VIRUSES IN THE NORTH SEA:

VIROMICS AND PROPHAGE GENOMICS

by

Alexa Garin-Fernandez

a Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in Marine Microbiology

Thesis Committee:

Prof. Dr. Frank Oliver GLÖCKNER (chair)
Alfred Wegener Institute for Marine and Polar Research
Jacobs University Bremen

Dr. Antje WICHELS
Alfred Wegener Institute for Marine and Polar Research

Prof. Dr. Matthias ULLRICH
Jacobs University Bremen

Dr. Gunnar GERDTS
Alfred Wegener Institute for Marine and Polar Research

Date of Defense: October 30, 2019

Department of Life Sciences and Chemistry, Jacobs University Bremen

STATUTORY DECLARATION

Family Name, Given/First Name	Garin Fernandez, Alexa
Matriculationnumber	20331313
What kind of thesis are you submitting: Bachelor-, Master- or PhD-Thesis	PhD Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

……

Date, Signature

“*What is essential is invisible to the eye”*

Le Petit Prince (1943)

THESIS ABSTRACT

Viruses are the most abundant biological entity on Earth, dominating the marine environment. Despite their small size, viruses have an enormous influence on microbial population dynamics, due to lysis and horizontal gene transfer. Due the high abundance of their hosts, bacteriophages or phages comprise the majority of viruses and also provide the largest reservoirs of unexplored genetic diversity in marine environments. The rise of Next Generation Sequencing (NGS) techniques brings new opportunities to investigate the marine virus community. However, there is no current statutory pipeline applied in marine phage ecology. Therefore, this thesis proposes a virus-specific pipeline based on the integration of existing tools and state of the art techniques. The developed pipeline was applied to accomplish the two research aims of this thesis: (1) to analyze the virus community in the North Sea with viromics, and (2) to characterize lysogenic phages from potentially pathogenic *Vibrio* species.

For the first part of this thesis, the virus community of four sampling stations were described using virus metagenomics (viromics). The results show that the virus community is dominated by phages and they are not evenly distributed throughout the North Sea. In general, the coastal virus community was genetically more diverse than the open sea community. The influence of riverine inflow and currents, for instance the English Channel flow affects the genetic virus diversity with the community carrying genes from a variety of metabolic pathways and other functions. These results offer the first insights in the virus community in the North Sea using viromics and shows the variation in virus diversity and the genetic information moved from coastal to open sea areas. Although phages of emerging *Vibrio* species were identified in low percentage in the North Sea virome, is possible they are lysogenic phages integrated into the host genome.

The seawater temperature rise promotes the growth of potentially human pathogenic *Vibrio* species. In the North Sea, *V. parahaemolyticus* and non-O1/O139 *V. cholerae* have been isolated and characterized. These strains contain prophages that may contribute to the emergence of pathogenic strains in the marine environment (40 % of tested isolates). The genome structure and possible biological functions of four lysogenic phages were described in the second part of this thesis. The phages from *V. parahaemolyticus* (two tailed phages, one filamentous phage) and *V. cholerae* (one tailed phages) can integrate into their host genome and might have a role in pathogenicity. This study provides new insights with respect to the presence of lysogenic phages in environmental *Vibrio* strains, which might have a role in the emergence of new pathogenic strains in the North Sea.

This thesis represents an exemplary study of the virus community in the North Sea, with special emphasis on the marine phages. The settled virus-specific pipeline as well as the obtained insights will contribute to extend the study of the virus diversity dynamics in other marine areas, including to estimate the abundance and functional diversity of novel phages.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1. MARINE VIRUSES: PART OF THE MICRO-UNIVERSE

Viruses are the most abundant biological entity on Earth (Suttle, 2005). In the ocean, their concentration is estimated of ~ 10^7 virus-like particles (VLP) per milliliter of surface seawater and their abundance is over 10^{30} viral particles (Wommack and Colwell, 2000, Suttle, 2005). In other words, if one bottle of wine (~ 750 ml) is filled with seawater, it would contain as many viruses as humans in the whole world. Despite their small size (20 – 200 nm), viruses represent 94 % of all the nucleic acid-containing particles in the ocean and compose the second largest biomass, exceeded only by prokaryotes (Suttle, 2005, Fuhrman, 1999).

Marine viruses have been identified in every investigated habitat so far. In general, viruses are not evenly distributed in the marine environment, and their abundance is highest in euphotic areas and then decreases with depth (Breitbart, 2012, Fuhrman, 1999). Furthermore, viral abundance is higher in coastal areas than offshores (Marchant et al., 2000). The dynamics of the virus communities is closely connected with the environmental parameters and host dynamics variations (Breitbart, 2012). Phages have an important role in marine ecosystems, due to horizontal gene transfer and release of dissolve organic matter (DOM) to the environment by host lysis (Breitbart et al., 2018). All cellular life forms are susceptible to viral infection, which applied both either to eukaryotes and to prokaryotes organisms (Fuhrman, 1999). Since bacteria are the most abundant cellular organisms, it is proposed that the most abundant infective viruses are bacteriophages or phages (Breitbart et al., 2018).

Phage particles consist of a single- or double-stranded DNA or RNA surrounded by a protein –in some also by a lipid– coat (Sime-Ngando, 2014). Phage genomes possess a mosaicism organization, composed in modules of structural cores with related function (Berard et al., 2016). This structure is the result of non-homologous recombination during the evolution of viruses in different taxa (King et al., 2012). Most of the described marine phages belong to the Caudovirales order, also called tailed phages (Ackermann, 2007). This order is divided in four families based on the tail morphology: Podoviridae (short noncontractile tail), Siphoviridae (long noncontractile tail), Myoviridae (long contractile tail), Ackermannviridae (long contractile tail, formerly known as the *Vi1virus* genus), and Herelleviridae

(long contractile tail, created after Spounavirinae reclassification) (figure 1.1A) (Krupovic et al., 2011, Adriaenssens et al., 2018).

Although most of the studied marine viruses are tailed phages, probably they are not the most abundant member in the environment (Ignacio-Espinosa and Fuhrman, 2018, Brum et al., 2013). New approaches applied to virus ecology bring new insights about virus groups that traditional methods may have overlooked in the marine environment. Examples of these groups are Mimiviruses or giant viruses (Fischer et al., 2010), virophages (Zhou et al., 2013), non-tailed phages from Autolykiviridae family (Kauffman et al., 2018b), RNA phages (Steward et al., 2012), and ssDNA phages as Inoviridae or filamentous phages (Tucker et al., 2011). In general, Inoviridae phages have a small genome size (~ 9 Kbps) with a filamentous virion structure (figure 1.1B). Members of the family Inoviridae present a chronic infection cycle that results in phage integration into the bacterial genome as a prophage (Day, 2011). In contrast to tailed phages, this process does not lyse the cell as lysogenic phages (section 1.2.2); instead, the infected cell can constantly release phage progeny in large amounts (Krupovic et al., 2011). Inoviridae prophages can alter host phenotype to express e.g. virulence traits. These filamentous phages contribute to bacterial virulence and therefore increase the bacterial survival in animal and plant hosts (Rakonjac, 2012). For instance, the filamentous phage CTXφ encodes major virulence factor (cholera toxin, CtxAB) in human pathogenic V. cholerae (McLeod et al., 2005, Waldor and Mekalanos, 1996). Inoviridae sequences have been identified in almost half of Vibrio genomes (45%, Castillo et al., 2018). Moreover, several environmental and harmless Vibrio genomes actually contain virulence trains from filamentous phages (Castillo et al., 2018). Despite the high prevalence of filamentous phages in several bacterial genomes, the presence of Inoviridae in metagenomic datasets have been overlooked (Roux et al., 2019). New approaches have been applied to identify Inoviridae sequences in marine metagenomes, showing a global prevalence in marine environments and high genomic diversity (Roux et al., 2019). Genome structure of Inoviridae phages and other characteristics are detailed in chapter 4.

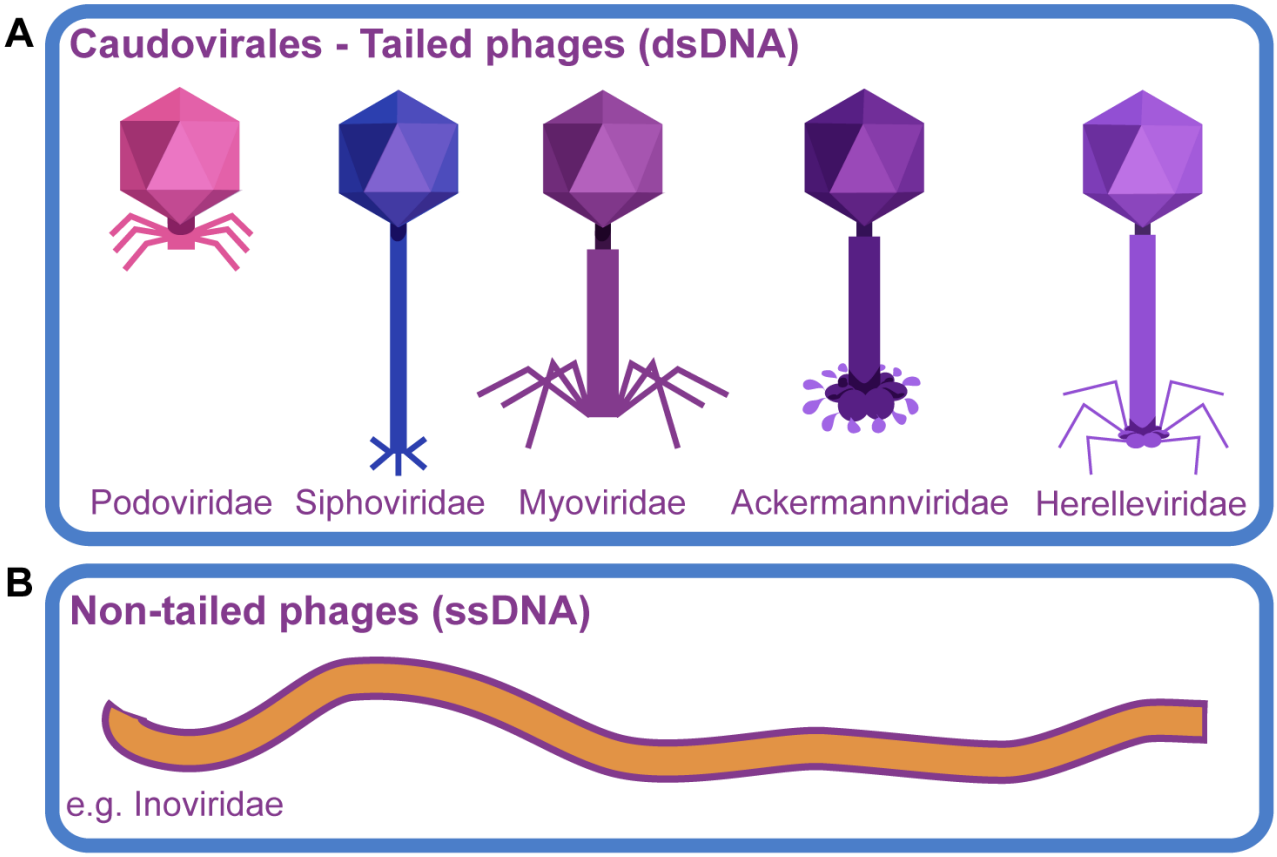


FIGURE 1. 1. Phage morphotypes. Caudovirales (tailed phages), including its taxonomic families (A). Example of non-tailed phage families (B). Modified and updated from (Clokie and Kropinski, 2009).

1.1.1. MARINE PHAGES AND OTHER MICROBIAL INHABITANTS FROM THE NORTH SEA

The North Sea is a semi-enclosed basin part of the northwest European shelf (Otto et al., 1990). This region is strongly influenced by the riverine inflow, as well the Atlantic ocean current from both the northern current inflow and the English Channel (Otto et al., 1990, Mork et al., 1981, Sündermann and Pohlmann, 2011). The North Sea is characterized by a complex connection of oceanographic forces, physicochemical influences and human activities (Eisma et al., 1987). The average surface water temperature varies from 6 °C (winter) to 17 °C (summer), while the salinity increase from 31to 35 PSU (Eisma et al., 1987, Otto et al., 1990). The water mass exchange cause hydrographical variations, that affects directly the microbial community composition (Leterme et al., 2008, Rink et al., 2011). The differences in nutrient concentrations and salinity of the current inflows result in distinct microbial clusters throughout the North Sea region (Brandsma et al., 2013)

The natural variability of the North Sea has been monitored during long-term programs, including the dynamics of the microbial community (Wiltshire et al., 2010, Southward et al., 2004). Previous studies proposed a difference of the microbial planktonic community composition between the northern and southern North Sea (Handling editor: Ruth et al., 2018). The target organisms included dinoflagellates, ciliates, microalgae, bacteria, among others (Bils et al., 2019, Lucas et al., 2016, McQuatters-Gollop et al., 2009, Wiltshire et al., 2010, Brandsma et al., 2013). Most of the virus ecology studies in this region are focused on culture- and microscopy-based approaches (Duhaime et al., 2016, Duhaime et al., 2011a, Gerdts et al., 2004, Wichels et al., 1998, Brandsma et al., 2013). For instance, Moebus (1997b) developed key approaches to isolate and characterize marine phages from the North Sea region. Based on the enrichment method, marine phage–host systems (PHSs) of lytic and pseudolysogenic phages from the Caudovirales order were isolated (Moebus, 1997a, Moebus, 1997b). These studies show the spatio-temporal variability of marine phages based on plaque-forming units (PFU) and/or phage–host cross-reaction (PHCR) (Moebus, 1991, Moebus, 1992). The pioneer techniques developed by Moebus were the basis for future research of marine phages (described in section 1.4.1). Furthermore, the occurrence and distribution of *Pseudoalteromonas* phages isolated from the North Sea have characterized in greater detail (Duhaime et al., 2017, Duhaime et al., 2016, Duhaime et al., 2011a, Wichels et al., 1998). All these phages are lytic, and they belong to the Caudovirales order (Wichels et al., 1998, Wichels et al., 2002). Moreover, the occurrence of *Pseudoalteromonas* phages was restricted to a narrow geographical region of the German Bight, where includes the source area of the bacterial host isolate (Wichels et al., 2002). Therefore, Wichels et al. (2002) propose a geographical distribution of *Pseudoalteromonas* phages from the North Sea. In addition, Brandsma et al. (2013) describe similar patterns of the virus (as VLP/ml) and phytoplankton abundances, which are influenced by the characteristics of the different hydrographic regions. Despite the biogeographical patterns of the virus relative abundance, subpopulation differences were not observed with flow cytometry analyses (Brandsma et al., 2013).

In general, the virus community structure can follow the trends of its host community (Brum et al., 2015, Sunagawa et al., 2015), which might suggest that the virus community composition in the North Sea would follow a similar distribution pattern differentiated between the coastal and the open sea regions.

1.2. THE HOST HACKERS VS THE TIME BOMBS: PROLIFERATION CYCLE OF TAILED PHAGES

A virus depends upon its host in order to reproduce, which makes a host organism a crucial part of the virus environment (Abedon, 2008). This section introduces the main proliferation cycles of Caudovirales, known as tailed phages. The tailed phages were fundamental for early virus descriptions (Ackermann, 2003), used as the basis for the first guidelines in taxonomy classification.

In general, Caudovirales are characterized to inject their dsDNA into the host bacteria upon infection to follow the lytic or lysogenic cycle (figure 1.2) (Krupovic et al., 2011). Both cycles entail different effect on the fitness and development of the host cell population, as well as the effect on ecological processes in the marine environment.

1.2.1. LYTIC CYCLE, THE HOST HACKERS

The lytic cycle is a viral replication process highly controlled, where the dsDNA phage redirects the host metabolism to the production and release of new phages (Weinbauer, 2004). This cycle follows mainly the steps: attachment, expression and replication of the nucleic acid, assembly of new phages particles and host cell lysis to release of new phages (figure 1.2) (Goyal et al., 1987, Bondy-Denomy and Davidson, 2014).

After the phage is attached irreversibly with the host membrane receptor, the phage genome is injected into the host cytoplasm (Rohwer et al., 2014, Kutter and Sulakvelidze, 2004). Inside the host cell, the phage genome is in general circularized (e.g. by linear ends repetitions, sticky ends or terminal redundancies) to protect itself from exonucleases and restriction enzymes (Kutter and Sulakvelidze, 2004). If the genome is not degraded, the host metabolism is redirected to phage genome replication and toward the production of virions (Weinbauer, 2004). In tailed phage genomes, this process is regulated by genes organized in structural cores (Krupovic et al., 2011). In detail, the *immediate early genes* are transcribed first for protection of the phage genome, follow by *middle genes* to replicate the new phage DNA, and the *late genes* that encode virion assembly proteins (Kutter and Sulakvelidze, 2004). Finally, the mature virions accumulate into the cytoplasm, leading to lysis and release of progeny phage particles (Weinbauer, 2004, Breitbart et al., 2018, Rohwer et al., 2014) (figure 1.2).

1.2.2. LYSOGENIC CYCLE, THE TIME BOMBS

During the lysogenic cycle, the temperate phage integrates into the host genome (prophage) and replicates along with the host (lysogen) (Breitbart, 2012). In this period, the lysogen can survive and proliferate almost normally (Bondy-Denomy and Davidson, 2014, Erez et al., 2017). Many temperate phages have their own specific integration sites, such as the attachment (*att*) sites for insertion into the lysogen chromosome of phage lambda (King et al., 2012). Others, such as phage Mu, uses integrases to integrate randomly into the lysogen genome (Kutter and Sulakvelidze, 2004, Kamp, 1987). As a prophage, the phage genes are repressed until an environmental factor triggers the prophage to become lytic, called prophage induction (figure 1.2) (Breitbart, 2012).

In general, lysogeny presents a survival advantage for the phage during unfavorable conditions. The prophages are protected from stress factors like UV radiation, proteolytic digestion, and periods of low host abundance (Breitbart, 2012, Fuhrman, 1999, Paul, 2008). On the other hand, the lysogen becomes immune to further infection by closely related phages, and increase its fitness due to the acquisitions of new functional genes or suppression of unnecessary metabolic functions (Erez et al., 2017, Fuhrman, 1999, Paul, 2008). The decision between lysis-lysogeny may be driven by several factors, such as the relative phage and host densities (Dou et al., 2018, Jiang and Paul, 1998), or the release of a phage peptide to regulate the process similar to quorum-sensing (Erez et al., 2017). Other indirect factors are temperature, oligotrophy and/or anoxic level; all of them affect the distribution and occurrence of potential hosts in marine environments (Paul, 2008, Mcdaniel et al., 2006, Weinbauer et al., 2003, Williamson et al., 2002).

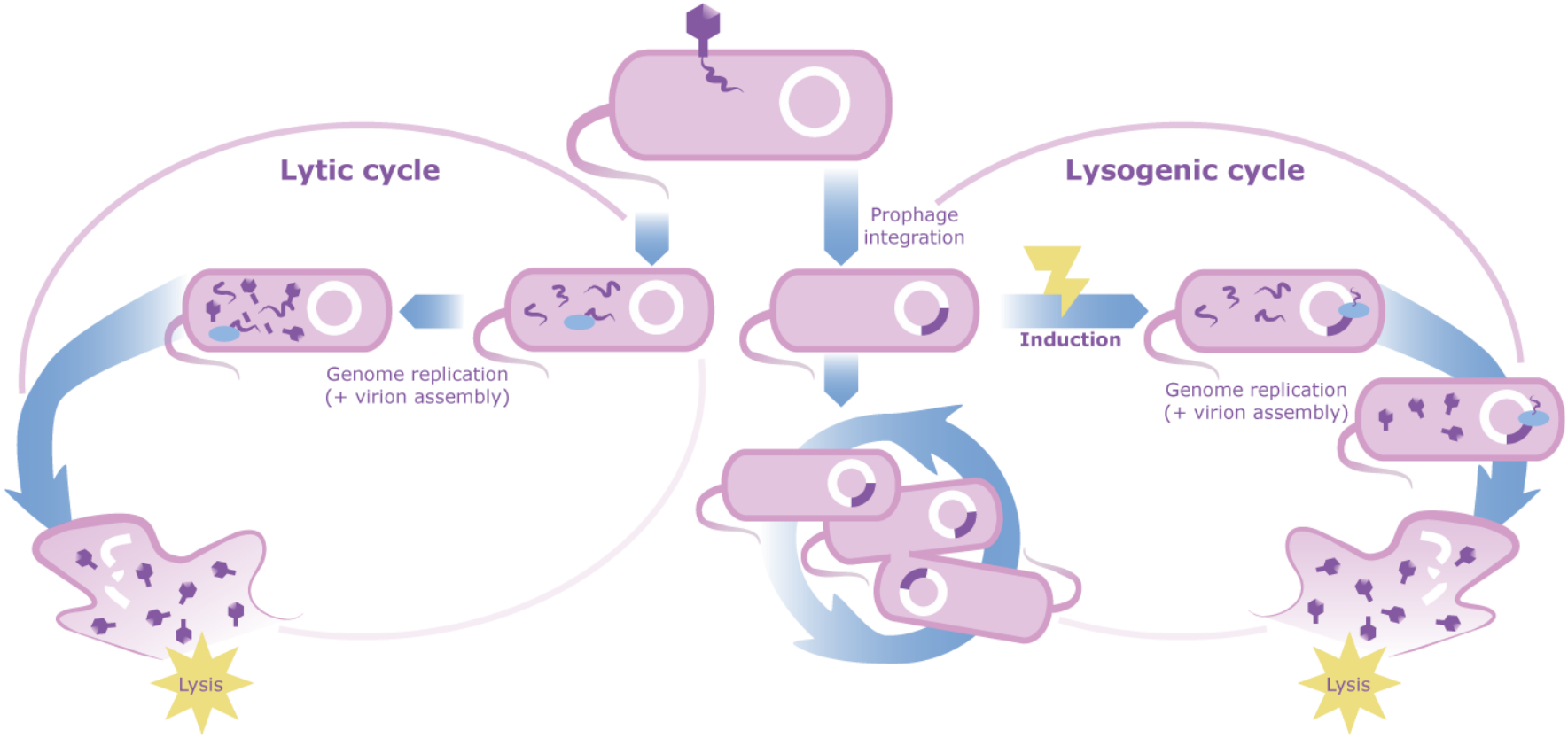


FIGURE 1. 2. Main proliferation cycles in tailed phages: lytic (*left*) and lysogenic cycle (*right*). Based on (Weinbauer, 2004).

1.2.3. THE *VIRAL* EFFECT:

ROLE ON THE ECOSYSTEM AND POPULATION DEVELOPMENT

Despite the small size, viruses have a tremendous role in ecological processes. In surface marine waters, phages are responsible for approximately half of the bacterial death per day (Breitbart et al., 2018, Breitbart and Rohwer, 2005). In contrast to grazers, the DOM is released and remineralized by bacteria within the microbial loop (Fuhrman, 1999). By this pathway named viral shunt, nearly 10 billion tons of dissolved carbon are released each day to the ocean (Wilhelm and Suttle, 1999). Moreover, up to a quarter of photosynthetically fixed carbon comes from the DOM released by viral lysis (Wilhelm and Suttle, 1999). Other nutrients, such as nitrogen, phosphorus and trace elements, frequently limit primary production are also released by phage lysis (Fuhrman, 1999). On the other hand, recent studies also have shown that viruses contribute to shuttle organic carbon from the surface to the deep ocean by the production of sticky lysate aggregations (Weinbauer, 2004, Laber et al., 2018). This process named viral shuttle, contributes to enhance the efficiency of the biological pump (Sullivan et al., 2016, Breitbart et al., 2018).

Phages can modify the community composition by selective infection (Maslov and Sneppen, 2017). According to kill-the-winner (KrW) dynamics, when a bacteria population becomes dominant, the abundance of its infectious phage will increase (Thingstad, 2000, Thingstad and Lignell, 1997). This mechanism promotes diversity in the population, and prevents the best bacterial competitors to take over the community (Thingstad and Lignell, 1997). Other models that include lysogenic phages have been proposed (Knowles et al., 2016). In the piggyback-the-winner (PtW) dynamics, the infective phages switch to lysogeny when the host population becomes dominant (Knowles et al., 2016, Weynberg, 2018) This mechanism avoids the lysogenic phage to compete for host cells or evade immune response (Weynberg, 2018). The PtW dynamic contrasts to classical models similar to KtW proposed to lysogenic phages. These models proposed that lysogeny is a survival strategy at low host abundance, whereas the lytic cycle is triggered at high host abundance (Weinbauer et al., 2003). In any of these cases, the phage infection yields a decrease in the population of the original dominant bacteria, opening a new niche. In a dynamic ecosystem, the newly dominant bacteria population can be infected by the progeny from the previous dominant phage, or by a different phage population that was present in low abundance (Breitbart et al., 2018).

Another mechanism involved in population development is horizontal gene transfer (HGT). In general, about 3 to 10 % of bacterial genomes have a phage origin (Brussow et al., 2004). Moreover, the prophage contribution to the host genome inventory can increase up to 20 % (Casjens, 2003, Fernández et al., 2018). Lytic phages can release complete host plasmids to the environment after lyse their host (Fernández et al., 2018, Keen et al., 2017). Also, temperate phages can transfer a gene cluster with an entire function to their host through HGT (Weinbauer, 2004). Among these functions are photosynthetic genes transferred by phages between cyanobacteria (Hurwitz et al., 2013b, Zeidner et al., 2005), antibiotic resistance genes (Lekunberri et al., 2017), spore formation (Fortier and Sekulovic, 2013), virulence factors and toxin production (Bondy-Denomy and Davidson, 2014), biofilm formation (Fortier and Sekulovic, 2013), and 16S rRNA genes transferred between different bacteria genera (Harrington et al., 2012, Beumer and Robinson, 2005).

The acquisition of new functional genes promotes adaptation and increases the diversity of the bacterial population (Weinbauer and Rassoulzadegan, 2004). In consequence, these genes might expand the genetic adaptability opening new ecological niches and increase the geographical distribution of species (Weinbauer, 2004). With emphasis to the genus *Vibrio*, HGT by phages can contribute to the development of new pathogenic bacteria strains in the marine environment through acquisition of virulence traits (Hazen et al., 2010, Fortier and Sekulovic, 2013).

1.2.4. INCEPTION OF INFECTION: PHAGE INFLUENCE IN THE DEVELOPMENT OF HUMAN PATHOGENIC *VIBRIO* STRAINS

Vibrio spp. are ubiquitously distributed in aquatic habitats worldwide, such as marine sediments, water column brackish water or associated with zoo- and phytoplankton (Thompson et al., 2004, Oberbeckmann et al., 2011a). Although this genus comprises more than 100 recognized species, only 11 are human pathogens (Miyoshi, 2013). For instance, *V. cholerae* from serogroups O1 and O139 have caused epidemics of cholera, while non-O1/O139 *V. cholerae* and *V. parahaemolyticus* can cause severe gastroenteritis (Chakraborty et al., 2000, Thompson et al., 2004). In general, *Vibrio* spp. can grow at warm seawater temperature (> 18 °C) and low salinity (< 25 ppt NaCl) (Vezzulli et al., 2013). Therefore, these pathogens have been commonly associated with tropical and subtropical regions (Baker-Austin et al., 2017). In addition, they exhibit a typical pattern of opportunistic blooms in coastal areas (Böer et al., 2013).

In the last decade, the seawater temperature raised significantly in over 70 % of worldwide coasts (Baker-Austin et al., 2017, Lima and Wethey, 2012). Several studies demonstrate that this climate change promotes the growth of potentially pathogenic *Vibrio* species throughout the North Sea, such as *V. parahaemolyticus* and *V. cholerae* (Martinez-Urtaza et al., 2018, Baker-Austin et al., 2017, Roux et al., 2015a, Baker-Austin et al., 2013, Bier et al., 2015, Vezzulli et al., 2016). In consequence, reports of

Vibriosis cases in the North Sea have been increased (Baker-Austin et al., 2017, Vezzulli et al., 2012). Several factors can be involved in the development of human pathogenic strains in the marine environment, such as the acquisition of virulence traits via HGT, as hypothesized in other environmental species (Chakraborty et al., 2000, Covacci et al., 1997).

The *Vibrio* genomes contain significantly more alien genes than other marine bacteria (Lin et al., 2018). The HGT by prophages is an important factor to the genetic diversity in *Vibrio* spp. (Fortier and Sekulovic, 2013). Among the pathogenic *Vibrio* strains, at least 65 genes are present exclusively in the pandemic strains compared with the non-pandemic strains, including several genes related to pathogenicity (Bastías et al., 2010). Those strains may acquire virulence traits by horizontal gene transfer in the environment (Hazen et al., 2010). For instance, the *V. cholerae* phage CTXφ was the first filamentous phage reported to transmit virulence traits, that results on the lysogenic conversion of the bacterium to produce toxins (Waldor and Mekalanos, 1996). Thus, the production of cholera toxin (*ctx*) in lysogen *V. cholerae* has been extensively studied as PHS model (Davis et al., 2000, Faruque and Mekalanos, 2012, McLeod et al., 2005, Rakonjac, 2012). Moreover, filamentous phages appear to have a widespread distribution as described by Castillo et al. (2018). Among 1,874 *Vibrio* genomes from several species, near a half contain a complete filamentous phage sequence, which encodes for the zonula occludens toxin (*zot*) described in *V. cholerae* (Castillo et al., 2018). For instance, marine *V. cholerae* from a coastal brackish pond contained the filamentous phage VCYϕ, this phage has a widespread distribution within its host population (Xue et al., 2012). Although it was not considered a virulence factor in these *V. cholerae* strains, the genome of VCYϕ contains the *zot* encoding protein and has genomic features similar to CTXφ (Xue et al., 2012). In this manner, the ability of these phages to transfer pathogenicity genes (e.g. *zot* or *ctx*) in marine *V. cholerae* strains represent a direct public health concern (Casas and Maloy, 2011, Casas et al., 2006). In addition, filamentous phages could play a role in transfer of pathogenic genes in other species, such as *V. parahaemolyticus* (Chang et al., 1998, Iida et al., 2001).

Besides the role of phages in host pathogenicity, lytic phages have been studied as therapeutic agents against *Vibrio* infections. For instance, the phages CK-2 and 153A-5 have been used in the treatment of systemic infections of *V. vulnificus* in mice (Cerveny et al., 2002). Furthermore, Srinivasan and Ramasamy (2017) isolated four phages (VV-1, VV-2, VV-3, and VV-4) for biocontrol of *V. vulnificus* in shrimps for aquaculture. These phages belong to the Tectiviridae family, characterized by no head-tail structure and non-enveloped icosahedral virion (Srinivasan and Ramasamy, 2017). In addition, tailed phages have been studied for phage therapy treatment against *V. cholerae* infections (Seed et al., 2011, Yen et al., 2017, Yen and Camilli, 2017). The ICP tailed phages isolated in Bangladesh, show efficiency and high specificity to prevent *V. cholerae* infections in infant mouse and rabbit models (Yen et al., 2017). Moreover, the epidemiological data from Dhaka (Bangladesh)

suggest that the *V. cholerae* population in that area might be reduced by phage infection during the peak of epidemics (Jensen et al., 2006).

In the case of tailed *Vibrio* phages, several members belong to Podoviridae, Myoviridae and Siphoviridae families have been characterized (e.g. Ackermann et al., 1984). Tailed phages have been described in pandemic and potentially pathogenic *Vibrio* species isolated from the marine environment, such as *V. cholerae* O1 Biotype ElTor strains (Sen and Ghosh, 2005), *V. parahaemolyticus* (Matsuzaki et al., 1992, Miller et al., 2003), *V. harveyi* (Baudoux et al., 2012), among others. Recently, the non-tailed phage family Autolykiviridae were characterized by Kauffman et al. (2018b). This novel family was identified during the sampling campaign of the Nahant Collection, a large-scale virus-host model system of cultivated and genome-sequenced bacterial and viral isolates focused in Vibrionaceae PHS (Kauffman et al., 2018a). This collection contains 251 dsDNA phages (Caudovirales and Autolykiviridae) infecting diverse marine Vibrionaceae (Kauffman et al., 2018a). As exemplary applied in the Nahant Collection, the characterization of marine *Vibrio* PHSs under a comprehensive approach can bring new insights to elucidate the ecological role of phages in the marine environment.

