically, and functionally. Culturing-dependent techniques englobe several challenges, from finding a host that can be infected by the phage and enabling its replication, to isolation of targeted viruses (Perez Sepulveda et al., 2016). In order to bypass the lack of a universally conserved viral gene (like the 16S rRNA gene in prokaryotes) researchers have used a combination of pulsed-field gel electrophoresis (PFGE) to separate viruses based on the genome size (Staley & Konopka, 1985) and randomly amplified polymorphic DNA (RAPD) PCR (Winget & Wommack, 2008).

The transmission electron microscopy also faces several challenges in the isolation and proper identification of viruses (Wilson et al., 2005). Still, this tool can be extremely useful by providing proof of the presence of viruses in histological parts in the study of host-virus interactions (Pascelli et al., 2018).

Studies that focus on metagenomic data are tied to the availability of viral sequences in public databases and analytical resources, which often results in a large amount of sequences with unknown taxonomic prediction (Hurwitz & Sullivan, 2013).

1.3 Viruses in the context of symbiosis

Marine viruses have been studied as disease agents in several different hosts, from fish (Crane & Hyatt, 2011) to crustaceans (Bateman & Stentiford, 2017) and even non-bacterial phytoplankton members (Koonin, Wolf, Nagasaki, & Dolja, 2008). In corals, viruses have also been targeted as a cause of pathogenesis, with bleached corals presenting a higher proportion of eukaryotic viruses (Messyasz et al., 2020). The prevailing argument is that in a situation of stress, due to eutrophication or shifts in temperature, triggers the viruses to destroy the coral tissues or

infect important symbionts, such as *Symbiodinium* spp, a group of dinoflagellate endosymbionts, predated mainly by phages of the *Mimiviridae* and *Phycodnaviridae* families (Wood-charlson, Weynberg, Suttle, Roux, & Oppen, 2015). This does not seem to be the only way viruses affect the corals, they may also perturb the flux of nutrients and therefore affect surrounding corals and spreading disease (Thurber, Payet, Thurber, & Correa, 2017). Viral families that showed increased abundances in bleached corals were the *Ascoviridae*, *Iridoviridae*, *Mimiviridae*, *Phycodnaviridae,* and *Pandoraviridae* (Messyasz et al., 2020). In contrast, the viral families that are widespread and abundant in healthy corals belong to the *Mimiviridae*, *Myoviridae*, *Retroviridae*, and *Siphoviridae* viral families (Cárdenas et al., 2020). There is also the case where no difference in the viral community was found between healthy and diseased corals, but even then, there is a prevalence of virulence and defense genes in the unhealthy animals (Soffer, 2016). Notwithstanding this, the bacterial viruses found in corals are supposed to contribute to the host's health, being so part of the coral's consortium (Peixoto, Rosado, Leite, Rosado, & Bourne, 2017). The lysogenic infection is theorized to prevent the lytic cycle, which can serve as another way to protect the important bacteria in corals (Sweet & Bythell, 2017). Bacteria can also incorporate key temperate viruses that pose a threat to other competitive bacteria that may arise and use this as an advantage against them. The prevalence of lysogeny can be verified by the high abundance of latency-associated genes and presence of prophage sequences (Weynberg et al., 2017).

The current knowledge of viruses in corals represents an important bridge to studying viral species that can infect the bacterial symbionts of marine sponges, as the two are very dependent on microorganisms to survive and sponges take a critical role in the coral reef structure and dynamics (Carballo, Cruz-Barraza, Vega, Nava, & Chávez-Fuentes, 2019).

Sponges (phylum Porifera) are invertebrates commonly known for their lack of complexity, early considered as the "most simple multicellular animals" (Koziol et al., 1998). While other creatures may have developed several visible adaptations that clearly improve the chance of survival, these sessile filter feeders were thought as stationary in time, leaving the scientific community wondering how they survived throughout time and what alternative strategies ensure their adaptation to a changing environment. For all these reasons, sponges became a tool to understand a big part of evolution. To answer some of the questions raised, one must take a close look into the inside of sponges.

Sponge's architecture is formed by distinct cell layers, each with differentiated cells with particular functions and morphologies. The pinacoderm is the outer layer formed by pinacocytes, epithelial cells that permeate the sponge. Underneath the pinacoderm there are chambers (choanoderm) made by flagellated cells, called choanocytes, which are responsible for filtering water and pumping it out. The particles and microorganisms filtered alongside the water are then brought to the mesohyl, an extracellular matrix that houses the archaeocytes where the digestion

takes place. These totipotent sponge cells responsible for the phagocytosis, engulf and digest particles in suspension captured by filtering (Taylor et al., 2007). Finally, the filtered water is released through the osculum to the open ocean. By filtering the water into their pores (ostia), many microorganisms are brought along.

Some of these may leave the sponge immediately through the osculum; some may even serve as carbon supply as part of the sponge diet (Maldonado et al., 2016), but others remain there, allowed to grow, and colonize the sponge in a process known as symbiosis.

Previous studies that focused on the structure of non-viral microbial communities in marine sponges found that such communities are often sponge species-specific and a result of convergent forces imposed by the poriferan host (Thomas et al., 2016). These sponge symbionts have been suggested to be mediators of several functional roles attributed to marine sponges (Slaby et al., 2017). In order for microbial symbionts to maintain the status of a member of the sponge holobiont (syn "metaorganisms"), they need to rely on different strategies to avoid the sponge immune response, and viruses may help them achieve this (Jahn et al., 2019).

1.4 A brief history of sponge virome studies

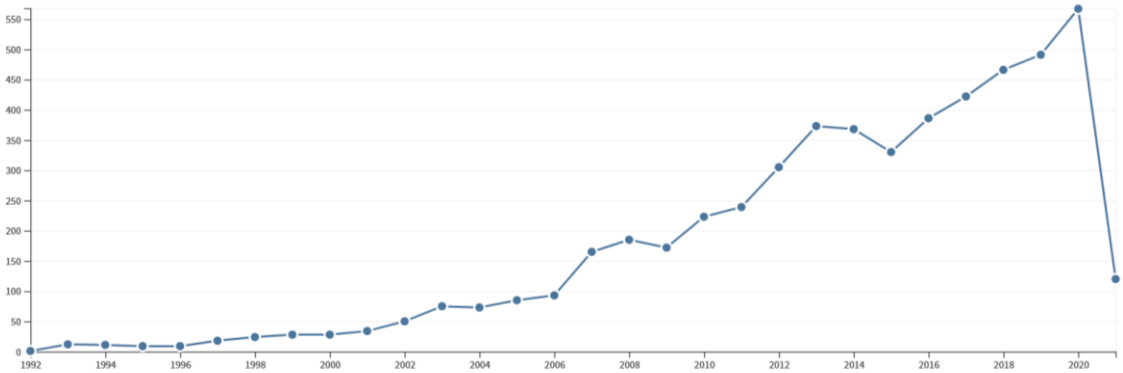


Figure 1 – Number of citations per year that refer to viruses and marine sponges. Search performed on the Web of Science, principal collection, using the search terms: (marine sponge) AND (phage OR virus*). Total number of 178 publications returned, and 5362 citations. Accessed on the 3 rd of march, 2021.

The study area of marine sponge viruses dates to approximately 25 years ago, with growing interest in the recent past (Figure 1). The few studies that focus on viruses discovered inside marine sponges concluded that the key differentiating aspects are mostly related with the viral life cycle proportions (Jahn et al., 2021), and especially the AMGs expressed (Nguyen et al., 2021; Pascelli et al., 2020). The viral communities examined exhibited inconsistency in the variability across samples, with cases of high intra-species similarity (Laffy et al., 2018), and high variability across samples (Nguyen et al., 2021). Although the viral composition varies greatly across sponge species, the commonly found viral families are the *Siphoviridae*, *Myoviridae*,

Podoviridae, *Phycodnaviridae*, *Poxviridae* and *Mimiviridae* families, with the first three families presenting higher abundances in several sponge species (Jamal et al., 2021).

In marine sponges that depend on the presence of symbiotic photosynthetic bacteria, herbicide-resistance genes may be provided to photobionts by their phages. These AMGs were found to be enriched in the *Synechococcus* phage belonging to the family *Myoviridae* that infects the *Synechococcus spongiarum*, a cyanobacterial symbiont of the sponge *Xestospongia testudinaria*, these can be transferred to cyanobacteria symbionts, and thus, improve its survivability (Laffy et al., 2018). Adding to this, genes encoding for Photosystem II proteins were found in cyanophages that infect cyanobacterial symbionts, and are thought to increase the bacterial host energy production upon infection (Nguyen et al., 2021; Pascelli et al., 2020). Another example of the hidden dynamics that phages have as facilitators of the bacteria-eukaryote coexistence is the discovery of phage-encoded ankyrin-domain-containing protein. Integration of these phages by symbiont bacteria allow the secretion of ankyrin proteins, that signal the sponge, resulting in reduced phagocytosis rates of bacteria expressing the ankyrin protein (Jahn et al., 2019). Other viral genes related with host interactions, whether microbial or eukaryotic, are associated with antibiotic biosynthesis (Nguyen et al., 2021; Pascelli et al., 2020), aegerolysin and toxin/antitoxin systems (Nguyen et al., 2021).

1.5 Objectives

Owing to the urgent need of improving our current knowledge of the structure and composition of prokaryotic communities in marine sponges, the objective of this study was to characterize the viral communities present in the marine sponge *Spongia officinalis* both taxonomically and functionally. This was accomplished by assessing the abundance, diversity, life cycle and functional profile of potential sponge-associated viruses using metagenomic and bioinformatic tools, comparing and elucidating key differences between the sponge-associated virome with free-living viruses in the ocean and investigating the functional profile of viral communities present in *S. officinalis* samples, including the presence of specific functional genes or pathways. Finally, to contribute to a better understanding of the ecological, evolutionary, and metabolic importance of marine viruses in marine sponge ecosystems and the connection with the sponge symbiotic communities.

CHAPTER 2. MATERIALS AND METHODS

Raw metagenome sequences obtained from *Spongia officinalis* and its environmental surroundings (seawater and sediments) established the starting point for this work, previous steps from sampling to sequencing were led by Karimi *et al.* (Karimi et al., 2017) and are summarized below.

2.1 Sampling and Sequencing

The collection of *S. officinalis*, seawater and sediment was executed in May 2014 in the coast of Pedra da Greta (36◦ 580 47.2N;7◦ 590 20.8W), Algarve, southern Portugal, at a depth of 20 m (+1 m for seawater collection and -1 m for sediment samples). Following the abiotic measurements (pH 8.13, temperature 18◦C, and salinity 36.40h) the individual samples collected via SCUBA diving were gathered, *in situ*, separately in ziploc® bags with seawater and then transferred to cooling boxes, as described in a prior similar work (Hardoim et al., 2012). *In vitro*, sponge samples were submitted to an identification procedure that included macro- and microscopic observations of the sponge morphology, followed by phylogenetic inference of the subunit I of the mitochondrial cytochrome oxidase (CO1) gene.

The sequencing method used the was HiSeq Illumina 2500 on the total metagenomics community. The approach delivered c. 15 million 100 bppaired end reads from each sample (4 × officinalis, 3 × seawater, 3 × sediments). Further information regarding the metagenomic DNA extraction and samples preparation is available in (Karimi et al., 2017).

2.2 Quality control and assembling

Raw reads were submitted to a quality control step before assembly, where a conjunction of Trimmomatic (Bolger et al., 2014) (v. 0.36, in the kbase enviroment) and FastQC was used to visualize and remove low quality ends caused by sequencing. Assembling of the reads was performed by metaSPAdes (Nurk et al., 2017) (v. 3.13.0 in k-base) with a minimum contig length defined to 1000 bp (base pair). In this step the sediment samples were discarded due to low number of assembled contigs.

2.3 Identification of viral genomes

Assembled reads were piped into VirSorter (v.1.0.6) (Roux et al., 2015) to identify viral sequences, selecting the Viromes database (all bacterial and archaeal virus genomes in Refseq, as of January 2014, plus non-redundant predicted genes from different viral metagenomes, including seawater). First, VirSorter detects circular sequences and predicts protein sequences. These sequences are compared against PFAM and Viromes using hmmsearch (Eddy, 2011) with a minimum score of 40 and maximum E-value of 10−05, and blastp (Altschul et al., 1997) with a minimum score of 50 and maximum E-value of 10−03.

Then, Virsorter attributes to each sequence a category from the more to less confident predictions, ranging from 1 to 3. Category 1 includes at least one hallmark viral gene and an enrichment in genes with hits from Viromes database; category 2 sequences meet one of the two conditions referred above, coupled with at least one other calculated metric (depletion in PFAM affiliated genes, enrichment in short or uncharacterized genes and depletion in strand switching, i.e. change of coding strand between two consecutive genes); category 3 sequences have at least two of the metrics mentioned and no hallmark viral genes nor enrichment in genes with hits in Viromes. Categories 1 and 2 represent the most probable viral sequences and were the ones selected for downstream analysis. Additionally, VirSorter identifies prophage sequences and categorizes them in a similar matter, attributed in descending quality. Category 4 was not attributed to any of the contigs, and category 5 was selected for further analysis, discarding category 6.

2.4 Taxonomic annotation

Taxonomic prediction was performed with VCONTACT2 using default settings. Viral sequences were grouped into viral clusters (VCs). Cytoscape (v.3.8.2) was used to visualize the gene-sharing network. First, after merging all categories 1 and 2 from each sample type (*S. officinalis* and water) and following the initial import, a network including only sponge samples and another with water samples was created, duplicated edges were removed, and the most prevalent viral families annotated with a colour scheme. Then a network including both water and sponge VCs was used to represent bacterial viruses annotated in each "biome" (that is, the biotope, or sample type, from which a given viral sequence comes from). By removing the duplicated edges and selecting according to biome, unique sponge and water cluster members (unique node and edge pair attributes) were identified and isolated. To aid in this process, a circular view of the networks was often used. This allowed for a clear division between major clusters and reduction of the processing power used. Data manipulation and statistical analysis of the VCs was performed using RStudio (v.1.2.5042).

For each category (1, 2 and 5) of each sample (officinalis 1-4 and water 1-3) and each VC, the taxonomy attributed was based on the most prominent viral family found in that VC and added the frequency (total of members in the VC), building an abundance matrix with absolute and relative abundances. This procedure was replicated to all different taxonomic ranks, this includes order, family, genus and species. This method of summarizing the VCs was compared with a standard selection and count of the taxonomic predictions (Figure 2).

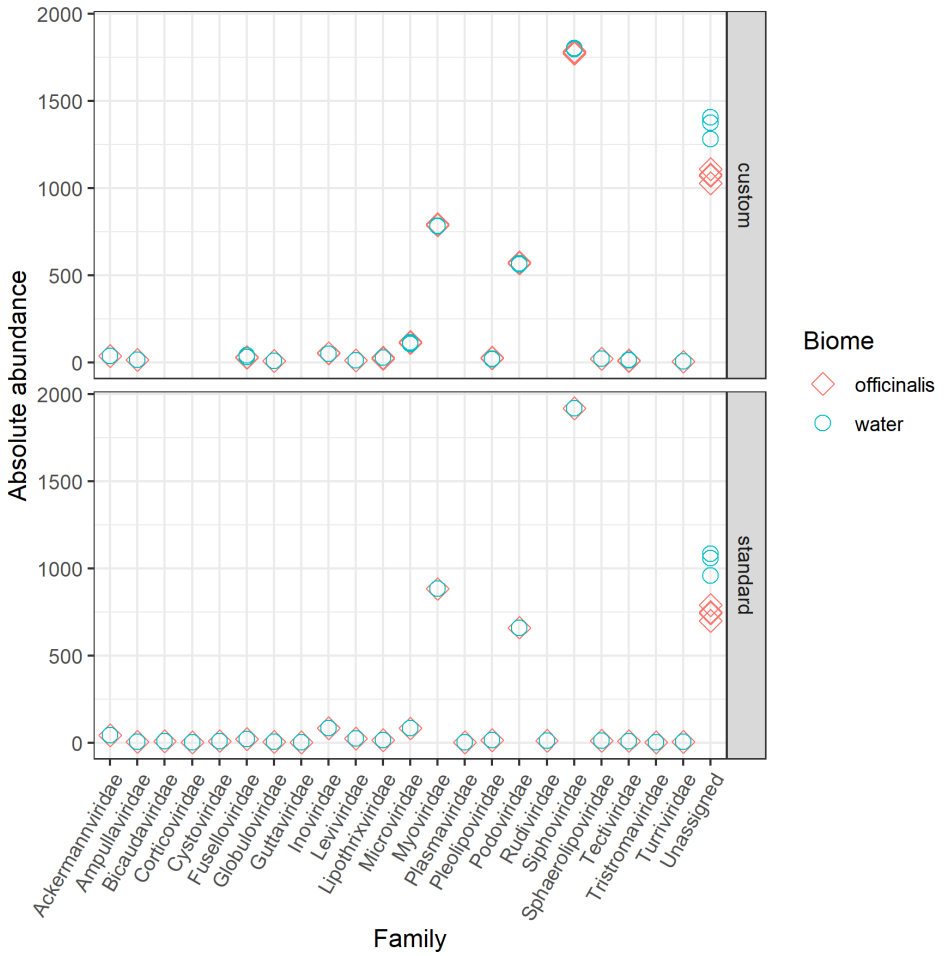


Figure 2 – Representation of the differences between using a custom approach for calculating the most abundant virus in a viral cluster, and simply using all individuals in a viral cluster.

Calculation of several diversity and richness indices was made by using the vegan package in RStudio. The aim of the use of multiple indices is to provide a complete summary of the structure of the viral taxonomy, where sensibility to lesser represented viral families and varies. This way it was possible to compare seawater and sponge samples and answer if differences in diversity occur at the most common or less represented viral families. Dependence of sample size is also a metric that varies between calculated indices and with great relevance for our case study. Based on (Kvålseth, 2015; Okpiliya, 2012) indices calculated were the following,:

i. Margalef's diversity index

$$D_{Mg} = \frac{S - 1}{\ln(N)} \quad (1)$$

Where S corresponds to the number of species (richness) and N equals to the total number of individuals.

ii. Menhinick diversity index

$$D_{Mn} = \frac{S}{\sqrt{N}} \quad (2)$$

iii. Shannon diversity index

$$H' = -\sum_{i=1}^{S} (p_i \times \ln p_i) \quad (3)$$

Where p_i is the probability to find $n_i = N_{pi}$ individuals in the ith species (i assumes values from one to S). The number of individuals in species i corresponds to n_i and is called the abundance of species.

iv. Simpson's index

$$\lambda = \sum_i \frac{n_i(n_i-1)}{n(n-1)} \quad (4)$$

The number of individuals in species i corresponds to n_i.

v. Pielou index

$$J = \frac{H'}{\ln (S)} \quad (5)$$

vi. Hills eveness for Shannon and Weiner

$$E_{a:b} = \frac{e^{H'}}{S} \quad (6)$$

$E_{a:b}$ corresponds to the ratio of diversity numbers between a and b orders.

vii. Hills eveness for Simpson's Index

$$E_{a:b} = \frac{1/\lambda}{S} \quad (7)$$

viii. Hills diversity numbers for Shannon and Weiner

$$E_{a:b} = e^{H'} \quad (8)$$

ix. Hills diversity numbers for Simpsons dominance index

$$E_{a:b} = \frac{1}{1-\lambda} \quad (9)$$

The statistical analysis using the absolute matrix calculated for viral families to investigate differences in biomes was conducted by multivariate analysis (PCoA/MDS). Additionally, a PCA model was built, using the factoextra package, first with a subset of sponge and seawater and predicting the placement of the samples "officinalis 4" and "water 3" (excluded from the model built). Then a model including all samples was tested with data from the Tara Oceans Expedition Station 36 surface, in the Indian Ocean (Northwest Arabian Sea, ~5 m depth) that was subjected to an identical bioinformatic pipeline by Benjamin Bolduc on k-base (available on the narrative (kbase.us/narrative/75811).

The species taxonomic rank was crossed with VirusHostdb to differentiate archaeal from bacterial hosts, and the viral life cycle obtained by matching the species (refseq id) with the 962 temperate and 1371 lytic phages predicted by presence/absence of integrase and parA

conserved domains. The prediction of the phage life cycle was performed by Mavrich and Hatfull (2017) (Mavrich & Hatfull, 2017) and available in (Song, 2020).

2.5 Functional Annotation

Functional annotation of the viral contigs was performed by DRAM-v (Shaffer et al., 2020), in the high-performance computer at the University of Algarve, under the supervision of Dr. Cymon Cox. A total of 26 threat were used and a custom script that runs over all samples was built, available in github.com/bbica/DRAM-v_automated. No custom flags predicted by the program were removed. Contigs bellow 2500bp (base pairs) were removed and Prodigal (Hyatt et al., 2010) (metagenome mode) was used to detect the open reading frames and respective amino acid sequences. These were then searched against the following databases: KEGG (Kanehisa et al., 2017); Uniref90 (Suzek et al., 2015); MEROPS (Rawlings et al., 2010); Pfam (El-Gebali et al., 2019); VOGDB (vogdb.org) and dbCAN (H. Zhang et al., 2018).