1.3. YOU CAN JUDGE A PHAGE BY THE SCOPE IT'S SEEN: APPROACHES APPLIED IN MARINE PHAGE ECOLOGY

As only a small fraction of the microbial diversity can be cultivated (approximately 1 %), the marine microbial abundance was drastically underestimated in early ecological studies (Rohwer et al., 2014, Jannasch and Jones, 1959). In consequence, these studies reflected a *'desert'* ocean, with an extremely low diversity of marine phages (Brum and Sullivan, 2015, Rohwer et al., 2014, Bergh et al., 1989, Farrah, 1987, Moebus, 1987). Such was the case, that Zobell (1946) proposed that marine phage could not be ecologically significant, due to the low abundance of a possible host or prey. Between 1940 to 1970s, direct microscopy approaches contrasted with indirect culture-based enumerations (Bergh et al., 1989, Jannasch and Jones, 1959, Rohwer et al., 2014, Torrella and Morita, 1979, Zobell, 1946). During the 1980s, Moebus (1980) was a pioneer to develop new approaches to isolate marine phages from mixed communities. After further additional evidence of high abundance of marine viruses (Bergh et al., 1989, Torrella and Morita, 1979), the interest of marine phage ecology was increasing till nowadays. At present, several techniques have been developed to study the marine virus community under different scopes.

1.3.1. METHODS TO COLLECT VIRUS BIOMASS

One of the critical steps to estimate the abundance of marine viruses is the virus biomass collection. The traditional method is the large-scale tangential flow filtration (TFF), in which the sample directed tangentially across the filter, while a cross-membrane pressure is applied through the filter membrane pores (Alonso et al., 1999, Clokie and Kropinski, 2009). Although this prominent method allows to concentrate the sample up to > 100-fold VLP in the retained water sample, this method requires large sample volume (> 100 l) and can be highly time-consuming (Alonso et al., 1999, John et al., 2011, Clokie and Kropinski, 2009). Therefore, other methods have been developed to concentrate viruses, e.g. adsorption-elution method and pelleting virus particles with ultracentrifugation. However, selective adsorption of viruses to treated filters and low virus recovery are important technique biases with these methods (Fuhrman et al., 2005, John et al., 2011, Percival et al., 2004). Additionally, the iron chloride flocculation method has several advantages as compared to the traditional TFF method (Hurwitz et al., 2013a, John et al., 2011). Briefly, the seawater sample is pre-filtered and the 'virus fraction' is incubated with iron chloride to form a Fe-virus complex, which can be collected onto a polycarbonate filter and resuspended with Ascorbate- or Oxalate-EDTA buffer for further analyses (John et al., 2011). The iron chloride flocculation method is easy to implement, and has high efficiency, proven by concentrate recovery (> 90 %) in tested seawater samples with tailed phages (John et al., 2011). Moreover, this advantage accounts for the requirement of relatively low sample volume (~ 20 l), one reason for the favor of this chemistry-based concentration method during oceanographic sampling campaigns, as well as for freshwater studies (Duhaime and Sullivan, 2012).

1.3.2. METHODS TO ESTIMATE ABUNDANCE AND DIVERSITY

The improvements of the analyses of marine virus abundance are in close relationship with the development of direct count methods for bacterial enumeration with microscopy-based techniques (Wommack and Colwell, 2000). Due the develop of epifluorescence microscopy (EpM) in the late 1970s, the ocean turns out to be an environment massively inhabited by microorganisms (Rohwer et al., 2014, Hobbie et al., 1977). Since then, methods to estimate the virus abundance with EpM were developed (Bergh et al., 1989, Frank and Moebus, 1987, Noble and Fuhrman, 1998, Torrella and Morita, 1979).

EpM appeared as a fast method to quantify VLP from both isolates and direct field samples (Rachel and Jed, 1998, Clokie and Kropinski, 2009). The traditional method applies a small pore size membrane filter to collect the fixed VLP for staining with a fluorescent dye, such as DAPI, Yo-Pro-1, SYBR Green and SYBR Gold (Clokie and Kropinski, 2009, Thurber et al., 2009, Rachel and Jed, 1998). Although this technique allows to quantify the total count of VLP in the sample, no deeper information regarding their infectivity and further classification is possible. On the other hand, the flow cytometry is a fast and precise approach to distinguish virus groups from total counts based on the scatter and fluorescence

obtained after stain the virus particles (Weinbauer, 2004, Clokie and Kropinski, 2009, Wommack and Colwell, 2000). In addition, these techniques can be complemented with other methods to obtain more information about virus infectivity. For instance, the 'viable' counts can be obtained as plaque-forming units (PFU) on a double agar overlay plaque assay (section 1.4.1) or in culture broth as most-probable-number (MPN) assay (Weinbauer, 2004, Clokie and Kropinski, 2009). However, the vast majority of marine microorganisms remains uncultivated and 'viable' counts yield strong underestimation of the total abundance (Rappé and Giovannoni, 2003, Brum and Sullivan, 2015, Weinbauer, 2004). Therefore, a comprehensive approach must be applied to reduce each technique bias to analyze the virus population structures.

The marine virus diversity can be estimated at the level of virion morphology, growth parameters in infection, and nucleic acid sequence (Duhaime, 2010). Electron microscopy (EM) was used to visualize phages from marine samples to estimate the abundance of total counts as well as the diversity of virus particles (Weinbauer, 2004). Based on the virion morphology (e.g. virion size, tail structure), the phages from isolates or direct samples can be classified at the family level (Ackermann, 2011, Ackermann and DuBow, 1987). Due to the information outcome from this method, virion morphology classification is currently one of the minimal requirements to characterize a phage taxon (Fauquet and Martelli, 2013). More information about EM application in phage morphology characterization is detailed in section 1.4.2. In addition, other features used for characterization include the proliferation cycle, host range, plaque morphology, latent period and burst sizes from isolated PHSs (Duhaime, 2010, Weinbauer, 2004).

In contrast to cell organisms, phages lack a universal gene to be used as molecular markers to establish phylogenetic relationships (Tolstoy et al., 2018, Alavandi and Poornima, 2012). In consequence, protein encoding genes such as large subunit terminase, major capsid, major tail proteins, and DNA polymerases can be used to build phylogenetic trees from specific phage families (Tolstoy et al., 2018). Thus, molecular-based techniques (e.g. PCR amplification) have been used to analyze the virus community structure (Williamson et al., 2012). Although PCR is a fast technique to identify the presence of specific virus groups from environmental samples, the high mutation rate and mosaicism nature of genomes give several problems for the applicability of this technique (Clokie and Kropinski, 2009). For this reason, PCR-based approaches are restricted to small-scale projects focusing on specific genes and on well-known phage genera (Clokie and Kropinski, 2009).

Due to the constantly growing number of virus genomes available in the databases, the advance of metagenomic-based studies appeared as a reliable alternative to overcome the cultivation bottleneck and virus gene marker methods (Brum and Sullivan, 2015, Thurber et al., 2009). The application of massive nucleic acid sequencing to characterize the virus community composition is detailed in section 1.5.

1.4. METHODS APPLIED TO TAILED PHAGE'S CHARACTERIZATION

The study of tailed phages (Caudovirales) was fundamental for the current understanding of the nature and function of nucleic acids and proteins (Grose and Casjens, 2014). Several phage strains from 16 Caudovirales genera have been isolated from the 'model' bacteria *Escherichia coli* (Grose and Casjens, 2014). These phage–host model systems were the basis to discover the proliferation cycle of Caudovirales phages (Krupovic et al., 2011, King et al., 2012). Particularly, the model T-phages (phages T1 to T7, Demerec and Fano, 1945) facilitated the comparability of results between laboratories, as well as the definition of key phage-related concepts (Salmond and Fineran, 2015). The study of these model phages was the basis to define the main shared criteria of phage taxonomic classification.

The ability to identify the members of an ecosystem is one of the first steps to understand the community structure and function (Systems et al., 1995). In ecological studies, the taxonomical identification of the community composition also permits the identification of the taxa with most ecological importance (Systems et al., 1995). In the case of viruses, the viral taxonomy is a discipline that constantly refining definitions of a virus species and their classifications (Morgan, 2016). As most of the phage classification was primarily based on the particle morphology, other properties such as replication, protein characterization and genome organization are described only in a small percentage of well-studied species (King et al., 2012). Phage taxonomy has evolved from that morphology-based classification to a comprehensive approach with multiple criteria (Tolstoy et al., 2018). However, several discrepancies in phage classification cause a limited number of higher hierarchies groups (Desselberger, 2018). For instance, the high rate of horizontal gene transfer (HGT) and the mosaicism nature of genomes give problems to classify phages with highly similar genomes but with distinct morphology (King et al., 2012, Van Regenmortel et al., 2013). Moreover, the taxonomic system of the International Committee on Taxonomy of Viruses (ICTV) comprises only cultivated viruses, and recently a 'consensus statement' propose the incorporation of uncultivated viruses (Desselberger, 2018, Simmonds et al., 2017). Thus, there are no current universal criteria in Caudovirales classification and they can vary between different genus (King et al., 2012, Fauquet and Martelli, 2013).

1.4.1. METHODS FOR ISOLATION OF PHAGE–HOST SYSTEMS

Phage–host systems (PHSs) have been used to study the interaction of phages and their hosts, including co-evolution and gene transfer (Weinbauer, 2004, Brum and Sullivan, 2015). The PHS enables to propagate phages to analyze their growth parameters, to integrate methods for temperate phages, as well as cross-reaction tests against other potential isolated hosts (Moebus, 1983, Weinbauer, 2004). The majority of the gained knowledge is based only on a narrow group of isolated phages, which represents only a small fraction of the total phage diversity (Brum and Sullivan, 2015, Holmfeldt et al., 2013). Thus, the isolation of novel hosts and their viruses is crucial to improve our understanding of the interaction of marine phages and their hosts.

Once the potential host is isolated, lysis-based screenings are typically used to identify infective phages (Clokie and Kropinski, 2009). Firstly described by Adams (1959), viable phages can be detected via plaque formation in the double agar overlay assay (DAOA). For DAOA, the phage sample is mixed with the host into a soft agar lawn (Clokie and Kropinski, 2009, Adams, 1959). After incubation, a phage plaque is a clearing spot in a bacterial lawn (figure 1.3A), formed by consecutive phage infections on adjacent host cells (Clokie and Kropinski, 2009). Modified methods, such as spot test assay (STA), allow the fast screening of several phage samples against one single potential host (figure 1.3B).

In the case of lysogenic phages, induction is needed prior to phage isolation. In laboratory conditions, these phages can be induced from lysogen bacteria isolates with mitomycin C or UV radiation (Zhao et al., 2010, Jiang and Paul, 1998). Inducing agents found in the marine environment, are sunlight (due to UV radiation), temperature, pressure, and hydrocarbons (Jiang and Paul, 1996). The induced phages can be concentrated for further host screening with DAOA or STA (Clokie and Kropinski, 2009).

The phage isolation is based on the consecutive propagation of one single plaque. Once isolated, a high titer of lysate elution can be stored as maintenance stock for further analyses, such as electron microscopy or genome sequencing (Clokie and Kropinski, 2009). Besides, the plaques can be counted as plaque forming units (PFU) to obtain the concentration of ‘viable’ counts (Clokie and Kropinski, 2009, Weinbauer, 2004). Together with DAOA, the one-step growth permits to determine the burst size (number of phages produced per infected bacterium), as well as the phage growth parameters from a PHS (Clokie and Kropinski, 2009). In addition, phage–host cross reaction tests can be used to determine host range and interaction PHSs (e.g. Moebus, 1983).

A deep characterization of novel PHSs is based on a comprehensive approach that applies these culture-based methods together with molecular-based methods. Culture-based methods are the basis to obtain data for in-depth characterization of phages, such as the description of the phage genome, virion structure, proliferation cycle, gene transfer and role of auxiliary metabolic genes (Breitbart et al., 2018, Brum and Sullivan, 2015).

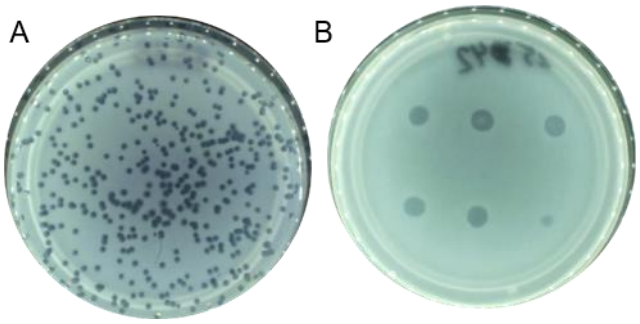


FIGURE 1. 3. The appearance of plaques formed by *Vibrio* phages with the double agar overlay assay (A), and spot test assay (B).

1.4.2. METHODS TO CHARACTERIZE MORPHOLOGY

Electron microscopy (EM) became a basic technique for phage classification (Clokie and Kropinski, 2009). Transmission EM compresses different techniques, the most used is the negative staining method (Clokie and Kropinski, 2009). The principle of the negative staining is to mix the phage particles with an electron-dense solution of a metal salt of high molecular weight and small molecular size, frequently as uranyl acetate (Ackermann and DuBow, 1987, Clokie and Kropinski, 2009). However, the preparation and execution of this technique can be time-consuming and require expensive equipment (Rachel and Jed, 1998). This method allows the analysis of free phage particles from PHSs and environmental samples (Proctor, 1997). In general, the characterization of a virion is based on digital electron microscopical images from the purified phage elution, which give measures for the capsid, the tail as well determination of presence and sizes of tail fibers (in nm scale), envelope, and other morphological characteristics (Clokie and Kropinski, 2009).

1.4.3. METHODS TO CHARACTERIZE A GENOME

Currently, the most efficient method to determine the genomic sequence of a DNA phage is the whole shotgun sequencing. The bioinformatic analysis of the phage genome sequence can give the information of the genome size and the guanine and cytosine content (GC content or GC %), as well as to predict the topology of the genome, packaging method, and protein encoding genes (Clokie and Kropinski, 2009, Brüssow and Hendrix, 2002, Hendrix, 2003). These genomic features are crucial for genomic characterization of new PHSs (Clokie and Kropinski, 2009). In addition, the identification of key genes or GC content differences can be used to detect horizontal gene transfer events and prophage sequences (Zhang and Zhang, 2004).

The development of Next Generation Sequencing (NGS, revised bellow) improved the number of phage genomes available in the public databases and the access to detailed characterization of PHSs (Brister et al., 2015, Clokie et al., 2018). The accessibility to novel technologies creates dynamism to understand the nature of marine viruses under different glances.

1.5. NGS APPLIED IN MARINE VIRUS (META)GENOMICS

In 2005, several high-throughput technologies known as Next Generation Sequencing (NGS) have been developed to sequence high-quality nucleotide sequences in large scales, reducing significantly the cost per sequenced base pair (Koboldt et al., 2013). Examples of NGS technologies are the pyrosequencing by 454 Life Science (now part of Roche) and Solexa (acquired by Illumina in 2007) (Metzker, 2009, Vincent et al., 2017). Due to its cost-benefit relation, the Illumina-based technology (e.g. MiSeq and Hiseq) dominates the current NGS market (Vincent et al., 2017).

The NGS access enables the production of genome sequencing from isolated microorganisms, as well as complete genomes from environmental samples (metagenomics) (Vincent et al., 2017). The deposited bacterial and phage genomes in publicly available databases increased significantly after the emerge of NGS technologies. In consequence, new challenges appear to manage this *big data* outcome. The development of new bioinformatic tools to analyze this *big data* provides unprecedented opportunities to address fundamental questions in phage ecology (Ibrahim et al., 2018).

1.5.1. PHAGE (META)GENOMICS

Virus metagenomics (hereafter viromics) is environmental sequence analysis of viral assemblages, commonly based on the genome reconstruction *de novo* of uncultivated DNA viruses without previous isolation (Martínez et al., 2014, Roux et al., 2018). Since 2016, this technique allowed to identify 750,000 uncultivated virus genomes (UViGs), five times more than the total number of genomes referred to virus isolates (Roux et al., 2018, Paez-Espino et al., 2016, Roux et al., 2016). Unfortunately, the general trend of virus genomes from isolates (iVGs) did not increase after the appearance of NGS technologies (Perez Sepulveda et al., 2016). Besides the work of Mizuno et al. (2013), the average number of marine iVGs is around 10 submitted genomes in the EMBL-EBI database from ENA per year (Perez Sepulveda et al., 2016). While most of the UViGs are probably derived from free lytic viruses, the characterization of lysogenic viruses remains underexplored (Howard-Varona et al., 2017).

Approximately, half of the marine bacterial isolates contain prophages (Jiang and Paul, 1998, Mcdaniel et al., 2006). Moreover, prophage-like elements were identified in almost a half of marine bacteria genomes (Bondy-Denomy and Davidson, 2014, Paul, 2008). In the case of poly-lysogeny (> 1 prophage), this percentage can increase to almost 70 %, as identified in certain genomes, such ain s marine *Vibrio* species (Castillo et al., 2018). However, the predicted prophage sequences might be underestimated, because the experimental prophage induction was demonstrated in some marine bacteria where no prophages could be identified on their genome (Zhao et al., 2010). In contrast, there are also cases in Gram-negative bacteria that identified prophage sequences does not produce functional virions (Summer et al., 2004), such as the Mu-like prophages from *Haemophilus influenzae* (FluMu), *Neisseria meningitidis* (Pnm1), *Deinococcus radiodurans* R1 (RadMu) and

Shewanella oneidensis (MuSo1 and MuSo2) (Heidelberg et al., 2002, Morgan et al., 2002, Summer et al., 2004). Despite the efforts for prophage prediction, these sequences are mostly underreported as part of the bacterial genome, or prophage-like sequences might be overreported when the prediction is not confirmed by e.g. microscopy or annotation analysis (Canchaya et al., 2003). In addition, metadata of published phage genomes contains terminology inconsistencies (e.g. temperate or lysogenic lifestyles) or missing information, that misleads to an underrepresentation of lysogenic phages in the public databases (Clokie et al., 2018). Thus, a comprehensive approach and data curation are key parts for UViGs and iVGs analyses.

Currently, UViGs represent the majority (≥ 95 %) of the taxonomic diversity of viruses in public databases (Roux et al., 2018). In this way, the low number of characterized phage isolates still limits the information available to characterize virome datasets and predicts possible phage–host interactions in the marine environment (Perez Sepulveda et al., 2016, Roux et al., 2015c). In addition, most of phage genomes are poorly annotated, and there is a limited range of computational tools focused on phage analysis (Ibrahim et al., 2018). In consequence, these viromes are dominated by unknown sequences (60 – 95 % of the virome), which show low or none similarity against in-deep characterized virus genomes (Krishnamurthy and Wang, 2017, Roux et al., 2015c, Perez Sepulveda et al., 2016). This suggests the incredible diversity of marine viruses (Edwards and Rohwer, 2005). In addition, novel approaches have been developed to analyze the spatial and temporal variability of virus diversity with viromics (Hurwitz et al., 2016). For instance, the predicted gene sequences have been used to compare the virus population diversity via virus protein clustering (Bolduc et al., 2016, Paez-Espino et al., 2017). This approach can be settled from the sequence comparison between each microbial population to determine the diversity of protein sequences of each sample (e.g. Lema et al., 2014, Zehr et al., 1995), as well as the sequence comparison using metagenomic virus contigs from the tool database (Paez-Espino et al., 2017). However, there is no general database of all available virus sequences (Ibrahim et al., 2018).

1.5.2. THE NEEDLE IN THE HAYSTACK:

PHAGE GENOMES IN THE DATABASES

The available marine virus datasets are currently deposited across diverse database repositories, such as the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genome/viruses), and the European Nucleotide Archive (www.ebi.ac.uk/ena/ and https://www.ebi.ac.uk/genomes/phage.html) (Bolduc et al., 2016). During the last decade, the number of UViGs in the databases increased drastically. In consequence, new databases of virus genomes have become available for virus analysis. The JGI IMG/VR database (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi) currently represent the largest sequence repository of UViGs from virome and metagenomic studies (Roux, 2019). This database permits to visualize the UViGs' dataset across the human body and the habitats around the world (Paez-Espino et al., 2016, Pati et al., 2013). In addition, the iMicrobe platform (www.imicrobe.us) unites datasets and tools from different sources to bioinformatic analysis in the web-based platform of CyVerse (https://de.iplantcollaborative.org/de, Goff et al., 2011, Hurwitz, 2014). This platform is also connected with the iVirus (www.ivirus.us), a user-friendly resource for viromic analysis. Near half of the total projects available in iMicrobe (44.44%) involves virus identification as UViGs, and around the third part of these studies are from marine environments (31.25 % of virus-related projects available until April 2019). Several virome datasets are available in this platform, including the first published marine virome (Angly et al., 2006), the Tara Ocean Virome (TOV) (Brum et al., 2015), and the Pacific Ocean Virome (POV) (Hurwitz and Sullivan, 2013). The lack of integrated databases creates usability problems, and often local datasets are created (Bietz and Lee, 2009). In consequence, the analysis of marine UViGs for comparative metagenomics can be laborious and time-consuming (Ibrahim et al., 2018). Hence, the selection of the database has a crucial effect on the genome characterization outcome.

1.5.3. VIRUS (META)GENOMIC TOOLS

The diversity of uncultivated viruses in the databases by far exceeds those of uncultivated bacteria (Breitwieser et al., 2017). In addition, the classification of these viruses is far more difficult to conduct than in cell organisms and requires different methods from bacteria (Breitwieser et al., 2017). Thus, a defined strategy of appropriated virus-specific database and tools is critical to analyze phage genomes and viromes. This section describes tools developed for virus sequence analyses, with emphasis on phage-based approaches. Examples of bioinformatic tools and databases are present in table A.2.1, which can vary in database size, accuracy, and the number of citations.

As the methods applied in viromic analyses are still in development, there is no state of the art routine approach to study viromes in an integrated workflow (Ibrahim et al., 2018). Due the increase of viromic studies, the development of flexible and integrative bioinformatic approaches are necessary to allow rapid and accurate phage classification (Ibrahim et al., 2018, Chibani et al., 2019). Despite the increase of published tools for viromics, e.g. Bolduc et al. (2016) emphasize the lack of reliable tools for phage ecologist without deep knowledge in the programming language.

In general, the analysis of virome reads follows three main steps: (1) sample preparation, focus on biomass collection and concentration (section 1.3.1); (2) data preparation, focus on quality control for high-quality virome contigs; and (3) data analysis, basis for genome characterization which includes the predicted genome features, as well as identification of taxonomic and functional data from the nucleotide and predicted protein sequences (Hurwitz et al., 2016, Clokie and Kropinski, 2009). Although the overall steps are similar to the pipelines for other metagenomic analyses (e.g. Tragin and Vaulot (2018) and https://github.com/genomewalker/microbeco2015/wiki), viromics have critical methodological challenges including steps from sample collection to read assembling (Krishnamurthy and Wang, 2017, Clokie and Kropinski, 2009, Hurwitz et al., 2016). Virome data is characterized by reads with high proportions of repeat regions, hypervariable genomic regions associated with host interaction, high mutation rates by strain variation, and limited read coverage due to low DNA biomass (Sutton et al., 2019, Roux et al., 2017, Warwick-Dugdale et al., 2019). These characteristics can result in fragmented virome assemblies (García-López et al., 2015). Thus, the selection of filtration and assembling tools developed for short hypervariable reads are essential for high-quality viromes (Sutton et al., 2019). After filtration and decontamination of the virus reads, the application of *de novo* assemblers generates local sample-specific databases that avoid reference genomes biases, such as low diversity due to small PHS databases (Coutinho et al., 2017, Brum et al., 2015b, Dutilh, 2014, Sunagawa et al., 2015). Currently, the most effective tools for assemble viromes are SPAdes, Megahit, and IDBA-UD (Roux et al., 2017, Sutton et al., 2019). In addition, SPAdes has been implemented in other frequent used viromic approaches such as iVirus (Bolduc et al., 2016). The Newbler assembler was developed for 454 pyrosequenced reads, which was implemented in the V-GAP assembly pipeline (Nakamura et al.,

2016). Furthermore, comparison of assembled contigs per virome has been applied to study the diversity of viral populations (Warwick-Dugdale et al., 2019).

The genome annotation is a key process to identify and characterize virus genomes from assembled contigs during the data analysis (Ibrahim et al., 2018, Hurwitz et al., 2016). Annotation of phage genomes includes BLAST (BLASTx, tBLASTx, or BLASTP) similarity search against NCBI databases (nr and/or refseq), also HMMER searches against Pfam (Altschul et al., 1997, Henn et al., 2010, Bateman et al., 2004, Eddy, 1998). Due to the limited representation of viral genomes in these datasets, only a small fraction of the virome can be annotated (Hurwitz et al., 2016). Therefore, several auto-annotation tools can be applied to analyze the assembled contigs for further curation before submission (Clokie et al., 2018). Besides, other tools have been developed to identify prophage-like sequences from single bacterial genomes or metagenomic sequences (table A.2.1). Due to the high versatility and reliability, VirSorter is currently one of the most widely used approaches to study the dsDNA viruses (Zheng et al., 2019). VirSorter (from iVirus) allows to detect prophage regions, as well as virus contigs for functional and taxonomic characterization (Bolduc et al., 2016, Paez-Espino et al., 2017, Roux et al., 2015b). The iVirus is connected with the iMicrobe platform, and contains several available virome datasets, including the first published marine virome (Angly et al., 2006), the Tara Ocean Virome (TOV) (Brum et al., 2015), and the Pacific Ocean Virome (POV) (Hurwitz and Sullivan, 2013). In addition, other tools have been developed for single phage genomes analyses have been developed. For instance, the taxonomy of tailed phages can be defined based on neck protein encoding genes and genome context with Virfam Caudovirales classifier (Lopes et al., 2014), and the phage method of genome packaging can be predicted with PhageTerm (Garneau et al., 2017). Moreover, other prediction tools have been developed to identify important linkages, e.g. CRISPR identification (Clustered Regularly Interspaced Short Palindromic Repeats), horizontal gene transfer, and potential host prediction (Roux, 2019). Unfortunately, most of these tools are based on reference genomes and the results are not optimal for novel genomes.

On first sight, these bioinformatic tools might be a solution to identify a larger portion of uncharacterized genomes. However, the functional data in the viromes is still largely unknown. The tools applied in virus (meta)genomics still require important methodological and conceptual breakthroughs to characterize the large percentage of unknown marine viruses (Ibrahim et al., 2018). The collaboration between virologists and bioinformaticians is necessary to improve and interconnect the existing tools (Roux, 2019). Although the current tools are powerful and impressive (Ibrahim et al., 2018), the interconnection between these tools and databases is highly necessary. Therefore, the application of a comprehensive approach using the existing bioinformatic tools can generate a more detailed characterization of novel marine phages. Moreover, the analysis of exemplary samples must be included to develop a virus-specific pipeline.

1.6. MOTIVATION AND RESEARCH AIMS

Despite the enormous ecological and evolutionary importance of marine phages, much of their biodiversity remains uncharacterized (Coutinho et al., 2017). Emerging NGS techniques have accelerated the expansion of the viral-universe, which imply new challenges but also yield opportunities to address fundamental questions in phage ecology (Ibrahim et al., 2018). Despite that new bioinformatics tools appears almost every month and the accelerated increase of marine virus studies, there is no statutory pipeline applied in marine phage ecology. Thus, this research is focused on the integration of existing tools to develop virus-specific pipelines. This approach was applied in exemplary samples to study marine phages from the North Sea. Therefore, this thesis is based on two research aims:

1. **To analyze the virus community with viromics** (chapter 2).

Metagenomics has become a powerful tool to characterize the microbial community in greater detail (Coutinho et al., 2017). The application of viromics (virus metagenomics) can be used to compare the taxonomic composition and the carried genes of viruses from the North Sea. The estimated occurrence, diversity and distribution of the virus community can be used for comparison of areas with distinct oceanographic features, such as the coast and open sea. To accomplish this research aim, a method pipeline was developed to ensure high-quality output for taxonomic, functional and comparative analyses. For this, exemplary seawater samples from the North Sea were first selected for virus biomass collection and further DNA extraction; the second part was *in silico* preparation of the genomic data to obtain high-quality reads. To reduce possible bias caused by reference genome databases from cultivable strains, this pipeline is based on *de novo* assembling to produce sample-specific databases.

Addressed questions:

How distinct is the virus community composition between the coast and the open sea?

How is their geographic distribution in relation to environmental parameters?

What is the distribution of environmental relevant genes carried by phages
(e.g. auxiliary metabolic genes or pathogenicity genes)?

2. **To characterize lysogenic phages from potentially pathogenic *Vibrio* species** (chapter 3 and 4).

Lysogenic phages may contribute to the emergence of disease-causing strains from environmental populations. Considering the increasing abundance of potentially pathogenic *Vibrio* species in the North Sea, the second research aim is focused on the characterization of novel phages from *V. parahaemolyticus* and *V. cholerae* isolated from the North Sea. For this, the method pipeline set-up must be performed to isolate potentially pathogenic *Vibrio* from water samples, follow by the induction of their prophages for DNA extraction. Exemplary isolates were selected for lysogenic phages characterization. The data preparation and *de novo* assembling are critical steps to assemble a complete phage genome separated from the– non-sequenced– host genome.

Addressed questions:

How is the distribution of lysogen *Vibrio* in the North Sea?

What are these *Vibrio* phages?

Can these phages be related to pathogenicity in these strains?

The research aims and the overview of the contribution of this thesis is represented in figure 1.4.

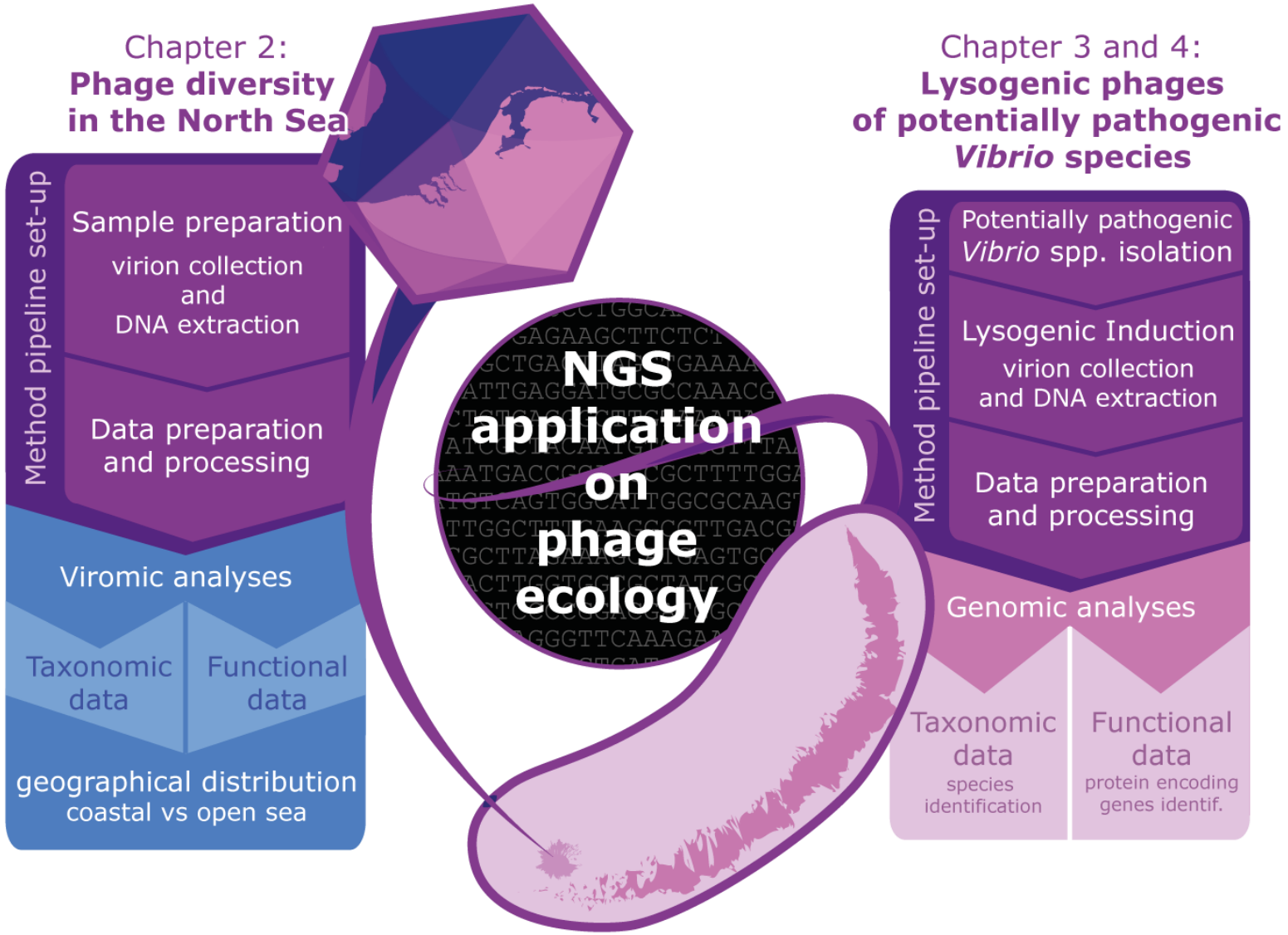


FIGURE 1. 4. Overview of the components of this thesis, including the main approach to accomplish the main research aims.

1.7. CONTENT OVERVIEW

The present thesis consists of an introductory overview, three chapters based on the achievements of the addressed research aims, and a final general discussion. A short overview of the research articles follows.

Chapter 2

The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages

Authors: Alexa Garin-Fernandez, Emiliano Pereira-Flores, Frank Oliver Glöckner, Antje Wichels

Published in: *Marine Genomics*. 1 Jun 2018.

This manuscript describes the implementation of virus metagenomics (viromes) to analyze the virus community in the North Sea. The influence of environmental parameters on the occurrence and diversity of phages is also presented. The planning, experiment designs and genomic data submission were conducted by Alexa Garin-Fernandez with the assistance of Antje Wichels. Statistical analyses and the comparison of predicted genes distribution were performed by Alexa Garin-Fernandez together with Emiliano Pereira-Flores. Data analyses, evaluation and manuscript writing were carried out by Alexa Garin-Fernandez under the guidance of Antje Wichels and Frank Oliver Glöckner.

Chapter 3

Looking for the hidden: Characterization of lysogenic phages
in potential pathogenic *Vibrio* species from the North Sea

Authors: Alexa Garin-Fernandez, Antje Wichels

Submitted to: *Marine Genomics*. 28 August 2019

This manuscript presents the identification of lysogenic phages from potentially human pathogenic *Vibrio* species, and the genomic characterization of Caudovirales phages. This study applies culture and molecular techniques to analyze non-model systems, such as *V. parahaemolyticus* and *V. cholerae*. The genome structure and the predicted biological functions of three novel phages are also described. The planning, experiment designs and genomic data submission were conducted by Alexa Garin-Fernandez with the assistance of Antje Wichels. Data analyses, genomic characterization and manuscript writing were carried out by Alexa Garin-Fernandez under the guidance of Antje Wichels.

Chapter 4

Genomic characterization of filamentous phage vB_VpaI_VP-3218,
an inducible prophage of *Vibrio parahaemolyticus*

Authors: Alexa Garin-Fernandez, Frank Oliver Glöckner, Antje Wichels

Submitted to: *Marine Genomics*. 28 August 2019

Based on the outcome from chapter 3, this manuscript presents the characterization of one filamentous phage and its potential role on pathogenicity. This study complements prophage induction and bioinformatic analysis applied to non-model *V. parahaemolyticus* strains. The planning, experiment designs and genomic data submission were conducted by Alexa Garin-Fernandez with the assistance of Antje Wichels. Data analyses, genomic characterization and manuscript writing were carried out by Alexa Garin-Fernandez under the guidance of Antje Wichels and Frank Oliver Glöckner.

In addition, contributions to the following paper were given.

Chapter A.1

Dangerous hitchhikers? Evidence for potentially pathogenic *Vibrio* spp. on microplastic particles

Authors: Inga V. Kirstein, Sidika Kirmizi, Antje Wichels, Alexa Garin-Fernandez, Rene Erler, Martin Löder, Gunnar Gerdts

Published in: Marine Environmental Research. Sep 2016.

This manuscript presents the occurrence of potentially pathogenic *Vibrio* spp. attached to microplastics from Baltic and the North Sea. This study integrates the sample analyses from both cruises HE409 and HE430. Sidika Kirmizi (cruise HE409, 2013) and Inga V. Kirstein (cruise HE430, 2014) collected and analyzed the microplastic and surface water samples. Isolation of potentially pathogenic *Vibrio* spp. during the cruise HE430 was conducted by Inga V. Kirstein, supported by Sidika Kirmizi and Alexa Garin-Fernandez. Identification of *Vibrio* isolates with MALDI-TOF MS were carried by Rene Erler (cruise HE409, 2013) and Alexa Garin-Fernandez (cruise HE430, 2014). Microplastic identification was carried by Martin Löder. Data evaluation and manuscript writing were carried out by Inga V. Kirstein and Sidika Kirmizi under the guidance of Antje Wichels and Gunnar Gerdts.

Chapter 2

NORTH SEA GOES VIRAL:

Occurrence and Diversity of North Sea Viruses

Alexa Garin-Fernandez[a,b,*], Emiliano Pereira-Flores[b,c], Frank Oliver Glöckner[b,c], Antje Wichels[a]

[a] Alfred-Wegener-Institute Helmholtz-Center for Polar and Marine Research, Biologische Anstalt Helgoland, Helgoland, Germany
[b] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[c] Jacobs University Bremen gGmbH, Bremen, Germany
[*]To whom correspondence should be addressed.
Tel: +49(4725)819-3233; Fax: +49(4725)819-3283; Email: alexa.garin@awi.de

Received 4 January 2018; *Revised* 19 April 2018; *Accepted* 19 May 2018
Marine Genomics. 2018 Jun 1. pii: S1874-7787(17)30379-3. doi: 10.1016/j.margen.2018.05.004

Keywords: Metagenome / Virome / Marine phages / Virioplankton

Abstract: Marine viruses are dominated by phages and have an enormous influence on microbial population dynamics, due to lysis and horizontal gene transfer. The aim of this study is to analyze the occurrence and diversity of phages in the North Sea, considering the virus-host interactions and biogeographic factors. The virus community of four sampling stations were described using virus metagenomics (viromes). The results show that the virus community was not evenly distributed throughout the North Sea. The dominant phage members were identified as unclassified phage group, followed by Caudovirales order. Myoviridae was the dominant phage family in the North Sea, which occurrence decreased from the coast to the open sea. In contrast, the occurrence of Podoviridae increased and the occurrence of Siphoviridae was low throughout the North Sea. The occurrence of other groups such as Phycodnaviridae decreased from the coast to the open sea. The coastal virus community was genetically more diverse than the open sea community. The influence of riverine inflow and currents, for instance the English Channel flow affects the genetic virus diversity with the community carrying genes from a variety of metabolic pathways and other functions. The present study offers the first insights in the virus community in the North Sea using viromes and shows the variation in virus diversity and the genetic information moved from coastal to open sea areas.

2.1. INTRODUCTION

Viruses are the most abundant biological entity in world's oceans as well as the second largest component of biomass after prokaryotes (Suttle, 2007, Breitbart, 2012). Despite their small size, they have an enormous influence on oceanic ecosystems. They do not only affect the global biogeochemical cycles, but also many oceanic food webs and marine population dynamics, due to lysis, reprogramming of host metabolism and horizontal gene transfer (Puxty et al., 2014, Breitbart, 2012, Fuhrman, 1999). Bacteriophages or phages – viruses that infect bacteria –probably comprise the majority of viruses and also provide the largest reservoirs of unexplored genetic diversity in marine environments (Mizuno et al., 2013).

Most marine viruses lack both cultivated representatives (Holmfeldt et al., 2013) and a universal marker for viral diversity; thus, virus metagenomics (hereafter viromes) currently represent the best approach to analyze the viral genomic diversity in terms of taxonomic composition and their putative functions (Edwards and Rohwer, 2005, Rosario and Breitbart, 2011, Tangherlini et al., 2016). Viromes generate a large amount of sequence data to investigate the unknown marine phage diversity from various environments (Simmonds et al., 2017), such as marine ecosystems.

The microbial community in marine ecosystems is shaped by environmental parameters such as salinity (Lozupone and Knight, 2007) and temperature (Fuhrman et al., 2008). The different biogeochemical properties in the open sea and coastal environments affect the microbial community composition. In the North Sea, the southern estuary regions are characterized by low salinity and high productivity, due to the high load of organic and inorganic matter introduced by rivers and weirs (Emeis et al., 2015). The southeastern coastal zone of the North Sea is connected to the English Channel through the Dover Strait and is characterized by a higher salinity, strong tidal currents, and high storm activity (Guillou et al., 2009, Otto et al., 1990). The northern North Sea is influenced by Atlantic Sea currents; therefore, its water masses have a higher salinity in contrast of the lower latitude zones and are considered low productive zones (Emeis et al., 2015). These different hydrographical regions and the high nutrient concentration from land are mixed in the central zone, consequently this area in the North Sea is prone to anthropogenic pollution (Emeis et al., 2015). The mean water circulation of this semi-enclosed basin is dominated by the Atlantic inflow from the northern North Sea, the English Channel exchange flow, and the Norwegian Coastal Current outflow (Otto et al., 1990). Due to the natural variability of the North Sea, biotic and abiotic parameters have been monitored for more than five decades around Helgoland Island in the German Bight (Wiltshire et al., 2010). The inflow current from the transition zone off the southern coast of the UK in the English Channel have been monitored by the Western Channel Observatory, which has one of the largest collection data series in the world (Southward et al., 2004).

Previous studies were conducted in the North Sea to understand how environmental parameters shape the microbial communities. These studies were focused on temporal and spatial variability, applying different methods like cultured based isolation, DGGE, metatranscriptomes, and 16S rRNA gene tag sequencing (Gerdts et al., 2004, Kopf et al., 2015b, Lucas et al., 2015, Wiltshire et al., 2010). However, only a few studies have focused on virus diversity, and all of them were based on virus-like particles (VLP) counting or culture dependent techniques (Duhaime et al., 2017, Duhaime et al., 2016, Duhaime et al., 2011b, Gerdts et al., 2004, Wichels et al., 1998).

Like microorganisms, the marine virus community structure is not homogenously distributed, delineating that the occurrence and distribution of marine viruses are affected by both virus-host interaction and biogeographic patterns at a local and global scale (Chow and Suttle, 2015, Roux et al., 2016). The distribution dynamics of microbial communities in the North Sea are driven by environmental factors (Brandsma et al., 2013, Lucas et al., 2015, Wiltshire et al., 2008), which might be reflected in the occurrence and distribution of viruses, especially by dominant bacteriophage families.

Based on these findings, the aim of this study is to analyze the occurrence and diversity of phages in the North Sea. Four viromes were analyzed to (1) identify the phage diversity and the main phage taxa present, (2) compare the phage diversity among different hydrographical zones, and (3) study the distribution of the genes carried by phages.

2.2. EXPERIMENTAL PROCEDURES

2.2.1. SAMPLE COLLECTION

Viral concentrates were collected from water samples from August 5 to 9, 2014 during the RV Heincke HE430 in four sampling stations in the North Sea (figure 2.1) (Kirstein et al., 2016). Three sampling sites were located in the southern North Sea, near to the coastal region of the Netherlands and Belgium (station 15), in Dover Strait, near to the English Channel (station 18), and near to the Dogger Bank (station 20), and one sampling site was located in the northern North Sea (station 24). The physical oceanographic data acquired during this research cruise have been deposited in the Data Publisher for Earth & Environmental Science PANGAEA (Gerdts and Rohardt, 2016).

Surface water samples were taken with a rosette sampler (SBE 911 plus, Sea-Bird Electronics, US). Approximately 40 l of seawater per station were pre-filtered in succession with decreasing mesh sizes (10, 3 and 0.2 µm pore size filters, 142 mm diameter, polycarbonate membrane). The 'viral fraction' seawater was concentrated using iron chloride flocculation method (John et al., 2011). Briefly, each sample was incubated with iron chloride (1 mg Fe l^{-1} filtrate) for 1 h before filtration recovering (1 µm pore size filter, 142 mm diameter, polycarbonate membrane). The filters were stored dark at 4 °C until further processing. Each 'viral fraction' seawater was subsampled and fixed with formaldehyde (2% final) to determinate the virus concentration. For this, each sample was filtered (0.02 µm pore size Anodisc filter, 25 mm) and the filter was stained with SYBR Gold prior VLP counting by epifluorescence microscopy (Patel et al., 2007).

2.2.2. DNA EXTRACTION AND SEQUENCING

The virus concentrates from each filter were resuspended overnight using modified ascorbate-EDTA buffer (0.1 M Mg_2EDTA, 0.2 M ascorbic acid, pH 6–7) according to established procedures (John et al., 2011). The resuspended solution was concentrated by centrifugation (4000 ×*g*; Sorvall RC-26 Plus, DuPont, with a GSA rotor) using Vivaspin 30 Centrifugal Concentrator (Sartorius), followed by ultracentrifugation (136,000 ×*g*, 2 h at 4 °C; Optima™ TL, Beckmann with a TLA 100.4 fixed angle rotor). The virus pellet was resuspended with 1 ml modified SM buffer (23.3 gl^{-1} NaCl, 4.93 gl^{-1} $MgSO_4$ × $7H_2O$, 50 mM Tris HCl, pH 7.5) (John et al., 2011), which was used for further DNA extraction.

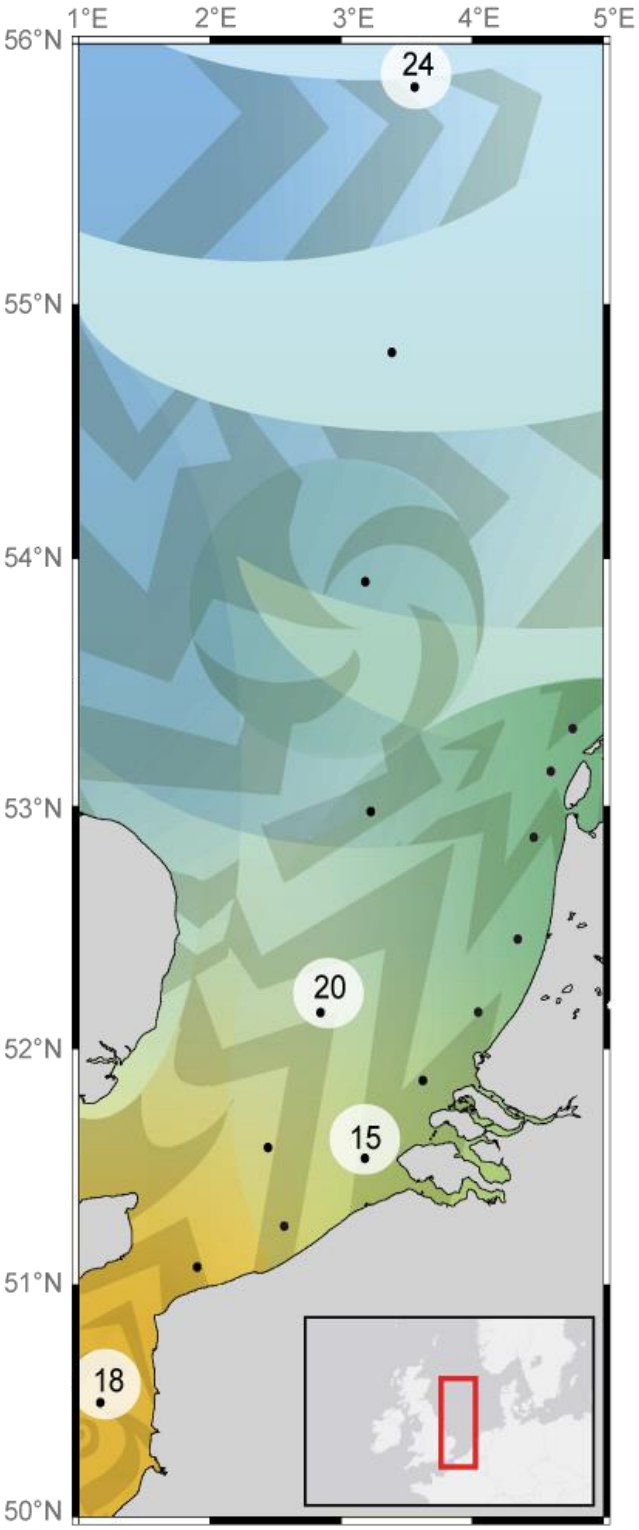


FIGURE 2. 1. Sampling site map for the four stations in the North Sea, showing the surface circulation of water masses currents (grey vectors), influence of English Channel flow (yellow), north Atlantic inflow (blue), and mixed flows zone (green). Map modified from (Kirstein et al., 2016)

DNA was extracted from the concentrated viral particles using modified CTAB method combined with phenol:chloroform method (Williamson, 2011). For this, each sample was incubated (1 h, 55 °C) with Proteinase K (100 $\mu g ml^{-1}$ final) and sodium dodecyl sulphate (0.5% final), followed by addition of 5 M NaCl and incubation (10 min, 65 °C) with NaCl/CTAB solution (70 mM NaCl 1% cetyltrimethylammonium bromide final). Subsequently, the DNA was phenol:chloroform extracted, followed by isopropanol precipitation and resuspension in 1× TE buffer (2 M Tris base, 0.2 M ethylenediaminetetraacetic acid, pH 8.5) (Thurber et al., 2009). DNA quantity and quality were determined photometrically with a Tecan Infinite©M200, NanoQuant microplate reader (Tecan, Switzerland) using the Invitrogen Quant-iT PicoGreen® dsDNA Reagent (Carlsbad, CA, USA) according to the manufacturer's instructions. Library preparation and sequencing of the virus dsDNA from each sample were done by LGC Genomics (LGC Genomics GmbH, Berlin, Germany). The total DNA per sample (50–200 ng each) was first sheared by ultrasonication (Covaris S220), follow by purification and concentration using MinElute Spin Columns (Qiagen). Library constructions were prepared using Encore Rapid DR Multiplex System (NuGen) according to the manufacturer's instructions. In detail, each library was constructed from 40 to 100 ng sheared genomic DNA, including end-repair blunt end ligation to indexed adaptors, followed by a final repair step, and purification with Ampurebeads. The purified libraries were amplified by PCR for 8 cycles using Illumina primers (5PE/3PE) with MyTaq DNA Polymerase (Bioline) and BioStabII PCR Enhancer (Sigma). The samples were sequenced on an Illumina MiSeq sequencer using 2 × 300 bp chemistry.

2.2.3. VIROME SEQUENCE ANALYSIS

All sequence reads were processed following a modified protocol established primarily by the Ocean Sampling Day (OSD) analysis group (https://github.com/MicroB3-IS/osd-analysis/wiki/Sequence-Data-Pre-Processing) (Kopf et al., 2015a). In detail, the obtained reads from each virus metagenome were first processed using BBtools (https://sourceforge.net/projects/bbmap/) to trim Illumina adaptor sequences and remove low quality and short reads with ambiguous bases. First, the bbduk tool was used under the defined set values of k-mer size (k = 23), minimum k-mer size (kmin = 11), right-trimming (ktrim = r), quality score (trimq = 20), minimum length (minlen = 50) values and trim both ends (qtrim = rl) to remove low quality reads without ambiguous nucleotides (maxns = 0); follow by bbmerge tool (default parameters). Reads were de novo assembled using Megahit (version 1.0.2, default generic metagenome parameters) under minimum contig length (−min-contig-len) set to 300 (Li et al., 2015), and IDBA-UD (version 1.1, default generic metagenome parameters) with minimum contig length (−min_contig) set to 300 (Peng et al., 2012). The assembly quality was evaluated using MetaQUAST (Mikheenko et al., 2016), based on the best recommended values of assembled contigs, contig legth and number of predicted genes per contig. After evaluation under the mentioned parameters, the assembled Megahit contigs were selected for downstream analyses. To identify and classify potential viral sequences and avoid possible host contamination, the predicted Megahit contigs were analyzed using

the VirSorter tool (default settings, Virome Decontamination mode) from iVirus (https://de.iplantcollaborative.org/de/) (Bolduc et al., 2016). VirSorter can identify viral sequences from metagenome datasets, which contigs were classified under different 'confidence' categories and also predicted their potential proteins gene sequences (Bolduc et al., 2016, Roux et al., 2015b). The viral contigs predicted with VirSorter were mapped using BWA-MEM (version 0.7, default parameters) (Li and Durbin, 2009) and the mapped sequences were converted from SAM to BAM files using SAMtools (version 1.2) (Li et al., 2009). The mean coverage per contig was calculated using genomeCoverageBed from Bedtools (version 2.23, default parameters) (Quinlan and Hall, 2010). The virus contig statistic values were calculated using BBtools (stats.sh, default parameters). All the protein sequences predicted as genes with VirSorter were used as potential virus amino acid sequences for the following analyses.

2.2.4. VIRUS GENE COMPARISON

To compare the overall virus gene diversity present in each virome, all the predicted viral protein sequences with VirSorter (section 2.2.3) were clustered into operational protein units (OPU) using UCLUST software (ID = 0.6, default parameters) (Edgar, 2010). To compare the total predicted virus protein distribution through each virome, the OPU abundance table was build using the UCLUST software followed by the visualization using the VennDiagram Package (version 1.6.0 for R version 3.3.2, default parameters) (Chen, 2013).

2.2.5. TAXONOMIC IDENTIFICATION OF VIRUS CONTIGS

To identify the viral and functional gene diversity from each virome, the predicted viral protein sequences were compared against the GenBank nonredundant protein (nr) database (January 2017) using BLASTP version 2.2.30+ (Altschul et al., 1997). BLAST hit alignments of genome sequences with e-value below 0.001 and a similarity > 50% were considered significant. The taxonomic classifications were assigned only for predicted virus contigs larger than 3 Kbps and with at least one significant BLAST hit matched a virus or phage sequence. Virus taxonomy is still under development, and the virus classification methods for metagenomic data are constantly improved (Adams et al., 2017, Simmonds et al., 2017). Thus, the taxonomy classification for each contig was done using a semi-automated approach. First, the identified genes in each contig were classified using the NCBI taxonomy ID of each BLAST hit with home written scripts. Then, the contigs larger than 3 Kbps and with at least one virus gene were selected and classified to the highest possible taxonomic level (e.g. family or genus level), based on the taxonomic information shared by the majority of genes in each virus contig. In case a contig matched several taxa and the taxonomic classification was not conclusive, the contig was manually assigned to its respective next lower taxonomic level, either family or class.

Additionally, the virus contigs were classified in groups based on their potential prokaryotic hosts. The identification was based on either the manual comparison of the virus genus names which contain host names, or on related relevant literature from each genome sequence entry if available (Mihara et al., 2016).

The contig mean read coverage (section 2.2.3) was used to estimate proportion and abundance indices of virus taxa within each virome. Rare taxonomic groups were arbitrarily defined as having a frequency below 1% within each station. The relative abundance visualization as heatmap was built using Vegan (version 2.4–4) (Oksanen et al., 2017) and Gplots Package (version 3.01) (Warnes et al., 2005).

2.2.6. IDENTIFICATION OF GENES CARRIED BY PHAGES

To analyze the virus genes and their possible effect on the host, all the virus-related proteins previously identified (section 2.2.5) were functionally classified using the UniProt Knowledgebase (UniProtKB) database (Magrane and Consortium, 2011). The functional classification was based on their ontology's biological or molecular function from comparable reference organism annotated in Swiss-Prot or TrEMBL.

We analyzed the distribution of the proteins identified in the virus was created using main coverage of the contigs in which protein was identified. The resulting matrix was z-score normalized and just values with a maximum relative abundance higher than 0.5 per sample were retained. The dissimilarity matrices of sampling sites and protein functions were computed using the Bray Curtis dissimilarity index. Subsequently, these two matrices were used to create a two-way clustering using the average hierarchical clustering method. All data analyses were performed using R Package Vegan (version 2.4–4) (Oksanen et al., 2017) and the heatmap was built using ggplot2 package (version 2.2.1, default parameters) (Wickham, 2009).

2.2.7. NUCLEOTIDE SEQUENCE ACCESSION NUMBERS

The virome sequence data have been deposited under the INSDC accession number PRJEB21210 in the European Nucleotide Archive (www.ebi.ac.uk/ena/) (Toribio et al., 2017) using the data brokerage service of the German Federation for Biological Data (www.gfbio.org/) (Diepenbroek et al., 2014). The sequence associated contextual (meta) data are Minimal Information about any (X) Sequence (MIxS) compliant (Yilmaz et al., 2011).

2.3. RESULTS

2.3.1. NORTH SEA WATER SAMPLES

Surface water samples were obtained from four North Sea stations (figure 2.1): Station 15 was located in the coastal area with riverine influence, station 18 was near to the English Channel exchange flow, station 20 near to mixing zone from Dogger Bank, and station 24 was located in the open sea of the northern North Sea, which is indirectly influenced by the Atlantic inflow (Emeis et al., 2015). The mean VLPs concentration was 2.48×10^6 VLP/ml (SD ± 0.79) and it was statistically not different between the stations. Samples, metadata and metagenomic descriptive statistics are summarized in table 2.1.

Altogether, the four viromes analyzed (> 36 million reads in total) represent diverse pelagic ocean features, which vary in salinity and temperature correlated with their proximity to land (table 2.1). The proteins from all predicted virus contigs were grouped into 114,768 OPUs (operational protein units). Their distribution decreased from station 15 and 18 in the coastal area (mean = 56,378 OPUs, SD ± 8,891.5) to station 24 in the open sea (21,060 OPUs) (figure 2.2).

From the total OPUs, 2.06% were omnipresent in all analyzed samples, 61.18% were distributed exclusively in coastal areas (station 15 and 18), 18.10% were exclusively present in the intermediate or open sea (station 20 and 24) and the remaining 18.66% of all OPUs were present in three of the four stations (figure 2.2). Around half of the OPUs were unique for each station (mean = 52.21%, SD ± 2.78), with an exception of 32.88% from station 20. Station 20 represents a mixing area between the coastal and open sea, which together with its closest station 15 (coastal area) comprise 66,268 OPUs and 15.70% of them were shared by both stations (10,402 OPUs).

In case of stations 20 and 24 (intermediate area and open sea), they comprise together 44,555 OPUs, and 12.77% of them were shared by both stations (5689 OPUs) (figure 2.2).

TABLE 2.1. . Description of virome samples and associated metadata

Station No.	Sampling date	Latitude N	Longitude E	Temperature (°C)	Salinity (PSU)	Site location	# reads	Viral contigs > 3 kbps predicted by VirSorter										
								# viral contigs	Mbps	Longest contig (Kbps)	Avg. contig length (Kbps)	N50 (Kbps)	# cat 1	# cat 2	# cat 3	# cat 4	# cat 5	# cat 6
15	08/05/2014	51.5395	3.1823	20.70	32.75	Coastal	8,904,876	719	3.66	26.18	5.10	240	90	505	124	0	0	0
18	08/06/2014	50.4967	1.1655	18.34	35.00	Coastal	16,898,592	711	3.53	34.06	4.97	246	97	498	115	0	1	0
20	08/07/2014	52.1498	2.8427	18.38	34.87	Intermediate	7,235,850	369	1.77	22.81	4.81	369	50	243	75	0	1	0
24	08/09/2014	55.8355	3.5624	18.7	34.2	Open sea	3,224,142	197	0.94	19.57	4.76	69	37	141	18	0	1	0

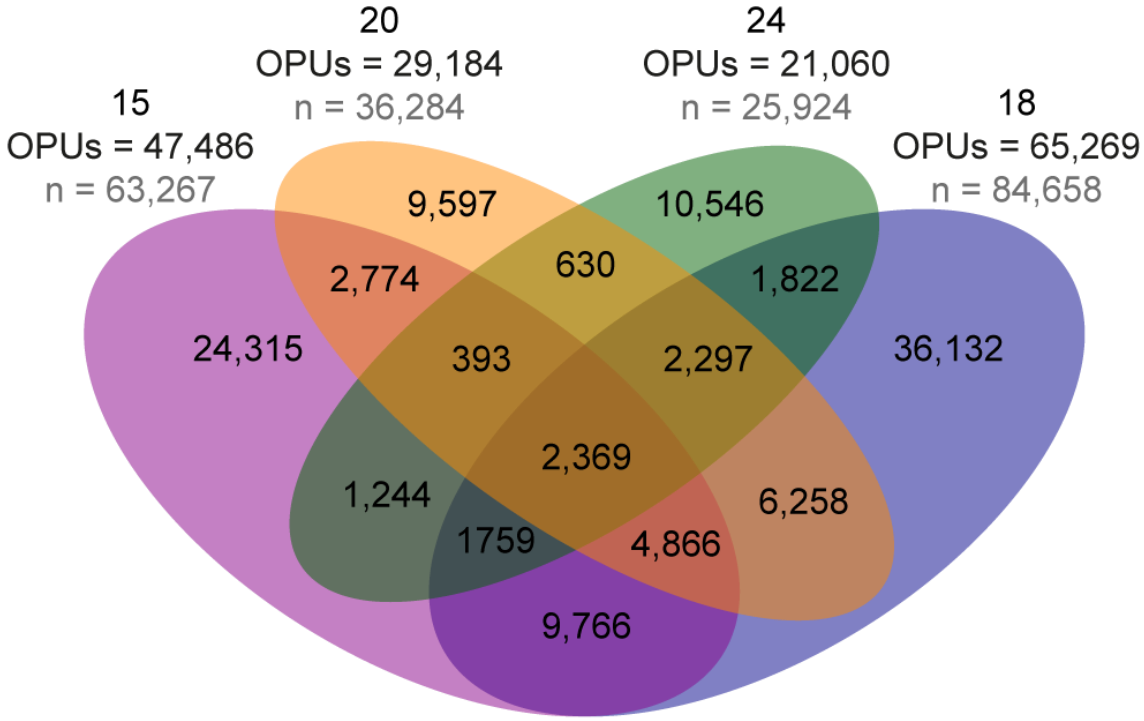


FIGURE 2.2. Distribution of predicted viral proteins according to their sample origin. The predicted virus genes per contigs (n) using VirSorter were clustered (ID=0.6) and the operational protein unit (OPU) of each predicted amino acid sequence was used to compare the virus genes distribution.

2.3.2. TAXONOMIC COMPOSITION OF NORTH SEA VIROMES

The predicted viral contigs from North Sea viromes were selected and compared against the GenBank nr database. From the 929,666 assembled contigs, 8,195 were predicted by VirSorter in one of the probabilistic categories (table A.3.1). Contigs larger than 3 Kbps were retained for further taxonomic analysis, representing 1,996 putative partial virus genomes in size range of 3 to 34.06 Kbps (Table 2.1). The distribution of these larger contigs decreased from coastal stations to open sea, where stations 15 and 18 presented more predicted viral contigs (mean = 715 contigs) than stations 20 (369 contigs) and 24 (197 contigs) (table 2.1). From them, 24.43% were unidentified and 58.49% were classified as virus contigs (table 2.2). In general, station 24 demonstrated the highest proportion of virus contigs per station and station 20 the lowest. Almost all virus contigs per station originated from bacteriophages and were dominated by unclassified phage group (mean = 62.45%, SD ± 4.56), followed by Caudovirales order (mean = 32.28%, SD ± 4.24) (table A.3.2).