In RStudio, entries with auxiliary score of 5 (least certain) and duplicated predictions were removed and added counts for each category and respective biome or sample type. KEGG pathways were retrieved utilizing the given id with keggGet function from the KEGGREST package (v 1.38.0.) (Tenenbaum D & Maintainer B, 2022).

Aiming to facilitate visualization and interpretation, functional predictions were fitted into custom functional tags (viral ontology terms), with major and minor functional roles assigned to each annotation. This manually built ontology system was made with the intent to only suit the data examined here, as more expressive or functionally relevant annotations were given priority and the ecological context considered. Instead of approaching this task using ontology based on viral life-cycle terms (Hulo et al., 2017), a more empirical and comprehensive criterion was used that potentially increases the coverage of created ontology terms. Each minor function was built as a vector object containing one or more strings with the maximum length. These were then used to find matches in the standardized annotations, using make_call_names function from janitor package. All minor functions found were categorized into the following major functions, with ordering dictating the sequential functional assignment: genome, energy and metabolism, structure, infection, peptidase and other function. In the case of relevant annotations marked with a major function but not fitting into a minor function or are the only unique annotation present, the original annotation was kept.

Hypothetical proteins were identified by containing strings such as “putative” or “uncharacterized”, then the selected hits were removed if no prior function was attributed. For example, putative DNA polymerase still was considered a DNA polymerase. Heatmaps with the grouped annotations for each category were built using pheatmap package.

As a final step, the Geneious Prime software (v. 2022.2.2) was used to perform the Geneious Alignment (with "automatically determine direction and build guide tree via alignment" enabled) on selected sequences with functional interest and distributed among replicated samples. The alignment tree was built using Geneious Tree builder with Tamura-Nei distance model and Neighbour-joining built method, with no outgroups.

2.6 Aiming for a reproducible work

The main manipulations performed in Rstudio were converted into function objects that can include parameters and receive different input data. These functions were then compiled into a newly created package entitled MetaViral, resorting to the devtools package to build the necessary foundations. The MetaViral package was then adapted into an application (shinyapp) with the same name. The interactive interface makes replicating the analysis of viral data possible for a larger audience, regardless of coding experience and knowledge of Rstudio. The sample data provided is composed of a small subset of the samples from this study that is meant to aid in the understanding of the manipulations that are being performed and provide guidelines of how the data must be structured.

For even easier access and sharing potential, the MetaViral application was deployed to shinnyapps.io available in (bbica.shinyapps.io/MetaViralApp), turning the requirements to use the application as few as possible, without even the need to have Rstudio installed. More information regarding the utility and features of the created package and application can be found in the GitHub page (https://github.com/bbica/MetaViral).

CHAPTER 3. RESULTS AND DISCUSSION

Following the process of assembling the sequences and contig identification, a summary was constructed with standard sequence metrics (Table 1 and 2). These metrics provide insights into the quality and quantity of the sequencing data and are essential for downstream data analysis and interpretation.

Table 1 – Summary of the sequence's metrics and percentage of viral contigs found. These metrics include the total sequence length, the percentage that was dropped in the trimming step and characterization of the assembled reads.

Sample	Total length of reads (bp)	Dropped by Trimmomatic (%)	Total length of contigs (bp)	No. of contigs	N50	L50	%GC	N's per 100 kbp	% of viral contigs
officinalis 1	1,557,352,207	0.3	101,418,310	24645	8365	2369	58.95	181.17	0.527
officinalis 2	1,577,602,520	0.3	95,498,364	22813	8655	1777	60.56	433.29	0.359
officinalis 3	1,409,120,468	0.3	91,041,560	21421	8688	1976	59.95	233.49	0.420
officinalis 4	1,704,648,653	0.4	101,418,310	24408	8235	2094	60.75	407.37	0.147
water 1	1,464,778,541	0.6	51,843,358	11857	8823	1127	41.83	313.57	3.559
water 2	1,513,344,606	0.4	48,193,315	11746	8451	1023	41.88	243.17	3.380
water 3	1,524,364,585	0.5	41,728,822	9652	10495	756	41.69	347.26	3.067

Table 2 – Summary of the length and number of viral contigs for each confidence category predicted by VirSorter. Categories represented reflect certainty in the viral identity of contigs, with 1 being the most confident followed by category 2. Category 5 represents the predicted prophage samples.

Sample	Category	Total length of viral contigs (bp)	No. of viral contigs
officinalis 1	cat 1	64205	10
	cat 2	846838	117
	cat 5	45098	3
officinalis 2	cat 1	23936	8
	cat 2	371408	74
	cat 5	0	0
officinalis 3	cat 1	36216	11
	cat 2	506402	77
	cat 5	20261	2
officinalis 4	cat 1	12807	3
	cat 2	149136	33
	cat 5	0	0
water 1	cat 1	567859	65
	cat 2	3225366	356
	cat 5	7505	1
water 2	cat 1	504503	61
	cat 2	2808134	334
	cat 5	91971	2
water 3	cat 1	300988	47
	cat 2	2116860	248
	cat 5	15104	1

Viral sequences identified represent 0.36% to 0.53% in the case of sponge samples, and 3.07% to 3.56%, for water samples (Table 1). In contrast, metagenome sequences belonging to the sponge biome exhibited approximately two times more sequences than water and an overall higher sequence length (Table 1). The percentage of viral sequences retrieved from the sponge, although arguably small, is comparable with previous studies also investigating viruses in sponge species (Nguyen et al., 2021; Pascelli et al., 2020).

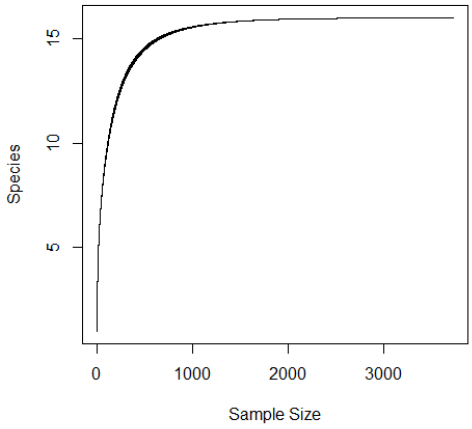


Figure 3 – Rarefaction curves. All samples of both biomes, seawater and sponge, are represented but overlapped. The x axis represents the number of sequences sampled while the y axis measures the number of viral families detected.

Rarefaction curves revealed an adequate sampling effort since a *plateau* can be seen for all samples (Figure 3). All samples exhibited the same number of viral families (richness), this is corroborated by the almost identical results for the species richness measures, namely the Menhinick's and Margalef's indices that are independent of relative abundances within the different viral families (Figure 4).

Relative abundances of viral sequences vary between samples with an overall higher abundance for the water biome and less representation for sponge samples, especially "officinalis 4" (Figure 4). Species diversity indices consider both richness and dominance/evenness of viral families. The commonly used Shannon-Wiener index revealed little difference in the diversity across all samples, which can be justified by the weight given to species proportional abundance, given by $\log_e$, see (3) (Strong, 2016). The Simpson index, which is less sensitive to low represented families, showed a general higher diversity for sponge samples. The measure of evenness was given by Pielou evenness and Hill's ratios (for both Shannon-Wiener index and Simpson's index). Evenness was consistently superior in sponge samples, indicating less variation in the relative abundance (or lower dominance) of different families in the sponge biome. The effective (true) diversity was calculated by true Shannon-Wiener index and true Simpson's index, also known as diversity numbers. Interestingly the true Simpson's index attributes for the first time a higher diversity to the water samples.

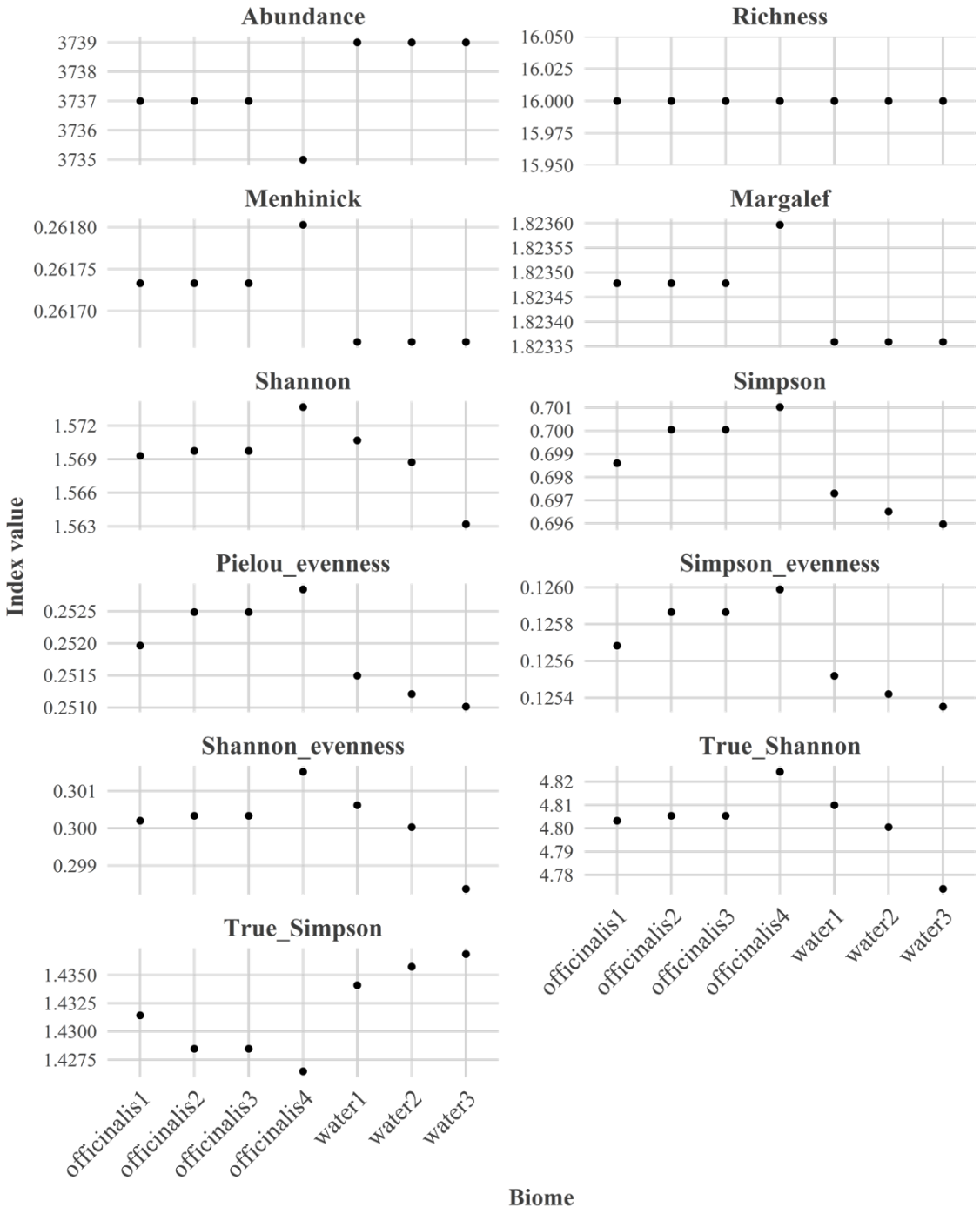


Figure 4 – Diversity, relative abundance and evenness metrics obtained for viral families detected in sponge and seawater samples.

Caudovirales order (dsDNA) prevailed in all samples (mean = 86.33%, SD = ± 4.5 for *S.officinalis* and mean=86.57%, SD = ± 2.52 for seawater) with *Siphoviridae* dominating over other viral families, followed by *Myoviridae* and *Podoviridae* (Figure 5). *Ligamenvirales* was the only other order found (mean = 0.50%, SD = ± 2 for *S.officinalis* and mean = 0.53%, SD = 0).

The *Megavirales* order, that infects mainly green algae and amoeba (Colson et al., 2013), has been consistently reported in sponges species from the Great Barrier Reef and the Red Sea (Laffy et al., 2018; Nguyen et al., 2021; Pascelli et al., 2020), however the absence of this order from samples of this study is related with the limitation of VirSorter in identifying eukaryotic viruses, since it lacks the databases for this group (Roux et al., 2015).

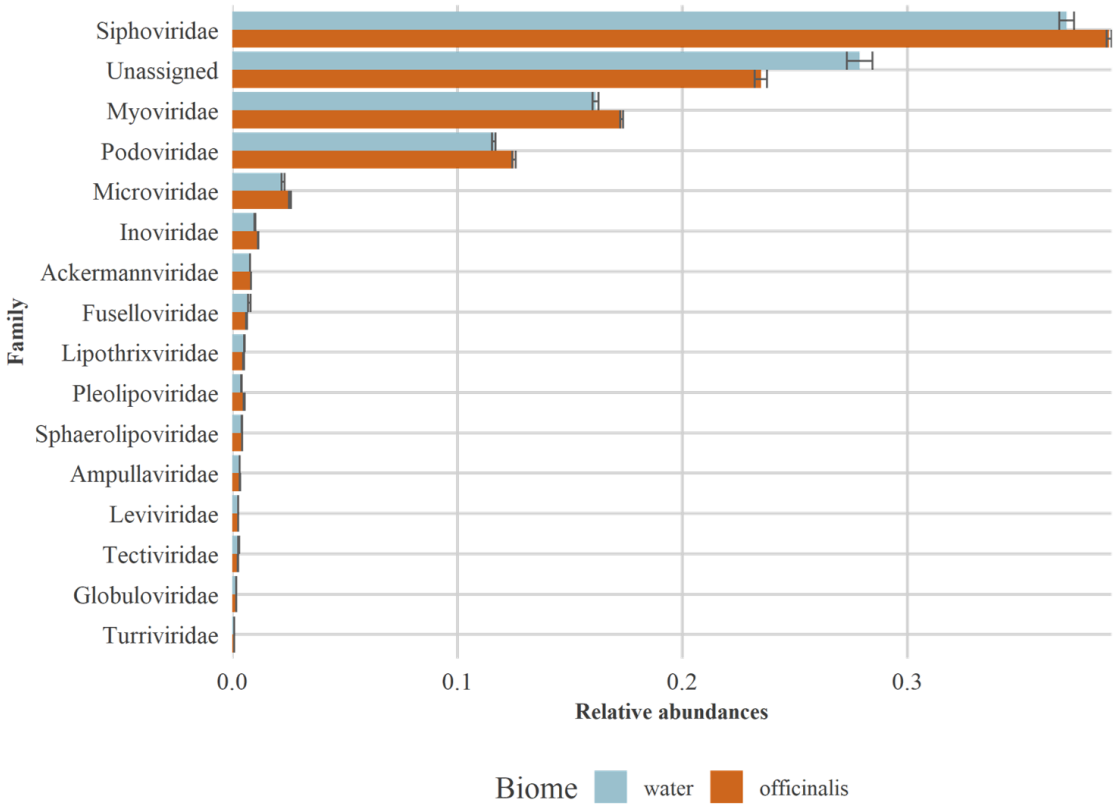


Figure 5– Mean relative abundances for all viral families present in each biome. Error bars correspond to the mean associated error.

Unassigned clusters were also predominant with 12.9% more representation in the water samples. The remaining families presented relative abundances inferior to 0.02 for both biomes. Across samples of the same biome there are very few notable variations, and between biomes, sponge samples presented higher abundances for the most dominant viral families.

The relative abundance matrix generated to the viral families was used as input for multivariate statistics by means of Principal Components Analysis. The PCA graph revealed a clear separation between biomes for the first principal component/dimension, that explained 84.8% (PCA) of the difference (Figure 6A). Water samples showed less clustering for the second principal component, while sponge samples clustered together with more divergence in the first

component (Figure 6A). *Siphoviridae* family contributed the most to explain the differences among biomes, followed by *Podoviridae*, *Fuselloviridae* and *Microviridae* (ssDNA) (Figure 6B).

Prediction of the positioning of officinalis 4 and water 3 in a PCA model based on the remaining samples, reinforced the difference between biomes, since the placement was consistent with clusters from which they were originally taken from (Figure 6C). When applied to the TARA sample (Figure 6D), the difference between sponge and seawater was lost due to extreme divergence between the newly added data and samples from this work. This result suggests a strong influence of the surrounding environment in the composition of the viromes of the sponge and seawater found in the sponge, which makes sense when considering the filter-feeding strategy.

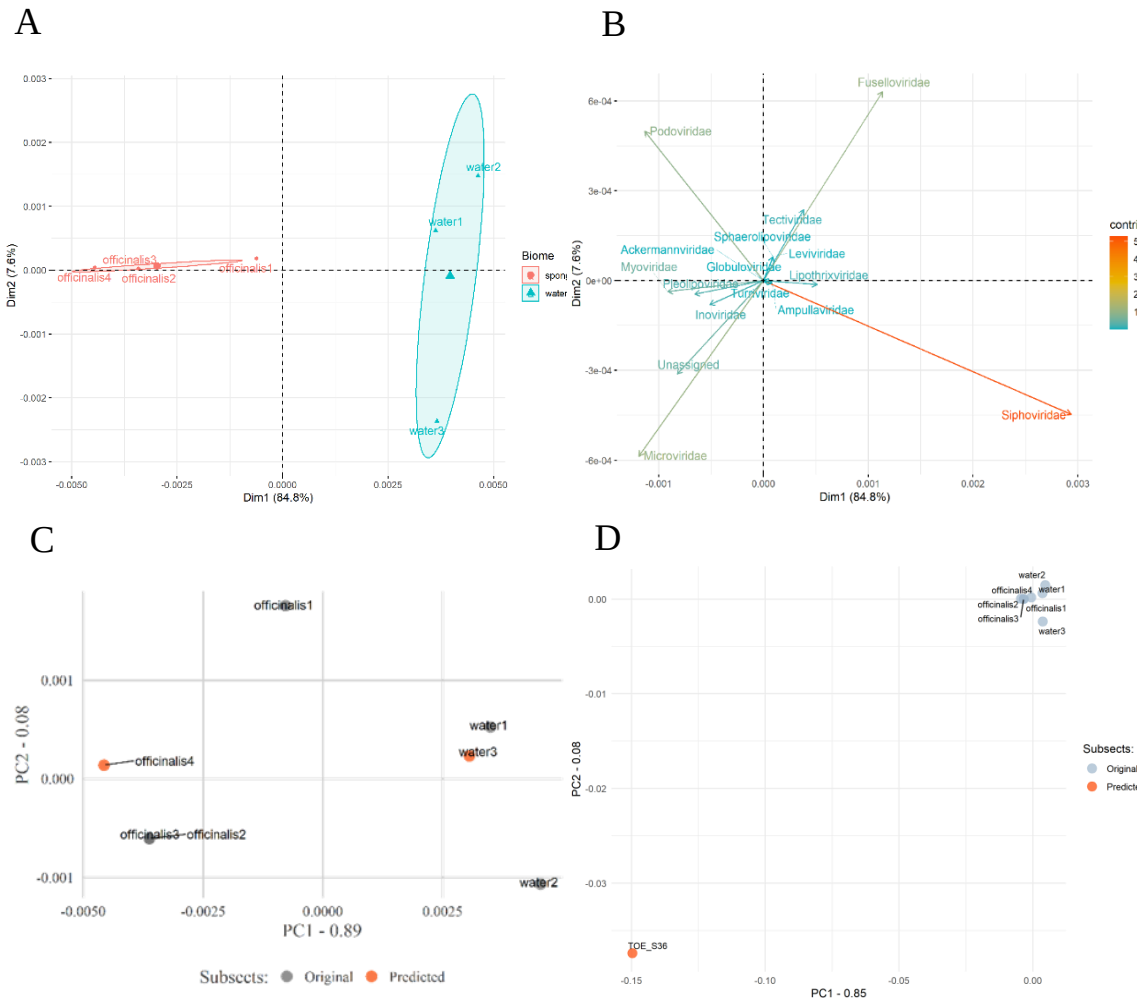


Figure 6 – Principal Components Analysis (PCA) of the viral families (A) and the contribution of each family (B). A PCA model was created and tested with a fraction of the samples from this work (C) and predicted the placement of samples water 3 and officinalis 4. This model (including all samples) was then applied to the Tara Oceans Expedition Station 36 surface (D), collected from the Indian Ocean.