To examine the taxonomic composition of the marine virus community of our samples in greater detail, we compared the abundance of classified virus contigs based on their best BLAST hit at family level (figure 2.3). Although unclassified phages were present in a high

proportion through all the stations, they were excluded for the analysis of taxonomic composition, due to their unclear taxonomy. Excluding the unclassified phages, the Myoviridae family dominated the total community in the North Sea (Myoviridae = 58.92 %; Podoviridae = 13.04 %; other viral family = 10.28 %; Siphoviridae = 8.21 %; unclassified Caudovirales = 5.66 %; Phycodnaviridae = 3.9 %). From each station, the virus community from station 15 was dominated by Myoviridae (67.49 %), followed by Phycodnaviridae, Siphoviridae, Podoviridae and other virus family in similar amounts (total mean = 7.01 %, SD ± 0.83). Stations 18 and 20 were dominated by Myoviridae (mean = 54.20 %, SD ± 3.73), followed by Podoviridae family (mean = 14.04 %, SD ± 3.08). The virus community at station 24 was dominated by Myoviridae family (45.74 %) and the Podoviridae family abundance increased to 30.84% (figure 2.3). Overall, the community distribution was not uniform throughout the North Sea. Myoviridae and Phycodnaviridae appeared to decrease from the coast to the open sea, and Siphoviridae showed differences between the station 24 (open sea) (3.61 %) and the rest of the stations (mean = 9.29 %, SD ± 1.53). On the other hand, the occurrence of Podoviridae viruses increased from 7.33 % at station 15 (coastal area) to 30.84 % at station 24 (open sea) (figure 2.3).

TABLE 2. 2. Percentage of contig's BLAST hit per station and in the North Sea virome at all stations. The contigs were first analyzed using VirSorter tool from iVirus prior BLAST comparison against GenBank nr database. The contigs larger than 3 Kbps were classified as Virus, Bacteria or other organism (Archaea or Eukaryote), contigs without BLAST hit was classified as NA (not available).

	BLAST hit	**St. 15**	**St. 18**	**St. 20**	**St. 24**	**Total station**
		Contigs (%)	Contigs (%)	Contigs (%)	Contigs (%)	Contigs (%)
virus	Unclassified phage	34.17	39.69	36.92	39.40	37.12
	Caudovirales	20.21	15.02	17.65	25.59	18.34
	Other virus family	2.88	3.17	2.40	4.42	3.03
	Bacteria	13.50	18.18	15.22	18.06	15.9
	Other organism	1.57	1.44	0.49	0.00	1.18
	NA	27.67	22.50	27.32	12.53	24.43

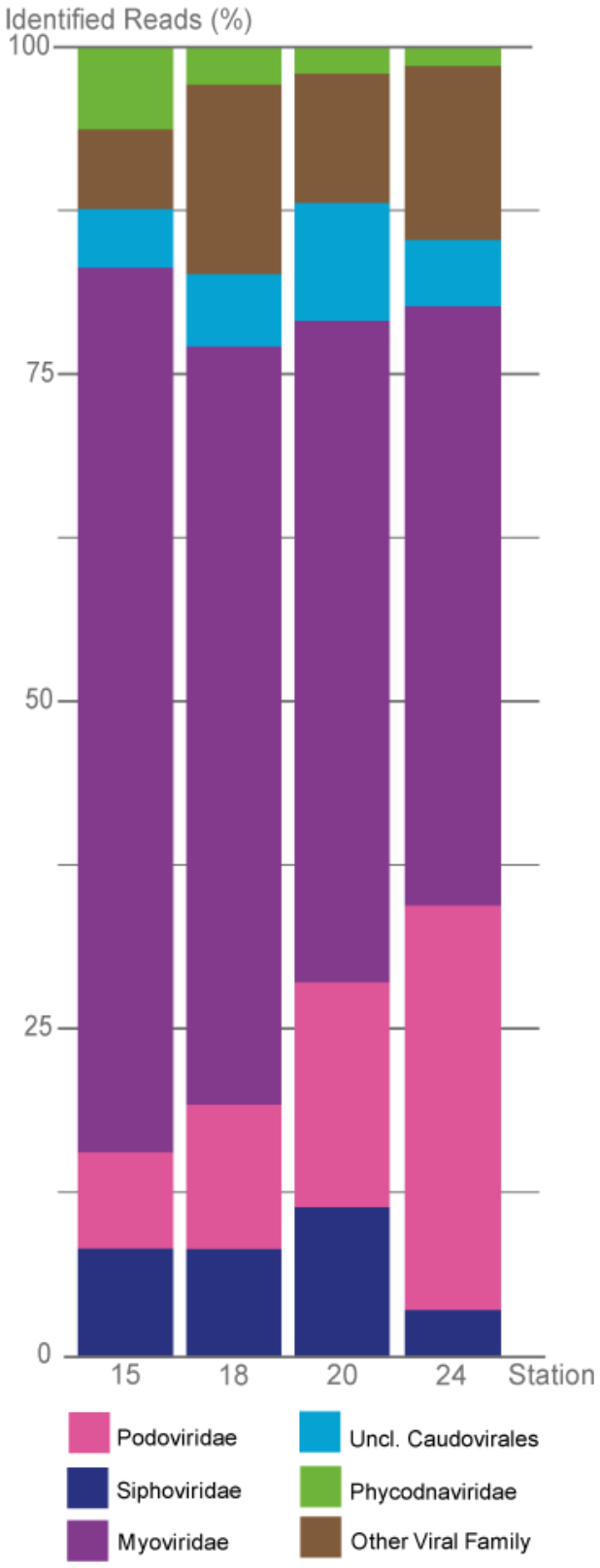


FIGURE 2.3. Percentage abundance of taxonomical classification from selected virus contigs identified by alignment-based (BLASTp) analysis (Unclassified phage excluded).

The selected virus contigs were additionally classified in 95 virus groups, based on their highest identified virus taxonomic level (Figure 2.4). Overall, the virus community at stations 15 and 18 (coastal area) was richer than station 24 (open sea), which presented the highest Margalef index values (mean = 6.75, SD ± 0.35) in comparison of station 24 (3.89) (Table A.3.3). This trend was reflected by a higher number of virus groups (Figure 2.4). In terms of abundance of virus group members, phages (unclassified phages and order Caudovirales) dominated the North Sea virome, and Myoviridae was the most diverse family of the whole virus community (Figure 2.4). The dominant virus group at all stations was the Uncultured Mediterranean Phage uvMED (mean = 47.33 %, SD ± 4.70) (figure 2.4), identified for the first time from environmental samples of the deep chlorophyll maximum in oligotrophic open ocean waters (Mizuno et al., 2013). Most of the virus taxa at all stations belonged to the rare taxonomic groups (< 1 % per station), while 27 virus groups were present in a higher proportion in at least one station (yellow-red shading, figure 2.4). Most of the identified potential phage hosts belonged to Cyanobacteria (47.85 %) or Proteobacteria Phylum (45.04 %). Cyanophages from all Caudovirales families and unclassified phages were present in the viromes, unclassified Myoviridae cyanophages were present in high proportion all over the stations. Cyanophages S-CAM8 (Myoviridae), cyanophage S-SKS1 (Siphoviridae) and unclassified Caudovirales cyanophages were present at intermediate and coastal areas (stations 15 to 20), whereas they were not identified at station 24 (open sea) (figure 2.4). Phages from all Caudovirales families and unclassified phages seem to be able to infect Gamma- and Alphaproteobacteria hosts (figure 2.4). Betaproteobacteria phages were identified as *Deftia* phage (Myoviridae) and *Methylophilales* phage (unclassified phages) groups at stations

15 and 18. At these stations, Epsilonproteobacteria phages were identified only as *Campylobacter* phage (Myoviridae and unclassified phages). The Phycodnaviridae was the eukaryotic virus family with the highest number of taxonomic groups identified, which occurrence decreased from the coast to the open sea (figure 2.4).

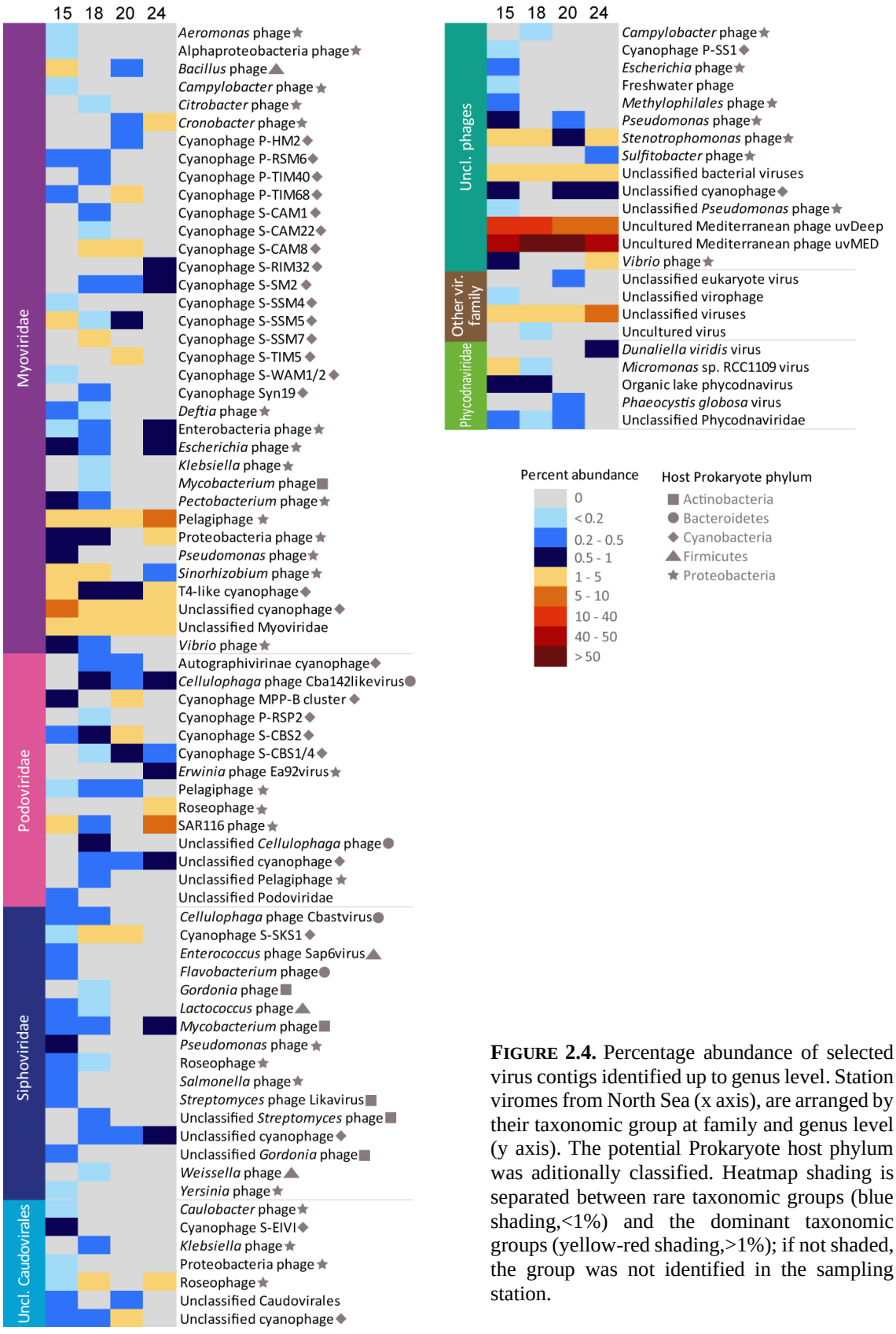


FIGURE 2.4. Percentage abundance of selected virus contigs identified up to genus level. Station viromes from North Sea (x axis), are arranged by their taxonomic group at family and genus level (y axis). The potential Prokaryote host phylum was aditionally classified. Heatmap shading is separated between rare taxonomic groups (blue shading,<1%) and the dominant taxonomic groups (yellow-red shading,>1%); if not shaded, the group was not identified in the sampling station.

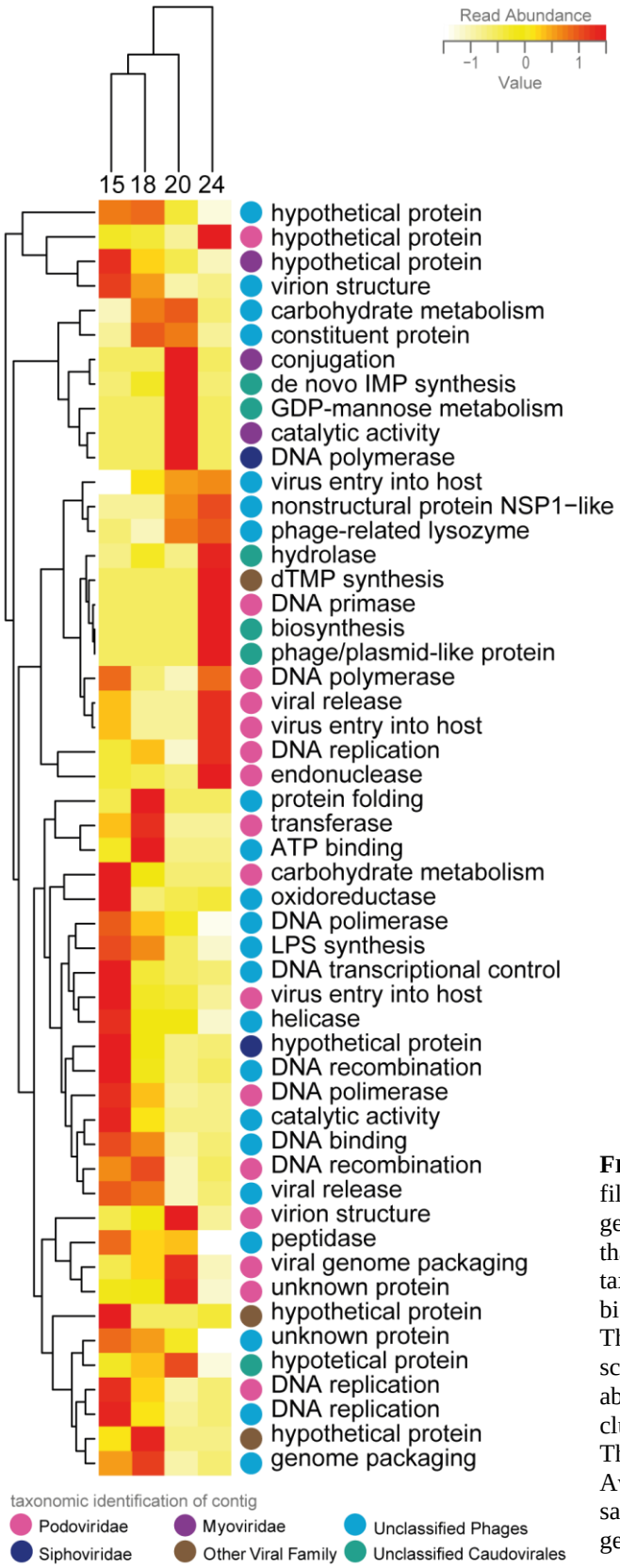


FIGURE 2.5. Heatmap comparison of the filtered protein functions from virus-encoded genes. Identified genes from contigs larger than 3 Kbps were classified in base of their taxonomical family and molecular or biological function of the encoded protein. The gene abundance was transformed to z-score, the values with a maximum relative abundance higher than 0.5 were retained for clustering by Bray Curtis dissimilarity index. The Each dendrogram represents the Average Hierarchical analysis of the sampling stations (x axis) and the classified gene function (y axis).

2.3.3. IDENTIFICATION OF GENES CARRIED BY PHAGES

To explore the distribution and function of genes carried by phages, we classified the identified virus proteins based on the taxonomy of the contig and the molecular or biological function. Contigs larger than 3 Kbps and identified as virus were selected to classify their carried genes. Overall 3,662 genes were classified in 320 groups, differentiated by their protein function and contig taxonomic group. Protein functions less abundant were omitted from further analyses (section 2.2.6). The resulting subset comprises 52 protein functional groups (figure 2.5). Most of these proteins were classified as hypothetical proteins from Myoviridae, Podoviridae and unclassified phages. All the other proteins were related to virion production processes. Other identified protein functions were related to host fitness and metabolism, which includes carbohydrate metabolism, de novo IMP synthesis and LPS synthesis (figure 2.5). These identified protein functions were more abundant at stations 15, 18 and 20.

Overall, the protein functions from stations 15 and 18 (coastal area) were more similar than station 24 (open sea) (figure 2.5). Low abundance functions related to host fitness and metabolism were evenly present at all stations (figure A.3.2), which includes genes related to photosynthesis, polysaccharide synthesis, pathogenicity, and biofilm formation.

2.4. DISCUSSION

The hydrographic features in the North Sea affect the planktonic diversity, including viruses and their hosts. This study offers the first characterization of the phage community in the North Sea by virome analysis. It reveals highly dynamic virus distribution and occurrence patterns distinguishing the coast from the open sea.

2.4.1. GEOGRAPHIC DISTRIBUTION OF VIRUS GENES

To expand our knowledge on the virus community, we first compared all predicted virus contigs from the North Sea viromes regardless of their predicted identity. This approach assumes that similar phage communities share more genetic features than dissimilar communities (Angly et al., 2006). This avoids the potential bias by relying the identification of viruses using only previously characterized sequences (Thurber, 2009). As there were potential host related reads in our assemblies, only the predicted virus contigs were used for our analysis. Additionally, we performed an analysis of all virus predicted contigs regardless of their length or identification, thereby contributing informative output to compare the stations. However, the detection of BLAST homologs using translated amino acid proteins was significantly low for short-reads sequences from natural viral communities (Wommack et al., 2008). Thus, all the predicted virus contigs were used to compare the overall virus gene distribution and only the larger (> 3 Kbps) predicted virus contigs were compared against the GenBank database to identify the viral and functional gene diversity from each virome.

The environmental variability and its impact on the marine viral and host community have been reported previously in other (meta) genomic studies (Brown et al., 2012, Hurwitz and U'Ren, 2016, Jiang et al., 2012). The predicted virus genes were translated and grouped in operational protein units (OPUs), which were not evenly distributed throughout the North Sea viromes. The OPUs distribution showed an abundance gradient from coast to open sea and only a small proportion of them were found at all stations (figure 2.2). This trend is also reflected in species and functional richness, which have been previously reported in surface water viromes (Hurwitz and Sullivan, 2013) and planktonic microorganisms (Fuhrman et al., 2008). The North Sea station with the highest number of OPUs was station 18, where the influence of the English Channel increases the temperature and salinity concentrations in the Southern Bight (Mathis et al., 2015). Station 20 revealed the lowest percentage of unique OPUs (32.88 %), which might result from the mixing of currents in the intermediate North Sea (figure 2.1).

2.4.2. TAXONOMIC CLASSIFICATION OF NORTH SEA VIRUSES

The North Sea virome exhibited high richness with a large number of distinctly identified virus groups. Although almost a quarter of the sequences appear to be uncharacterized, the remaining identified virus sequences indicated clear differences between the stations. The identified virus communities from the coastal areas possess a higher diversity and richness than the open sea community (table A.3.3). Although we discarded 15.9 % of the contigs due to probable host contamination, we obtained a mean coverage of 11.36 × of virus contigs. This is a reasonable coverage for the reliable assembling of virome reads (Roux et al., 2017). The assembly of long contigs allow a more accurate open reading frame detection and better sensitivity for similarity searches by BLAST (Wommack et al., 2008). The reduction of bacterial DNA, to increase the number of samples and their sampling size would reduce the potential technical bias.

The virus contigs from all stations were dominated by phages (Unclassified phage and order Caudovirales = 94.82 % total virus contigs, table A.3.2), which is comparable to viromes from the Atlantic Ocean (92.3 %), Mediterranean Sea (83.6 %) (Winter et al., 2014), and Chesapeake Bay (> 90 %) (Bench et al., 2007). The virus family assigned groups presented a distinct distribution pattern, which might be explained by each oceanographic feature (figure 2.3). This trend was reported in previous studies of marine viruses in surface waters (Roux et al., 2016, Suttle, 2016), where some identified taxa were represented across the surface ocean, and others seemed to be persistently rare in particular environments (Clokie et al., 2011, Roux et al., 2016). As described in other marine environments (Rastrojo and Alcamí, 2017), most of the classified viruses in our viromes belonged to the Caudovirales order (tailed phages) and were generally dominated by the Myoviridae (contractile tailed phages) family at all stations. Myoviridae have been previously identified as dominant Caudovirales family in other environments (Rastrojo and Alcamí, 2017), and in high abundance in nutrient-poor waters of the tropics (Thurber, 2009, Williamson et al., 2008) as well as in surface waters of the Pacific Ocean (Hurwitz and Sullivan, 2013).

The abundance of Siphoviridae (long noncontractile tailed phages) decreased from station 15 (coastal area) to station 24 (open sea), while Podoviridae (short tailed phages) increased from coast to open sea (figure 2.3). The Podoviridae family has been identified as the most abundant member of Caudovirales of the Baltic Sea (Zeigler Allen et al., 2017).

In general, a small number of taxa dominate the community while the low abundant community is highly diverse and represented by several different taxa (Sogin et al., 2006). This supports the hypothesis of the presence of an unknown community of low abundant phages in the North Sea
(< 1 % virome). Previous studies suggest that prokaryotes exhibit a preference for particular environments, which in turn means that phages can also develop biogeographic patterns (Clokie et al., 2011). However, there is a debate about the existence of such biogeographic patterns in phages (Dennehy, 2014, Thurber, 2009). Although phage abundance and richness were higher in the coastal area, the diversity was similar throughout the North Sea (figure 2.4, table A.3.3). Interestingly, a richness and diversity gradient were observable in Phycodnaviridae and Siphoviridae families from lower to higher latitudes. The difference in virus community richness seems to reflect the oceanographical conditions of the region and the decrease in host abundance from southern to northern North Sea regions (Brandsma et al., 2013).

The abundant viruses might be permanently active in the ecosystem and the rare taxonomic groups, which carry a diverse gene pool, might be activated under certain environmental and host conditions (Breitbart and Rohwer, 2005, Jousset et al., 2017). The virus community analyzed in this study represents a snapshot of the community in time, where these rare viruses (< 1 % virome) might have a different adaptive strategy to proliferate or present a restrictive response caused by unfavorable environmental conditions (Skopina et al., 2016). Besides the large number of rare taxonomic groups in the North Sea viromes there were a few taxonomic groups with high abundance (figure 2.4). High abundant viruses are active as long as the host population is susceptible to infection (Breitbart and Rohwer, 2005). Thus, the virus community in the North Sea would be especially dynamic and further studies are needed to elucidate how lytic and lysogenic viruses interact with their hosts under a spatial-temporal scale.

Throughout the North Sea, cyanophages were evenly present in high proportions in all phage families at each station (figure 2.4). Most known cyanophages are lytic and affect the host diversity over monthly timescales (Rosario and Breitbart, 2011), which is ecologically important due to the role of their host as predominant primary producers in oligotrophic ocean areas and their role in global carbon cycling (Clokie et al., 2011). The dominant cyanophage family in the North Sea was Myoviridae, followed by Podoviridae. These cyanophage families are ubiquitous and present in high abundance in several coastal water and open ocean environments (Huang et al., 2015). In contrast to these studies (Huang et al., 2015, Sabehi et al., 2012), cyanophages S-CBP2, S-TIM5 S-SSM5 and P-SSM7 did not occur at all stations, and cyanophage P-SSM7 was only detected at station 15 (figure 2.4). Podoviridae cyanophage MPP-B group was detected in coastal and intermediate zones.

The cyanophage S-EIVI represents a distinct evolutionary group of cyanophages, which is widespread and relatively abundant in aquatic systems (Chénard et al., 2015). The polar freshwater cyanophage S-EIVI was only detected at station 15 (figure 2.4), which is the area with the highest riverine influence.

The major clades of heterotrophic bacteria in surface marine environments include *Roseobacter*, SAR 11, SAR116 clades, among others (Giovannoni et al., 1990, Luo and Moran, 2014). Phages able to infect these abundant marine bacteria were also identified. *Roseobacter* clade is present in marine environments worldwide and represents an abundant component of the bacterial community in the North Sea (Giebel et al., 2010, Luo and Moran, 2014). Roseophages (Siphoviridae, Podoviridae and unclassified Caudovirales) were identified in coastal and open sea areas (figure 2.4). Roseo- and pelagiphages have been identified in high abundance in other marine viromes (Hwang et al., 2016). SAR11 phages as pelagiphages (Myoviridae and Podoviridae) were identified at all stations (figure 2.4). The SAR11 phages are highly represented in several marine viromes (Zhao et al., 2013). Phages of SAR116 clade were identified both in coastal and open sea areas (figure 2.4). These abundant Podoviridae phages were formerly found in ocean surfaces (Kang et al., 2013). *Cellulophaga* phages (Podoviridae and Syphoviridae) were identified in low abundance throughout the North Sea, as reported in other marine ecosystems (e.g. Holmfeldt et al., 2013). Previous studies of the virus diversity in the North Sea were performed with culture-dependent methods (Duhaime et al., 2016, Wichels et al., 1998). Wichels et al. (1998) identified cultivable *Pseudoalteromonas* phages within several years in a narrow geographical region of the North Sea. Although stations 20 and 24 were near the sampling cruise routes of that study, we did not identify these phages in the North Sea viromes. This supports the hypothesis of natural barriers causing a reduced geographical distribution of *Pseudoalteromonas* phages (Wichels et al., 2002). In general, viral and bacterial communities are highly dynamic (Breitbart, 2012). Further investigations are needed in order to have a better overview of the spatial and temporal variability of the North Sea viruses.

2.4.3. IDENTIFICATION OF GENES CARRIED BY VIRUSES

Until now, at least half of predicted phage genes in viral (meta) genomes are predicted as hypothetical proteins without an assigned function (Clokie et al., 2011, Hurwitz et al., 2014, Klumpp et al., 2013). Therefore, despite the growth of databases it is still difficult to predict a molecular or ecological function (Clokie et al., 2011). Further genome and proteome characterization would be necessary to elucidate the function of these phage proteins. Additionally, a high proportion of identified virus-encoded genes in our viromes were involved in genome packaging and virion production of different phage groups. A phage population composed by a higher diversity of virion production might increase the opportunities to recognize more host lineages, which is reflected of a higher proportion of these functions in variable genome regions from viromes (Mizuno et al., 2014).

Horizontal gene transfer by viruses is widespread among microorganisms, significantly affecting organism survival, competition, interspecific interactions, and even the community assembly (Dennehy, 2014). These foreign genes might represent critical steps in host metabolism such as photosynthesis processes. Their expression can affect the success of certain marine phages (Breitbart, 2012). Thus, the possible ecological implications of virus gene transfer on the host and virus community are important to analyze (Dennehy, 2014). The virus communities in coastal areas (stations 15 and 18) carry the most similar virus protein functions (figure 2.5), which might be related to a closer exchange of water masses between these stations. The indirect influence of the Atlantic inflow at station 24 might cause the observed differences concerning the other stations. In case of identified low abundance virus genes, foreign genes involved in host metabolism and fitness as well as genes related to host virulence as biofilm formation, LPS synthesis, pathogenesis and antibiotic response were identified (figure A.3.2). These virulence genes can be carried to the host, which can be major factors in the emergence of new diseases (Dennehy, 2014). However, there is still a debate about the role of phages in transferring antibiotic resistance to their hosts in clinical and environmental strains (Anand et al., 2016, Enault et al., 2017, Subirats et al., 2016). More information is needed before drawing conclusions in the North Sea and other marine environments. Other protein functions, for instance photosynthesis, have been identified in Myoviridae, Podoviridae and unclassified phages (figure A.3.2). These genes come from cyanophages and their presence in surface seawater has been reported previously in other marine environments (Hurwitz et al., 2013b). Probable cyanobacterial hosts such as *Synechococcus* play a major role in carbon fixation in marine environments and their phages acquire important genes to adapt to new environments (Sullivan et al., 2005), which boost the host metabolism and subsequent viral propagation in surface waters (Hurwitz and U'Ren, 2016). The transfer of these functional

genes would increase the fitness of low abundant host taxa and might open novel metabolic pathways.

Further studies are necessary to extend the knowledge not only of the gene inventory of the virus community, but also of the relationship with their hosts possibly due to the expression of their carried genes. The application of other approaches based on metatranscriptomics and -proteomics or culture dependent techniques would complement the virome analyses to increase our knowledge of the ecological role of marine phages in the North Sea.

2.4.4. CONCLUDING REMARKS

Despite the high ecological and evolutionary importance of phages, the methodological approaches are still under development and fundamental questions of virus ecology are difficult to investigate (Hurwitz et al., 2014). Viromes have significantly contributed to expand our understanding of phage–host interactions and characterize environmental virus communities. Although a large sequence fraction still remains functionally unknown (Hurwitz et al., 2016, Hurwitz et al., 2014), this provides a basis for comparisons with future (meta)genome studies in marine environments.

With this study, we were able to show that the virus community in the North Sea is dominated by phages, with significant variations in different hydrographical zones. Despite the low sampling number, a gradient between coastal and open sea communities was observed. The present study offers the first insights into the virus community of this highly dynamic area. The environmental differences seem to influence occurrence and diversity of viruses as shown by patterns of different taxonomic groups as well as their functional genes.

A small fraction of the identified phage community appears in high abundance, which might be active and evenly present throughout the stations. However, the richness of each area is driven by a large number of low abundant virus taxa, which carry a diverse pool of genes. These results represent a snapshot of the virus community in time, and further studies are needed to investigate the spatiotemporal variation of the North Sea viruses.

Supplementary data to this article can be found online and in chapter A.3.

Chapter 3

LOOKING FOR THE HIDDEN:

Characterization of Lysogenic Phages in Potential Pathogenic *Vibrio* Species from the North Sea

Alexa Garin-Fernandez[a,b*], Antje Wichels[a]

[a] Department of Microbial Ecology, Biologische Anstalt Helgoland, Alfred-Wegener-Institute Helmholtz-Center for Polar and Marine Research, Helgoland, Germany
[b] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[c] Jacobs University Bremen gGmbH, Bremen, Germany
[*]To whom correspondence should be addressed.
Tel: +49(4725)819-3233; Fax: +49(4725)819-3283; Email: alexa.garin@awi.de

Submitted 28 August 2019 to
Marine Genomics

Keywords: *Vibrio parahaemolyticus* / *Vibrio cholerae* / Prophage / Temperate phage / Caudovirales / Tailed phage / Vibriophage

Abstract: The incidence of potentially pathogenic *Vibrio* species is correlated with the increase of the seawater temperature in Europe. Despite their importance, little is known about the trigger factors of potential outbreak-causing strains in this region. As prophages may compose a major reservoir of virulence traits in marine ecosystems, this study aims to identify and characterize the genomes of lysogenic *Vibrio* phages exemplarily from the North Sea. Therefore, 31 isolates from potentially pathogenic *Vibrio* species from the North Sea were screened for inducible prophages with mitomycin C. From them, one *V. cholerae* isolate and 40 % *V. parahaemolyticus* isolates carried inducible prophages. Three lysogenic phages were selected for genomic characterization. The phage vB_VpaM_VP-3212 (unclassified Myoviridae) has a genome with a length of 36.81 Kbp and 55 CDS were identified. This lysogenic phage of *V. parahaemolyticus* contains genes related to replicative transposition mechanism, such as transposase and mobile elements similar to Mu-like viruses. The phage vB_VpaP_VP-3220 (Podoviridae, unclassified *Nona33virus*) has a genome length of 58,14 Kbp and contains 63 CDS. This *V. parahaemolyticus* phage probably uses a headful (pac) packaging replication mechanism. The phage vB_VchM_VP-3213 (unclassified Myoviridae) has a genome with a length of 41 Kbp and 63 CDS were identified, including integrase and Xer system for lysogenic recombination. This lysogenic phage of *V. cholerae* has similar genomic features as lambdoid phages. Although none pathogenicity genes were identified, their similarity among other phage genomes indicates that these phages can affect the development of pathogenic *Vibrio* strains in marine environments.