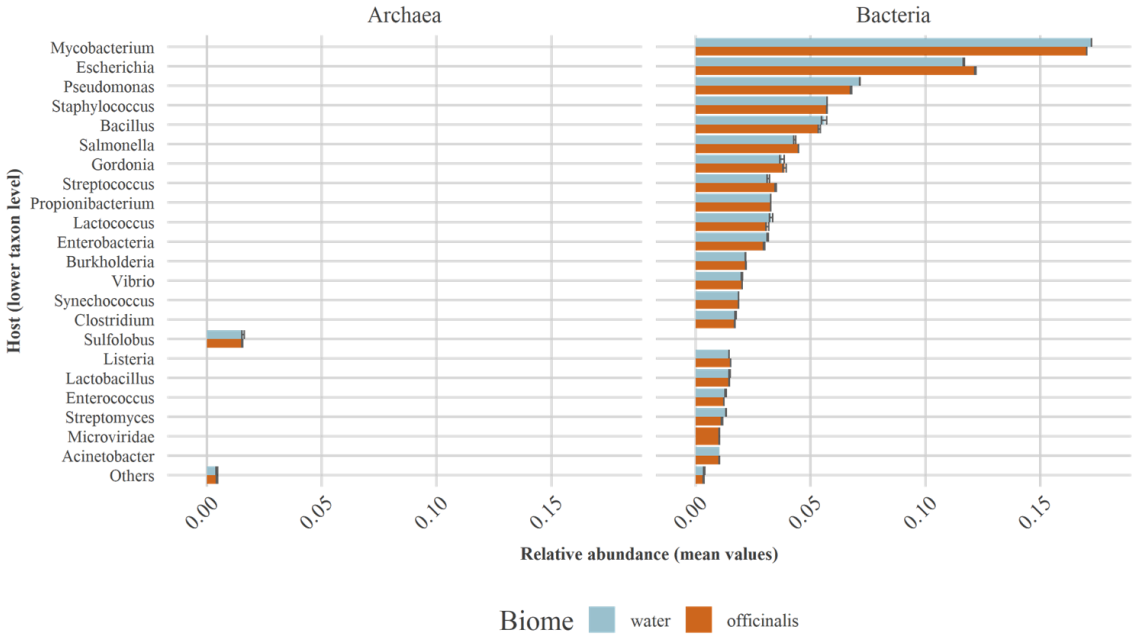


Figure 7– Mean relative abundances of host-specific viruses at lower host taxonomy levels (host) for each biome, in each prokaryotic domain. The "Others" category includes the viral species (lower taxon level) for relative abundances less than 0.01. Error bars describe the mean error.

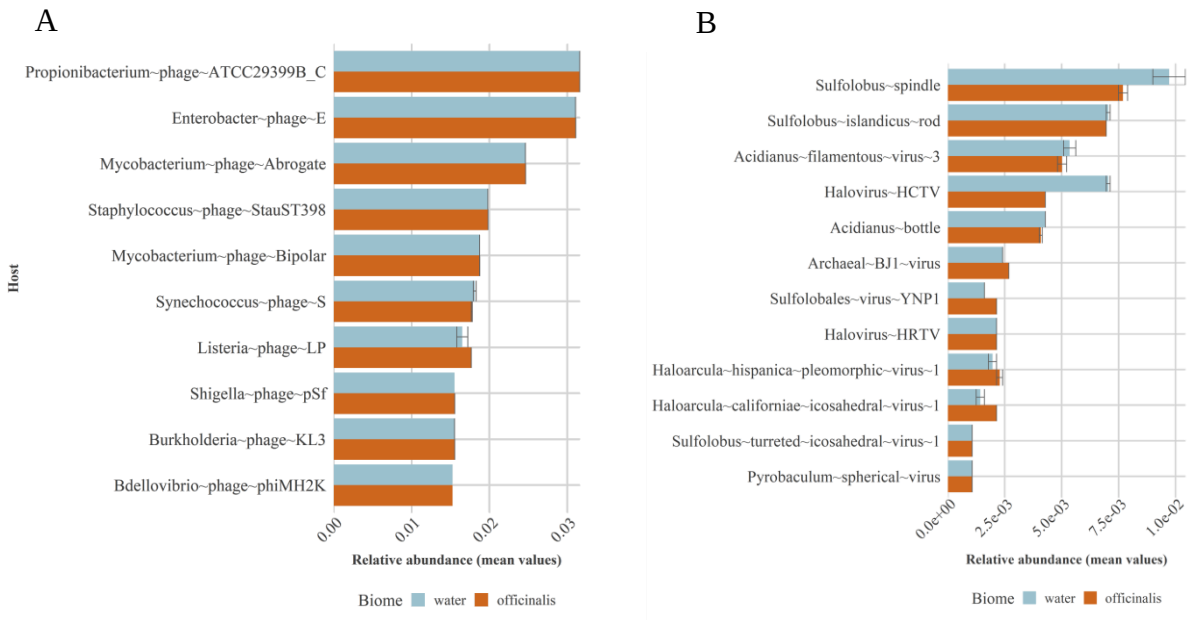


Figure 8 – Relative abundances of the ten most abundant bacterial-infecting viral species (A), and all archaea infecting viral species (B). Error bars correspond to the error associated with the mean.

Analysis at the species level revealed Mycobacterium phages to be the most dominant of all classified viruses, represented by *Mycobacterium~phage~Abrogate* and *Mycobacterium~Bipolar*. The most abundant individual species was *Propionibacterium~phage~ATCC29399B_C* followed by *Enterobacter~phage~E* (includes both *Enterobacter~phage~E1* and *Enterobacter~phage~E2* viral species). This is consistent for both biomes with little observed differences between individual samples.

Attribution of the host domain covered all data apart from *Deep-sea~thermophilic~phage*. One virus was identified as infecting a eukaryotic host, the Chimpanzee~faeces~associated~microphage~1, given the biological and ecological context this was treated as a mismatch or a contamination. Archaeal viruses were more predominant at low abundances (<0.001), with the exception of viruses belonging to the *Sulfolobus* group (Figure 7). The most accentuated difference between biomes in the archaea domain was the *Halovirus~HCTV* (Figure 8B).

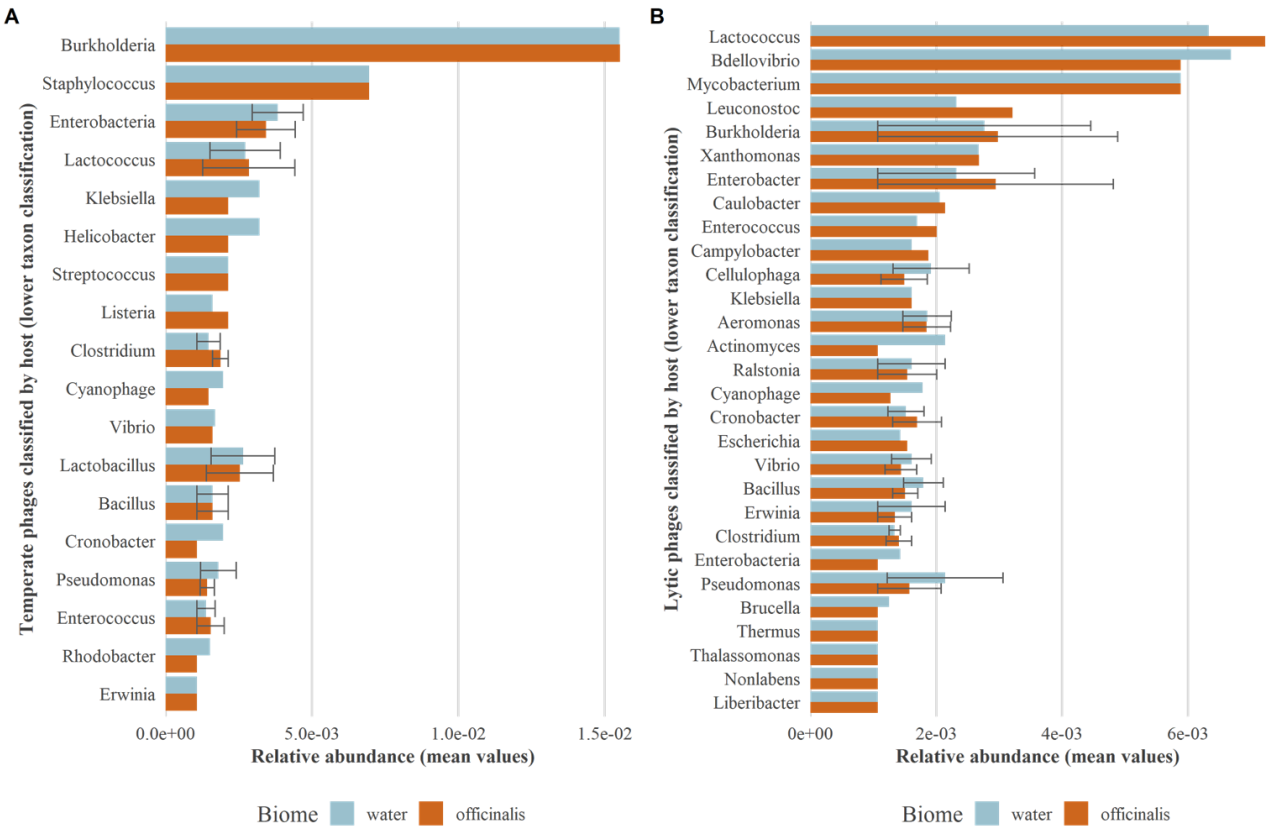


Figure 9 – Relative abundances of viral species according with temperate or lysogenic (A), and lytic (B) life cycle. Prediction of the viral life cycles for categories 1 and 2, for each biome. Error bars correspond to the standard error associated with the mean.

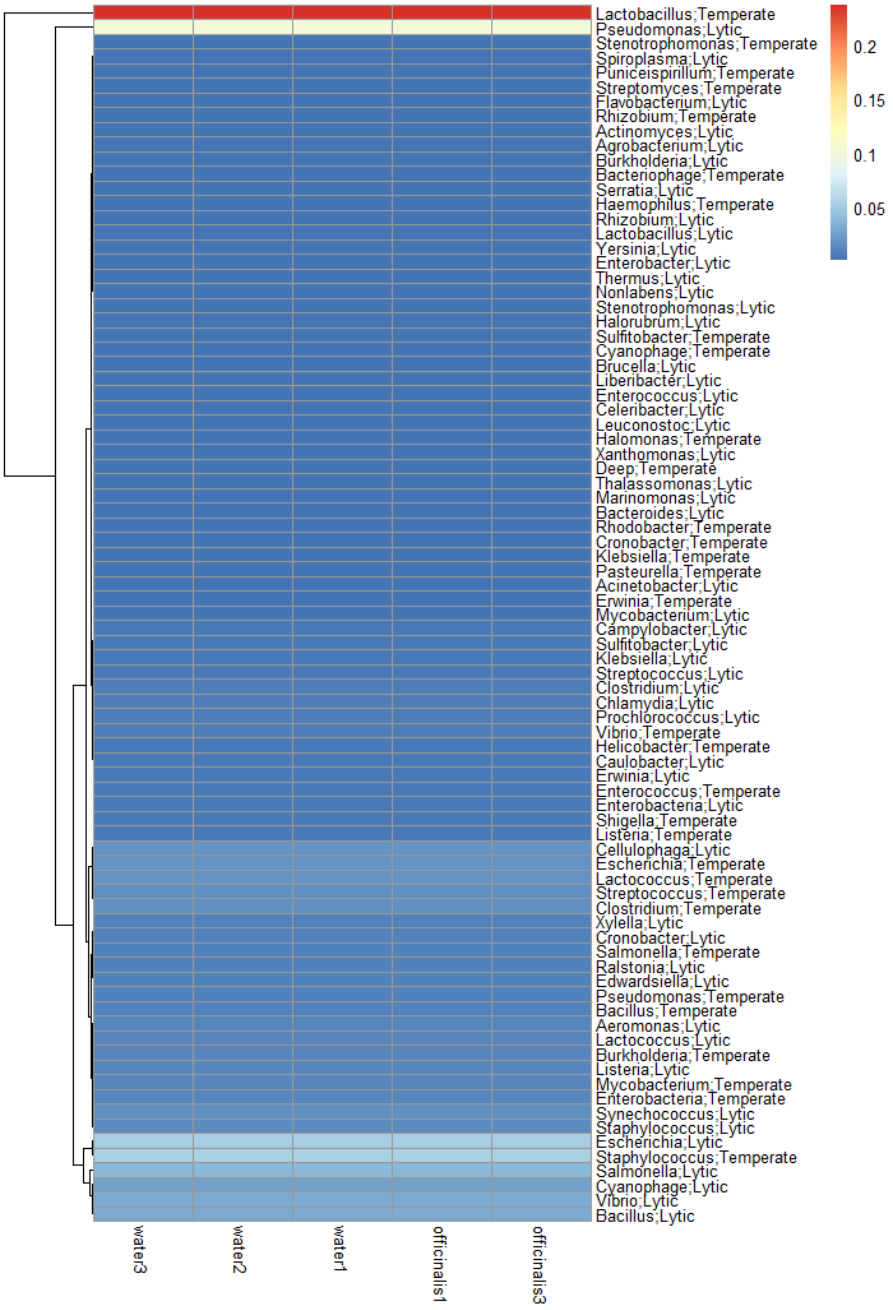


Figure 10 – Heat map representing the predicted life cycle for category 5. Viral life cycle predicted to viral species (only represented the lower taxon level), separated by a semicolon.

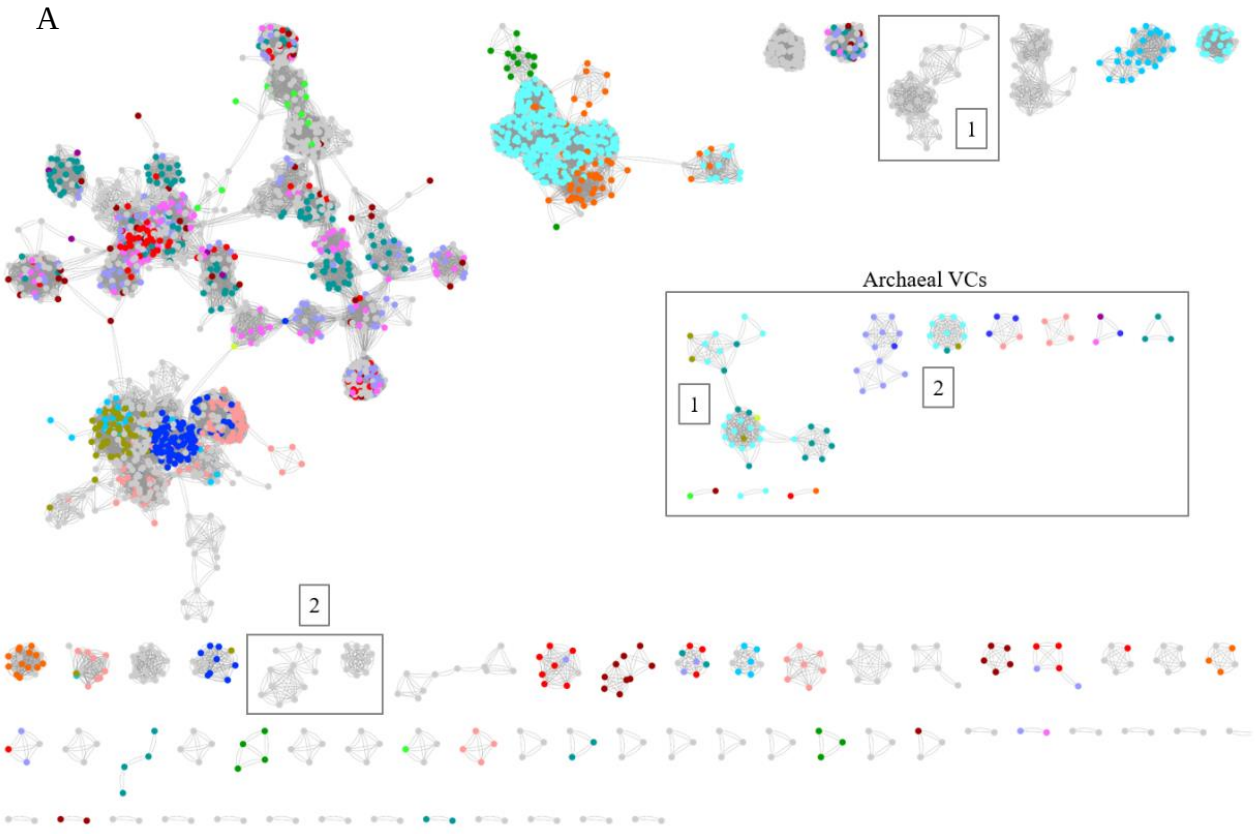
Viral life cycle was successfully assigned to 38% of viral species (70 out of 253 unique viral species) with 46 lytic and 24 temperate viruses identified (Figure 9). No significant difference was detected between biomes for the assigned life cycle (Kruskal-Wallis chi-squared=0.0010473, df=1, p-value=0.9742). This contrasts with previous studies where temperate viruses were found to be more expressive in the sponge (Jahn et al., 2021). Prophage sequences

given by category 5 were not returned for the sponge samples number 2 and remaining samples were identical regardless of the biome, both in species returned and prophage densities.

Global networks created for both biomes (Fig.11) revealed little to no observable differences, with a visual overlap for most of the largest viral clusters. Three major groups of clusters were found, the largest network includes a variety of viruses such as the *Pseudomonas* phage, *Salmonella* phage, *Enterobacteria* phage, *Streptococcus* phage and cyanophages; the second group connects *Mycobacterium* phage, *Gordonia* phage and *Streptomyces* phage; the third is constituted by only archaeal viruses.

Driven by the observation of similar networks across biomes, isolation of the clusters that are unique to each biome resulted in one VC belonging only to the seawater and three separated groups of VCs found only in *S.officinalis*. The seawater isolated cluster was composed by *Gordonia~phage~Phinally*, connected to eleven different Mycobacterium phage species (Fig.12A). VCs belonging exclusively to *S.officinalis* (Fig.12B and Fig.12C), divided into three separate groups, reflected complex interactions between different taxonomic groups but also individual non clustered viruses.

Comparison with findings from Elham Karimi's work (Karimi et al., 2017) in this sponge species, with the same raw sequences, revealed that the bacterial genus under expressed in *S.officinalis* match, to a certain degree, with members of the unique VCs isolated from the sponge network. This correlation was most notable in the *Flavobacterium* family.



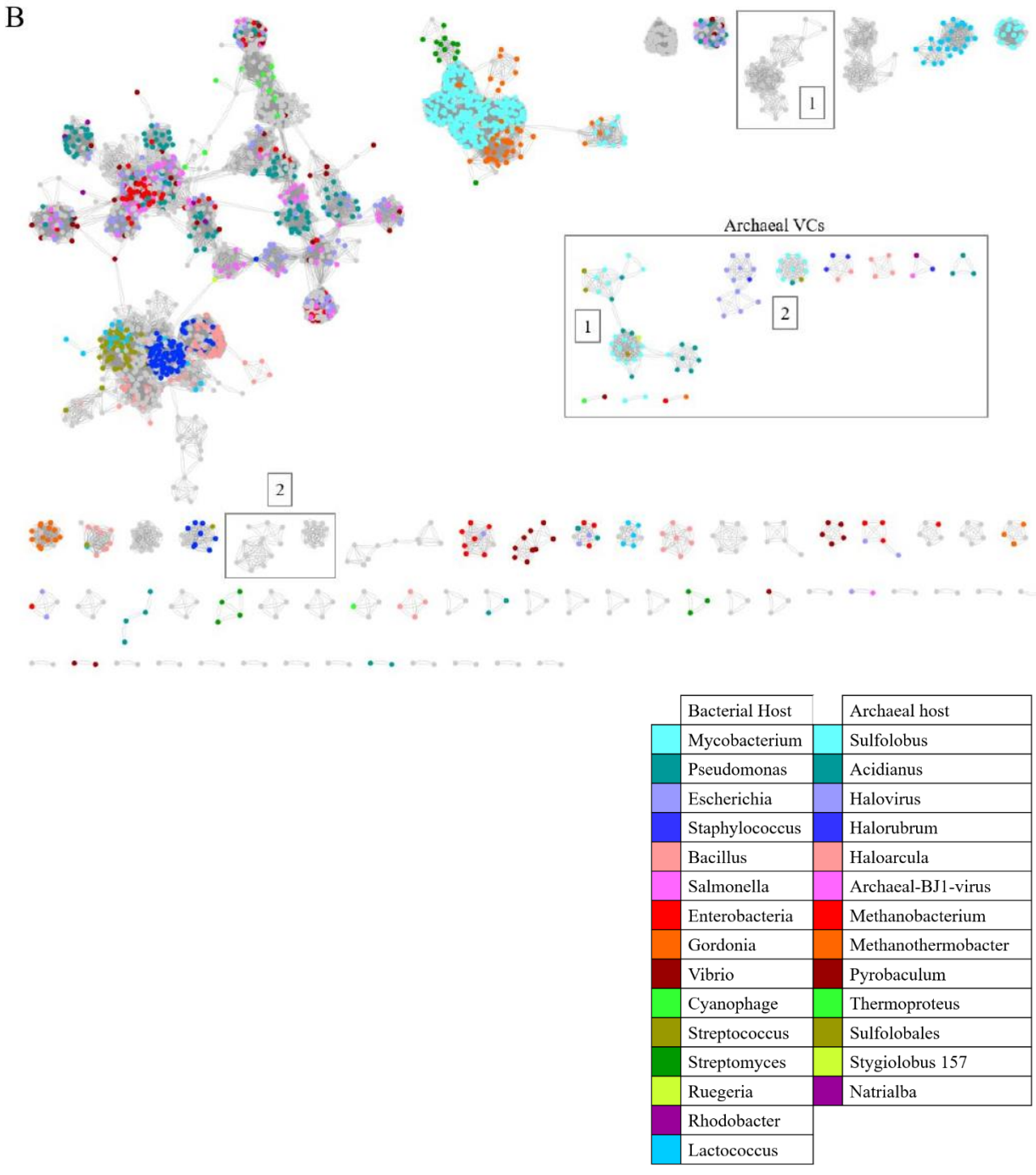


Figure 11 – Global networks for category 1 and 2 of sponge samples (A), and seawater samples (B). The top 15 bacteria-infecting viral clusters are represented for each biome and coloured according with taxon classification. Viral clusters of archaeal viruses were isolated and attributed a colour scheme.