3.1. INTRODUCTION

Tailed phages are probably the most abundant viruses, comprising the majority of the marine virus community (Wommack and Colwell, 2000, Ackermann, 2007). Tailed phages compose the order Caudovirales, which is divided into five families: Podoviridae (short noncontractile tail), Siphoviridae (long noncontractile tail), Myoviridae (long contractile tail), Ackermannviridae (long contractile tail, formerly known as the *Vi1virus* genus), and Herelleviridae (long contractile tail, created after Spounavirinae reclassification) (Krupovic et al., 2011, Adriaenssens et al., 2018, Barylski et al., 2018). Caudovirales are characterized to inject their dsDNA into the host bacteria upon infection to follow the lytic or lysogenic cycle (Krupovic et al., 2011, King et al., 2012). In the lytic cycle, the phage redirects the host metabolism to produce new particles and lyses the host cell to release them (Weinbauer, 2004, Bondy-Denomy and Davidson, 2014). In the lysogenic cycle, the phage genome is integrated into the host genome as a prophage, and the host cell replicates as a lysogen. Prophage genes are repressed to maintain cell viability until the prophage induction triggers the lytic cycle (Breitbart, 2012, Bondy-Denomy and Davidson, 2014). Prophage infection can alter the lysogen phenotype to increase the bacterial fitness (Paul, 2008, Breitbart, 2012). These changes include the production of toxins, biofilm formation, sporulation, and antibiotic resistance (Chen et al., 2015, Fortier and Sekulovic, 2013). Therefore, prophages can modify the physiological properties of the lysogen bacteria, which can develop on new pathogenic strains in marine environments (Fortier and Sekulovic, 2013, Breitbart, 2012, Paul, 2008). Approximately half of the sequenced bacterial genomes present prophage-like elements (Paul, 2008), which can increase to almost 70 % of sequenced genomes from certain species, as described in *Vibrio* genomes (Castillo et al., 2018).

Vibrio is a diverse bacterial genus that comprises more than 100 species, only 11 of them are human pathogens (Miyoshi, 2013). Representative pathogenic species of gastrointestinal diseases are *V. parahaemolyticus* and *V. cholerae* (Miyoshi, 2013, Thompson et al., 2004). Strains of *V. parahaemolyticus* and non-O1/O139 *V. cholerae* can cause severe gastroenteritis, while serogroups O1 and O139 of *V. cholerae* have caused epidemics of cholera (Chakraborty et al., 2000, Thompson et al., 2004). Due to the increase of the sea surface temperature, their niche is expanding to new areas (Baker-Austin et al., 2016, Baker-Austin et al., 2013). In Northern Europe, the number of *Vibrio* infections is increasing rapidly, including human illnesses caused by *V. parahaemolyticus* and non-O1/O139 *V. cholerae* (Martinez-Urtaza et al., 2010, Baker-Austin et al., 2017, Roux et al., 2015a). Potentially pathogenic *Vibrio* species in the North Sea have been monitored recently and previous studies demonstrated that the abundance of *V. parahaemolyticus* increase during the summer in the German Bight, caused by higher water temperature and low salinity conditions (Oberbeckmann et al., 2012, Oberbeckmann et al., 2011b). Several *Vibrio* species have been identified in the water column as well as attached to microplastic particles found in the North Sea (Kirstein et al., 2016).

Due to the increase of reported *Vibrio*-associated illnesses in Europe, these potentially pathogenic *Vibrio* species are of enhanced importance (Hartnell et al., 2019, Roux et al., 2015a).

While tailed phages with lytic cycle have been characterized in several potentially pathogenic *Vibrio* species (Hazen et al., 2010), the knowledge of *Vibrio* prophages with a lysogenic strategy is still limited. For instance, the lysogenic phage CP-T1 of *V. cholerae* carry a pathogenicity island and can infect both classical and El Tor biotypes of *V. cholerae* (Comeau et al., 2012, Guidolin et al., 1984, O'Shea and Boyd, 2002). Furthermore, other lysogenic phages have been characterized in *V. parahaemolyticus*, which are able to infect several potentially pathogenic strains of this species (Alanis Villa et al., 2012, Lan et al., 2009). Although lysogenic phages have been identified in several pandemic and clinical *Vibrio* strains (Hazen et al., 2010, Faruque and Mekalanos, 2012, Sakib et al., 2018, Bastías et al., 2010, Iida et al., 2001, Lan et al., 2009), in particular potentially pathogenic *Vibrio* strains from marine environments have not been studied deeply.

As *Vibrio* prophages may compose a major reservoir of virulence traits in marine ecosystems (Castillo et al., 2018), there is a need to investigate lysogenic phages as one possible factor affecting the development of human pathogenic *Vibrio* species in the marine environment. Hence this study aims to identify and characterize the genomes of lysogenic *Vibrio* phages exemplarily in marine environmental strains of the North Sea. For this, potentially pathogenic *Vibrio* species isolated from the North Sea were screened for inducible prophages. Two lysogenic phages of *V. parahaemolyticus* and one lysogenic phage of non-O1/O139 *V. cholerae* were selected to apply a novel pipeline to characterize their genome without sequencing the lysogen bacterial genome. In this study, we present the genome structure and the predicted biological functions of the novel Caudovirales phages vB_VpaM_VP-3212, vB_VpaP_VP-3220 and vB_VchM_VP-3213.

3.2. MATERIAL AND METHODS

3.2.1. ISOLATION AND IDENTIFICATION OF LYSOGEN *VIBRIO*

For the isolation of *V. parahaemolyticus*, seawater samples were taken from the North Sea from July 31st to August 5th, 2014, during the RV Heincke HE430 (Gerdts and Rohardt, 2016). Filtered seawater and particles collected with a Neuston net were analyzed as described in detail by Kirstein et al. (2016), to isolate planktonic and synthetic polymer attached *Vibrio* strains. Additionally, *V. parahaemolyticus* isolates were obtained with the raw-seawater enrichment method. This method was primarily established by Moebus (1980) and modified for potentially pathogenic *Vibrio* species. The surface seawater samples were collected with a rosette sampler (SBE 911 plus, Sea-Bird Electronics, US). Approximately 1 l of raw seawater per station was incubated with alkaline peptone water (10 g/l peptone, pH 8.5, 10 PSU final) at 37 °C in the dark. After overnight incubation, the selective enrichment was plated on CHROMagar™Vibrio (MAST Diagnostica GmbH, Germany) (Di Pinto et al., 2011) with a Spiral-plater (easySpiral® Dilute; Interscience, France), and incubated at 37 °C for 24 h in the dark. The grown colonies were compared with the colony coloration typical for *V. parahaemolyticus* according to the manufacturer's instruction. All mauve colored single colonies were transferred onto marine broth agar (Oppenheimer and Zobell, 1952) with reduced salinity (MB-50% = 16 PSU) and serially re-inoculated to pure cultures.

All selected *Vibrio* isolates were further subjected to MALDI-TOF analysis and to PCR amplification of species-specific and virulence-associated genes for conclusive identification at the species level. MALDI-TOF is a rapid and reliable technique to distinguish species of the genus *Vibrio* (Dieckmann et al., 2010, Hazen et al., 2009). MALDI-TOF spectrum of each isolate was obtained and analyzed as described by Erler et al. (2015). Each spectrum was compared against the Bruker mass spectra and VibrioBase libraries (Erler et al., 2015), as well as the *V. parahaemolyticus* strains obtained in the cruise HE430.

Since the differentiation value of branch clusters is not defined for *Vibrio* species, this cut-off value was arbitrarily defined (Malainine et al., 2013). Thus, the species cut-off of the principal component analysis (PCA) from MALDI-TOF spectra comparison was defined for *V. parahaemolyticus* at a distance level below 850 (Cheng et al., 2015).

For the PCR identification method, the bacterial DNA of each isolate was extracted with lysozyme/SDS lysis and phenol/chloroform, followed by precipitation with isopropanol (Oberbeckmann et al., 2011a). Species-specific PCR for *toxR* genes, as well as PCR screening for thermostable direct hemolysin (*tdh*) and *tdh* related hemolysin (*trh*) in assigned *V. parahaemolyticus*, and cholera toxin gene *ctxA*, serotypes O139 and O1 for assigned *V. cholerae*, were performed as described by Kirstein et al. (2016).

V. parahaemolyticus RIMD 2210633 (German Collection of Microorganisms and Cell Cultures, DSMZ) was used as a PCR positive control. After identification, the *V. parahaemolyticus* isolates were stored in liquid nitrogen (VibrioNet, Erler et al., 2015).

3.2.2. LYSOGENIC PHAGE INDUCTION

To identify inducible prophages from *V. parahaemolyticus* and *V. cholerae*, the isolates were screened with the mitomycin C induction method (Raya and H'bert, 2009, Clokie and Kropinski, 2009) modified for potentially pathogenic *Vibrio spp.* For this, the *V. parahaemolyticus* strains were incubated in 96-well plates (flat bottom with cover; Corning™Costar™) with MB-50% broth supplied with 0.1 % agarose, at 30-37 °C and shaking. The culture was monitored using a TECAN infinite M200 reader (Switzerland), and mitomycin C (0.25 $\mu g ml^{-1}$ final) was added on each tested well after 2 h of incubation. Each prophage induction was performed in eight replicates of 200 µl each, negative controls were performed in duplicate with MB-50% broth instead of mitomycin C. After 6 h of incubation, all replicates of each strain were collected by filtration (0.2 µm, 25 mm diameter, polycarbonate membrane syringe filter, Corning™) and stored at 4 °C in the dark. In order to identify possible phage particle release, the induced samples were fixed with formaldehyde (2 % final), and the virus-like-particles (VLP) were counted using a wet-mount method with SYBR Gold and epifluorescence microscopy (Cunningham et al., 2015).

The *V. parahaemolyticus* strain VN-3212, VN-3220 and *V. cholerae* strain VN-3213 were selected to characterize their inducible prophages. Each isolate was incubated in 600 ml MB-50% broth at 37 °C and shaking. As performed in the first screening, mitomycin C (0.25 $\mu g ml^{-1}$ final) was added at an optical density of 0.1 – 0.2 (600 nm), monitored using a Biophotometer (Eppendorf). Upon prophage induction, the culture was first filtrated (0.2 µm pore size, 47 mm diameter, polycarbonate membrane), followed by concentration of phage particles with centrifugation (4,000 ×*g*; Sorvall RC-26 Plus, DuPont, with a GSA rotor) with a Vivaspin 30 Centrifugal Concentrator (Sartorius). This concentrate was subjected to DNAse I treatment using Baseline-ZERO™ DNase (Epicenter) according to the manufacturer's instructions, and final ultracentrifugation (136,000 ×*g*, 2 h at 4 °C; Optima™ TL, Beckmann with a TLA 100.4 fixed angle rotor). The concentrated phage elution pellet was resuspended with 500 µl of modified SM buffer (23.3 $g l^{-1}$ NaCl, 4.93 $g l^{-1}$ $MgSO_4 \times 7H_2O$, 50 mM Tris HCl, pH 7.5) (John et al., 2011), and stored at 4 °C in the dark. The strain VN-3218 without induction was used as a negative control, *V. cholerae* strain AC53 together with its lytic phage ICP1 (Seed et al., 2011), and the lysogenic *Vibrio* phage PV94 induced from the *V. vulnificus* strain VN-0094 (Pryshliak et al., 2014) were used as positive experimental controls.

3.2.3. DNA EXTRACTION AND SEQUENCING

The DNA from each lysogenic induction was extracted with a modified CTAB and phenol:chloroform extraction method (Williamson, 2011, Garin-Fernandez et al., 2018). For this, the concentrated phage elution was incubated with proteinase K (100 µgml^{-1} final) and sodium dodecyl sulfate (0.5 % final) for 1 h at 55°C, terminated by addition of sodium chloride (5 M final), and incubated with NaCl/CTAB solution (70 mM NaCl, 1 % cetyltrimethylammonium bromide final) for 10 min at 65 °C. The DNA was phenol:chloroform extracted, precipitated with isopropanol and resuspended in 1× TE buffer (2 M Tris base, 0.2 M ethylenediaminetetraacetic acid, pH 8.5). The concentration and purity of the extracted DNA were determined photometrically via a Tecan Infinite©M200, NanoQuant microplate reader (Tecan, Switzerland) with Invitrogen Quant-iT PicoGreen® dsDNA Reagent (Carlsbad, CA, USA) according to the manufacturer's instructions.

Library preparation and sequencing were performed at LGC Genomics (LGC Genomics GmbH, Berlin, Germany). The DNA was sheared by ultrasonication (Covaris S220), purified and concentrated with MinElute Spin Columns (Qiagen). The library was constructed with the Encore Rapid DR Multiplex System (NuGen) according to the manufacturer's instructions. The purified libraries were amplified by PCR for 17 cycles using Illumina primers (5PE/3PE) with MyTaq DNA Polymerase (Bioline) and BioStabII PCR Enhancer (Sigma). Sequencing was performed on an Illumina MiSeq V3 sequencer using 2 × 300 bp chemistry.

3.2.4. GENOME SEQUENCE ANALYSIS

The sequence reads were first processed using BBtools (version 35.14, https://sourceforge.net/projects/bbmap) to decontaminate and discard low-quality reads before assembling them into contigs. In detail, the bbduk tool was used to define k-mer size (k = 23), hamming distance (hdist = 1), right and left trimming (qtrim = rl), screen in reverse-complement sequences (rcomp = t), remove low quality (trimq = 20, maq = 20), short reads (minlen = 100), trim Illumina adapter and *Escherichia virus* Phi-X174 genome (NC_001422.1) reads. The bbnorm tool was used to error-correct and normalize the over-coverage reads from the phage genome by selection of reads with an average depth of 100 × (target = 100) and discard low coverage reads (min = 5). The resulting reads were assembled with SPAdes (version 3.11.1, --only assembler) (Nurk et al., 2013). The contigs were then filtered using VirSorter tool (default settings, Virome Decontamination mode) from iVirus (https://de.iplantcollaborative.org/de/) (Bolduc et al., 2016), followed by identification of (pro)phage sequences with PHASTER (https://phaster.ca) (Arndt et al., 2016, Zhou et al., 2011). The phage genome sequence was mapped using BWA-MEM (version 0.7, default parameters) (Li and Durbin, 2009) and converted to BAM files with SAMtools (version 1.2) (Li et al., 2009). Coverage values of the genome were analyzed using the total reads with Qualimap (version 2.2.1) (García-Alcalde et al., 2012, Okonechnikov et al., 2016). The annotation was firstly performed using the annotation software Rast (http:// http://rast.nmpdr.org/rast.cgi, default parameters) (Aziz et al., 2008, Overbeek et al., 2014), followed by manual comparison with Prokka (viruses and bacteria databases, default parameters) (Seemann, 2014), and Metavir from VirSorter tool (default parameters) (Roux et al., 2014, Bolduc et al., 2016). The manual curation and software outcome comparison was performed with Artemis (version 17.0.1) (Rutherford et al., 2000) to obtain the final annotated genome in flat format. Genome organization, visualization, and GC content were performed with Artemis (version 17.0.1) (Rutherford et al., 2000). Overview of the data analysis is provided in figure 3.1.

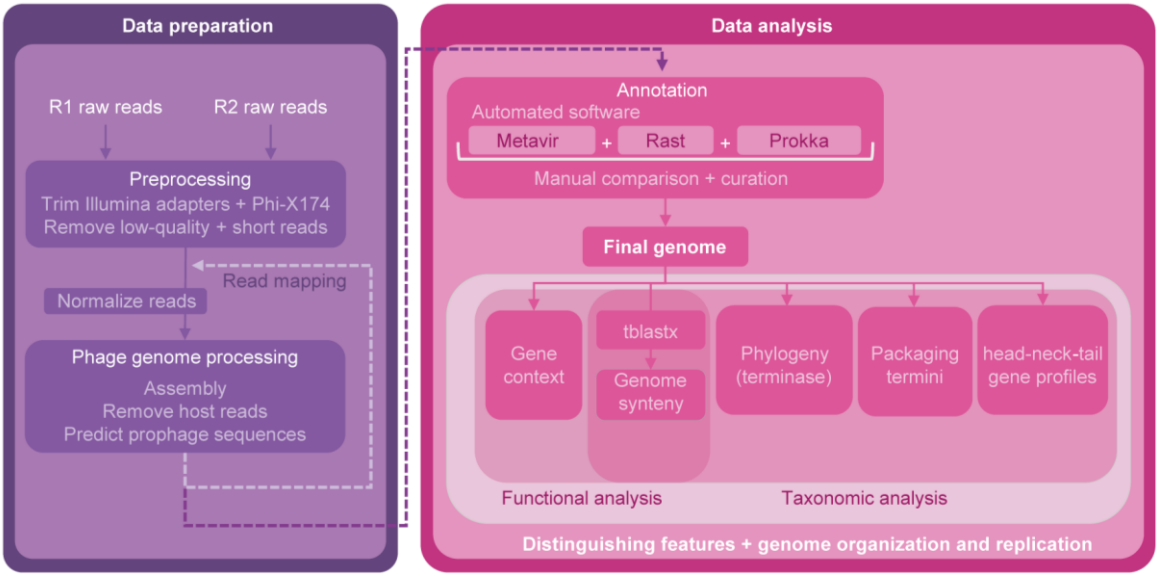


FIGURE 3. 1. Workflow overview of data preparation and analysis for genome characterization used in this study. Details of each step are described in section 3.2.4.

For the taxonomic classification, the head-neck-tail modules of the phage genomes were first analyzed using Virfam Caudovirales classifier (http://biodev.extra.cea.fr/Virfam), which compares the protein profiles and gene context against representative tailed bacteriophages of its database to classify them into Caudovirales family clusters (Lopes et al., 2014). Furthermore, the phage genomes were compared with tBLASTx (version 2.2.30+) against the virus genomes available in the National Center of Biotechnology Information (NCBI, refseq database). Genome comparison with the Sequence Demarcation Tool (SDT version 1.2) was performed to pairwise genetic identity determine species or genus level (Muhire et al., 2014). If applicable, the genus level was based on the demarcation criteria of the most similar genus (King et al., 2012). As recommended by the International Committee on Taxonomy of Viruses (ICTV), the genus demarcation criteria were performed with EMBOSS Stretcher (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher) (McWilliam et al., 2013) and CoreGenes 3.5 (http://binf.gmu.edu:8080/CoreGenes3.5) (Zafar et al., 2002). Due to discrepancies between the ICTV virus taxonomy 2018 release (https://talk.ictvonline.org/taxonomy), and the taxonomy databases of the NCBI (https://www.ncbi.nlm.nih.gov/taxonomy) and European Nucleotide Archive (ENA, www.ebi.ac.uk/ena/), the taxonomic names used in this study correspond to the nomenclature used by NCBI/ENA. Synteny comparison within the most similar virus genomes was visualized with Easyfig (version 2.2.3) (Sullivan et al., 2011). The phage termini and packaging mechanism were determined with PhageTerm (Garneau et al., 2017). Additionally, the identified terminase encoding genes were selected to analyze their phylogenetic relationship. For this, the product sequences identified as terminase were aligned together with the terminase large subunit sequences of *Vibrio* phages and model lysogenic phages (> 138 amino acids, refseq database, MUSCLE algorithm), which corresponds to genome sequences of Podoviridae (YP_001648943.1, BAT31856.1,

NP_049511.1, BAT32029.1, BAA94156.1, YP_024418.1, YP_009140132.1, YP_009152754.1, YP_009153029.1, YP_008239661.1), Myoviridae (NP 536651.1, YP_008766840.1, Q9T1W6.1, YP_007877524.1, YP_004251070.1) and unclassified phages (YP 007877435.1). The phylogeny was inferred using Neighbor-joining (uniform rates among sites, pairwise deletion of gaps, 1,000 bootstrap replicates) in MEGA (version 10.0.1) (Kumar et al., 2018). All used databases were updated to their latest version from May 2018, and BLAST hit alignments with e-value below 0.001 and a similarity greater than 50 % were considered significant.

3.2.5. NUCLEOTIDE SEQUENCE ACCESSION NUMBERS

Lysogenic phages were named based on Kropinski et al. (2009) nomenclature proposal. The nucleotide sequences and their associated (meta)data were submitted with the data submission service from the German Federation for Biological Data (www.gfbio.org/) (Diepenbroek et al., 2014). The associated (meta)data were based on the Minimal Information about any (X) Sequence (MIxS) compliant (Yilmaz et al., 2011).

The phage genomes vB_VpaM_PN-3212, vB_Vpa_PN-3220 and vB_VchM_PN-3213 have been deposited in ENA (Toribio et al., 2017) under the INSDC accession number PRJEB33505.

3.3. RESULTS AND DISCUSSION

The incidence of reported *Vibrio* illnesses in northern Europe is correlated with the increase of the seawater surface temperature in this region (Vezzulli et al., 2016). Despite the importance to follow up the increase of human pathogenic *Vibrio* in Europe, little is known about the trigger factors of potential outbreak-causing strains in this region (Vezzulli et al., 2016, Hartnell et al., 2019, Roux et al., 2015a). Most ecological studies of pathogenic bacteria have focused on the presence and activity of these bacterial strains *in situ*, which might overlook the role of their phages in the disease persistence and spread (Breitbart et al., 2005). The interaction within prophages can also influence the emergence of pathogenic clones, as described in *V. cholerae* (Faruque and Mekalanos, 2003, Banerjee et al., 2014). As lysogenic phage infection is known as a key factor for the development of new pathogenic strains in marine environments (Fortier and Sekulovic, 2013, Breitbart, 2012, Paul, 2008), lysogenic phages of potential pathogenic *Vibrio* species from the North Sea were characterized in this study. The genomic characterization of these inducible phages can provide valuable information concerning the potential to transfer and carry relevant machinery of bacterial pathogenesis in the marine environment.

3.3.1. IDENTIFICATION OF *VIBRIO* AND INDUCTION OF PROPHAGES

During the cruise HE430, 30 *V. parahaemolyticus* isolates and one *V. cholerae* strain were obtained near the coastline of the North Sea (figure 3.2A, station 1 to 15) in a temperature range from 19.64 to 21.53 °C and at salinities between 30.62 and 33.51 PSU (Gerdts and Rohardt, 2016). All but two isolates were classified to the secure species level by MALDI-TOF spectra comparison against the VibrioBase database (Erler et al., 2015 and table A.4.1). Two isolates were identified to a probable species level (table A.4.1). All species identification results of *V. parahaemolyticus* were additionally verified by positive species-specific PCR (section 3.2.1). Virulence-associated genes encoding *tdh* and *trh* hemolysins were not identified in any of these isolates (table A.4.1). The *V. cholerae* isolate was classified to the secure species level by MALDI-TOF, and positive for the species-specific PCR. This isolate was neither tested positive for cholera toxin gene (*ctxA*) nor positive for the pandemic serotypes O139/O1 (table A.4.1). Despite all these 31 *Vibrio* isolates do not carry virulence-associated genes, these isolates were identified as species related to human pathogen strains. Thus, these *V. parahaemolyticus* and *V. cholerae* strains isolated from the North Sea were considered as potentially pathogenic *Vibrio* species.

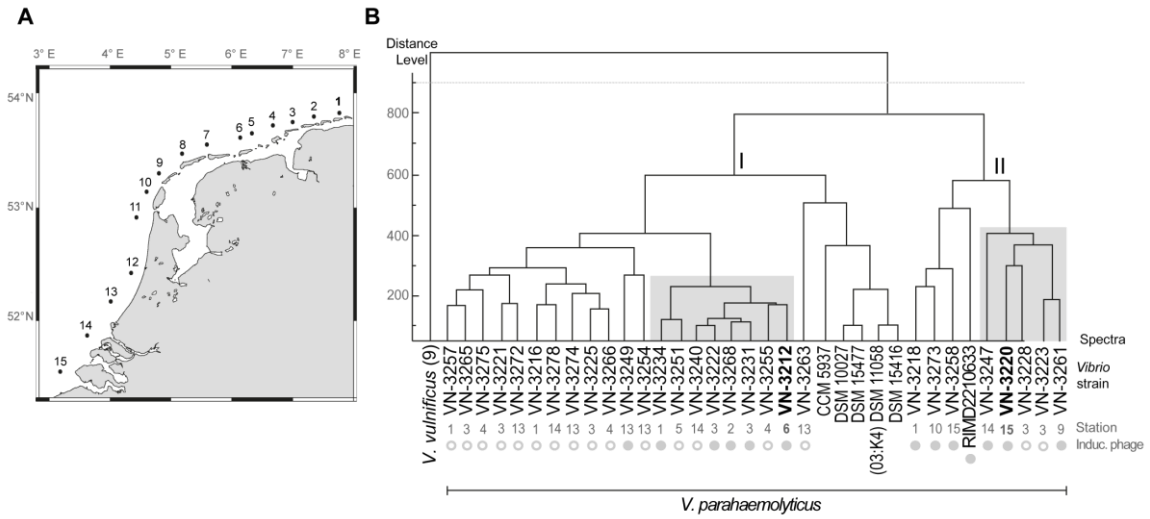


FIGURE 3. 2. Sampled site coastal map from research cruise HE430 (A) and principal component analysis (PCA) dendrogram generated by MALDI Biotyper mass spectra for all the *V. parahaemolyticus* strains isolated in this campaign and type strains from the Czech collection of microorganisms (CCM), German collection of microorganisms and cell cultures (DSM) available in the Bruker mass spectra library (B). The grey shadowed clusters correspond to the isolate VN-3212 and VN-3220 cluster groups. The black dots (●) represents successful prophage induction with mitomycin C, while the empty dots (○) represents non-lysogen strains under the tested conditions.

Based on PCA MALDI-TOF spectra, the *V. parahaemolyticus* isolates (distance level < 850) were differentiated into two separated clusters (distance level < 600) (figure 3.2B). The first cluster is composed of three main branches, which include five type strains from the Czech and German collection of microorganisms. The second cluster which includes the representative *V. parahaemolyticus* strain RIMD2210633 is composed of two main branches. The PCA dendrogram did not show a clear differentiation at the subspecies level. No pattern was obvious related to either sample site, isolation method, or presence of inducible prophage.

The identified *Vibrio* isolates were tested for inducible prophages using mitomycin C (section 3.2.2). From the 31 tested isolates, 12 *V. parahaemolyticus* (40 % of this species, figure 3.2B) and one *V. cholerae* carried inducible prophages (41.94 % of total, table A.4.1). Most of lysogen *Vibrio* isolates were from stations 1 to 4, and station 13 and 15 (table A.4.1). The induction rate was in line with other marine bacteria (reviewed by Breitbart, 2012) and *in silico* identification of prophages in *V. parahaemolyticus* genomes (~ 40.3 %, Castillo et al., 2018). In contrast, other studies have shown a much lower proportion of lysogen strains, such a former screening study of *V. parahaemolyticus* isolates from food poisoning patients (9.2 %, Muramatsu and Matsumoto, 1991). Besides, a recent *in silico* study of 1,567 *Vibrio* genomes identified at least one complete prophage in more than 70 % of *V. cholerae* genomes (Castillo et al., 2018). Therefore, it was not surprising that the only *V. cholerae* isolated during the HE430 (station 15) also carries one inducible phage (table A.4.1).

The induced *Vibrio* prophages from the North Sea were tested against 153 potential *Vibrio* host strains, including environmental and clinical strains from *V. parahaemolyticus, V. cholerae, V. mimicus and V. vulnificus*. However, the search for a potential host showing lytic response failed (data not shown). The lack of suitable phage propagation has been observed already in other studies investigating inducible tailed phages, including *V. campbellii* and potentially pathogenic *V. vulnificus* (Lorenz et al., 2016, Pryshliak et al., 2014). Presumably this might either be due to the general narrow host range of Caudovirales phages from *Vibrio* species (Matsuzaki et al., 1992) or additional prophages that protect the lysogen bacteria from superinfection (Lorenz et al., 2016).

The prophages from *V. parahaemolyticus* VN-3212 (cluster I, station 6), *V. parahaemolyticus* VN-3220 (cluster II, station 15), and *V. cholerae* VN-3213 (station 15) were induced with mitomycin C (section 3.2.2). As expected, the bacterial cell number decreased after induction in the tested replicates from each isolate (figure A.4.1) compared to the control replicates (VN-3212 = 40.43 %; VN-3213 = 13 %; VN-3220 = 34.64 %). The cell density of the *V. parahaemolyticus* induced cultures decreased significantly (Mann-Whitney test, VN-3212 P = < 0,001; VN-3220 P = 0,003). This difference was triggered by prophage induction, identified by the presence of VLP in the induced culture (VN-3212 = 1.8×10^8 VLP/ml; VN-3220 = 3.8×10^8 VLP/ml). The decrease in cell density in the induced *V. cholerae* cultures was not significant (VN-3213 P = 0,066), but VLP were identified in the induced culture (VN-3213 = 6×10^6 VLP/ml). Each of these *Vibrio* strains carry a single complete prophage sequence (table 3.1). The *V. parahaemolyticus* VN-3212 and the *V. cholerae* VN-3213 carried a Myoviridae phage, while the strain VN-3220 carried one Podoviridae phage. The complete DNA sequence of the lysogenic phages was determined with a shotgun sequencing approach and *de novo* assembling (section 3.2.4).

TABLE 3. 1. Summary of the features from phage genomes analyzed in this study. The tailed phages were induced from potentially pathogenic *Vibrio* from the North Sea. Features include lysogen *Vibrio* strain and station source from cruise HE430, assembling coverage of genome (×), sequence length (Kbp), GC-content (GC %), number of annotated CDS (# CDS), possible packaging mechanism, possible integration mechanism identified in the phage genome, taxonomic classification, as well as most similar phage genus (synteny, refseq genome).

Phage name	Lysogen strain (source)	Assembling coverage (×)	Length (Kbp)	GC (%)	# CDS	Packaging mechanism	Integration lysogen chromosome	Family	Closest genus
vB_VpaM_VP-3212	*V. parahaemolyticus* VN-3212 (station 6)	3,038.95 (SD ± 1,548.24)	36.81	45.47	55	Possibly Mu-like	Transposase	Myoviridae	*Muvirus*
vB_VpaP_VP-3220	*V. parahaemolyticus* VN-3220 (station 15)	3,297.3 (SD ± 837.91)	58.14	43.56	63	headful packaging mechanism (Headful pac)	Integrase	Podoviridae	*Nona33virus*
vB_VchM_VP-3213	*V. cholerae* VN-3213 (station 15)	6,204.75 (SD ± 998.04)	41	46.25	53	Headful pac; Possible Lambdoid phage	Integrase | recombinase	Myoviridae	-

3.3.2. *VIBRIO* PHAGE VB_VPAM_VP-3212

The genome of vB_VpaM_VP-3212 has a sequence coverage of 3,038.95 × (SD ± 1,548.24), with a length of 36,805 bp and GC content of 45.47 %. In the genome, 55 CDS were identified, of which 53 are located on the forward strand and the remaining 2 on the reverse strand (figure 3.3A). From the 55 identified CDS, 22 encode for hypothetical proteins, 4 encode proteins for transcription regulation, 24 encode proteins for virion assembly and lysis, 2 encode proteins for mobile elements, the remaining 3 CDS encoded proteins with DNA replication functions (figure 3.3A, table A.4.2). Host-pathogenicity related genes were not identified in the lysogenic phage genome. Annotation descriptions, as well as the CDS length and transcription positions, are summarized in figure 3.3A, and a detailed description of each gene is present in table A.4.2.

The phage vB_VpaM_VP-3212 shows homology with the *Vibrio* phage martha 12B12 (NC_021070.1, 66.6 % pairwise identity) and Enterobacteria phage Mu (NC_000929.1, 47.3 % pairwise identity) (Morgan et al., 2002) (figure 3.4A). The packaging system was not identified with PhageTerm, the general tool for fast and accurate determination of phage termini and packaging mechanism (Garneau et al., 2017). However, the function can be assumed at least since the coverage mapping spectrum was similar to other Mu-like read mapping patterns (Kauffman et al., 2018a). As the host termini of Mu is asymmetric, the short read length from Illumina sequencing can bring problems to determine the short end fragment size (~ 50 bp) systematically (Garneau et al., 2017). In addition, the small number of well-studied Mu-like phage species (King et al., 2012) and the lower similarity within the reference Enterobacteria phage Mu genome (47.3% pairwise identity) might result in a method bias to analyze novel phage genomes with reference-based tools. Based on the similarity between the packaging system and coverage mapping spectrum of Mu-like phages, the phage vB_VpaM_VP-3212 might amplify its linear genome through replicative transposition. Annotated virion structure proteins in the vB_VpaM_VP-3212 genome include major capsid (CDS37), portal (CDS32), terminase (CDS31), adaptor (CDS40), tail competition (CDS42), neck (CDS41), and sheath (CDS44) proteins (table A.4.2). Due the presence of sheath protein encoding genes, the Myoviridae phages can be distinguished from other Caudovirales families (Lopes et al., 2014, King et al., 2012). Moreover, the genome context and their assigned function of the virion structure proteins were used to identify the phage with Virfam Caudovirales classifier as Myoviridae of Type1 (cluster 8, figure A.4.2A). Virfam classification delineates four types of morphological subfamilies, the most abundant Type1 yields 10 clusters (Lopes et al., 2014). The Type1 cluster 8 contains Siphoviridae and Myoviridae phages, differentiated by absence and presence of a sheath protein, respectively (Lopes et al., 2014). This cluster is composed by several phages related to Mu (Lopes et al., 2014). In addition, the terminase protein phylogeny

shown that the terminase encoding gene of the phage vB_VpaM_VP-3212 clustered together within terminase from Mu-like phages (figure 3.5).

The vB_VpaM_VP-3212 genome contains protein encoding genes related to genetic recombination and gene repression, similar to the lysogenic lifestyle of Mu-like viruses (Oakey et al., 2002) These protein encoding genes include transposase (CDS03), mobile elements (CDS04), and repression genes such as transcriptional regulators (CDS01-02, table A.4.2). Several protein encoding genes from Mu-like related sequences of Gram-negative bacteria were also identified (table A.4.2). In addition, the vB_VpaM_VP-3212 genome contains 34 annotated genes similar to prophages genes of *Shewanella oneidensis* MR-1. The bacteria *S. oneidensis* MR-1 contain two phylogenetically distinct Mu-like prophages, which share syntenic regions with the reference Enterobacteria phage Mu and other Mu-like phages from *Haemophilus influenzae* (Heidelberg et al., 2002). The *H. influenzae* Rd genome contains a Mu-like prophage sequence denominated FluMu (Fleischmann et al., 1995, Morgan et al., 2002). The prophage FluMu contains only one identified tail fiber gene and lacks an invertible segment, these features differ with the described in the Enterobacteria phage Mu (Morgan et al., 2002). Interestingly, this structure is present also in the vB_VpaM_VP-3212 genome (table A.4.2) and puts this phage in closer correlation to prophage FluMu. Moreover, five protein encoding genes similar to prophage FluMu were annotated in the vB_VpaM_VP-3212 genome (table A.4.2). Besides, a few annotated genes were similar to genes from *Shigella flexneri* bacteriophage V (unclassified Myoviridae), *Escherichia* phage D108 (unclassified *Muvirus*), Bacteriophage K139 (Myoviridae, *Hp1virus*), and *Vibrio* phage Kappa (Myoviridae, *Hp1virus*) (table A.4.2). Although no pathogenicity genes were identified, further approaches are necessary to analyze the role of this transposable phage on development of virulence traits in infected bacterial strains.

Although the close similarities within the *Muvirus* genus, several genomic features cannot achieve the overall demarcation criteria to consider the phage vB_VpaM_VP-3212 as part of this genus. The Mu-like phages have a lysogenic cycle, integrate into the host chromosome at random locations (replicative transposition), have a linear genome of size around 40 Kbp, and lack of DNA polymerases (Maniloff and Ackermann, 1998). The pairwise identity value between the vB_VpaM_VP-3212 and the Enterobacteria phage Mu genomes (47.3 %) was much lower than between the two current *Muvirus* members (Enterobacteria phage Mu and Enterobacteria phage SfMu, 87.5 % pairwise identity). Moreover, the characteristic structure of protease/scaffolding/major capsid encoding genes (*I*, *Z* and *T* genes) present in the Enterobacteria phage Mu genome (Morgan et al., 2002), differ from the protease/major capsid encoding gene structure present in the vB_VpaM_VP-3212 genome (CDS35-36, figure 3.3A). Although this is a characteristic structure, this is not conserved in all Mu-like phages e.g. *Vibrio* phage martha 12B12 (Hulo et al., 2015). Although the *Vibrio* phage martha 12B12 can be considered as a Mu-like phage, several key genomic features differ between them (Hulo et al., 2015). The phage vB_VpaM_VP-3212 share several features within the *Vibrio* phage martha 12B12, including the mentioned protease/major capsid encoding gene structure (figure 3.4A). Although *Vibrio* phage martha 12B12 is taxonomically identified as an "Unclassified Myoviridae", no further information except for the genome annotation is available. Currently, *Vibrio* phage martha 12B12 is part of a proposal for the new virus family Saltoviridae, composed by more than 20 virus species (Hulo et al., 2015, Kropinski et al., 2018).

Based on the overall shared genomic properties, we assume that phage vB_VpaM_VP-3212 is an unclassified Myoviridae phage. This phage might be involved in the enrichment of the genetic diversity of marine *V. parahaemolyticus* due horizontal gene transfer, as described in other lysogenic Myoviridae phages from the pandemic O3:K6 strain (Lan et al., 2009).

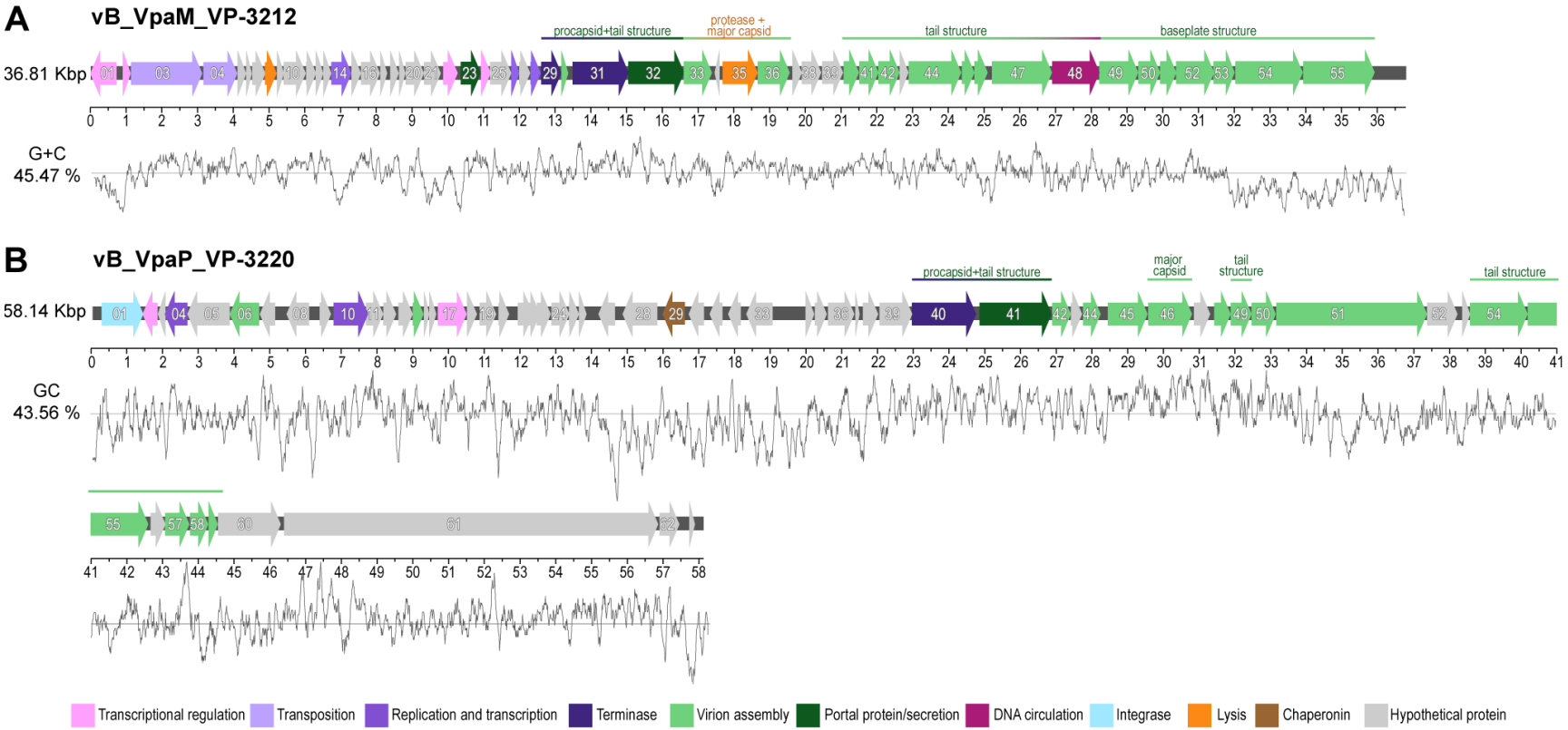
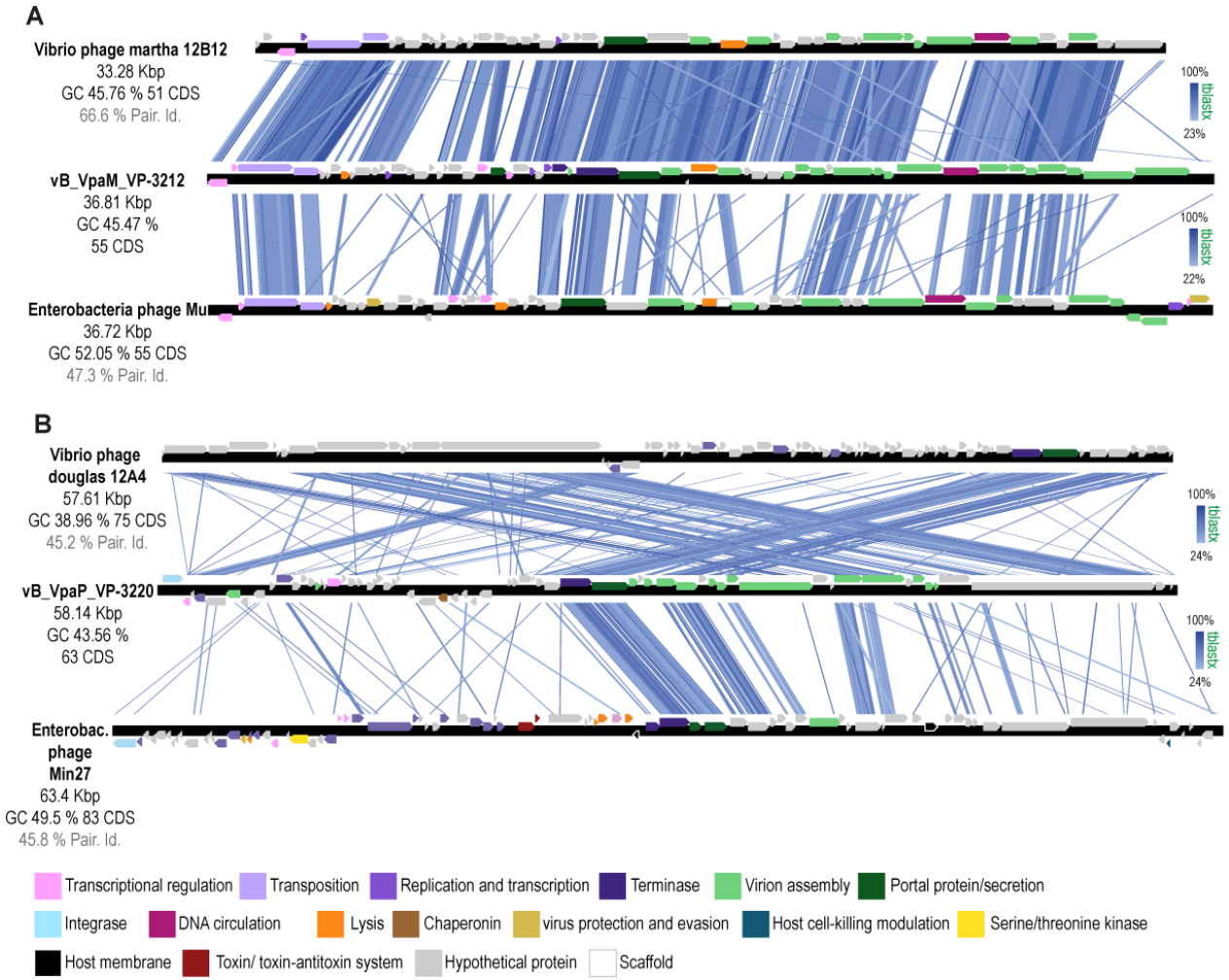


FIGURE 3. 3. Diagram of vB_VpaM_VP-3212 (A) and vB_VpaP_VP-3220 (B) phage genomes. Predicted CDS are shown with arrows pointing in the direction of transcription, the colors were assigned accordingly to the possible protein function, as shown in the figure. Sequence length (Kbp), GC-content (GC %) are shown above each genome profile.



.

FIGURE 3. 4. Genome synteny plot. Full-genome comparison based on tBLASTx similarity between vB_VpaM_VP-3212 (A) and vB_VpaP_VP-3220 (B) and their two most similar phage genomes. To facilitate the genomic comparison, the genomes of *Vibrio* phages martha 12B12 and douglas 12A4 are visualized in reverse direction. The sequence length (Kbp), GC-content (GC %), number of predicted CDS (# CDS), and percentage of pairwise identity between the studied and reference phage genome comparison (Pair. Id.), are shown above each genome profile. The colors of the protein encoding genes are marked accordingly to their annotated replication functions, as shown in the figure.

70

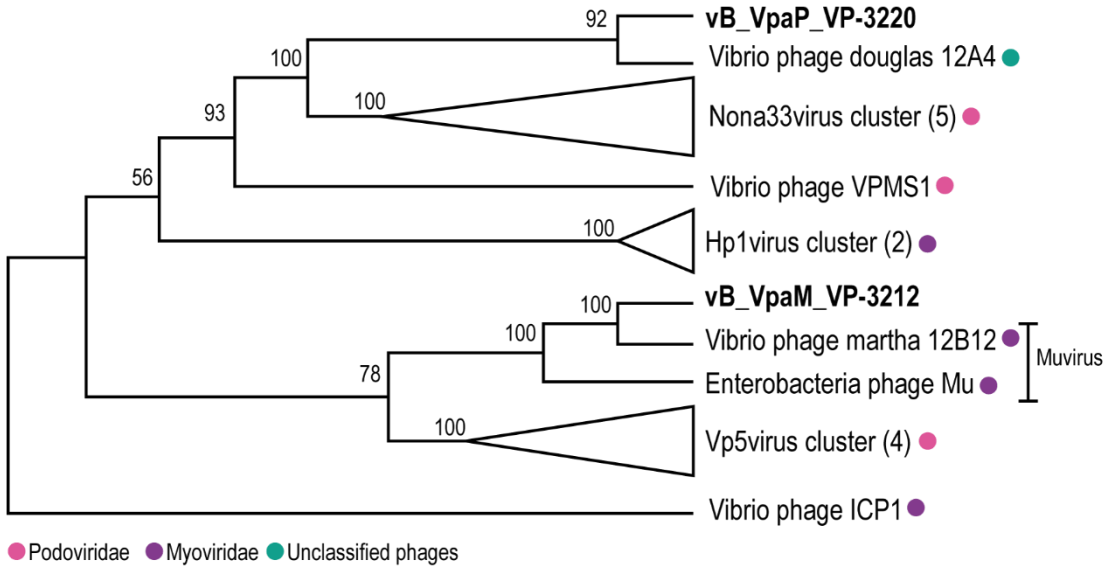


FIGURE 3. 5. Neighbor-joining tree of large terminase subunit amino acid sequences. The consensus tree was built with 1,000 bootstrap replicates. Numbers at the nodes represent percent bootstrap support where unlabeled nodes had bootstraps of > 50 %. Terminase sequences of vB_VpaP_VP-3220 and vB_VpaM_VP-3212 were aligned with similar terminase sequences of Podoviridae (pink) Myoviridae (violet) and unclassified phages (green).

3.3.3. *VIBRIO* PHAGE VB_VPAP_VP-3220

The genome of vB_VpaP_VP-3220 has a sequence coverage of 3,297.3 × (SD ± 837.91), with a length of 58,137 bp and GC content of 43.56 %. Within the genome 63 CDS were identified, of which 49 are located on the forward strand and the remaining 14 on the reverse strand (figure 3.3B). From the 63 identified CDS, 40 encode for hypothetical proteins, 2 encode proteins for transcription regulation, 17 encode proteins for virion assembly, 1 encodes for integration into the bacterial genome, the remained 3 encode proteins with DNA replication functions (figure 3.3B, table A.4.3). Host-pathogenicity related genes were not identified in the lysogenic phage genome. Annotation descriptions, as well as CDS length and transcription positions, are summarized in figure 3.3B, and a detailed description of each gene is present in table A.4.3.

The phage vB_VpaP_VP-3220 shows similarity with the *Vibrio* phage douglas 12A4 (NC_021068.1, 45.2 % pairwise identity) and Enterobacteria phage Min27 (NC_010237.1, 45.8 % pairwise identity) (figure 3.4B). Based on PhageTerm analysis, the phage vB_VpaP_VP-3220 probably uses a headful (pac) packaging mechanism during its replication. This mechanism is characteristic for linear genomes but circularly permuted structure (King et al., 2012). The presence of headful (pac) and integrase have been also described in other lysogenic Podoviridae, such as Enterobacteria phage P22 (Tye et al., 1974, Groth and Calos, 2004). Besides, virion structure proteins used for taxonomic classification with Virfam include major capsid (CDS46), portal (CDS41), and terminase (CDS40) proteins (table A.4.3). Two head-completion proteins were identified as adaptor protein Ad3 (CDS49),

and the head closure protein Hc3 (CDS54, table A.4.3). An additional gene involved in tail structure was identified. This gene encodes a phage host specificity protein J of unclassified Lambda-like phages, including members of *Nona33virus* (CDS55, table A.4.3). This protein is related to virion attachment to the host membrane and further DNA injection (Roessner and Ihler, 1984). Procapsid and tail protein encoding genes were identified in the same genome context as found in the Enterobacteria phage Min27 genome (CDS40-41, figure 3.4B). The genome synteny comparison (figure 3.4B) and terminase protein phylogeny (figure 3.5) showed strong relationship between vB_VpaP_VP-3220 and *Vibrio* phage douglas 12A4 (unclassified phage), as well as *Nona33virus* phages. Although *Vibrio* phage douglas 12A4 is taxonomically identified as an "Unclassified phage", no further information except for the genome annotation is available. Lopes et al. (2014) classified the *Vibrio* phage douglas 12A4 genome (Virfam name 12A4) as Podoviridae of Type3 with Virfam Caudovirales classifier (figure A.4.2B). The P22-like phages clustered together as Virfam Type3, which includes members of *Nona33virus* genus (Lopes et al., 2014). In similar manner, the genome context and the virion structure encoding genes of the vB_VpaP_VP-3220 genome is similar to phages from Podoviridae of Type3 cluster (figure A.4.2B, table A.4.3).

Most of the genomic features from the phage vB_VpaP_VP-3220 are similar to Podoviridae phages, specifically within the *Nona33virus* genus. The Podoviridae phages are characterized by virions with short and non-contractile tail structure (King et al., 2012). This tail structure is formed by the portal, and two head-completion (adaptor and head-closure) proteins (Lopes et al., 2014), which were identified in the vB_VpaP_VP-3220 genome (figure A.4.2B, green). Due the integrase encoding gene identified in the vB_VpaP_VP-3220 genome (CDS01, table A.4.3), this lysogenic phage might recombine within the bacterial chromosome as described in integrases from lysogenic tailed phages (Groth and Calos, 2004). Additionally, other genes related to lysogeny, such as prevent host-death, were annotated in the vB_VpaP_VP-3220 genome (CDS34, table A.4.3). In the annotated vB_VpaP_VP-3220 genome, 10 structural protein encoding genes were similar to *Nona33virus* phage genes (table A.4.3). In detail, three annotated protein encoding genes were similar to genes from bacteriophage 933W (CDS40-41, CDS46), one gene was similar to Enterobacteria phage Min27 (CDS40), and nine genes were similar to Stx phage genes (CDS01, CDS04, CDS40, CDS45-46, CDS50, CDS54-55, CDS57, table A.4.3). Although Shiga toxin encoding genes were not identified in the vB_VpaP_VP-3220 genome, its similarity with Stx phages (Shiga toxin encoding phages) may indicate a potential role in pathogenicity.

Five approved species of *Nona33virus* genus are included in the ICTV virus taxonomy 2018 release (renamed as *Traversvirus*). The species number increases to more than 50 in the NCBI taxonomy database, including the type species bacteriophage 933W (Adriaenssens et al., 2017, Plunkett et al., 1999), Enterobacteria phage Min27 (Su et al., 2010), and several Stx phages (Sato et al., 2003). The genome alignment comparison between unclassified *Nona33virus* (e.g. Stx2-converting phage 86, NC_008464.1) and bacteriophage 933W (NC_000924.1) have a similarity below 45 % pairwise identity and 70 % shared proteins (data not shown), which is lower than the proposed for this genus (above 80 %). The pairwise similarity of unclassified *Nona33virus* and other genomic features are comparable to the phage vB_VpaM_VP-3212.

Based on the overall genomic properties, the phage vB_VpaP_VP-3220 is assigned as unclassified *Nona33virus*. The phages from *Nona33virus* are closely related to the develop of pathogenicity of *E. coli*, which contribute to the production of Shiga toxin and antibiotic resistance genes in nonpathogenic strains (Gamage et al., 2003, Colavecchio et al., 2017). Although no pathogenicity genes were identified in vB_VpaP_VP-3220, this phage is a potential contributing factor to disseminate these genes in potentially pathogenic *Vibrio* strains.

3.3.4. *VIBRIO* PHAGE VB_VCHM_VP-3213

The genome of vB_VchM_VP-3213 has a sequence coverage of 6,204.75 × (SD ± 998.04), with a length of 40,999 bp and GC content of 46.25 %. Within the genome 53 CDS were identified, of which 48 are located on the forward strand and the remaining 5 on the reverse strand (figure 3.6). From the 53 identified CDS, 24 encode for hypothetical proteins, 2 encode proteins for transcription regulation, 19 encode proteins for virion assembly and lysis, 2 encode for mobile element proteins, 1 encodes for integration into the bacterial genome, 3 encode for proteins with ATPase activity, the remaining 2 encode proteins with DNA replication functions (figure 3.6, table A.4.4). In addition, the genome context and similarity of 39 annotated proteins with Rast show a close relation with *Vibrio vulnificus* CMCP6 chromosome I (NC_004459) (table A.4.4). Host-pathogenicity related genes were not identified in the lysogenic phage genome. Annotation descriptions, as well as CDS length and transcription positions, are summarized in figure 3.6, and a detailed description of each gene is present in table A.4.4.

The phage vB_VchM_VP-3213 shows similarity with the Myoviridae phages *Burkholderia* phage phiE202 (NC_009234.1, 45.1 % pairwise identity) and *Vibrio* phage VHML (NC_004456.1,
45.2 % pairwise identity) (Oakey et al., 2002) (figure 3.7). The phage vB_VchM_VP-3213 probably uses Headful (pac) packaging during its replication. This system is present in Myoviridae phages with linear but circularly permuted structure, such as phages P1 and T4 (Garneau et al., 2017, King et al., 2012). The virions structure proteins identified in the vB_VchM_VP-3213 genome include major capsid (CDS03), portal (CDS49), tail competition (CDS09), tail tube (CDS18) neck (CDS08), and sheath (CDS17) proteins (table A.4.4). In detail, one encoding region related to tail and baseplate structure was identified (CDS08-23, figure 3.6). This region encodes for four tail structure proteins similar to P2-like phages (CDS10, CDS13-14, CDS17, CDS22), including a similar host specificity protein J from Lambda-like phages (baseplate assembly protein CDS13, table A.4.4). Several protein encoding genes were similar to the *Vibrio* phage VHML.
For instance, structural proteins encoding genes showed similarities to genes from phage VHML and other Myoviridae *Vibrio* phages (CDS03, CDS13, CDS17), as well as phage protein were similar to the ORF51 from *Vibrio* phage VHML (CDS36), and portal protein from Myoviridae phages (CDS49, table A.4.4). The virion assembly region in the vB_VchM_VP-3213 genome shows similarity within several virion assembly genes in both *Burkholderia* phage phiE202 and *Vibrio* phage VHML (green, figure 3.7). Despite their similarity, the genomic context is not completely conserved and some CDS are not shared between these genomes (e.g. CDS15-16, table A.4.4). Besides, Virfam Caudovirales analysis has shown that

the genome context and assigned function of these structural proteins are similar to Myoviridae of Type1 (cluster 6, figure A.4.2A). This cluster includes Siphoviridae lambda phage, Myoviridae from the *Hapunavirus* and *Vhmlvirus* genus, among others (Lopes et al., 2014). Despite their morphological differences, lambda phages have a genome organization which is closely related to other Myoviridae and Podoviridae phages (King et al., 2012). The presence of the sheath protein encoding gene is the main feature to distinguish lambdoid phage genomes from the Myoviridae family (Lopes et al., 2014), which was identified in the vB_VchM_VP-3213 genome (CDS17, table A.4.4).

The vB_VchM_VP-3213 genome contains an integrase encoding gene (CDS41, table A.4.4), and two systems that mediate lysogenic recombination. Two tyrosine recombinases from Xer system were identified as CDS30 (*XerC*) and CDS43 (*XerD*, table A.4.4). Although *Xer* system is a conserved feature in most bacteria, the *Xer* system based on two tyrosine recombinases *XerC* and *XerD* are unusual and vary with the substrate (Rajeev et al., 2009). This system is used by the *Vibrio* phage CTX, VGJ, and TLC (Inoviridae, filamentous phages) to integrate into the *V. cholerae* chromosome I (*diff1* site) (McLeod and Waldor, 2004, Das, 2014). Moreover, most of the filamentous phages that infect *V. cholerae*, can integrate at the *dif1* site (Das, 2014). The *V. cholerae XerC/D* sites flank the bacterial genome attachment, which is required for recombination within the phage CTX genome attachment site (*attP*) (McLeod and Waldor, 2004). In contrast to the described genomes, the phage vB_VchM_VP-3213 carries this *Xer* system (CDS30, CDS43). Further culture-based studies are necessary to define if this system may play a role in the lysogenic phage competition between other *Vibrio* phages. Also, the genome of the lysogenic host strain VN-3213 would provide the necessary information to define the integration sites in the bacterial chromosome. Besides, annotation results shown that 39 of 53 protein encoding genes (73.58 %) are similar to *V. vulnificus* CMCP6 (Rast, table A.4.4). Most of them are hypothetical proteins and/or with an unknown function (table A.4.4), and other identified functions such as lysogenic recombination (CDS30, CDS43) and integrase (CDS41, table A.4.4). The clinical strain *V. vulnificus* CMCP6 is an opportunistic pathogen with no plasmids (Kim et al., 2003, Kim et al., 2011). This similarity might suggest a connection between *V. cholerae* phage vB_VchM_VP-3213 and potential pathogenic *V. vulnificus*. Although *Vibrio* phages are characterized by a narrow host range, some lysogenic phages can infect closely related *Vibrio* species (Muramatsu and Matsumoto, 1991). Although none of the *V. vulnificus* isolated during the cruise HE430 showed the potential as lytic host of the phage vB_VchM_VP-3213 (data not shown), these isolates probably possess other mechanisms (e.g. carried prophage in its genome) to protect against infection of the phage vB_VchM_VP-3213, as described in *V. campbellii* phage-host systems (Lorenz et al., 2016).

Although phage vB_VchM_VP-3213 were most similar to phage genomes from *P2virus* (renamed by ICTV as *Peduovirus*) and *Vhmlvirus* genera, this phage has genomic features that differ from these genera. The *P2virus* genus is characterized by lysogenic phages with linear and non-permuted genome, cohesive ends, size around 34 Kbp, and 40 protein encoding genes (King et al., 2012). On the other hand, the *Vhmlvirus* genus is composed by *Vibrio* phages that can exist as linear prophage plasmids, these phages have a linear genome with cohesive ends, size around 40 Kbp and 60 encoding-genes (Kropinski et al., 2015, Oakey et al., 2002). P2, VHML-like and also Lambda-like phages carry a *cos* site related to the DNA packaging mechanism based on cohesive ends (Garneau et al., 2017, Oakey et al., 2002, King et al., 2012). These key genomic features are not present in the phage vB_VchM_VP-3213, which probably uses Headful (pac) packaging system. Moreover, the pairwise similarity values between the vB_VchM_VP-3213 genome, and the *Burkholderia* phage phiE202 (*P2virus*) and *Vibrio* phage VHML (*Vhmlvirus*), are too low to be considered as part of these genera (~ 45 % each) (Kropinski et al., 2015). Moreover, synteny comparison also reflects these differences by the lack of conserved structural regions between the vB_VchM_VP-3213 genome and these phage genomes (green, figure 3.7).

Based on the overall shared genomic properties, the phage vB_VpaM_VP-3212 is an unclassified Myoviridae. Further studies are necessary to define if this phage belongs to a known taxon of Lambda-like phages. Despite the identification of lysogenic tailed phages in pathogenic *V. cholerae* strains, their contribution to virulence has not been characterized as deeply as in filamentous phages (Faruque and Mekalanos, 2003). Most of the prophages identified in *V. cholerae genomes* are filamentous phages with associated toxin genes (Castillo et al., 2018). Closer phages, such as *Vibrio* phage VHML can encode a toxin component responsible of virulence conversion in *V. harveyi* (Oakey et al., 2002, Oakey and Owens, 2000). Although no pathogenicity genes were identified in vB_VchM_VP-3213, the similarity of the overall genome with other toxin encoding phages may indicate the potentiality to transfer pathogenicity genes. Further studies are necessary to analyze the role of this phage in infected hosts from the North Sea, such as host integration site or effects on the phenotype.

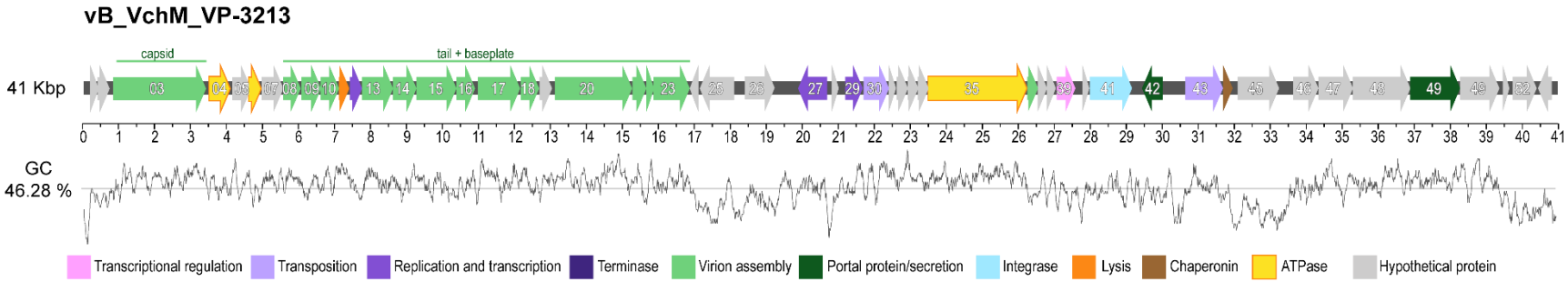


FIGURE 3. 6. Diagram of the vB_VchM_VP-3213 phage genome. Predicted CDS are shown with arrows pointing in the direction of transcription, the colors were assigned accordingly to the possible protein function, as shown in the figure. Nucleotide base pairs (Kbp) and the GC % are shown above each genome profile.