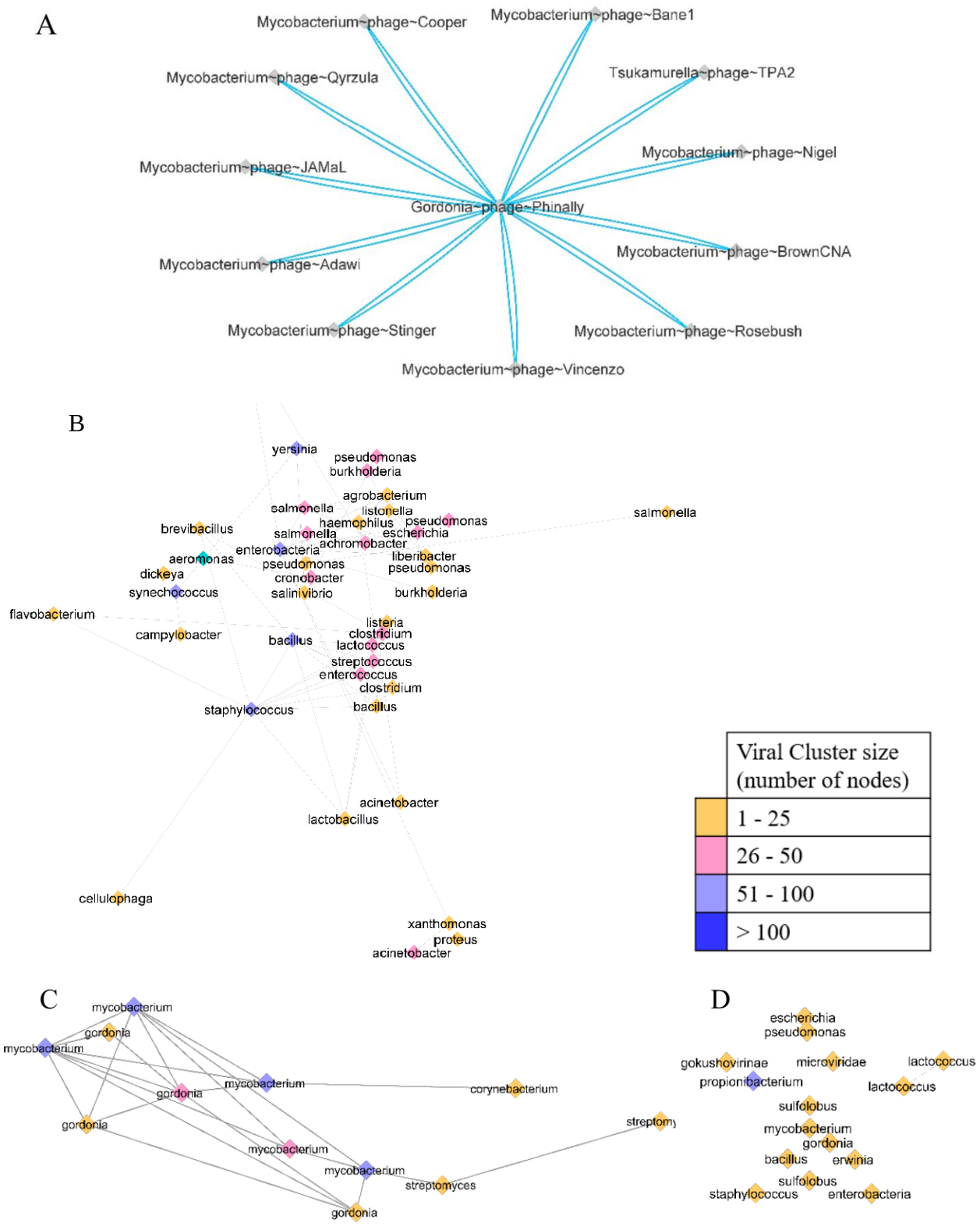


Figure 12 – Unique networks summarised for each biome. Networks obtained by removing the duplicated edges. (A) represents the unique edges found in the seawater biome, represented by the blue edges; (B-D) corresponds to the viral clusters unique to the sponge biome, in descending order of complexity, coloured according to size (number of nodes that are part of each cluster in a network). Summary networks (B-D) were obtained with MCODE App, under Cytoscape, with a network scoring based on the degree cut-off of 2, and a cluster node score cut-off of 0.2, with a K-Core of 2 and max. depth from seed of 100.

Functional analysis conducted by DRAM-v attributed a functional prediction to 12916 viral genes for the seawater samples, and 3107 genes for *S.officinalis* samples (19.3% of found genes) (Figure 13). A total of 11169 genes were found with no predicted function, with 22.4% belonging to sponge samples. Initial functional summary reflects again a higher gene count for seawater samples across all functional groups.

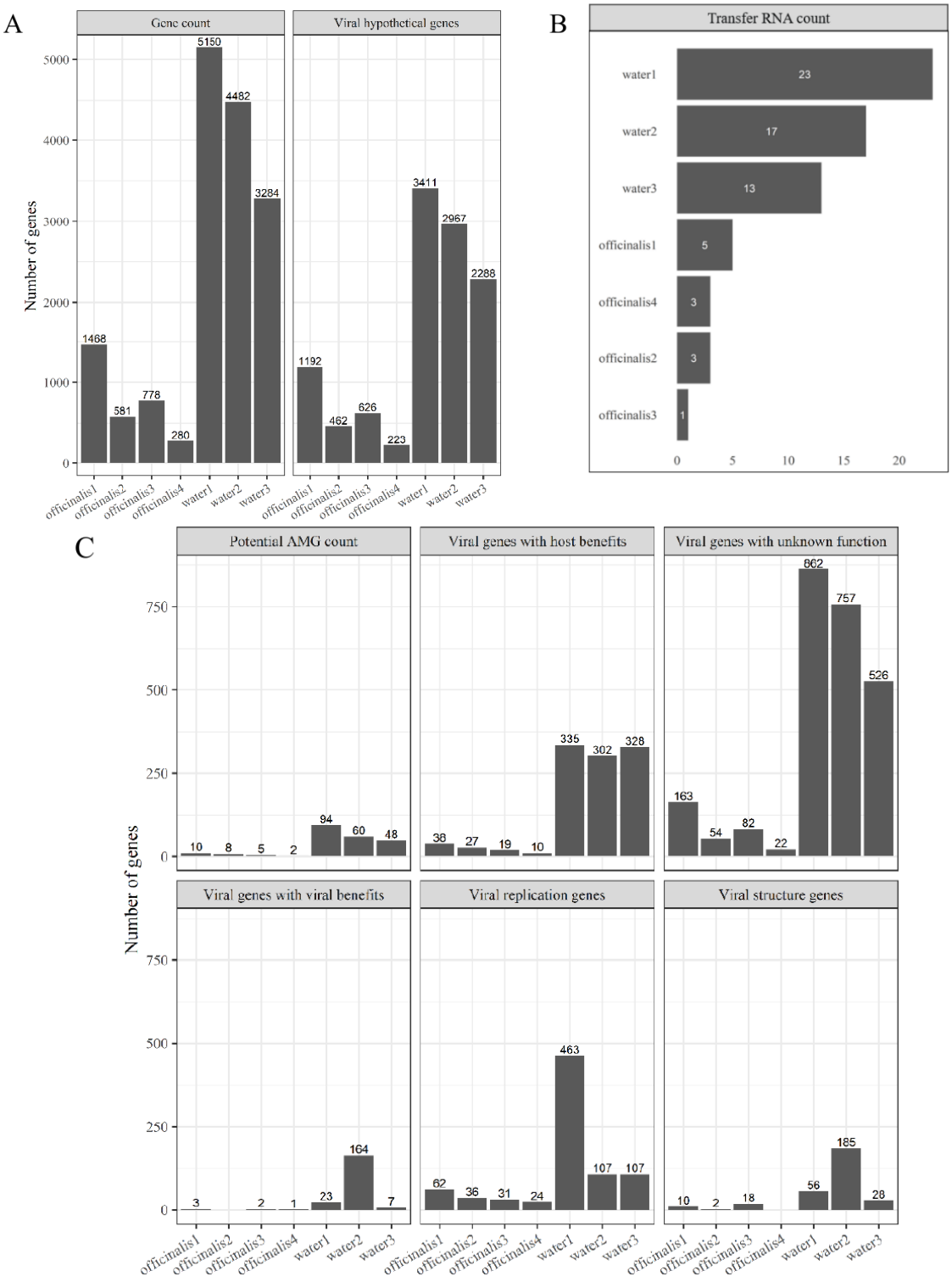


Figure 13 – Preliminary summary given by DRAM-v. Number of genes found and number of predicted viral genes (A); number of transfer RNA (tRNA) predicted (B) and functional summary of viral genes predicted (C).

Only 12 transfer RNAs were found in the sponge samples and a total of 53 in water samples, size varied between 67 and 86 bp (Figure 13B).

After removal of duplicated predictions for the same sequences and genes without function, 2439 annotations remained, with 38% found by Refseq, 30.9% by Pfam, 20.7% by VOGDB, 7.6% by KEGG, 1.8% by MEROPS and 1% by cazy. KEGG pathways obtained for 27.1% (247 total pathway counts) of the KEEG predictions, differed significantly between biomes (Krustal-Wallis chi-squared=34.721, df=1, p-value=3.805^{-9}). The viral replication pathway accounted for 41.7% (103 total counts) of all pathways returned and was only found in seawater predictions (Figure 14).

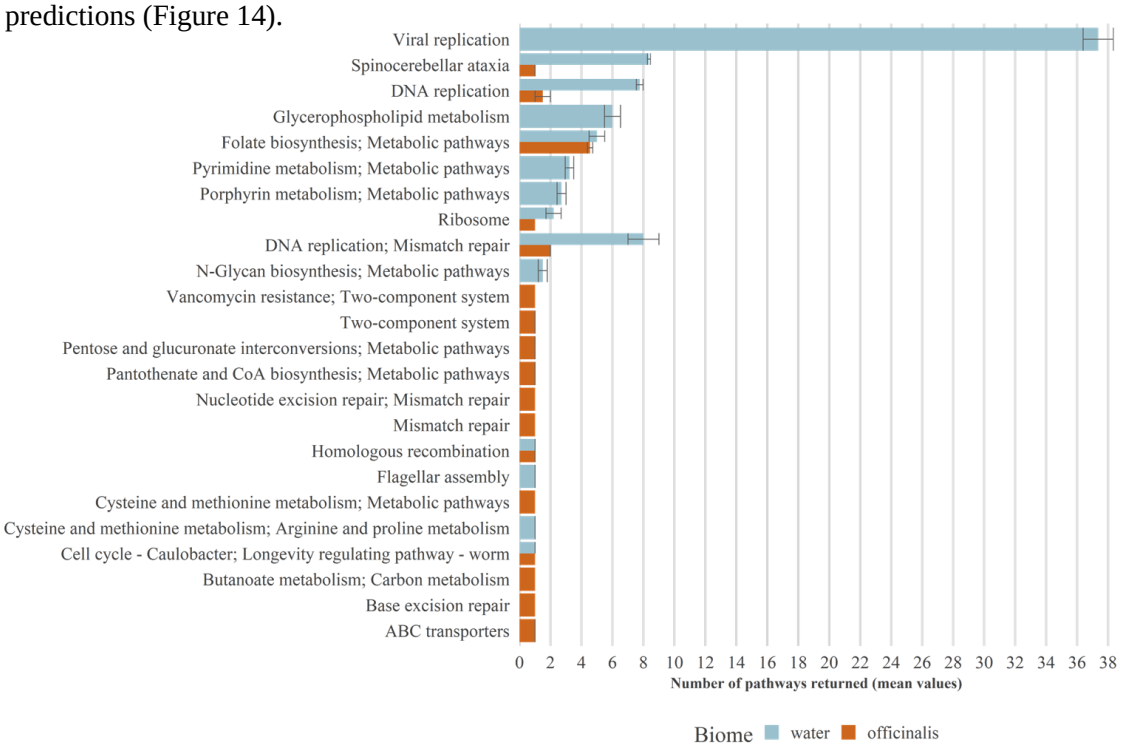


Figure 14 – KEGG pathways identified for the KEGG predictions. KEGG pathways were obtained by using with keggGet function from KEGGREST package. When a second pathway (more specific) was returned, it was added and separated with a semicolon. Error bars correspond to the error associated with the mean values.

Excluding 240 entries marked as hypothetical genes and 398 with no function attributed, custom functional tags were attributed to a total of 2199 annotations. Consistent with the initial functional summary, marked annotations showed an overall higher medium count for seawater across all major and minor functions (Fig.15). The only exception to this was the very late expression factor 1, SNF2 family N-terminal domain, ABC transporters and integrases, with higher expression in *S.officinalis* than in seawater. Surprisingly no annotations were found in any sponge samples related with translation, translation/transcription repressors, clamp proteins or cyclases for the genome major function. The very late expression factor 1 is an integrase that was present in great numbers across sponge samples. An alignment with all contigs with this predicted function (Figure S2, S3) shows distinct coding regions and coverages, that separate the officinalis

viral contigs from de water viral contigs. Additionally, a tree grouping the different contigs revealed genetic similarity between viral sequences of officinalis samples.

Several proteins that are associated with viral infection were also absent, such as type VIII secretion system (T8SS), curlin proteins, cell wall hydrolases and spanin proteins (both inner and outer membrane subunits), some of which can be indicative of a preference for the lysogenic state (Song, 2020).

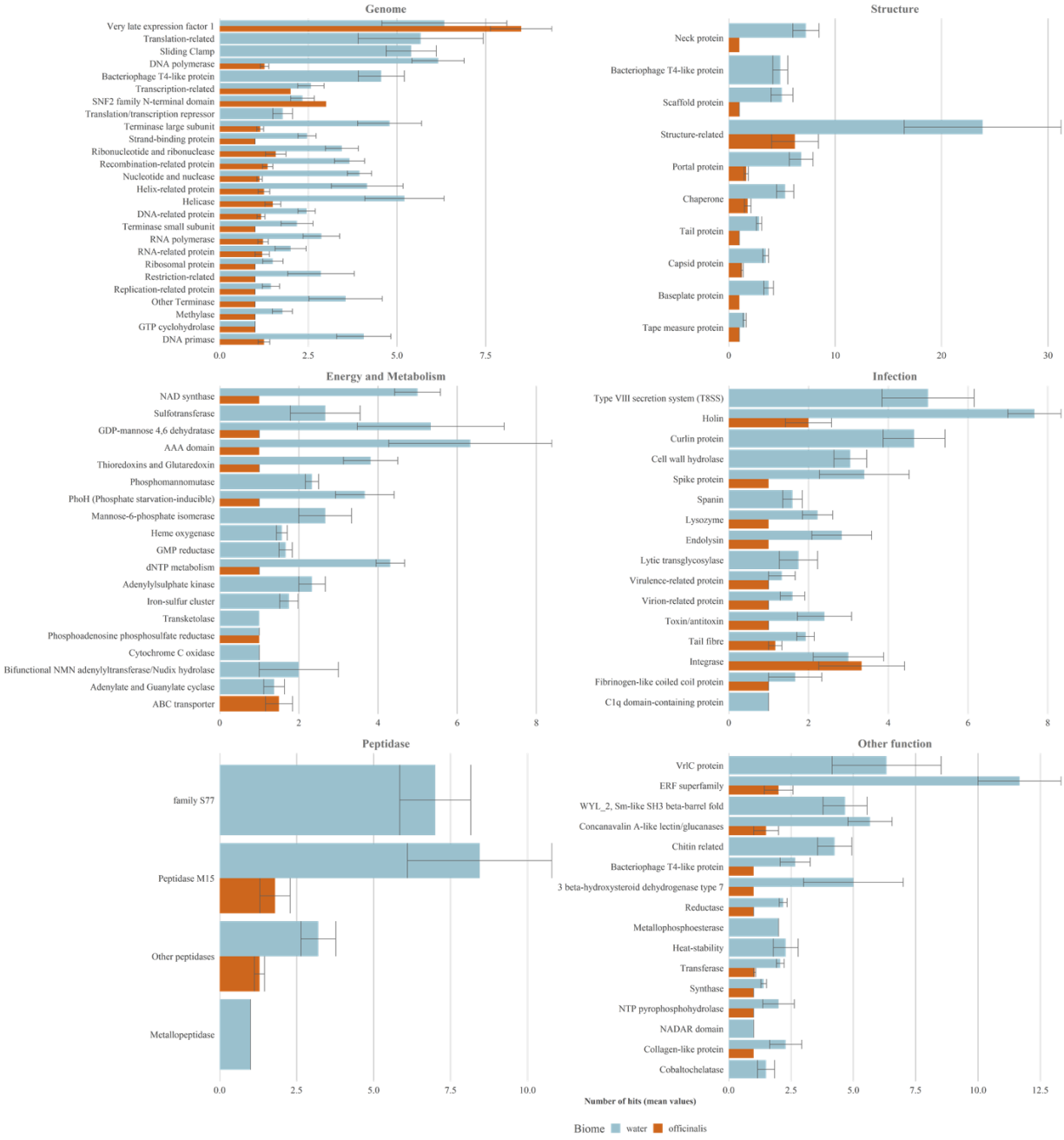


Figure 15 – Summary of functional predictions returned by DRAM-v. Functional predictions were classified according to custom ontology terms as follows: Genome, Structure, Energy and Metabolism, Infection, Peptidase and Other function. The bar plots represent the mean count values, colored according to the biome, with error bars associated with the mean error.

Functional annotations related with the "Genome" were divided into polymerases (Fig.16) and Restriction-modification system (RM) (Fig17). Polymerase predictions included the sigma-70 factor that was encountered exclusively in sponge samples, with presence of region 1.2, region 2 and ECF subfamily. RM was obtained by grouping the minor functions that harbor restriction-related proteins, helicases, endonucleases and methylases. Distinct hits only present in the sponge samples for RM include the nucleotidyl transferase, that was present across all sponge samples, tRNA nucleotidyltransferase in officinalis 2 sample and adenine-specific DNA methylase.

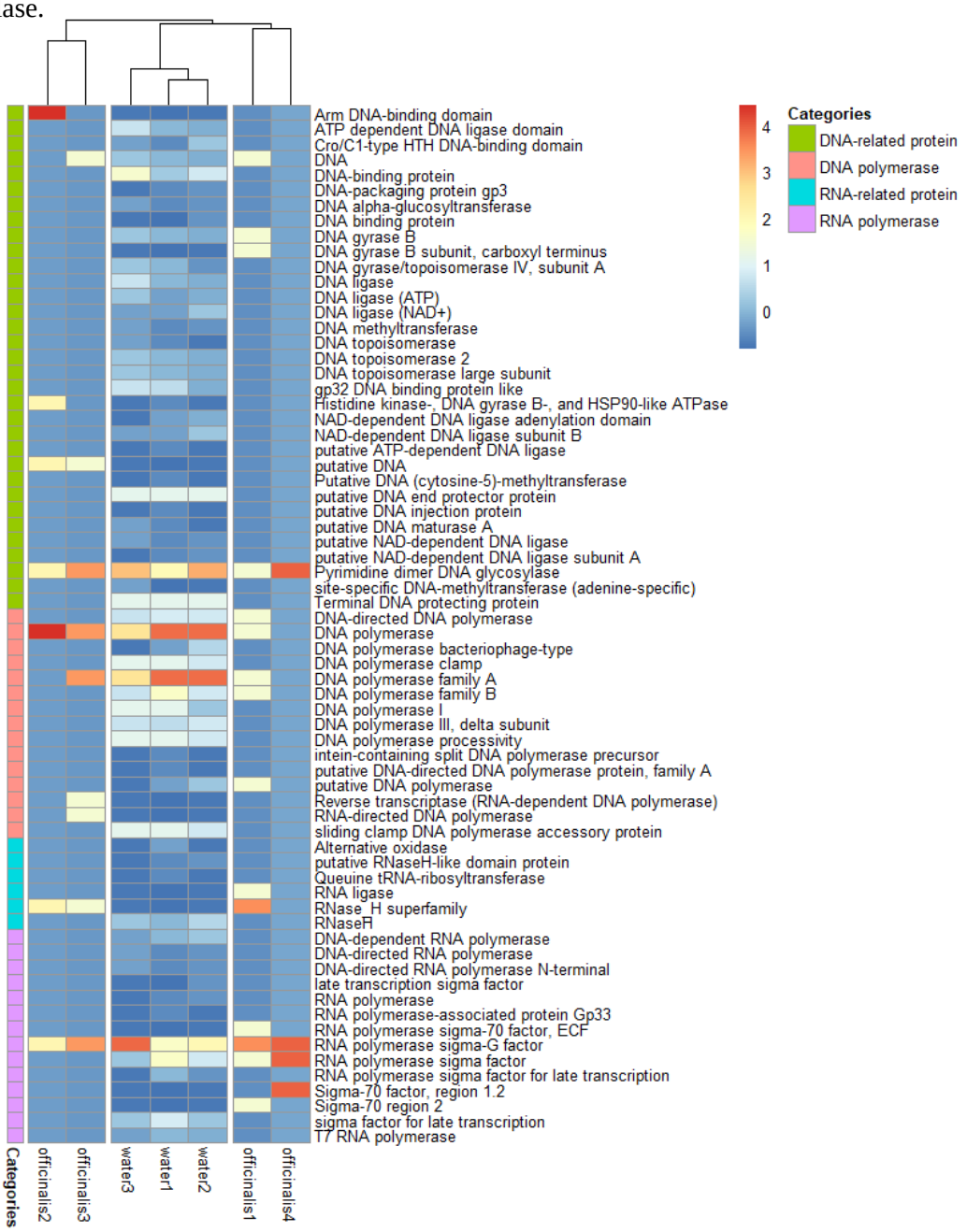


Figure 16 – Heat map of functional predictions classified as polymerases and DNA/RNA related proteins. The data prior to plotting was centered (subtracted the mean before scaling) and divided by the standard deviation when scaling. Categories (at the left) are colored and labeled according with the custom functional ontology terms created.

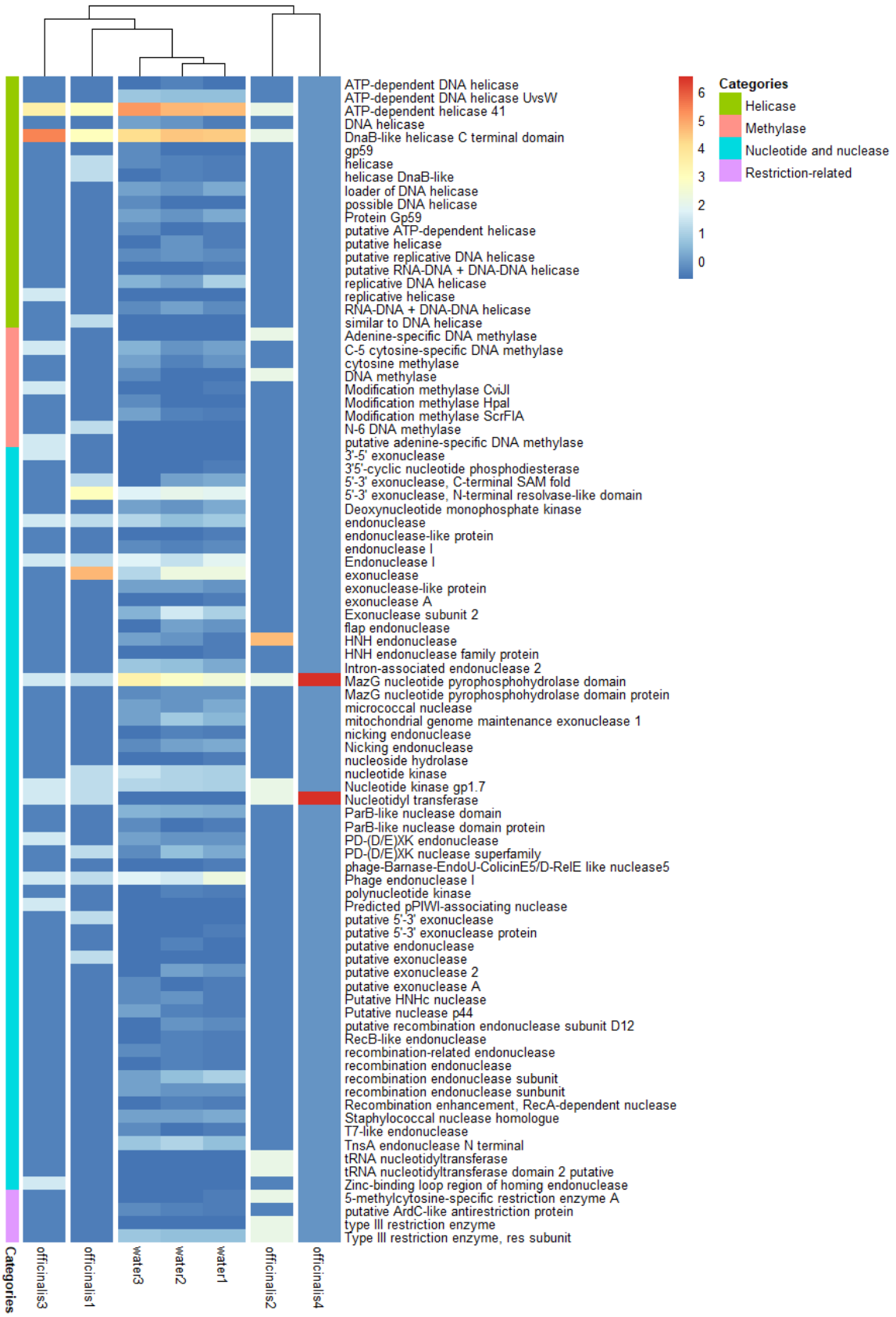


Figure 17 – Heat map of predictions classified as Restriction-modification systems (RM). RM predictions under the Genome major function include helicases, methylases, nucleotides, nucleases and restriction-related proteins. Scaling according with Figure 15

Energy and metabolism functional category (Fig.18) harbored interesting differences between biomes. Intermediates involved in sulfur metabolism were lacking from sponge samples, apart from one hit to phosphoadenosine phosphosulfate reductase family. Iron-sulfur (Fe-S) clusters (K13628) only present in seawater, can be involved in a myriad of conserved roles such as electron transfer, oxygen-iron sensing and enzyme catalysis (Hurwitz et al., 2015). Additionally, Fe-S clusters can be detrimental for the function of the tail tip complex, and this way, enable the conformational changes or guide DNA injection (Tam et al., 2013). Two proteins related to iron-sulfur clusters were also exclusively present in the seawater. The first, Glutaredoxin (PF00462.25; PF00268.22), is a protein required for Fe-S cluster biosynthesis, assembly and involved in sensing of cellular iron (Rouhier et al., 2010). Second, the polyamine S-adenosylmethionine decarboxylase (K01611), can be incorporated into Fe-S clusters to manipulate the host response to stress (Hurwitz et al., 2015). The robust presence of Fe-S clusters and related proteins in seawater raises the question of why these were totally absent in the *S.officinalis*, given that iron, as a micronutrient, is essential for microorganisms and it's availability in the ocean is very limited (Lauderdale et al., 2020; Sun et al., 2021).

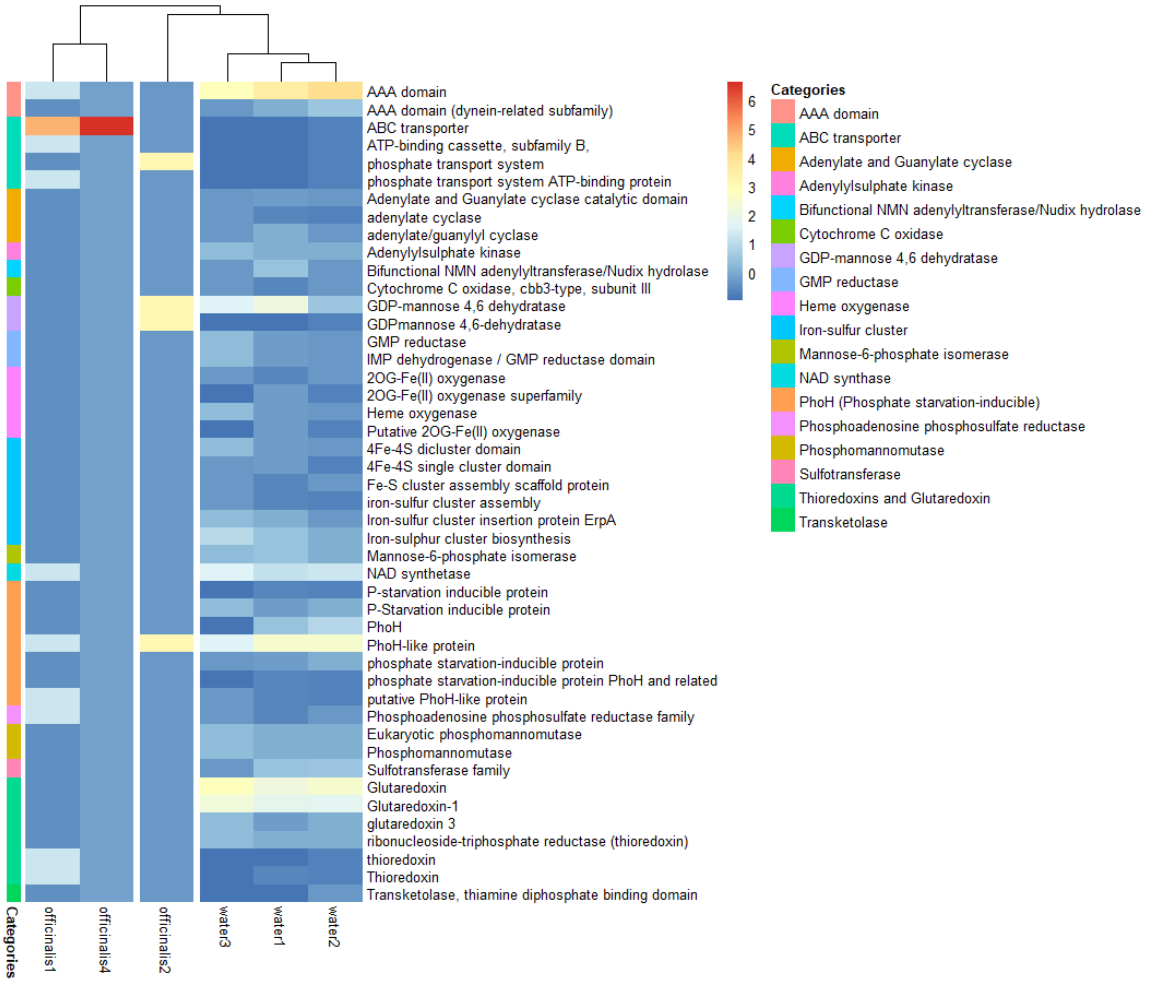


Figure 18 – Heat map of predictions classified as Energy and Metabolism. These include proteins related with nutrient uptake pathways and energy molecules. Scaling according with Figure 15

A possible explanation may be related with the ability of this sponge species to retain and accumulate iron on the spongine fibres within lepidocrocit grains (up to 7.5% of dry weight in polluted areas) (Vacelet et al., 1988). This is reinforced by the fact that *S.officinalis* is often referred to as a biomarker for metal contaminated areas, demonstrating high tolerance to heavy metals (Berthet et al., 2005; Roveta et al., 2020). To corroborate this, there is a study that reports bacterial symbionts present in *S.officinalis* that show high tolerances to metal contamination (Bauvais et al., 2015). The hypothesis is that a portion of this iron is made available to the both the microbial symbionts of the sponge, and, consequently, to phages that infect the sponge - associated bacteria, and use this element to sustain their growth. Additionally, lysis of microbial communities that are pumped into the sponge will also be an important source of micronutrients, including iron.

The global AMG pool outside the sponge, would this way evolve to offer multiple ways for microbial life to extract iron from their surrounding environments. This grants viruses an opportunity to exploit this "weak point" and interfere with these pathways by encoding Fe-S clusters, Glutaredoxin, S-adenosylmethionine decarboxylase and heme oxygenase. When the survival of the host is beneficial for the virus, these pathways can also ensure the host survival by providing the infected cell with the necessary genes to ultimately acquire iron.

In *S.officinalis*, where iron is not a limiting factor, the presence of pathways dedicated to iron sequestration would not be extremely necessary and so absent from bacteriophage genomes. However, this still means that there must be a way to solubilize, capture, and deliver this metallic compound to the cells. A possible solution is the use of ATP-binding cassette (ABC transporters) (PF00005.28, K06147, K02036) to transport iron complexes to the cytoplasm (Sandy & Butler, 2009), since they are found in higher frequencies in the sponge prokaryotic virome. Corroborating this is the fact that ABC transporters were also found to be abundant in *S. officinalis* bacterial symbionts (Karimi et al., 2017). These integral membrane proteins were found enriched in viruses sampled from the aphotic zone along with nucleotide synthesis (Coutinho et al., 2017). Whether there is a connection to iron is up to debate and further investigation is required, however ABC transporters encoded by viruses seem to have a clear role in facilitating the nutrient uptake of bacteria, with the intention of providing the necessary conditions to benefit the synthesis of new viral particles.

This hypothesis crosses paths with the the Ferrojan Horse Hypothesis, that states that the use of iron by phages, incorporated into the tail fibers, can facilitate the entry of the phage into the host cell, by recognition of the host sidephore-bound iron receptors (Bonnain et al., 2016). This use explains the interest for the phage, in acquiring remineralized iron upon cell lysis. By using this limited micronutrient as a phage strength in the arms race against bacteria, viruses greatly influence the biogeochemical cycling of iron in the ocean. Inside the sponge the

availability of the dissolved iron may prevent phages from following this path. Viruses greatly depend on iron for infection by competing with siderophore-bound iron present on the surface of the host cell (Bonnain et al., 2016). Interestingly, reports show that bacteria isolated from marine sponges that have no siderophore production (Guan et al., 2001). In this case, siderophores are hypothesized to have a role in signaling necessary compounds for growth and iron uptake for bacteria that don't produce siderophores and are under iron-limited conditions (Guan et al., 2001). The absence of these structures both in viruses and sponge symbiont bacteria reinforces the existence of a shared iron pool inside the marine sponge.

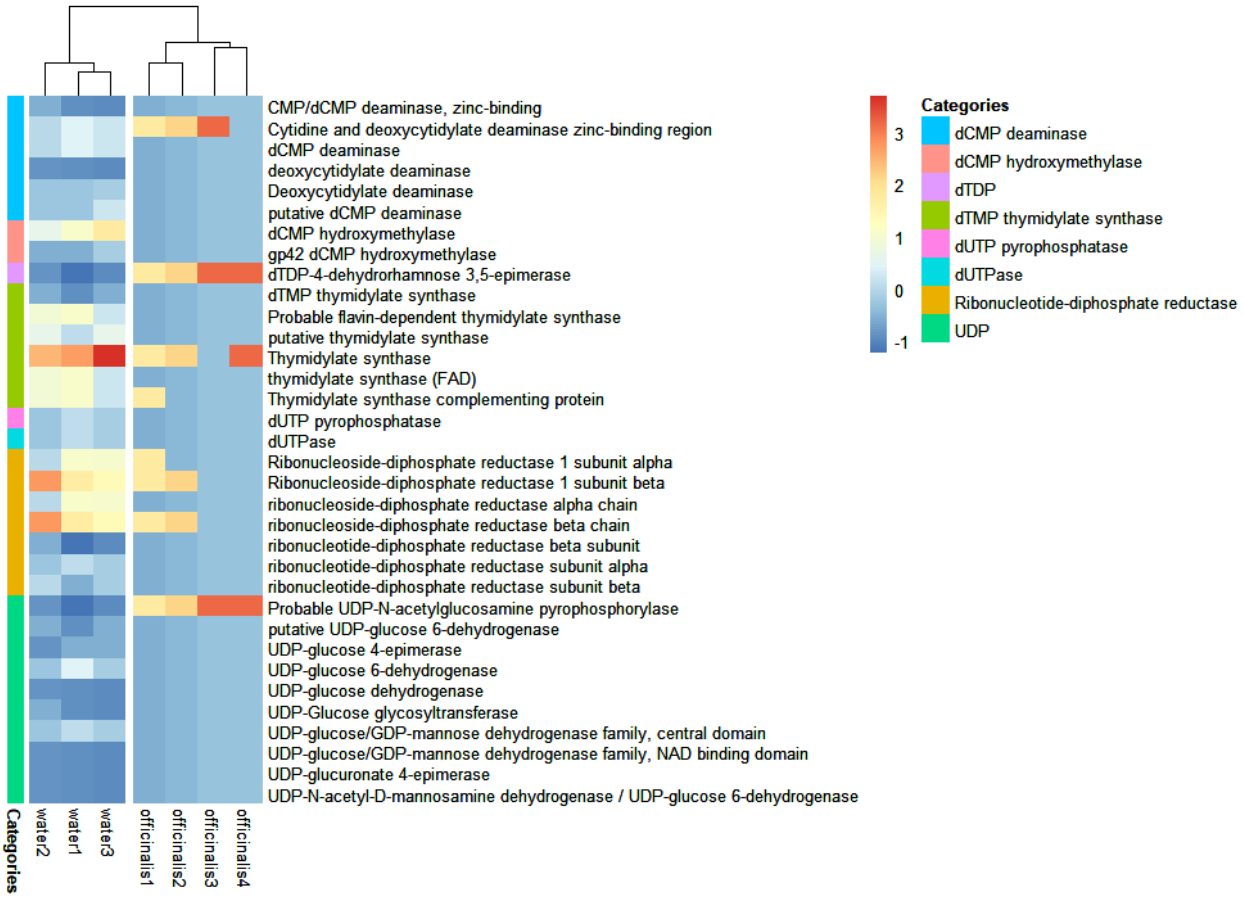


Figure 19 – Heat map of predictions classified as dNTP biosynthesis, under the Energy and Metabolism category. These include proteins related with nutrient uptake pathways and energy molecules. Scaling according with Figure 15.

The dNTP metabolism (Fig.19) revealed two unique annotations for *S.officinalis*, the dTDP-4-dehydrorhamnose 3,5-epimerase (PF00908.18) and the probable UDP-N-acetylglucosamine pyrophosphorylase (returned by VOGDB). Apart from these, the only intermediates found in the sponge were the thymidylate synthase, cytidine and deoxycytidylate deaminase (zinc-binding region) and different ribonucleoside-diphosphate reductase components, all more present in the seawater. Sequences containing the four counts found in the sponge (one per sample) for dTDP-4-dehydrorhamnose 3,5-epimerase were aligned together and ORFs predicted. This resulted in a complete alignment with 100% of similarity between sequences.

Together these findings suggest that synthesis of deoxyuridine monophosphate (dUMP), an essential intermediate for the de novo synthesis of dTTP, follows two different pathways in seawater. The first by deamination of dCMP to dUMP by dCMP deaminase, and the second by reduction of UDP to dUMP by dUTPase, both described in (Y. Zhang et al., 2007).

The ubiquitous presence of dTDP-4-dehydrorhamnose 3,5-epimerase (RmlC) is involved in the biosynthesis of rhamnose, which is a sugar molecule that is found in the cell walls of some bacteria, archaea, and viruses (Graninger et al., 1999). This metabolic process has been previously described in large DNA viruses, where it is proposed to be essential for virus replication and infection (Parakkottil Chothi et al., 2010), even though only one intermediate of the pathway was found. Interestingly dTDP-4-dehydrorhamnose 3,5-epimerase and related enzymes were recently described in three different species belonging to the *Nitrososphaerota* phylum (Archaea), isolated from sponge species of the Red Sea (Haber et al., 2021). The proposed explanation is that rhamnose may modify the surface layer glycoproteins that are characteristic of these archaeal species, and this way aid in the evasion of digestion by the porifera host. A similar case is described for sponge associated *Endozoicomonas* where rhamnose metabolism was proposed to be an alternative carbon source (Alex & Antunes, 2019).