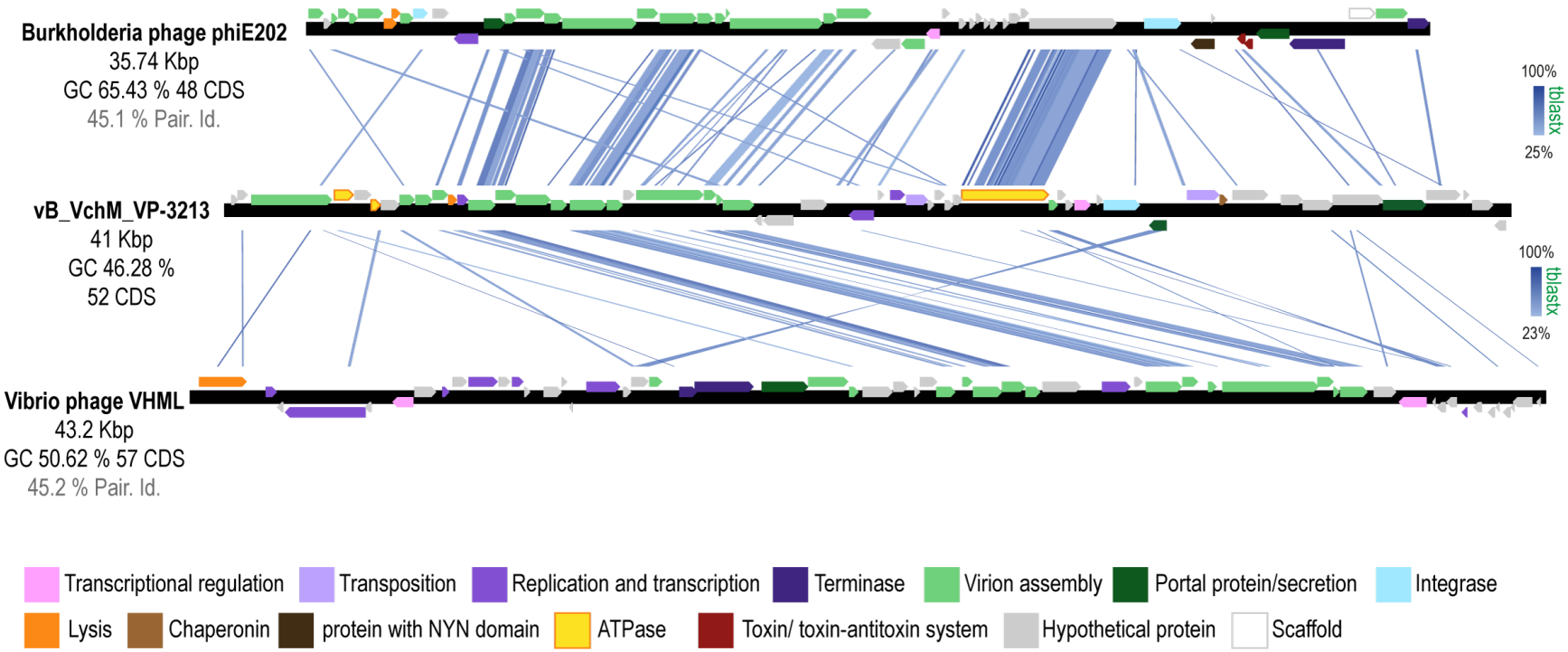


FIGURE 3. 7. Genome synteny plot. Full-genome comparison based on tBLASTx similarity between vB_VchM_VP-3213 and the two most similar phage genomes. To facilitate the genomic comparison, the *Burkholderia* phage phiE202 genome is visualized in reverse direction. The sequence length (Kbp), GC-content (GC %), number of predicted CDS (# CDS), and percentage of pairwise identity (Pair. Id.) between vB_VchM_VP-3213 and the reference genome comparison, are shown above each genome profile. The colors of the protein encoding genes are marked accordingly to their annotated replication functions, as shown in the figure.

3.4. CONCLUDING REMARKS

Despite the low number of characterized tailed *Vibrio* phage genomes, previous *in silico* studies have demonstrated that marine *Vibrio* genomes contain lysogenic phages with virulence-related traits, such as the Myoviridae phage VMHL (Castillo et al., 2018, Oakey et al., 2002). Others *Vibrio* phages contain pathogenicity islands transduced e.g. in *V. cholerae* by the phage CP-T1 (O'Shea and Boyd, 2002). Moreover, the increase of *Vibrio* abundance within the plankton-associated community has coincided with the unprecedented increase in infection cases caused by these bacteria in the North Sea (Baker-Austin et al., 2017, Vezzulli et al., 2012). Considering the influence of prophages on the development of pathogenic bacteria (Breitbart, 2012, Fortier and Sekulovic, 2013), it is crucial to characterize lysogenic phages of potentially pathogenic *Vibrio* in the marine environment.

Due to horizontal gene transfer, *Vibrio* species have been evolved in close relationship within their lysogenic phages. These phages can affect the development of human pathogenic *Vibrio* strains in marine environments, such as the *V. parahaemolyticus* and *V. cholerae* in the North Sea. In this study, three inducible *Vibrio* phages were exemplarily analyzed with a set of very powerful tools and based on a growing number of annotated virus genomes which are accessible. Although this approach was successful to characterize novel phage genomes, the application of reference-based tools were a limiting factor to define taxonomic relationship at genus or higher level. As in deep characterization of replication and genomic properties has been studied only in a small percentage of well-studied species (King et al., 2012), the lack of closely related taxa might be related to the limited number of characterized phage genomes in the public databases. Apart from these limitations, it was possible to classify them at family or at least genus level. The studied phage genomes corresponded to two phages of *V. parahaemolyticus* and one phage of *V. cholerae*. The *V. parahaemolyticus* phage vB_VpaP_VP-3220 (Podoviridae) has the longest genome (58.14 Kbp, 63 CDS), and the lowest percentage of GC-content (43.56 %, table 3.1). In contrast, the *V. parahaemolyticus* phage vB_VpaM_VP-3212 (Myoviridae) had the shortest genome (36.81 Kbp) and the *V. cholerae* phage vB_VchM_VP-3213 (Myoviridae) had the fewest CDS (53, table 3.1). In general, these phages have genomic features that differ from closer genus descriptions, indicates that the diversity of *Vibrio* phages is far from being explored completely. Although these phages did not present virulent traits, their similarity among other phage genomes indicates that these phages can affect the development of pathogenic *Vibrio* strains in the North Sea.

Supplementary data to this article can be found in chapter A.4.

Chapter 4

GENOMIC CHARACTERIZATION

of Filamentous Phage vB_VpaI_VP-3218, an Inducible Prophage of *Vibrio Parahaemolyticus*

Alexa Garin-Fernandez[a,b*], Frank Oliver Glöckner[c,d], Antje Wichels[a]

[a] Department of Microbial Ecology, Biologische Anstalt Helgoland, Alfred-Wegener-Institute Helmholtz-Center for Polar and Marine Research, Helgoland, Germany
[b] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[c] Alfred Wegener Institute Helmholtz Center for Polar and Marine Research, Bremerhaven, Germany
[d] Jacobs University Bremen gGmbH, Bremen, Germany
[*]To whom correspondence should be addressed.
Tel: +49(4725)819-3233; Fax: +49(4725)819-3283; Email: alexa.garin@awi.de

Submitted 28 August 2019 to
Marine Genomics

Keywords: Inoviridae / Vibrionaceae / Lysogenic phage / prophage / North Sea

Abstract: The seawater temperature rise promotes the growth of potentially human pathogenic *Vibrio* species. In the North Sea, *V. parahaemolyticus* strains have been isolated and characterized. These strains contain prophages that may contribute to the emergence of pathogenic strains in the marine environment. Here, we present the genome structure and possible biological functions of the inducible phage vB_VpaI_VP-3218, a novel filamentous phage carried by the *V. parahaemolyticus* strain VN-3218. Prophages of the strain VN-3218 were induced with mitomycin C and the DNA from the phage induction was sequenced. Two incomplete prophages were identified, only one complete phage genome with length of 11,082 bp was characterized. The phage vB_VpaI_VP-3218 belongs to the Inoviridae family and shows close homology to the *Saetivirus* genus. This phage can integrate into the chromosomal host genome and carries host-related regions absent in similar phage genomes, suggesting that this phage integrates within other host genomes. Furthermore, this phage might have a role in pathogenicity due to potential zonula occludens toxin genes. Based on genomic similarity, this phage is an episomally reversible integrated lysogenic phage. This study complements prophage induction and bioinformatic studies applied to non-model species of potentially pathogenic *Vibrio* species. The characterization of this phage provides new insights with respect to the presence of filamentous phages in environmental *V. parahaemolyticus* strains, which might have a role in the emergence of new pathogenic strains in the North Sea.

4.1. INTRODUCTION

Species of the genus *Vibrio* are ubiquitous aquatic bacteria present in riverine, coastal, and estuarine ecosystems (Vezzulli et al., 2013). This genus comprises more than 100 species, 11 of them negatively impact human health and cause serious infections (Miyoshi, 2013). One of them, *V. parahaemolyticus*, is a human pathogen, which primarily causes severe gastroenteritis, as well as wound infections through exposure to contaminated seawater (Thompson et al., 2004, Miyoshi, 2013b).

In the last decade, the seawater temperature raised significantly in over 70 % of coastal areas around the world (Baker-Austin et al., 2017, Lima and Wethey, 2012). This temperature rise promotes the growth of potentially pathogenic *Vibrio* species (Baker-Austin et al., 2013), and *Vibrio*-associated illnesses are increasing worldwide (Vezzulli et al., 2013, Baker-Austin et al., 2013). These outbreaks, e.g. caused by *V. parahaemolyticus* infections, are affecting areas where infections by *Vibrio* pathogens had little previous incidence (Baker-Austin et al., 2017, Baker-Austin et al., 2013). Reported pathogenic *Vibrio* outbreak areas in Europe include e.g. the pandemic *V. parahaemolyticus* O3:K6 spread in France (Quilici et al., 2005), *V. parahaemolyticus* O3:K6 and O4:K11 infections in Spain (Martinez-Urtaza et al., 2005, Martinez-Urtaza et al., 2018), and other heat wave-associated *Vibrio*-associated illnesses in Sweden and Finland (Baker-Austin et al., 2016). Recently, potentially human pathogenic *Vibrio* species have been reported in the North Sea (Oberbeckmann et al., 2011a, Kirstein et al., 2016), where the abundance of *V. parahaemolyticus* increased during the summer months (Oberbeckmann et al., 2011b).

Genomes of the genus *Vibrio* have a high plasticity and they contain significantly more alien genes than other marine bacteria genomes (Lin et al., 2018). These alien genes include pathogenic traits, which can be transferred horizontally in the environment by phages (viruses infecting bacteria). In *Vibrio spp.*, most lysogenic phages (prophages) are known to encode virulence factors (Castillo et al., 2018). Virulence and antibiotic resistance genes carried by prophages are widely distributed among environmental *Vibrio* populations, and they can contribute to the dissemination of virulence, niche adaptation, and emergence of pathogenic *Vibrio* strains (Castillo et al., 2018, Hazen et al., 2010).

The role of (pro)phages on virulence and bacterial evolution has been described for several important pathogens, e.g. for filamentous phages (Fortier and Sekulovic, 2013). Filamentous Inoviridae phages can be integrated as prophages without killing or affecting the life cycle of the host bacteria (Day, 2011). The virion is composed of a circular single-stranded DNA packed in a two-stranded helix as a replicative form (RF) (Rakonjac, 2012). The role of filamentous phages in virulence has been described in pathogenic *Vibrio* species, such as *V. cholerae* phage CTXφ, which encodes the cholera toxin (*CtxAB*) (McLeod et al., 2005); or the ORF8 region carried by the filamentous phage f237, which is related to the pandemic *V. parahaemolyticus* O3:K6 strain (RIMD2210633) (Chang et al., 2002, Nasu et al., 2000). Moreover, filamentous phages are present in nearly every *Vibrio* genome (Hazen et al., 2010, Day, 2011), and occur in pandemic and clinical *V. parahaemolyticus* strains (Nasu et al., 2000, Iida et al., 2001, Taniguchi et al., 1984). Although about 45 % of *Vibrio* genomes available in NCBI database carry a complete Inoviridae prophage-like element (Castillo et al., 2018), the occurrence and description of filamentous phages in potentially pathogenic environmental *Vibrio* strains have not been thoroughly studied so far.

Based on the possible contribution of filamentous phages on the emergence of pathogenic *Vibrio* species in marine environments such as the North Sea, the aim of this study is to characterize the genome of the phage vB_VpaI_VP-3218, a novel filamentous phage carried by the *V. parahaemolyticus* strain VN-3218.

4.2. MATERIAL AND METHODS

4.2.1. ISOLATION AND IDENTIFICATION OF *V. PARAHAEMOLYTICUS* VN-3218

The *V. parahaemolyticus* strain VN-3218 was isolated from the North Sea (station 1, figure 4.1A), during the RV Heincke HE430 on July 31st, 2014 (Gerdts and Rohardt, 2016). The strain VN-3218 was isolated using the raw-seawater enrichment method for potentially pathogenic *Vibrio* species (chapter 3) (Garin-Fernandez and Wichels, submitted). Briefly, approximately 1 l of surface seawater was collected, and the raw seawater was incubated with alkaline peptone water (10 g/l peptone, pH 8.5, 10 PSU final) at 37 °C in the dark. After overnight incubation, the selective enrichment was plated on CHROMagar™Vibrio (MAST Diagnostica GmbH, Germany) (Di Pinto et al., 2011) and incubated at 37 °C in the dark. The grown colonies were compared with the colony coloration typical for *V. parahaemolyticus* according to the manufacturers' instruction. The mauve colored single colonies were transferred and isolated onto marine broth agar (Oppenheimer and Zobell, 1952) with reduced salinity (MB-50% = 16 PSU). The isolate was further identified using MALDI-TOF, and PCR analyses, as previously described (Kirstein et al., 2016, Garin-Fernandez and Wichels, submitted). The MALDI-TOF spectrum from one isolated colony was compared against Bruker mass spectra and VibrioBase libraries (Erler et al., 2015), together with the *V. parahaemolyticus* isolates obtained in the cruise HE430 (chapter 3) (Garin-Fernandez and Wichels, submitted). The spectra comparison was analyzed with the Biotyper™ software (version 3.1), the outcome figure was processed with Adobe Illustrator CS6 (version 16.0.0). The PCR identification was performed using species-specific PCR for *toxR* genes (Kirstein et al., 2016, Oberbeckmann et al., 2011a). As thermostable direct hemolysin (*tdh*) and/or *tdh* related hemolysin (*trh*) play a crucial role in human pathogen *V. parahaemolyticus* strains (Thompson et al., 2004), their protein encoding genes were screened by PCR as previously described (Garin-Fernandez and Wichels, submitted, Kirstein et al., 2016).

After identification, the *V. parahaemolyticus* strain VN-3218 was stored in a culture collection from VibrioNet (Erler et al., 2015).

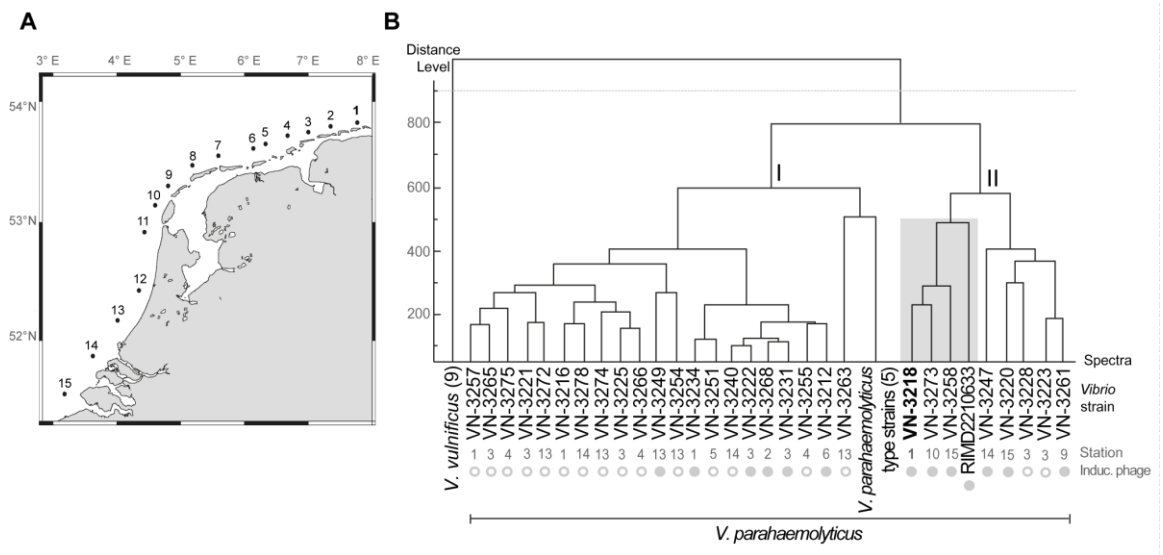


FIGURE 4. 1. Sampled site coastal map from research cuise HE430 (A) and principal component analysis (PCA) dendrogram generated by MALDI Biotyper mass spectra for all the *V. parahaemolyticus* strains isolated in this campaign (B). The grey shadowed cluster corresponds to the isolate VN-3218 cluster group. The black dots (●) represents successful prophage induction with mitomycin C, while the empty dots (○) represents non lysogen strains under the tested conditions. Modified from chapter 3 (Garin-Fernandez and Wichels, submitted).

4.2.2. INDUCTION AND PURIFICATION OF PROPHAGES

The prophages from *V. parahaemolyticus* were induced as previously described in chapter 3 (Garin-Fernandez and Wichels, submitted). Briefly, the strain VN-3218 was incubated in 600 ml MB-50% broth at 37 °C and shaking to further mitomycin C induction (0.25 µgml-1 final) at an optical density of 0.1 – 0.2 (600 nm). After induction, the bacteria DNA was reduced with DNAse I before concentration by ultracentrifugation (Pryshliak et al., 2014, Garin-Fernandez and Wichels, submitted). Virus-like-particles (VLP) after induction were sampled and fixed with formaldehyde (2 % final), and the virus-like-particles (VLP) were counted using a wet-mount method with SYBR Gold and epifluorescence microscopy (Cunningham et al., 2015).

The strain VN-3218 without induction was used as negative control, the positive experimental controls were the *V. cholerae* strain AC53 with the lytic phage ICP1 (Seed et al., 2011), and the lysogen *V. vulnificus* VN-0094 (Pryshliak et al., 2014).

4.2.3. DNA EXTRACTION AND GENOME ANALYSIS

The DNA from the lysogenic induction was extracted using a modified CTAB and phenol:chloroform extraction method as previously described in chapter 3 (Garin-Fernandez and Wichels, submitted). The library preparation was performed by LGC Genomics (LGC Genomics GmbH, Berlin, Germany) and the reads were sequenced on an Illumina MiSeq V3 sequencer using 2 × 300 bp chemistry (Garin-Fernandez and Wichels, submitted).

The sequence reads were processed as described in chapter 3 (Garin-Fernandez and Wichels, submitted). Briefly, the sequences were first decontaminated, and the over-coverage reads were further normalized using BBtools (version 35.14, https://sourceforge.net/projects/bbmap). The retained reads were assembled with SPAdes (version 3.11.1) (Nurk et al., 2013). The potential phage contigs were filtered with VirSorter from iVirus (https://de.iplantcollaborative.org/de/) (Bolduc et al., 2016) and PHASTER (https://phaster.ca) (Arndt et al., 2016, Zhou et al., 2011). In order to find the possible genome duplication in its replicative form (RF), the intact phage contig was aligned with its reverse complement sequence using Mega (version 10.0.1, default parameters MUSCLE algorithm) (Kumar et al., 2018). The genome was mapped against the decontaminated reads with BWA-MEM (version 0.7, default parameters) (Li and Durbin, 2009), the coverage values were analyzed with Qualimap (version 2.2.1) (García-Alcalde et al., 2012, Okonechnikov et al., 2016). Each phage genome was manually curated with DNA master (version 5.22.5) to obtain the final annotated genome, based on the results from Prokka (Seemann, 2014), Rast (http:// http://rast.nmpdr.org/rast.cgi) (Aziz et al., 2008, Overbeek et al., 2014) and Metavir (Roux et al., 2014, Bolduc et al., 2016) annotation software, as described in chapter 3 (Garin-Fernandez and Wichels, submitted). Genome organization, visualization, and GC content were performed with Artemis (version 17.0.1) (Rutherford et al., 2000). The outcome figure was processed in Adobe Illustrator CS6 (version 16.0.0).

To identify the genome on a taxonomic level, pairwise genetic identity was computed using a Sequence Demarcation Tool (SDT version 1.2, default parameters MUSCLE algorithm) (Muhire et al., 2014). The nucleotide sequence of the genome was compared against the Inoviridae complete genomes available in NCBI and the SDT Mastrevirus reference dataset. Species level demarcation was based on pairwise genetic identity values above 95 %, according to the ICTV species demarcation criterion for this family. Based on the genome annotation, at least a threshold of 40 % of the shared orthologous proteins was considered as the same genus (Duhaime et al., 2017, Lorenz et al., 2016).

The phage genome was compared against the phage genomes from GenBank reference sequence genomes (refseq) and nonredundant protein (nr) databases using tBLASTx and BLASTx version 2.2.30+, respectively (Altschul et al., 1997). Synteny comparison was performed within the three most similar complete phage genomes with Easyfig (version 2.2.3) (Sullivan et al., 2011) and processed in

Adobe Illustrator CS6 (version 16.0.0). All used databases were updated to their latest version from May 2018, and BLAST hit alignments with e-value below 0.001 and a similarity greater than 50 % were considered significant. Additionally, the gene encoding Zonula occludens toxin (*zot*) was identified and compared against *zot* gene encoding from *Vibrio* and phage genomes as previously described (Castillo et al., 2018). For this, the possible *zot* gene from the annotated genome was aligned together with the *zot* gene database from Castillo et al. (2018) (Clustal W algorithm), follow by the phylogenetic tree (maximum likelihood, 1,000 bootstrap replicates) using Geneious prime (version 2019.0.3) (Kearse et al., 2012).

4.2.4. NUCLEOTIDE SEQUENCE ACCESSION NUMBERS

The filamentous phage name was defined based on Kropinski et al. (2009) naming nomenclature proposal. The nucleotide sequence was submitted using the data brokerage service of the German Federation for Biological Data (www.gfbio.org/) (Diepenbroek et al., 2014). The sequence associated contextual (meta)data are Minimal Information about any (X) Sequence (MIxS) compliant (Yilmaz et al., 2011).

The genome vB_VpaI_VP-3218 has been deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena/) (Toribio et al., 2017) under the INSDC accession number PRJEB30510.

4.3. RESULTS

4.3.1. LYSOGEN *V. PARAHAEMOLYTICUS* VN-3218

As described in chapter 3, 30 *V. parahaemolyticus* isolates were obtained in the North Sea during cruise HE430 (Garin-Fernandez and Wichels, submitted). Amongst these *V. parahaemolyticus* isolates, the strain VN-3218 was isolated from surface seawater from station 1 (figure 4.1A), at a temperature of 19.6 °C and at a salinity of 31.7 PSU (Gerdts and Rohardt, 2016). Based on principal component analysis of the MALDI-TOF spectra (figure 4.1B), the lysogen VN-3218 clustered together with the reference strain RIMD 2210633 (figure 4.1B, cluster II) in a distinct group from the rest of the *V. parahaemolyticus* isolates at station 1 (figure 4.1B, cluster I). No virulence-associated genes encoding *tdh* and *trh* hemolysins were identified by PCR nor BLAST search (Garin-Fernandez and Wichels, submitted).

Potential prophages of the strain VN-3218 were induced with mitomycin C (section 4.2.2). After induction of VN-3218, the growth (OD) was reduced to 86.17 % compared to the control replicates (figure A.5.1). Despite the high variability between the induced replicates, the cell density of the induced cultures decreased significantly (Mann-Whitney test, $P = < 0.001$). This difference was triggered by phage induction, identified by the presence of VLP in the induced culture (mean = 1.4×10^8 VLP/ml). For FNA sequencing, larger amounts of phage biomass were obtained by induction of the strain VN-3218 in a larger volume (600 ml).

The strain VN-3218 contained two incomplete and one complete phage genome. Two incomplete phage genomes were identified with PHASTER and annotated with Prokka (section 4.2.3). The first incomplete prophage VP-3218_C38 has a sequence coverage of 68.17× (SD ± 41.14), length of 19,533 bp and GC content of 45.67 %. The second incomplete prophage VP-3218_C77 has a sequence coverage of 54.69× (SD ± 26.98), length of 17,348 bp and GC content of 46.46 % (figure A.5.2). Most of the identified coding sequences (CDS) of the incomplete prophage VP-3218_C38 (20 of 24 CDS) are similar to hypothetical proteins from several phages of Gram-negative bacteria, including *Psychrobacter* phage, *Mycobacterium* phage, and *Vibrio* phage (figure A.5.2A). In the same manner, most of the identified CDS of the incomplete prophage VP-3218_C77 (13 of 15 CDS) are similar to hypothetical proteins from *Sinorhizobium* phages (figure A.5.2B). Due to the lack of clear structural cores, these incomplete prophages were considered defective and they were discarded from further characterization.

4.3.2. *VIBRIO* PHAGE VB_VPAI_VP-3218

The complete genome (RF form) of the filamentous phage vB_VpaI_VP-3218 has a sequence coverage of 915.31× (SD ± 265.87), with a length of 11,082 bp and GC content of 44.15 %. Within the genome 14 CDS were identified, of which 11 are located on the forward strand and the remaining 3 on the reverse strand (figure 4.2A). The zero coordinate ATG site was defined by the genome structure of the closest genomes, and the fundamental cores and genes were identified in the genome sequence. Annotation description, as well as CDS length and transcription positions, are summarized in figure 4.2 and detailed description of each gene are present in table A.5.1.

The phage vB_VpaI_VP-3218 belongs to the Inoviridae family and is similar to the two members from *Saetivirus* genus. This filamentous phage shows high similarity with the *V. cholerae* phages VFJ (NC_021562.1, 64 % shared orthologous proteins and 61.8 % pairwise identity) (Wang et al., 2013) and FS-2 (NC_001956.1, 42 % shared orthologous proteins and 61.7 % pairwise identity) (Ikema and Honma, 1998), and with the *V. parahaemolyticus* phage VfO3K6 (NC_002362.1, 28 % shared orthologous proteins and 63 % pairwise identity) (figure 4.3A). These three phage genomes carry the zonula occludens toxin (*zot*) gene (Castillo et al., 2018), and the annotated *gI* gene of vB_VpaI_VP-3218 (CDS08, table A.5.1) shows close homology to *zot* gene from VFJ phage (orf361, figure A.5.3). However, all the compared genomes lack an insert region of 2,584 bp (figure 4.3A, orange shadowed region). This region is formed by genes encoded for a hypothetical protein with restriction endonuclease domain (CDS10), ATP/GTP phosphatase protein (CDS11, *ATP/GTP phos*) and a second hypothetical protein (CDS12) (details in table A.5.1). This region with unknown function is highly similar to the *V. parahaemolyticus* S105 (contig157, NZ_AWJT01000157.1). In detail, the encoding genes of CDS10 resembles the DUF2726 domain-containing protein (BLASTx, identity = 99 %), CDS11 and CDS12 are identical to genes encoded for hypothetical proteins (BLASTx, identity = 100 %) in *V. parahaemolyticus* S105.

Due to the high similarity to the VFJ phage, this filamentous phage is probably an episomally reversible integrated phage. The phage vB_VpaI_VP-3218 was analyzed to identify a 20 bp *att*-like sequence (*attP*) near to the transcriptional regulator cluster, similar to VFJ and VfO3K6 phages. Despite the difference in the last region, the *attP* sequence in our phage was more similar to VFJ than the VfO3K6 phage *att*-like sequence (figure 4.3B).

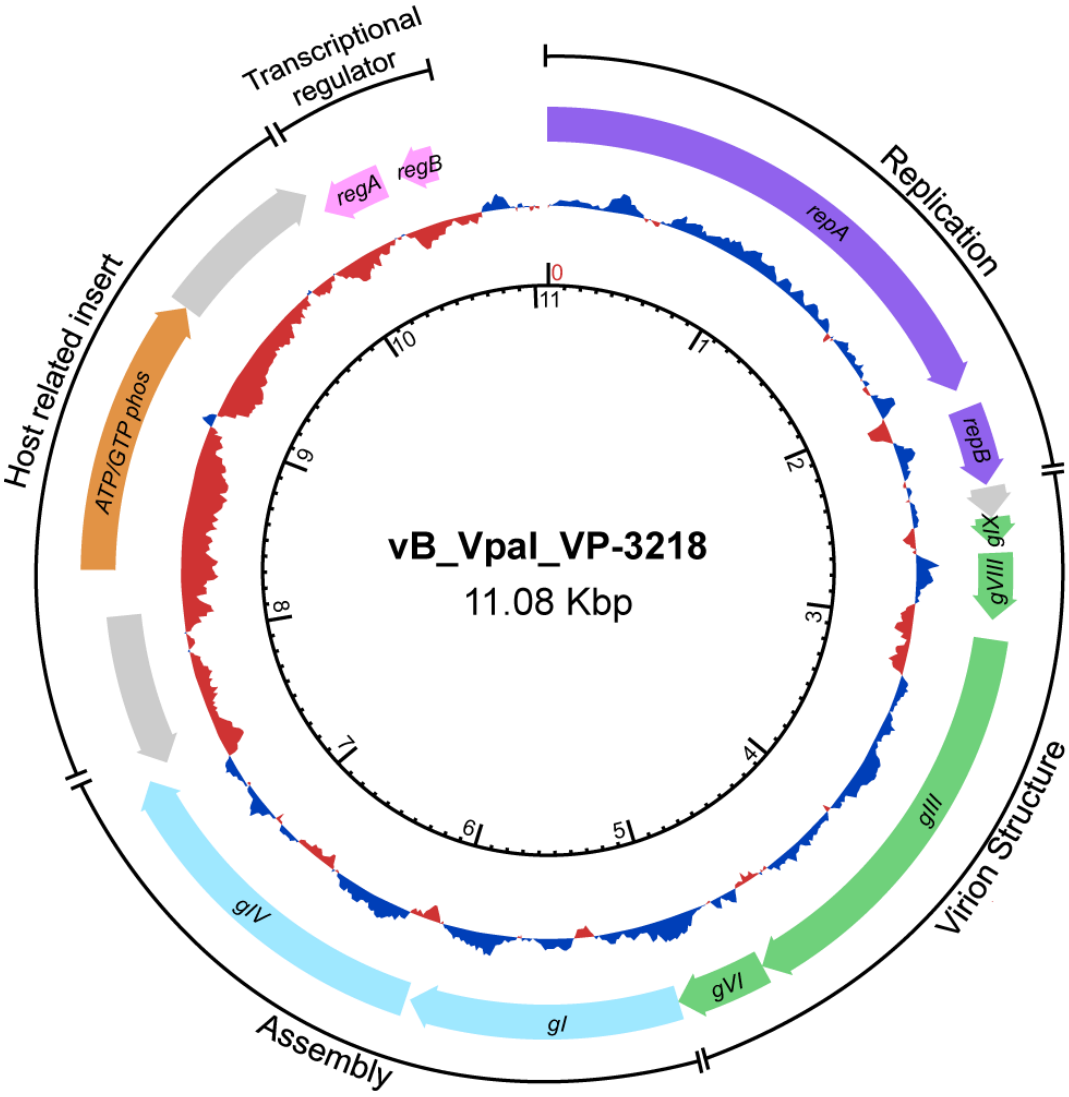


FIGURE 4. 2. Circular diagram of the vB_VpaI_VP-3218 phage genome and predicted CDS. The three main structural cores are marked in violet, green and light blue, transcriptional regulator encoding genes are marked in pink, host-related insert is marked in orange, and hypotetical protein encoding genes are marked in gray. Inner circle corresponds to nucleotide base pairs (Kbp) and middle circle corresponds to GC %, where blue and red represent the sequences above and below the GC average, respectively.