Alternatively, the absence of dTDP-4-dehydrorhamnose reductase from cyanobacterial symbionts of sponges was hypothesized to have a connection with recognition by the sponge host, and possibly phage resistance (Burgsdorf et al., 2015). Considering these studies, the presence of dTDP-4-dehydrorhamnose 3,5-epimerase in the *S. officinalis*-associated bacterial-viruses may suggest either a way for these viruses to mediate the immunity of bacterial and archaeal symbionts to sponge digestion, or have implications not yet described for phages as part of the sponge consortium.

The alignment performed on all the 4 viral sequences (officinalis samples), in the positions predicted for the dTDP-4-dehydrorhamnose 3,5-epimerase, returned 100% identity (Fig S1). This corroborates the fact that this intermediary is, in some degree, conserved in viral samples, and important to the possible interference of the rhamnose biosynthesis pathway of sponge bacterial symbionts.

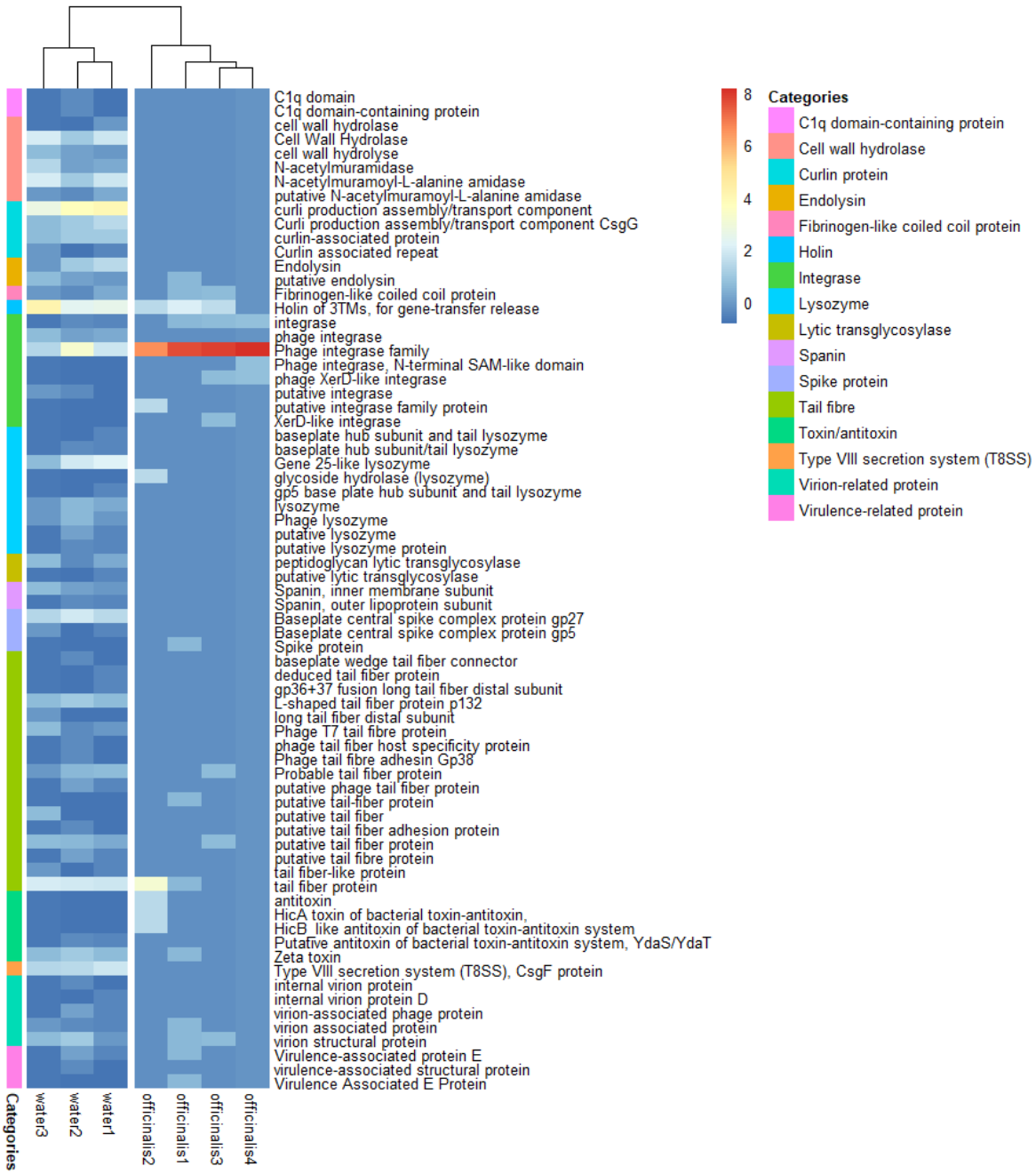


Figure 20 – Heat map of predictions classified as Infection. These include proteins related with integration of the viral genome upon infection, disruption of the bacterial components such as the cell wall and recognition of bacterial surfaces. Scaling according with Figure 15.

As previously mentioned, integrases are more predominant in sponge samples, although also present in seawater (Fig.20). Unique integrases found in the *S.officinalis* prokaryotic virome include the phage integrase (N-terminal SAM-like domain) and XerD-like integrase. In the recombination minor function, under the genome major function, five hits to XerD-like recombinases were also found in sponge samples and absent from the seawater. The XerD recombinase was recently described as essential for filamentous phages integration into the bacterial host (Huber & Waldor, 2002). This could suggest a hypothetical narrower infection range for sponge associated viruses, targeting the disease-causing bacteria.

The enrichment of integrases in the *S.officinalis* prokaryotic virome could also translate into an increase of the potential for lysogenic life cycle and consequently a higher number of prophages integrated in microbial genomes, as described in (Touchon et al., 2016) where integrases were identified as the most encoded phage-specific function (86%). The same study proposed that larger bacterial genomes have more neutral targets, and so, are more prone to viral integration (Touchon et al., 2016). Despite the prediction of prophage sequences by in category 5, the presence of this prophage marker is still relevant, and can indicate preference for integration into symbiont bacterial and archaeal genomes.

Peptidases returned mainly by the MEROPS database showed no genes belonging to sponge samples that codify to peptidases of the family S77, with a total of 21 hits found in seawater (Figure 21). Moreover, metallopeptidases (mainly zincin-like metallopeptidases) also were not found in *S. officinalis*. The only peptidase found in this biome belongs to the subfamily S49C non-peptidase homologue, and peptidase C26, excluding unassigned peptidases.

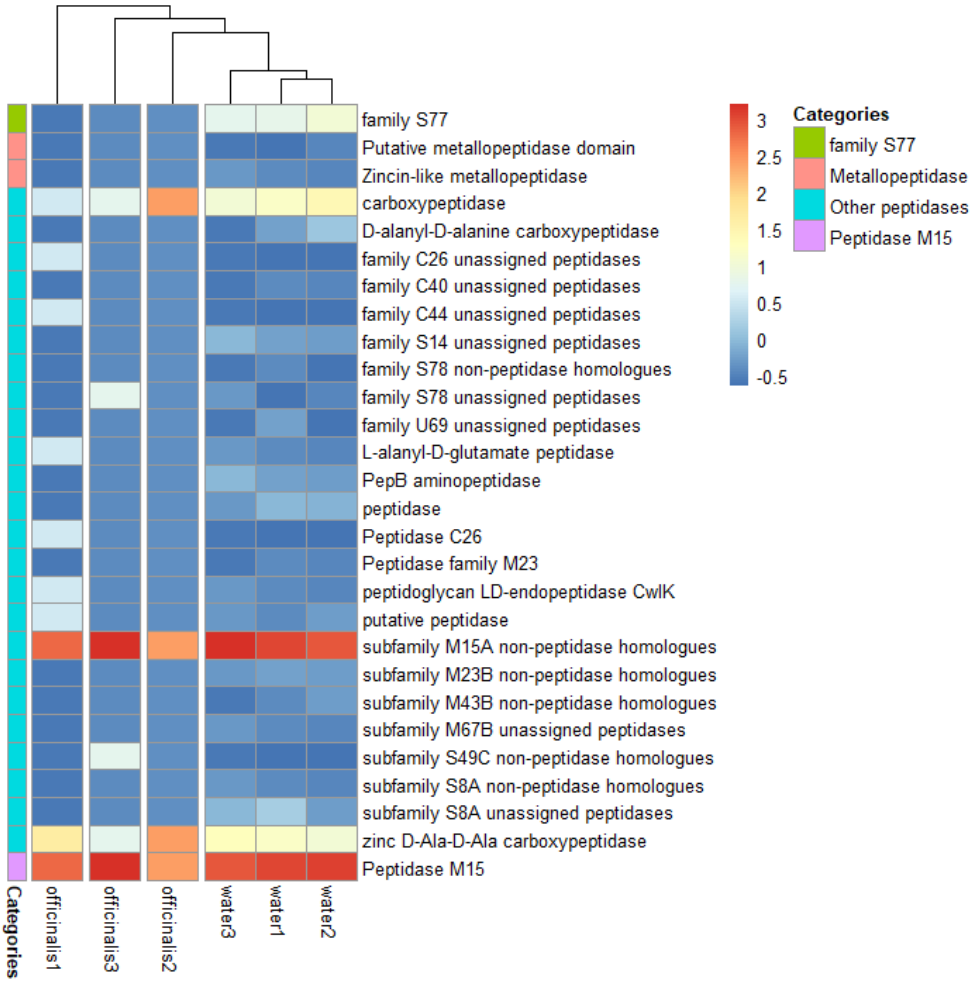


Figure 21 – Heat map of predictions classified as peptidases. Predictions returned mainly by the MEROPS database. Scaling according with Figure 15.

Several unique annotations tagged with “Other functions” (Figure 22) were only found in sponge samples, the most noticeable (more than one hit) was glutamine amidotransferase, predicted to be involved in cell wall modification (Iyer et al., 2009), and the 7-cyano-7-deazaguanine synthase and 7-carboxy-7-deazaguanine synthase, both involved in the Queuosine and archeosine synthesis pathways that allow viruses to modify their DNA and acquire resistance to digestion by restriction enzymes (Hutinet et al., 2019).

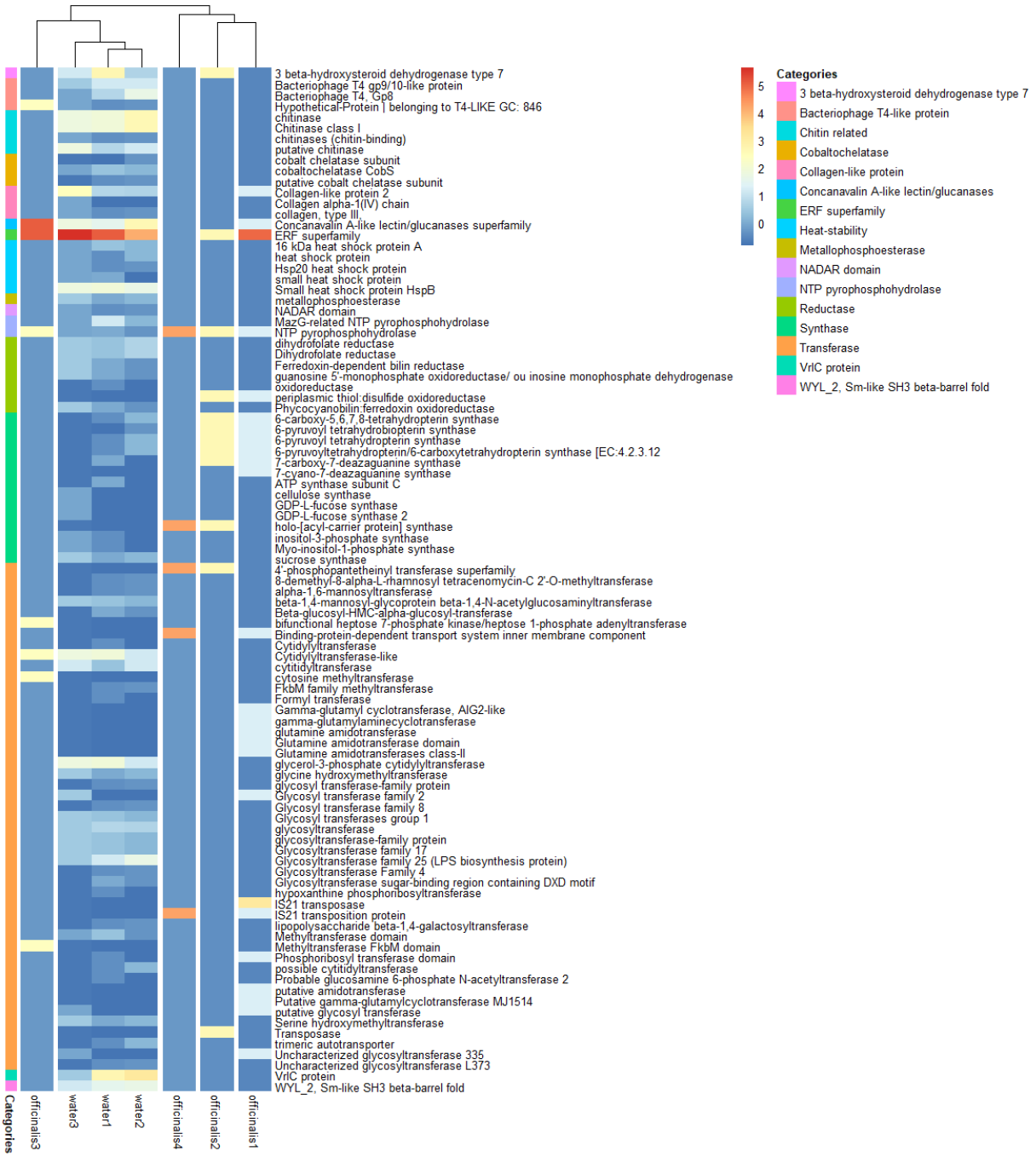


Figure 22 – Heat map of predictions classified as “Other functions”. Predictions returned mainly by the peptidase database. Scaling according with Figure 15

The functional predictions attributed to the Structure category shows a massive divergence of baseplates, head, capsid and tail related proteins (Figure S5 and S6). The less variety showed for the officinalis viral sequences, in this category and others (such as the Other function, Figure 22) hints for a more restricted viral gene pool. There is presence of functional predictions of most minor function (baseplate, head, capsid, neck and tail), and the adjusted numbers are comparable to viral sequences of water samples, however few annotations are present in the officinalis samples. This might be due to the fraction that was lost during the identification of viral contigs, where a lower threshold should be experimented; or reflects evolutionary forces that constrict the viral genetic freedom, due to close interaction with the bacterial symbionts of *S.officinalis.*

Annotations that did not match any of the custom tags were divided into gene product (gp) (Figure S8) and other non-identified annotations (Figure S9). For gene products a clear divergence between biomes can be seen, with a total of 14 different "gp" unique to sponge samples (Figure S7). Other annotations also revealed a contrast between sample types, namely D2 protein and S-adenosylmethionine decarboxylase stood out due to their widespread presence in seawater samples (2 hits per sample and 2 hits in both water 1 and water 2, respectively), and absence from all sponge samples.

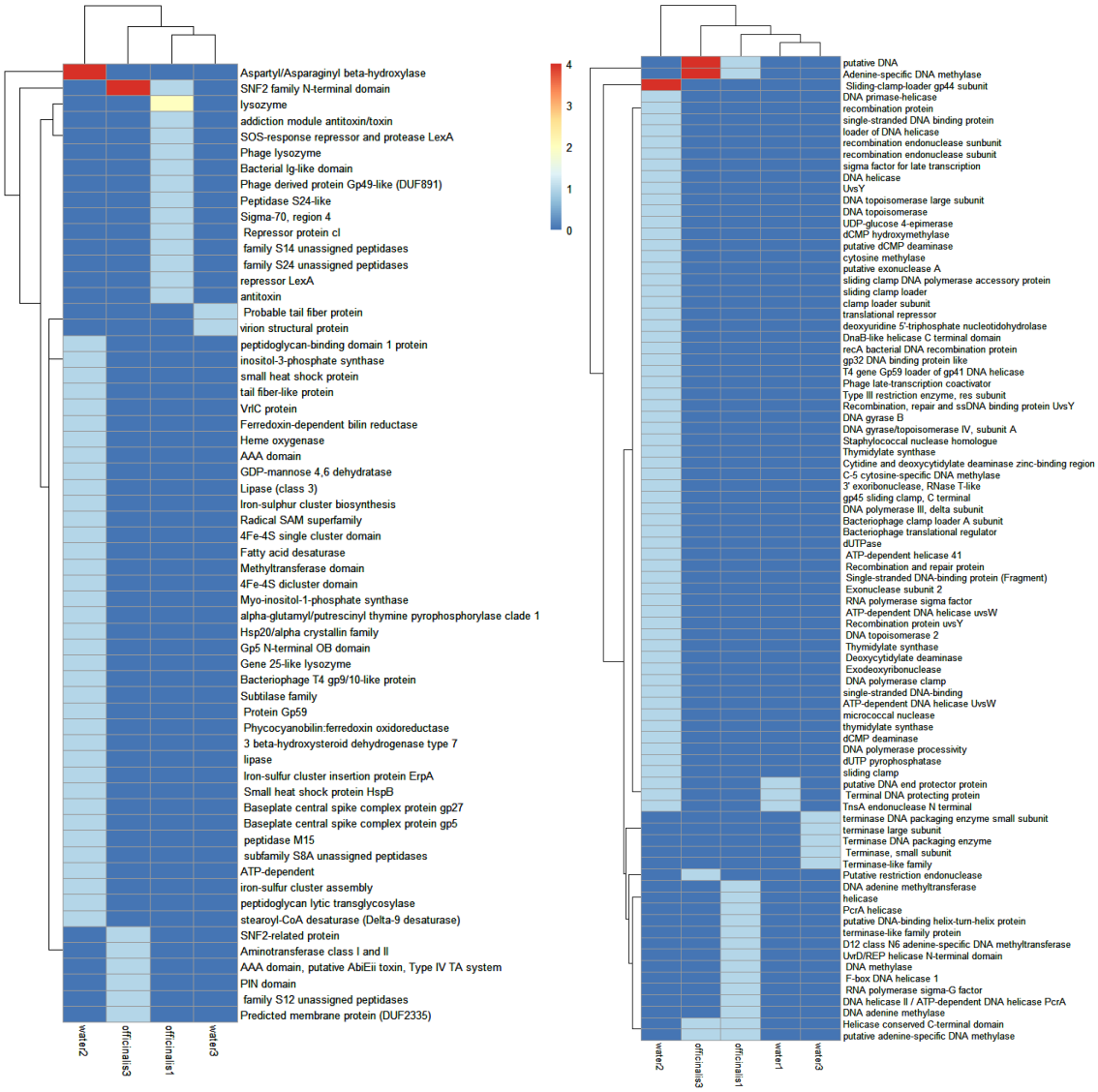


Figure 23 – Heat map of counts for all predictions for the category 5 (prophage). Predictions are divided into not related with "Genome" (B), and "Genome" related (B), based on the custom functional categories.

Prophage sequences demonstrated an absolute contrast between seawater and sponge (Figure 22), since no annotations were found present in both biomes, although some predictions use different names for the same function, mainly helicases. Seawater samples contained prophage genes coding for most functions related with "Genome" (Figure 23B) such as terminases, DNA polymerases and topoisomerases, recombination associated proteins, single-strand binding proteins and exonucleases. Sponge prophage genes coding for recombination associated proteins, DNA polymerases or exonucleases were not present. Instead, two RNA polymerases were found (sigma-70 factor region 4 and RNA polymerase sigma-G factor) along with the previously highlighted SNF2 family N-terminal domain and adenine-specific DNA methylase (including other non-specified adenine methylases). Both these polymerases may contribute for specific targeting of bacterial hosts, as the first remodels the chromatin and perform

specific alterations to the genome. The toxin /antitoxin systems present in these samples are described to play an important role in microbial symbionts of marine sponges by allowing the control of the horizontal gene transfer from external sources, which can interfere with the genome integrity (Webster & Thomas, 2016).

In this study, it was possible to conclude that *S. officinallis* harbors viruses that are different from the free living marine viruses, as reinforced by the multivariate analysis conducted (Figure 6A and 6B), and the distinct pathways associated with the functional profile. The viral life cycle investigated here also shows a contrast in the coding potential of sponge prophage samples that, together with the higher number of integrases predicted (Figure 19), and lesser prevalence of viral genes related with the viral replication pathway (Figure 13), this results lean towards a preference for temperate phages that undergo lysogenesis, as previously proposed for the sponge virome (Jahn et al., 2021; Nguyen et al., 2021). This question, would however, greatly benefit from the use of single-cell genomics-based analysis (Labonté et al., 2015; Martinez-Hernandez et al., 2017) to allow more confident correlations between virus and host.

The combination of shotgun and long-read viral metagenomics is also an improvement that could be made, that would allow for a greater coverage and error mitigation. This would also increase the chances of identifying niche-differentiating genomic features between the two sets of viruses.

The functional analysis using DRAM-v, from a bioinformatic standpoint, is to this date one of the most recent and robust tools, given the database resources it uses, allowing the identification of key differences between metagenomes.

CHAPTER 4. CONCLUSION

Altogether the results shown here reveal clear differences between the viruses belonging to *S.officinalis* and seawater samples. The analyses performed based on taxonomy predicted to viral clusters had several limitations in capturing a substantial difference between sample types. The multivariate analysis and functional prediction reflected the most divergence between the sponge and seawater viral samples.

Functional annotations showed a contrast between sample types at most defined functional categories.

Future studies in sponges should take into account the conclusions surfaced by this work, that need to be further analysed and be submitted to further testing.

CHAPTER 5. REFERENCES

Alex, A., & Antunes, A. (2019). Comparative Genomics Reveals Metabolic Specificity of Endozoicomonas Isolated from a Marine Sponge and the Genomic Repertoire for Host-Bacteria Symbioses. *Microorganisms*, *7*(12), 635. https://doi.org/10.3390/microorganisms7120635

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bar-On, Y. M., & Milo, R. (2019). The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell*, *179*(7), 1451–1454. https://doi.org/10.1016/j.cell.2019.11.018

Bauvais, C., Zirah, S., Piette, L., Chaspoul, F., Domart-Coulon, I., Chapon, V., Gallice, P., Rebuffat, S., Pérez, T., & Bourguet-Kondracki, M.-L. (2015). Sponging up metals: Bacteria associated with the marine sponge Spongia officinalis. *Marine Environmental Research*, *104*, 20–30. https://doi.org/10.1016/j.marenvres.2014.12.005

Berthet, B., Mouneyrac, C., Pérez, T., & Amiard-Triquet, C. (2005). Metallothionein concentration in sponges (Spongia officinalis) as a biomarker of metal contamination. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, *141*(3), 306–313. https://doi.org/10.1016/j.cca.2005.07.008

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonnain, C., Breitbart, M., & Buck, K. (2016). The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean. *Frontiers in Marine Science*, *3*. https://doi.org/10.3389/fmars.2016.00082

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

Bryan, M. J., Burroughs, N. J., Spence, E. M., Clokie, M. R. J., Mann, N. H., & Bryan, S. J. (2008). Evidence for the Intense Exchange of MazG in Marine Cyanophages by Horizontal Gene Transfer. *PLOS ONE*, *3*(4), e2048. https://doi.org/10.1371/journal.pone.0002048

Burgsdorf, I., Slaby, B. M., Handley, K. M., Haber, M., Blom, J., Marshall, C. W., Gilbert, J. A., Hentschel, U., & Steindler, L. (2015). Lifestyle Evolution in Cyanobacterial Symbionts of Sponges. *MBio*, *6*(3). https://doi.org/10.1128/mBio.00391-15

Cárdenas, A., Ye, J., Ziegler, M., Payet, J. P., McMinds, R., Vega Thurber, R., & Voolstra, C. R. (2020). Coral-Associated Viral Assemblages From the Central Red Sea Align With Host Species and Contribute to Holobiont Genetic Diversity. *Frontiers in Microbiology*, *11*. https://www.frontiersin.org/articles/10.3389/fmicb.2020.572534

Colson, P., Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D., Cheng, X.-W., Federici, B., Van Etten, J., Koonin, E., la Scola, B., & Raoult, D. (2013). "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archives of Virology*, *158*. https://doi.org/10.1007/s00705-013-1768-6

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

Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F., & Martiny, J. B. H. (2016). The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology*, *499*, 219–229. https://doi.org/10.1016/j.virol.2016.09.016

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Graninger, M., Nidetzky, B., Heinrichs, D. E., Whitfield, C., & Messner, P. (1999). Characterization of dTDP-4-dehydrorhamnose 3,5-Epimerase and dTDP-4-dehydrorhamnose Reductase, Required for dTDP-l-rhamnose Biosynthesis in Salmonella enterica Serovar Typhimurium LT2. *Journal of Biological Chemistry*, *274*(35), 25069–25077. https://doi.org/10.1074/jbc.274.35.25069

Guan, L. L., Kanoh, K., & Kamino, K. (2001). Effect of Exogenous Siderophores on Iron Uptake Activity of Marine Bacteria under Iron-Limited Conditions. *Applied and Environmental Microbiology*, *67*(4), 1710–1717. https://doi.org/10.1128/AEM.67.4.1710-1717.2001

Haber, M., Burgsdorf, I., Handley, K. M., Rubin-Blum, M., & Steindler, L. (2021). Genomic Insights Into the Lifestyles of Thaumarchaeota Inside Sponges. *Frontiers in Microbiology*, *11*, 622824. https://doi.org/10.3389/fmicb.2020.622824

Hardoim, C. C. P., Esteves, A. I. S., Pires, F. R., Gonçalves, J. M. S., Cox, C. J., Xavier, J. R., & Costa, R. (2012). Phylogenetically and Spatially Close Marine Sponges Harbour Divergent Bacterial Communities. *PLoS ONE*, *7*(12), e53029. https://doi.org/10.1371/journal.pone.0053029

Hobbs, Z., & Abedon, S. T. (2016). Diversity of phage infection types and associated terminology: The problem with 'Lytic or lysogenic.' *FEMS Microbiology Letters*, *363*(7), fnw047. https://doi.org/10.1093/femsle/fnw047

Huber, K. E., & Waldor, M. K. (2002). Filamentous phage integration requires the host recombinases XerC and XerD. *Nature*, *417*(6889), Article 6889. https://doi.org/10.1038/nature00782

Hulo, C., Masson, P., Toussaint, A., Osumi-Sutherland, D., de Castro, E., Auchincloss, A. H., Poux, S., Bougueleret, L., Xenarios, I., & Le Mercier, P. (2017). Bacterial Virus Ontology; Coordinating across Databases. *Viruses*, *9*(6), 126. https://doi.org/10.3390/v9060126

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

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

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

Hutinet, G., Kot, W., Cui, L., Hillebrand, R., Balamkundu, S., Gnanakalai, S., Neelakandan, R., Carstens, A. B., Fa Lui, C., Tremblay, D., Jacobs-Sera, D., Sassanfar, M., Lee, Y.-J., Weigele, P., Moineau, S., Hatfull, G. F., Dedon, P. C., Hansen, L. H., & de Crécy-Lagard, V. (2019). 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-13384-y

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

Ignacio-Espinoza, J. C., Ahlgren, N. A., & Fuhrman, J. A. (2020). Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nature Microbiology*, *5*(2), Article 2. https://doi.org/10.1038/s41564-019-0628-x

Iyer, L. M., Abhiman, S., Burroughs, A. M., & Aravind, L. (2009). Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: Synthesis of novel

metabolites and peptide modifications of proteins. *Molecular Biosystems*, *5*(12), 1636–1660. https://doi.org/10.1039/b917682a

Jahn, M. T., Arkhipova, K., Markert, S. M., Stigloher, C., Lachnit, T., Pita, L., Kupczok, A., Ribes, M., Stengel, S. T., Rosenstiel, P., Dutilh, B. E., & Hentschel, U. (2019). A Phage Protein Aids Bacterial Symbionts in Eukaryote Immune Evasion. *Cell Host & Microbe*, *26*(4), 542-550.e5. https://doi.org/10.1016/j.chom.2019.08.019

Jahn, M. T., Lachnit, T., Markert, S. M., Stigloher, C., Pita, L., Ribes, M., Dutilh, B. E., & Hentschel, U. (2021). Lifestyle of sponge symbiont phages by host prediction and correlative microscopy. *The ISME Journal*, *15*(7), 2001–2011. https://doi.org/10.1038/s41396-021-00900-6

Jamal, E. S., Ambalavanan, L., Iehata, S., & Zainathan, S. (2021). A review on marine viruses in sponges and seawater. *AACL Bioflux*, *14*, 1828–1854.

Jiao, N., & Zheng, Q. (2011). The Microbial Carbon Pump: From Genes to Ecosystems ▿ . *Applied and Environmental Microbiology*, *77*(21), 7439–7444. https://doi.org/10.1128/AEM.05640-11

Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W., & Weitz, J. S. (2014). The elemental composition of virus particles: Implications for marine biogeochemical cycles. *Nature Reviews Microbiology*, *12*(7), Article 7. https://doi.org/10.1038/nrmicro3289

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

Karimi, E., Ramos, M., Gonçalves, J. M. S., Xavier, J. R., Reis, M. P., & Costa, R. (2017). Comparative Metagenomics Reveals the Distinctive Adaptive Features of the Spongia officinalis Endosymbiotic Consortium. *Frontiers in Microbiology*, *8*, 2499–2499. PubMed. https://doi.org/10.3389/fmicb.2017.02499

Koziol, C., Kobayashi, N., Müller, I. M., & Müller, W. E. G. (1998). Cloning of sponge heat shock proteins: Evolutionary relationships between the major kingdoms. *Journal of*

Zoological Systematics and Evolutionary Research, *36*(1–2), 101–109. https://doi.org/10.1111/j.1439-0469.1998.tb00782.x

Kvålseth, T. O. (2015). Evenness indices once again: Critical analysis of properties. *SpringerPlus*, *4*, 232. https://doi.org/10.1186/s40064-015-0944-4

Laffy, P. W., Wood-Charlson, E. M., Turaev, D., Jutz, S., Pascelli, C., Botté, E. S., Bell, S. C., Peirce, T. E., Weynberg, K. D., van Oppen, M. J. H., Rattei, T., & Webster, N. S. (2018). Reef invertebrate viromics: Diversity, host specificity and functional capacity. *Environmental Microbiology*, *20*(6), 2125–2141. https://doi.org/10.1111/1462-2920.14110

Lauderdale, J. M., Braakman, R., Forget, G., Dutkiewicz, S., & Follows, M. J. (2020). Microbial feedbacks optimize ocean iron availability. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4842–4849. https://doi.org/10.1073/pnas.1917277117

Loc-Carrillo, C., & Abedon, S. T. (2011). Pros and cons of phage therapy. *Bacteriophage*, *1*(2), 111–114. https://doi.org/10.4161/bact.1.2.14590

Maldonado, M., Aguilar, R., Bannister, R., Bell, J., Conwa, K. W., Dayton, P., Diaz, M., Gutt, J., Kelly, M., Kenchington, E., Leys, S., Pomponi, S., Rapp, H., Rutzler, K., Tendal, O., Vacelet, J., & Young, C. (2016). *Sponge Grounds as Key Marine Habitats: A Synthetic Review of Types, Structure, Functional Roles, and Conservation Concerns* (p. 39). https://doi.org/10.1007/978-3-319-17001-5_24-1

Marantos, A., Mitarai, N., & Sneppen, K. (2022). From kill the winner to eliminate the winner in open phage-bacteria systems. *PLOS Computational Biology*, *18*(8), e1010400. https://doi.org/10.1371/journal.pcbi.1010400

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*, 17112. https://doi.org/10.1038/nmicrobiol.2017.112

Messyasz, A., Rosales, S. M., Mueller, R. S., Sawyer, T., Correa, A. M. S., Thurber, A. R., & Vega Thurber, R. (2020). Coral Bleaching Phenotypes Associated With Differential

Abundances of Nucleocytoplasmic Large DNA Viruses. *Frontiers in Marine Science*, *7*. https://www.frontiersin.org/articles/10.3389/fmars.2020.555474

Mietzsch, M., & Agbandje-McKenna, M. (2017). The Good That Viruses Do. *Annual Review of Virology*, *4*(1), iii–v. https://doi.org/10.1146/annurev-vi-04-071217-100011

Nguyen, M., Wemheuer, B., Laffy, P. W., Webster, N. S., & Thomas, T. (2021). Taxonomic, functional and expression analysis of viral communities associated with marine sponges. *PeerJ*, *9*, e10715. https://doi.org/10.7717/peerj.10715

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

Okpiliya, F. I. (2012). Ecological Diversity Indices: Any Hope for One Again? *Journal of Environment and Earth Science*, *2*, 45–52.

Parakkottil Chothi, M., Duncan, G. A., Armirotti, A., Abergel, C., Gurnon, J. R., Van Etten, J. L., Bernardi, C., Damonte, G., & Tonetti, M. (2010). Identification of an l-Rhamnose Synthetic Pathway in Two Nucleocytoplasmic Large DNA Viruses. *Journal of Virology*, *84*(17), 8829–8838. https://doi.org/10.1128/JVI.00770-10

Pascelli, C., Laffy, P. W., Botté, E., Kupresanin, M., Rattei, T., Lurgi, M., Ravasi, T., & Webster, N. S. (2020). Viral ecogenomics across the Porifera. *Microbiome*, *8*(1), 144. https://doi.org/10.1186/s40168-020-00919-5

Pascelli, C., Laffy, P. W., Kupresanin, M., Ravasi, T., & Webster, N. S. (2018). Morphological characterization of virus-like particles in coral reef sponges. *PeerJ*, *6*, e5625. https://doi.org/10.7717/peerj.5625

Paul, J. H. (2008). Prophages in marine bacteria: Dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal*, *2*(6), Article 6. https://doi.org/10.1038/ismej.2008.35

Perez Sepulveda, B., Redgwell, T., Rihtman, B., Pitt, F., Scanlan, D. J., & Millard, A. (2016). Marine phage genomics: The tip of the iceberg. *FEMS Microbiology Letters*, *363*(15), fnw158. https://doi.org/10.1093/femsle/fnw158

Rawlings, N. D., Barrett, A. J., & Bateman, A. (2010). MEROPS: The peptidase database. *Nucleic Acids Research*, *38*(Database issue), D227-233. https://doi.org/10.1093/nar/gkp971

Rouhier, N., Couturier, J., Johnson, M. K., & Jacquot, J.-P. (2010). Glutaredoxins: Roles in iron homeostasis. *Trends in Biochemical Sciences*, *35*(1), 43. https://doi.org/10.1016/j.tibs.2009.08.005

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *3*, e985. https://doi.org/10.7717/peerj.985

Roveta, C., Pica, D., Calcinai, B., Girolametti, F., Truzzi, C., Illuminati, S., Annibaldi, A., & Puce, S. (2020). Hg Levels in Marine Porifera of Montecristo and Giglio Islands (Tuscan Archipelago, Italy). *Applied Sciences*, *10*(12), Article 12. https://doi.org/10.3390/app10124342

Sandy, M., & Butler, A. (2009). Microbial Iron Acquisition: Marine and Terrestrial Siderophores. *Chemical Reviews*, *109*(10), 4580–4595. https://doi.org/10.1021/cr9002787

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883–8900. https://doi.org/10.1093/nar/gkaa621

Skurnik, M., & Strauch, E. (2006). Phage therapy: Facts and fiction. *International Journal of Medical Microbiology*, *296*(1), 5–14. https://doi.org/10.1016/j.ijmm.2005.09.002

Slaby, B. M., Hackl, T., Horn, H., Bayer, K., & Hentschel, U. (2017). Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *The ISME Journal*, *11*(11), 2465–2478. https://doi.org/10.1038/ismej.2017.101

Song, K. (2020). Classifying the Lifestyle of Metagenomically-Derived Phages Sequences Using Alignment-Free Methods. *Frontiers in Microbiology*, *11*. https://www.frontiersin.org/articles/10.3389/fmicb.2020.567769

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, *39*, 321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541

Strong, W. L. (2016). Biased richness and evenness relationships within Shannon–Wiener index values. *Ecological Indicators*, *67*, 703–713. https://doi.org/10.1016/j.ecolind.2016.03.043

Sun, Y., Debeljak, P., & Obernosterer, I. (2021). Microbial iron and carbon metabolism as revealed by taxonomy-specific functional diversity in the Southern Ocean. *The ISME Journal*, *15*(10), Article 10. https://doi.org/10.1038/s41396-021-00973-3

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

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & the UniProt Consortium. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739

Tam, W., Pell, L. G., Bona, D., Tsai, A., Dai, X. X., Edwards, A. M., Hendrix, R. W., Maxwell, K. L., & Davidson, A. R. (2013). Tail tip proteins related to bacteriophage λ gpL coordinate an iron-sulphur cluster. *Journal of Molecular Biology*, *425*(14), 2450–2462. https://doi.org/10.1016/j.jmb.2013.03.032

Taylor, M. W., Radax, R., Steger, D., & Wagner, M. (2007). Sponge-associated microorganisms: Evolution, ecology, and biotechnological potential. *Microbiology and Molecular Biology Reviews: MMBR*, *71*(2), 295–347. https://doi.org/10.1128/MMBR.00040-06

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J. R., Easson, C., Astudillo-García, C., Olson, J. B., Erwin, P. M., López-Legentil, S., Luter, H., Chaves-Fonnegra, A., Costa, R.,

Schupp, P. J., Steindler, L., Erpenbeck, D., Gilbert, J., Knight, R., Ackermann, G., Victor Lopez, J., … Webster, N. S. (2016). Diversity, structure and convergent evolution of the global sponge microbiome. *Nature Communications*, *7*(1). https://doi.org/10.1038/ncomms11870

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

Touchon, M., Bernheim, A., & Rocha, E. P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. *The ISME Journal*, *10*(11), 2744–2754. https://doi.org/10.1038/ismej.2016.47

Vacelet, J., Verdenal, B., & Perinet, G. (1988). The iron mineralization of Spongia officinalis L. (Porifera, Dictyoceratida) and its relationships with the collagen skeleton. *Biology of the Cell*, *62*(2), 189–198.

Webster, N. S., & Thomas, T. (2016). The Sponge Hologenome. *MBio*, *7*(2). https://doi.org/10.1128/mBio.00135-16

Wilson, W. H., Dale, A. L., Davy, J. E., & Davy, S. K. (2005). An enemy within? Observations of virus-like particles in reef corals. *Coral Reefs*, *24*(1), 145–148. https://doi.org/10.1007/s00338-004-0448-0

Winget, D. M., & Wommack, K. E. (2008). Randomly Amplified Polymorphic DNA PCR as a Tool for Assessment of Marine Viral Richness. *Applied and Environmental Microbiology*, *74*(9), 2612–2618. https://doi.org/10.1128/AEM.02829-07

Winter, C., Bouvier, T., Weinbauer, M. G., & Thingstad, T. F. (2010). Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: The "Killing the Winner" Hypothesis Revisited. *Microbiology and Molecular Biology Reviews*, *74*(1), 42–57. https://doi.org/10.1128/MMBR.00034-09

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*(W1), W95–W101. https://doi.org/10.1093/nar/gky418

Zhang, R., Li, Y., Yan, W., Wang, Y., Cai, L., Luo, T., Li, H., Weinbauer, M. G., & Jiao, N. (2020). Viral control of biomass and diversity of bacterioplankton in the deep sea. *Communications Biology*, *3*(1), 256. https://doi.org/10.1038/s42003-020-0974-5

Zhang, Y., Maley, F., Maley, G., Duncan, G., Dunigan, D., & Van Etten, J. (2007). Chloroviruses Encode a Bifunctional dCMP-dCTP Deaminase That Produces Two Key Intermediates in dTTP Formation. *Journal of Virology*, *81*, 7662–7671. https://doi.org/10.1128/JVI.00186-07

SUPPLEMENTAL MATERIAL

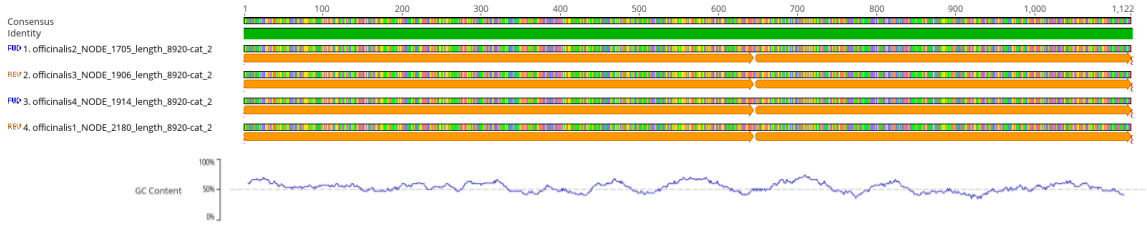


Figure S1 – Result of the alignment for the dTDP-4-dehydrorhamnose 3,5-epimerase, present in all sponge samples. The alignment returned a 100% identify between sequences. Predicted open reading frames in orange.

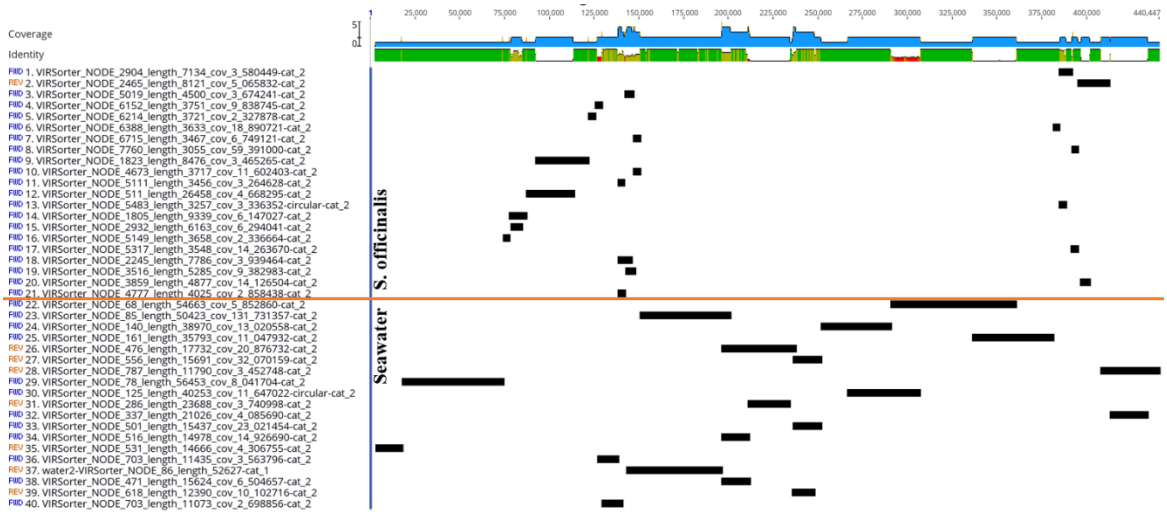


Figure S2 – Result of the alignment for the Very late expression factor 1, present in all samples. The orange line separates the sponge samples (above the line) and the seawater samples. Identity and coverage are represented by the green and blue sequences, respectively.

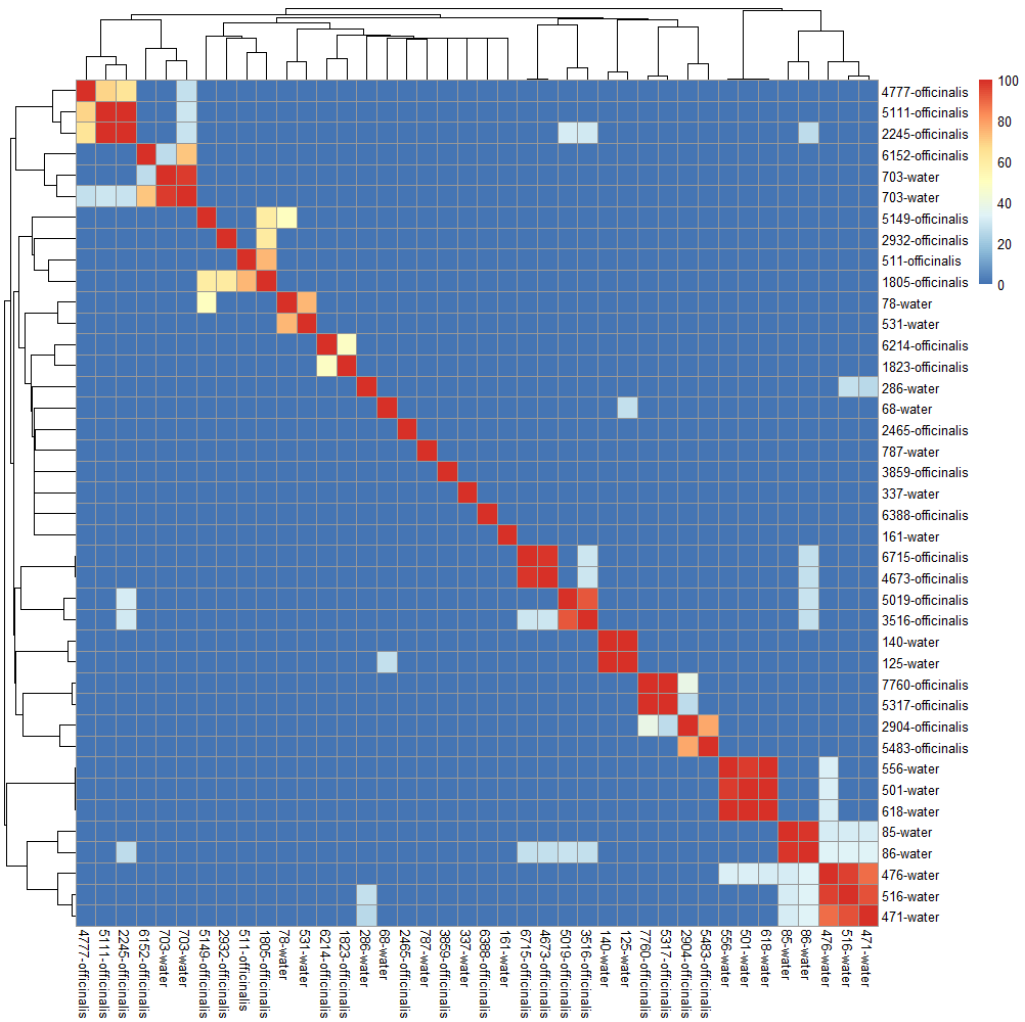


Figure S3 – Heat map of the alignment for the Very late expression factor 1, present in all samples.

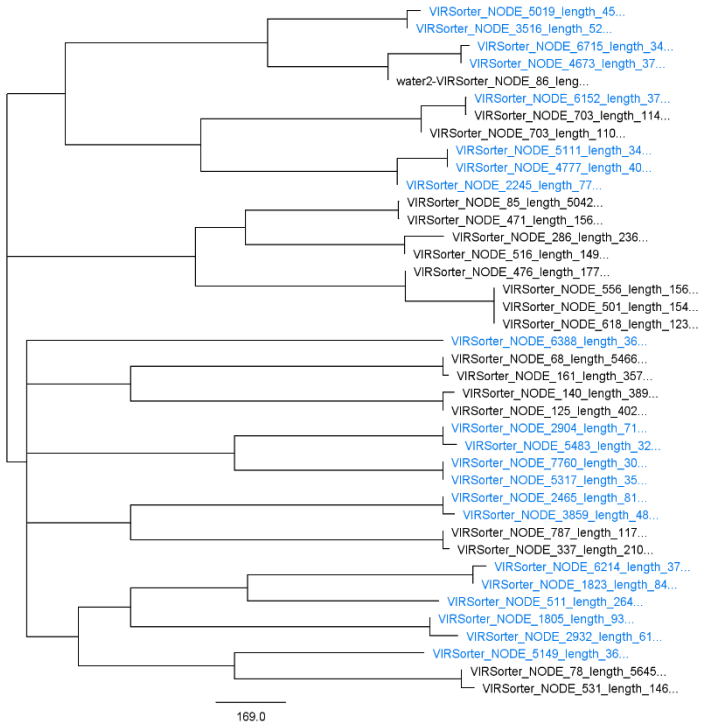


Figure S4 – Tree grouping of all the similarity between contigs for the Very late expression factor 1. The blue id represents the viral contigs from the sponge biome.


Base plate wedge protein 53
base plate wedge subunit
baseplate hub
baseplate hub assembly catalyst
Baseplate hub assembly protein gp51
baseplate hub subunit
Baseplate J-like protein
Baseplate protein gp9
Baseplate protein J
Baseplate puncturing device gp5.4
baseplate wedge
baseplate wedge initiator
baseplate wedge protein
Baseplate wedge protein gp10
Baseplate wedge protein gp25
Baseplate wedge protein gp53
Baseplate wedge protein gp6
baseplate wedge subunit
gp25 base plate wedge subunit
putative base plate wedge component
putative baseplate hub subunit
Bacteriophage head to tail connecting protein
capsid and scaffold
capsid and scaffold protein
Capsid assembly scaffolding protein
capsid decoration protein
capsid protease
capsid protein
capsid related protein
Caudovirus prohead serine protease
gp21 prohead core scaffold protein
gp4 head completion protein
head-proximal tip of tail tube tail completion + sheath stabilizer protein
head-tail adaptor protein
head-tail connector protein
Head-tail joining protein
head-to-tail joining protein
head completion
head completion protein
Head completion protein
Head completion protein gp16
Head completion protein Gp50
head morphogenesis protein
major capsid
major capsid protein
Major capsid protein
Major capsid protein Gp23
major head protein
Minor capsid protein
minor capsid protein 10
Minor head protein GP7
P22 coat protein
P22 coat protein - gene protein 5
P22 coat protein 5
Phage capsid family
Phage gp6-like head-tail connector protein
Phage Head-Tail Attachment
Phage head-tail joining protein
phage head maturation protease
Phage major capsid protein E
Phage T7 capsid assembly protein
portal vertex protein of head
precursor of major head subunit
Probable capsid maturation protease
Probable major capsid protein gp17
prohead assembly scaffold protein
prohead core protein
Prohead core protein protease
Prohead core protein serine protease
prohead core scaffold and protease
prohead peptidase
prohead peptidase gp21
prohead protease
putative coat protein
putative head-to-tail joining protein
putative major capsid protein
putative minor capsid protein
putative T7-like capsid assembly protein
scaffold prohead core protein
similar to major head protein
similar to portal vertex protein of head
Bacillus subtilis phage GA-1 neck appendage CIMCD self-cleaving protein
neck protein
Neck protein gp13
Neck protein gp14
Virus neck protein
6
5
4
3
2
1
0
Categories
Baseplate protein
Capsid protein
Neck protein
Tail protein

Figure S5 – Heat map for the structure functional category.

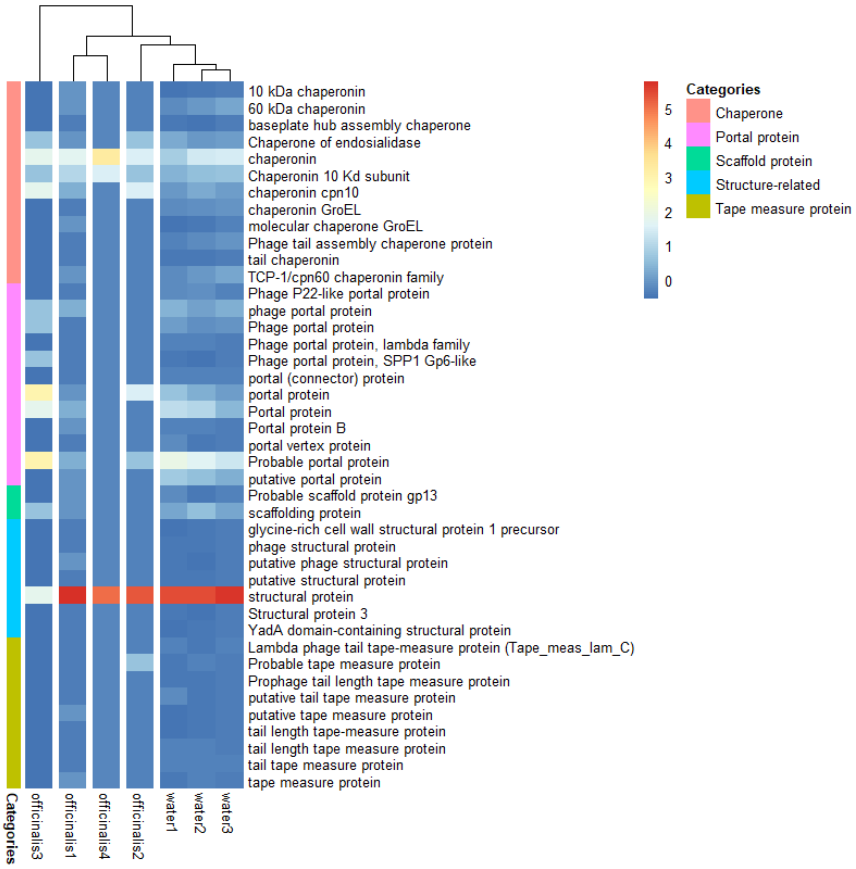


Figure S6 – Heat map for the structure functional category. Related with protein structure and transport.

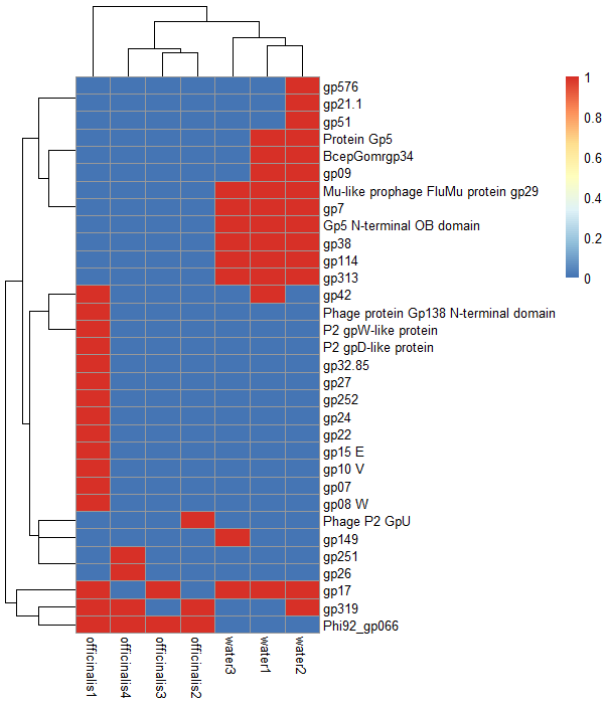
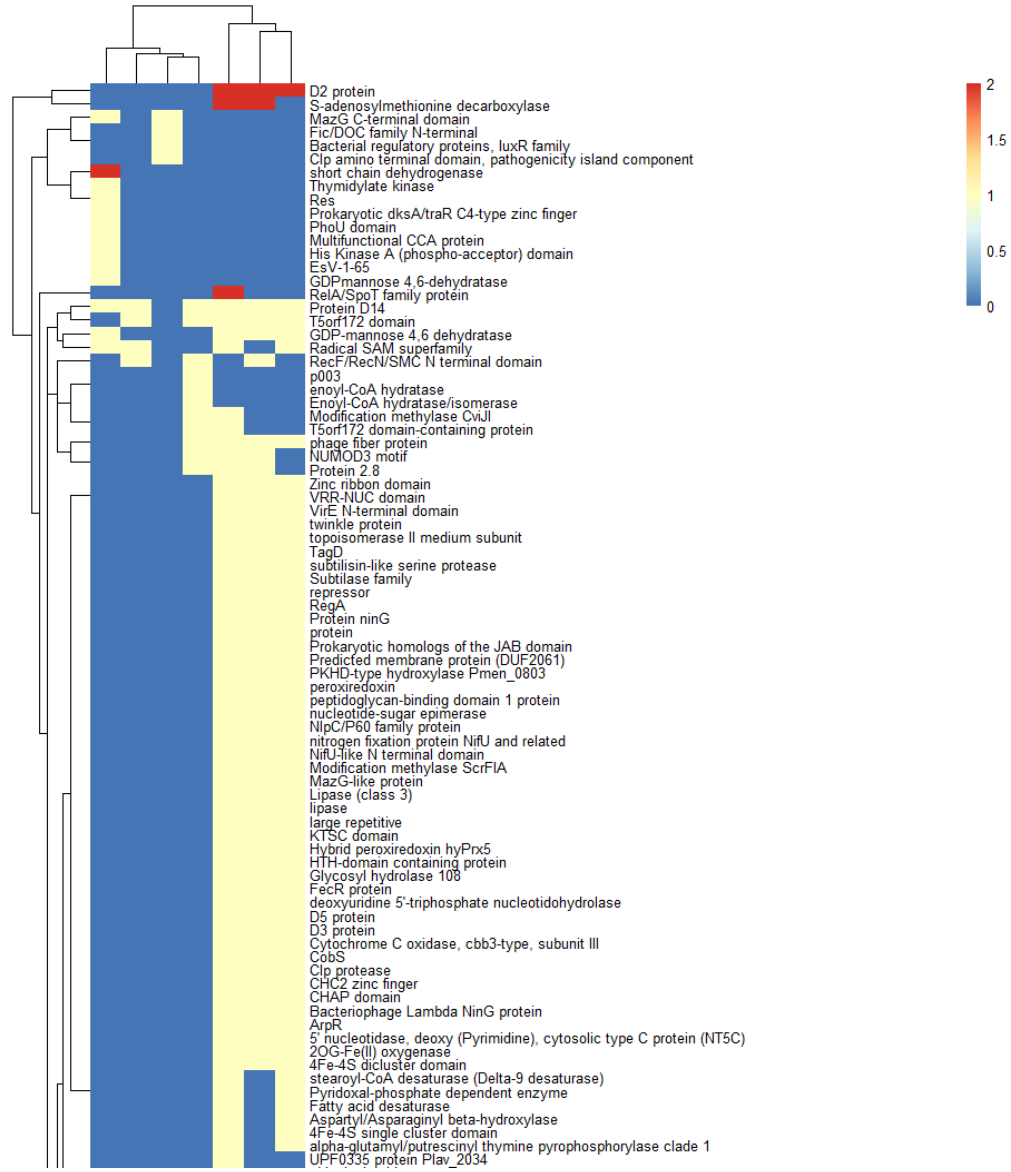


Figure S7 – Heat map for the gene product



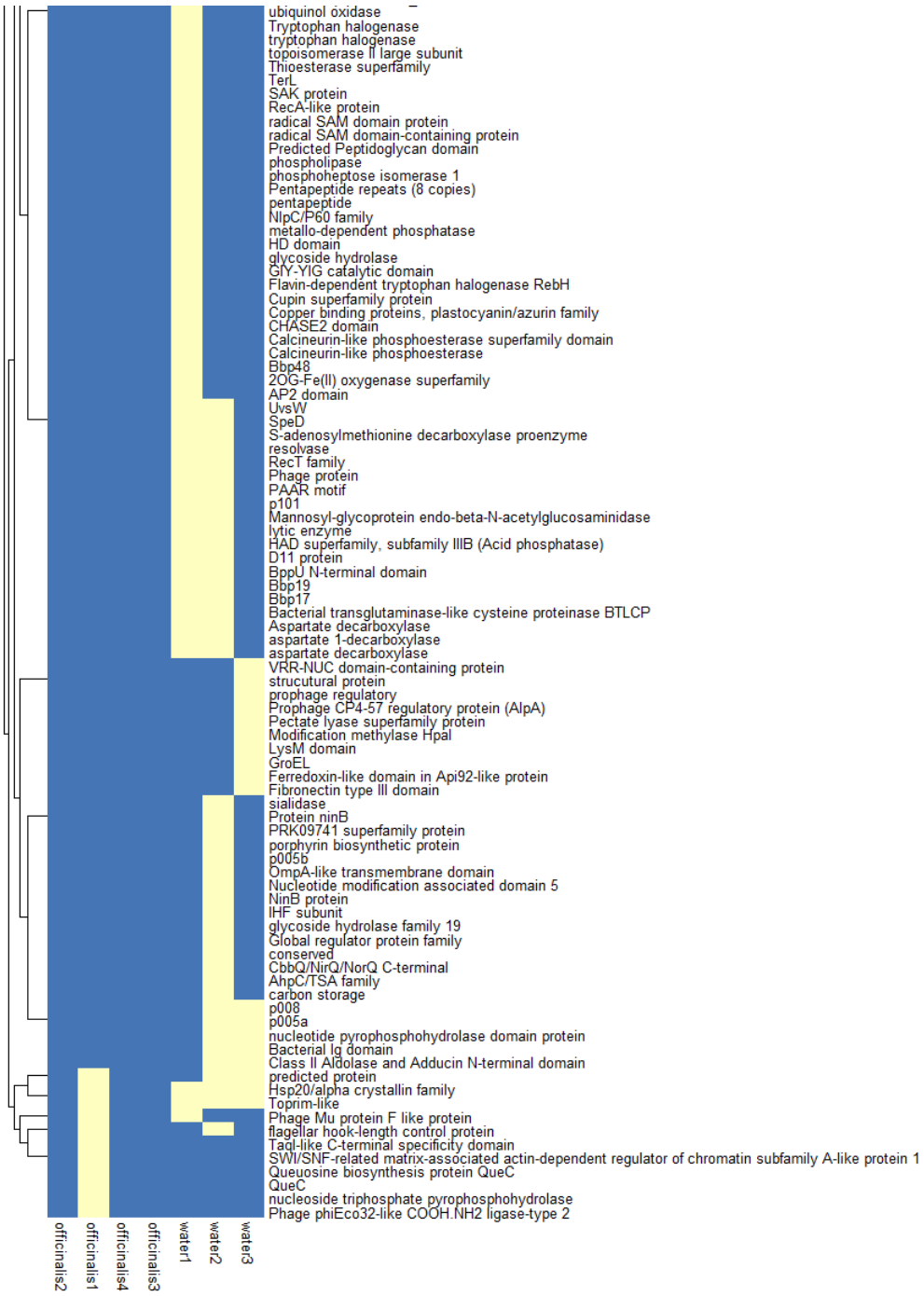


Figure S8 – Heat map for the predictions that didn´t fall into any of the custom functional categories.