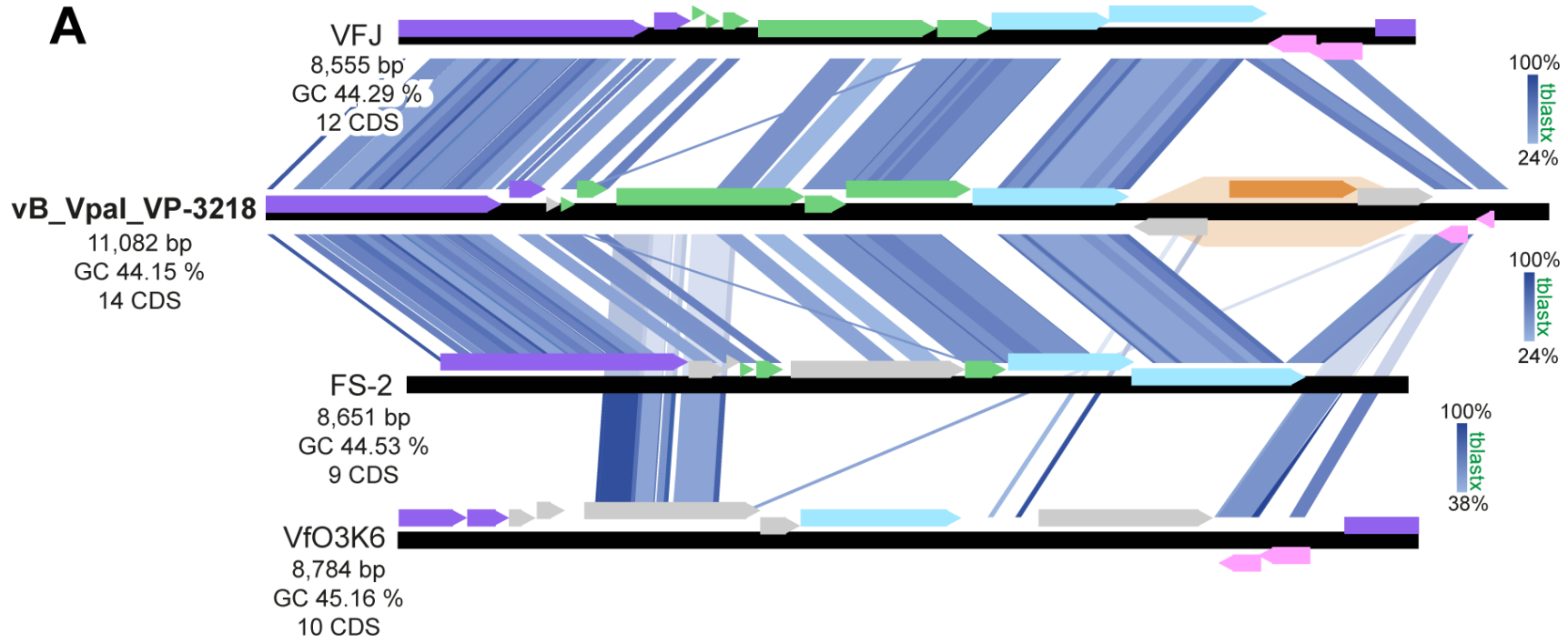


FIGURE 4. 3. Genome synteny plot. Full-genome comparison based on tBLASTx similarity between vB_VpaI_VP-3218 and the three most similar phage genomes: *V. cholerae* phages VFJ and FS-2, and *V. parahemolyticus* phage VfO3K6 (A). Sequence alignment of the *attP* region, part of the transcriptional regulator (gIV) from vB_VpaI_VP-3218 and compared with the analogous regions from VFJ and VfO3K6 phages (B). Protein encoding genes with replication functions are marked in violet, virion structure functions are marked in green, assembly functions are marked in light blue, transcriptional regulator functions are marked in pink, hypothetical protein functions are marked in gray, and the host-related insert is marked in orange.

4.4. DISCUSSION

The increase of *V. parahaemolyticus* outbreaks in Europe is triggered by the rise of temperature in European coastal waters. Therefore, it is crucial to study the possible factors of environmental *Vibrio* species to acquire virulence traits. Filamentous phages are known to play a crucial role in pathogenic *Vibrio* species, through infection and virulence traits transfer (Davis and Waldor, 2003, Iida et al., 2001, Nasu et al., 2000). In this study, we analyzed the potential pathogenic *V. parahaemolyticus* VN-3218 and its inducible prophage. The genome structure and possible biological functions of the inducible filamentous phage vB_VpaI_VP-3218 were analyzed in detail.

4.4.1. LYSOGEN *V. PARAHAEMOLYTICUS* VN-3218

Based on principal component analysis of the MALDI-TOF spectra, the strain VN-3218 was distinct from other strains isolated from the station 1 (figure 4.1B). The strain VN-3218 clustered together with other lysogen strains, including the reference strain RIMD2210633 (figure 4.1B). This reference strain corresponds to the O3:K6 serotype and it is the representative member of *V. parahaemolyticus*. In contrast to the strain VN-3218, the reference strain RIMD2210633 contains virulence-associated gene encoding *tdh* hemolysin (Park et al., 2004, Makino et al., 2003), carries the filamentous phage f237 (McLeod et al., 2005, Chang et al., 2002) and no incomplete prophages were identified in the reference genome using PHASTER. MALDI-TOF analysis is a fast and reliable approach to identify *Vibrio* at species level (Erler et al., 2015). If this method is applied together with molecular identification techniques, false-positive results at species level are reduced (Croci et al., 2007). In case of strain *V. parahaemolyticus* VN-3218, the additional identification of their prophages and other genomic features provides substantial information to define this strain as genetically distinct to the reference strain RIMD2210633.

The lysogen strain VN-3218 contained one complete filamentous phage and two incomplete prophages within its genome. The presence of additional incomplete prophages in the bacterial genome has been described already for other *Vibrio* species e.g. *V. cholerae* and *V. parahaemolyticus* (Castillo et al., 2018, Rakonjac, 2012). In our study, each of the incomplete prophages had a coverage greater than three times the mean coverage of the discarded contigs (mean 17.06×; mean SD ± 12.38). These discarded contigs did not have BLAST hit with virus genomes and may come from the lysogen *Vibrio* genome. Additionally, the incomplete prophages lack clear structural cores, and most of their genes remain uncharacterized. As incomplete prophages depend on inducible functional prophages for

packaging and infection (Hassan et al., 2010), we hypothesize that the incomplete prophages VP-3218_C38 and VP-3218_C77 may be produced at a low rate after induction and that they may use the machinery of the complete prophage vB_VpaI_VP-3218 to proliferate. Whole genome analysis of strain VN-3218 would provide details on these incomplete prophages, their possible topology, as well as their possible relationship with the filamentous phage.

4.4.2. *VIBRIO* PHAGE VB_VPAI_VP-3218

In general, filamentous phages are organized in three main structural cores: replication, virion structure and assembly (Rakonjac, 2012, Wang et al., 2013, Model and Russel, 1988). The predicted CDS of vB_VpaI_VP-3218 are organized in this structure, with an additional transcriptional regulator core (figure 4.2A). The replication core was composed by the CDS01 to CDS03, which encoded replication proteins *repA* and *repB* (table A.5.1). The predicted virion structure core was composed of CDS04 to CDS07 (table A.5.1). This region translates coat and attachment proteins necessary for the filamentous particle formation. The CDS03 encodes a hypothetical protein, which had a similar length and genome position to the minor coat *gVII* protein (figure 4.3A, table A.5.1). However, the predicted protein lacks similarity against other described proteins and further studies are necessary to elucidate its properties and its role in virion packaging. The assembly core was composed of CDS08 and CDS9, which encoded for morphogenesis proteins (table A.5.1). The transcriptional regulator core was composed of the CDS13 and CDS14 (table A.5.1), this region is common in chromosomally integrated filamentous phages and it is transcribed in the opposite direction of the rest of the genes (Wang et al., 2013). Between the transcription regulator and the replication core is a non-coding sequence, which includes an *att*-like sequence similar to VFJ phage (Wang et al., 2013).

The phage vB_VpaI_VP-3218 showed high similarity with *V. cholerae* phages VFJ and FS-2 phages, both comprise the *Saetivirus* genus (Adriaenssens et al., 2017). The VFJ phage produces high rates of virion particles and its infection has a direct effect on cell growth and phenotype. The phenotypic differences include lost cell mobility and resistance to ampicillin and kanamycin (Wang et al., 2013). The FS-2 phage has a wide host range among several different *V. cholerae* biotype strains, visualized by opaque plaque formation (Ikema and Honma, 1998). The FS-2 genome of FS-2 contains an intergenic homologous region involved in CTX phage integration into the host chromosome (Ikema and Honma, 1998); this region was not identified in the phage vB_VpaI_VP-3218. Furthermore, Castillo et al. (2018) identified zonula occludens toxin (*zot*) genes in the VFJ and FS-2 phage genomes, which are similar to other *V. parahaemolyticus zot* genes (figure A.5.3). Based on the phylogenetic analysis, the morphogenesis protein *gI* encoded by CDS08 in vB_VpaI_VP-3218 would have a *zot* toxin function highly similar to VFJ phage. Despite the high similarity with the members of the

Saetivirus genus, further experiments are necessary to define if the filamentous phage vB_VpaI_VP-3218 belongs to this genus. *Vibrio* phages are generally species- or even strain-specific, although a few isolates can have a broader host range to *Vibrio* and *Photobacterium* (Matsuzaki et al., 2000). Several genes from the phage genome vB_VpaI_VP-3218 were similar to *Photobacterium profundum* genes (table A.5.1), including capsid proteins, maturase and morphogenesis encoded genes. The strain *P. profundum* 3TCK contains four questionable prophages identified with PHASTER, and three of them are similar to other filamentous *Vibrio* phages. After lysogenic induction, the induced phage was tested against 153 potential *Vibrio* hosts. However, none of them showed inhibition (data not shown). As *Photobacterium* is closely related to *Vibrio* (Thompson et al., 2004), host screening against *Photobacterium* strains would help to investigate the relationship between this phage and other host genera.

Phage infection contributes to the emergence of disease-causing strains of the environmental population (Bastías et al., 2010, Hazen et al., 2010). Previous *in silico* studies showed that several *Vibrio* environmental and harmless species contain virulence traits acquired from filamentous phages of pathogenic donors (Castillo et al., 2018). Several factors affect the pathogenicity of *V. parahaemolyticus*, including the *tdh* and *trh* hemolysins production during infection (Thompson et al., 2004). Castillo et al. (2018) identified hemolysin encoding genes in prophage like elements of *V. parahaemolyticus*. Although these genes were not detected in the phage vB_VpaI_VP-3218, this phage carries host-related regions absent in other similar phage genomes. Similar to other non-hemolysin *V. parahaemolyticus* phages (Chang et al., 1998), this phage integrates into the chromosomal genome of the *V. parahaemolyticus* VN-3218, suggesting that this phage could interact within other host genomes. This possible role of gene transfer may include the spread of virulence-associated genes in marine environments, as suggested for other filamentous *Vibrio* phages (Castillo et al., 2018). An important factor in pathogenicity is the resistance to antibiotics, which was identified in *Vibrio* infected strains with VFJ phage (Wang et al., 2013). Although the VFJ phage does not carry antibiotic resistance genes, Wang et al. (2013) showed that the infected host strains acquired ampicillin and kanamycin resistance after phage infection. The portal proteins of VFJ phage, such as *gI*, *gIV*, and *gXI*, may be involved into membrane structures as drug resistance pumps or efflux systems (Wang et al., 2013, Rishovd et al., 1994). The sequence and genetic context of these VFJ phage's proteins were highly similar to vB_VpaI_VP-3218 encoding genes (CDS04 and CDS08-09, table A.5.1). Based on the high similarity of the vB_VpaI_VP-3218 phage and the *V. cholerae* phages VFJ and FS-2, further experimental analyses need to be done to elucidate the role of this phage onto pathogenicity in *Vibrio*, such as antibiotic resistance in infected cells and production of *zot*-like proteins.

In conclusion, the general characteristics of the *V. parahaemolyticus* phage vB_VpaI_VP-3218 are similar to other filamentous phages such as the VFJ and FS-2. Based on their genomic similarity, the phage vB_VpaI_VP-3218 is an episomally reversible integrated lysogenic phage. This phage can replicate as a plasmid and integrate reversibly in its RF into the genome of the *V. parahaemolyticus* VN-3218, due to the presence of an *attp* like site.

Filamentous phages have an important impact in environmental *Vibrio* species, such as marine *V. cholerae* populations, specifically regarding physiology and fitness (Xue et al., 2012). Most of the characterized lysogen *V. parahaemolyticus* genomes available in public databases correspond to clinical or pathogenic strains, therefore studies of phages of these *Vibrio* species in environmental populations are highly necessary to understand the development of new pathogenic *Vibrio* strains in the marine environment. In this context, bioinformatic studies are important and helpful to understand the potential role of filamentous phages in the pathogenicity of *Vibrio* species (Castillo et al., 2018). This study complements prophage induction and bioinformatic analysis applied to non-model species of potentially pathogenic *Vibrio* species, which can be applied in other filamentous *Vibrio* phages from marine environments. The characterization of this phage has provided new insights with respect to the presence of filamentous phages in environmental *V. parahaemolyticus* strains, which might have a role in the emergence of new pathogenic strains in the North Sea.

Supplementary data to this article can be found in chapter A.5.

Chapter 5

GENERAL DISCUSSION

The formerly assumed *'desert ocean'* is nowadays known for its diversity of microorganisms (Rohwer et al., 2014). The emerging Next Generation Sequencing (NGS) techniques have brought new opportunities to investigate the marine virus community, and thousands of newly discovered phages indicate that phages are genetically more diverse than once thought (Barylski et al., 2018, Ibrahim et al., 2018). While the NGS and related tools are advantageous to study the virus community composition, they can be limited also in scope and resolution (Roux, 2019). Therefore, the NGS-based techniques solely cannot be applied in phage ecology studies. The application of a comprehensive approach can reduce each technique bias and increase the scope of the research. As there is no current statutory pipeline applied in marine phage ecology, an important issue as well as one big challenge of this PhD thesis was to decide the most reliable and significant tools to develop a virus-specific pipeline for the analyses of novel viruses (meta)genomes (section 5.1). The fully developed pipeline was applied to accomplish these two research aims of this thesis: (1) to analyze the virus community with viromics, and (2) to characterize lysogenic phages from potentially pathogenic *Vibrio* species. This chapter discusses the settled pipeline in a general context based on the main results of this thesis research. Finally, further outlook and conclusions from this PhD thesis are emphasized.

5.1. VIRUS-SPECIFIC PIPELINE: FROM THE SEA TO THE GENE

The application of NGS technologies achieves tremendous advances in virus (meta)genomics and permitted to analyze the virus community composition under a broader scope (Koonin and Dolja, 2018, Ibrahim et al., 2018). Since currently there are a wide range of approaches applied in these studies, there is no a statutory pipeline for virus-specific analyses. This may cause several problems to obtain reliable (meta)genomic comparisons, which results in a limited analysis outcome. The establishment of a comprehensive approach that connects the existing tools and databases can provide a solution to this issue. In this thesis, a selection of available tools was settled to propose a virus-specific pipeline. The proposed pipeline comprises the following main stages:

- Sample preparation. Selection of methods applied from the virus biomass collection from seawater or laboratory samples, to obtain virus DNA for sequencing with Illumina MiSeq.

- Data preparation. *In silico* tools used for (meta)genome pre-processing, which includes read filtration and virus contig assembling. This stage permits to retain only high-quality virus contigs for further characterization.

- Data analysis. *In silico* analysis of virus contig(s) using automated tools together with manual or semi-automated curation.

The outcome from these analyses provides information regarding the taxonomic classification of the viruses as well as the functional genes carried by them. This information can be applied to suggest potential host taxa, lateral gene transfer and abundance in the sample.

This virus-specific pipeline can be used for viromic (section 5.1.1) and lysogenic phage genomics (section 5.1.2). The description of the tools and methods used in each approach are summarized in the following sections.

5.1.1. VIROMIC'S PIPELINE

1. Sample preparation.

For viromic analyses of planktonic viruses, the iron chloride flocculation method was used for virus biomass collection (chapter 2). Based on the chemical interaction between the virus particle and the iron chloride, this method provides several advantages compared to the TFF method (section 1.3.1, Hurwitz et al., 2013a, John et al., 2011). For instance, the iron chloride flocculation method has a high recovery rate (> 90 %), requires a low sample volume (~ 20 l), and it is easy to implement (John et al., 2011). However, this method can present several technical difficulties that must be considered (results discussed in section 5.2). After the biomass collection, the virus particles were resuspended in ascorbate-EDTA buffer and concentrated by centrifugation. The concentration with Vivaspin tubes and further ultracentrifugation allowed for the virus resuspension to be obtained in a small volume (up to 2 ml) for DNA extraction without major difficulties. In case the virus DNA contains rest of iron, additional purification steps can be implemented in viromic analyses (Henn et al., 2010, Hurwitz et al., 2013a).

In the course of this thesis, the virus biomass was collected from 10 stations during the RV Heincke HE430. However, due to different reasons, the concentration of a half of the DNA extraction was too low for sequencing (> 3 ng/ml, data not shown). The increase of sample volume (to 40 l), and optimization of the DNA recovery and extraction process (e.g. DNAse I addition, virus biomass recovery from the filter right after the end of the sampling campaign), may improve the quality of the virus DNA samples. In addition, the iron chloride flocculation method was implemented only with lytic Caudovirales phages (John et al., 2011), and the Fe-virus interaction with other viruses is not completely clear. For instance, there is no robust information regarding Fe-lipid virus complex in virion recovery. Besides, as Inoviridae phages have a long and flexible virus particle morphology (Brum and Sullivan, 2015, Roux et al., 2019), it is possible that these virions were not recovered from the iron chloride flocculation method. The TFF can be used to collect virus biomass without the biases faced with the iron chloride flocculation method (section 1.3.1). However, it must be considered that the TFF method is laborious, time-consuming and requires a large sample volume (Alonso et al., 1999, John et al., 2011). Despite the limitation of the techniques, the sample preparation allowed to obtain enough virus DNA for sequencing with Illumina MiSeq.

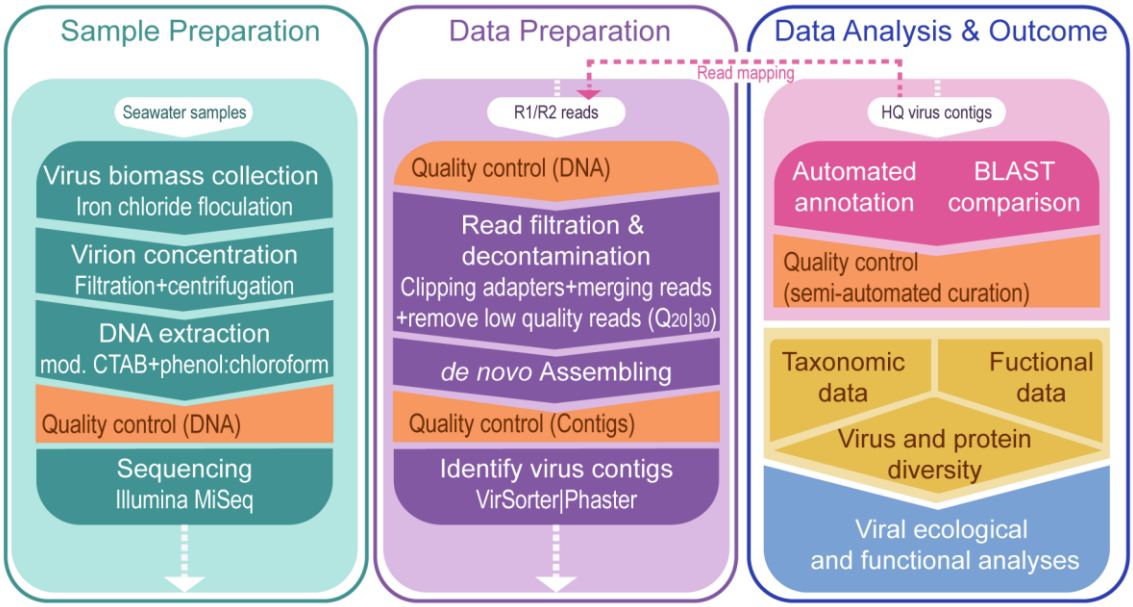


FIGURE 5. 1. General flow chart of the virus-specific pipeline applied in viromics. The pipeline comprises from the sample preparation of marine viruses, the main bioinformatic analyses of data preparation and data analysis, and expected outcome data to describe the virus and protein diversity.

2. Data preparation.

Once the forward (R1) and reverse (R2) reads are sequenced, the quality control with FastQC is applied to determine downstream quality control parameters (Hurwitz et al., 2016). The decontamination steps involve adapter clipping, merging reads, removing short reads, duplicates and low quality reads (figure 5.1) (Hurwitz et al., 2016, Clokie and Kropinski, 2009). The retained high-quality reads are used for contig assembly. The selection of the assembler tool is a critical step for virome preparation and more than one assembler should be tested before the final analysis (Sutton et al., 2019). Therefore, the MetaQUAST comparison tool provides the best overview before selection between these *de novo* assembler tools. MetaQUAST is a fast tool to visualize the number of assembled contigs, the contig length and the number of predicted genes per contig, among others (Garin-Fernandez et al., 2018, Mikheenko et al., 2016). The contigs selected for further analyses must be longer than 3 Kbps and have a coverage higher than 10 × (Roux et al., 2017, Roux et al., 2015b). Web-based tools such as VirSorter (table A.2.1) provide a fast and accurate approach to identify virus contigs as multifasta format. Despite rigorous virus DNA purification, host contaminant cannot be excluded from the total reads (Hurwitz et al., 2016). Therefore, the expect host contig contaminants can be used only for preliminary host gene screening, and discarded for virus community analyses. After the selection of virus contigs, a quality control is applied before data analysis.

3. Data analysis and outcome

In virus community analysis, several tools are available to identify the taxonomy and functional data from virus contigs. For instance, VirSorter automated annotation with Metavir provides a fast and solid outcome for virus contig sorting and identification of protein-encoding genes (Bolduc et al., 2016, Roux et al., 2015b). Subsequently, the predicted virus contigs can be compared with other virus (meta)genomes using BLAST and other comparison tools (table A.2.1). Since a high percentage of the virome sequences have no close matches in databases, there is a greater potential of misalignment and inaccurate taxonomic assignment (Hurwitz et al., 2016). Therefore, manual or semi-automated curation is necessary after annotation steps. Assuming one contig is one partial virus genome, this approach allows to define the taxa of each virus contig, as well as to identify predicted genes, and potential virus hosts (section 2.2). Additionally, to study the diversity of virus protein-encoding genes with unknown function, protein clusters for comparative analyses can be calculated to determine the virome protein richness (Hurwitz and Sullivan, 2013, Hurwitz et al., 2016). Based on their reliability, effectiveness, and open source accessibility, the tools for viromic analyses in this thesis were Metavir annotation from VirSorter, protein cluster comparison (as OPU), and BLAST comparison with curated databases (as local server) (Garin-Fernandez et al., 2018). After semi-automated curation, the annotated contigs can be analyzed with software for statistical computing (e.g. R, Matlab). To obtain further insights about virus-host linkage, the virus contigs highly similar to reference genomes can be analyzed using tools such as HostPhinder, WIsH, and PHACTS (table A.2.1).

In this last stage, the major limitation was the quality and online availability of reference-based tools. For instance, Metavir is an automated annotation tools for virus sequences available in VirSorter and also an independent web-based interface (Dudhagara et al., 2015, Roux et al., 2015b, Roux et al., 2014). However, the last maintenance of the Metavir website was in 2016 (http://metavir-meb.univ-bpclermont.fr), and it is updated only in VirSorter. Thus, the web-based tool VirSorter available in the Discovery Environment from Cyverse was reliable and immediately accessible to identify virus sequences.

5.1.2. GENOMIC'S PIPELINE

1. Sample preparation.

After isolation and screening of lysogen bacteria, the prophage induction method with mitomycin C was applied (section 1.4.1, figure 5.2). This method is widely used and can be complemented with another screening method, such as UV radiation (Zhao et al., 2010, Jiang and Paul, 1998). To confirm the presence of virus like particles (VLP) in the sample, a count method with SYBR Gold was applied prior to DNA extraction as described by Thurber et al. (2009). For all the cultured isolates, the phage induction in a larger volume (600 ml) produced enough virus DNA for further sequencing. The virions produced by the induced prophages are filtered and centrifuged. The membrane filtration (> 0.2 µm) removes cellular debris of the lysogen strain (Solonenko and Sullivan, 2013). Due to the low weight of virus particles, the final concentration by (ultra-)centrifugation is applied to concentrate the virus particles before DNA extraction (Clokie and Kropinski, 2009). As in the DNA extraction for viromic analyses, the modified CTAB and phenol:chloroform extraction method resulted in high-quality samples for sequencing with Illumina MiSeq (Williamson, 2011, Garin-Fernandez et al., 2018). In case the samples would need extra purification, methods such as density gradient (CsCl, Cs_2SO_4, sucrose, metrizamide) can be implemented (Clokie and Kropinski, 2009).

In general, the main difficulty was implementing the prophage induction screening of several samples with 96-well plates. The cultivation for long periods of time can cause water loss in the culture broth, which can be reduced with agarose addition (Hengstl et al., 2012). Overall, the sample preparation yielded enough virus DNA for further analyses.

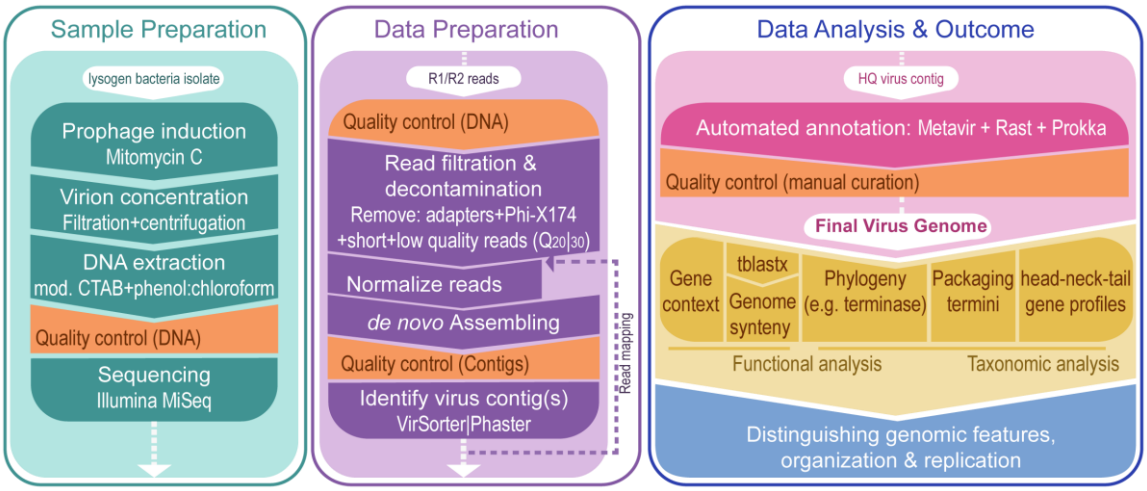


FIGURE 5. 2. General flow chart of the virus-specific pipeline applied in prophage genomics. The pipeline comprises from the prophage induction with mitomycin C in lysogen bacteria isolates, the main bioinformatic analyses of data preparation and data analysis for virus genome characterization, including the functional and taxonomic analyses.

2. Data preparation

After sequencing, the R1/R2 reads are filtrated and decontaminated to retain only the high-quality reads. The decontamination procedure is similar to the one described in the viromic pipeline. Nevertheless, some extra steps are necessary (figure 5.2, violet box). Since the *Escherichia virus* Phi-X174 is used for quality and calibration for sequencing runs in Illumina, the reads identical to Phi-X174 genome must be removed to avoid potential contamination (Mukherjee et al., 2015). In addition, most of the lysogen *Vibrio* genomes contain up to four prophage sequences (~ 80 %, Castillo et al., 2018). Therefore, the set of virus genomes is much smaller than in viromic analyses. In consequence, Illumina MiSeq might sequence hundreds of reads of the same phage genome, which generates overrepresentation in the resulting R1/R2 reads. Hence, an extra normalization step can be applied to retain only reads with coverage values between 5 and 100 ×, which can be reduced up to 30 × (Rihtman et al., 2016). In comparison to other tools, SPAdes provides the highest successful genome assemblies (Rihtman et al., 2016). Therefore, the SPAdes tool was selected for genome assembling of inducible prophages.

After assembling, the virus contigs can be searched using web-based tools such as VirSorter or PHASTER (table A.2.1). The VirSorter tool identifies possible viral sequences and classifies them in 'confidence' categories from 1 (highly probable phage sequence) to 6 (probable prophage sequence) (Roux et al., 2015b). This filtration can be followed by PHASTER screening. The PHASTER tool allows for a rapid identification and classification of the sequences as complete, partial or incomplete prophage (Arndt et al., 2016). After both filtrations, the virus contig(s) can be mapped against the decontaminated reads (prior normalization) to obtain the real coverage value.

3. Data analysis and outcome

The identified virus contig can be annotated using several tools. The MetaVir (from VirSorter) is an annotation tool connected with a curated virus-specific database, which permitted to obtain information about the predicted protein encoding genes as well as similarity within local virus protein clustering (Roux et al., 2015b, Roux et al., 2014). In addition, Prokka and Rast are commonly used for bacterial genome annotation (Loman et al., 2012, Motro and Moran-Gilad, 2017). Due to the set of predicted complete phage genomes, it was possible to implement manual curation prior to further analyses. The manual curation must include the comparison of the Open Reading Frames (ORFs) and predicted product of each encoding protein genes, as well as the identification of duplicated genome sequence due to e.g. replicative form (RF) (section 4.2.3).

The genome characterization is based on functional and taxonomic analysis (figure 5.2, yellow box). The BLAST comparison of the annotated phage genome against reference genomes (as tblastx) can bring information regarding the gene functions that might be overlooked by automated annotation tools. This functional data offers insights about the potential effect on the lysogen bacteria, as well as to identify phage genomic regions with a possible host source (e.g. vB_VpaI_VP-3218 phage, chapter 4). In addition, the identification of packaging mechanism and key genes involved in lysogeny (e.g. integrase, transposase) provide more details regarding the lysogenic cycle of the phage. This outcome together with the genome length, topology, and GC content, are part of the key genomic features necessary for taxonomic classification. This classification is based on demarcation criteria described in each virus genus (King et al., 2012), which includes the percentage of shared proteins, pairwise genetic identity value and shared genomic features. Besides, other analyses such as phylogeny and synteny comparison can be used to analyze the relationship between the characterized phage and members of the closest taxa.

As mentioned in the viromic pipeline, the major limitation in this step was the quality and online availability of reference-based tools. In this case, the Phage Classification Tool Set (PHACTS) is an open source software available as an online web interface and downloadable for local analyses (Bailey et al., 2012). However, the web server is temporarily unavailable (http://www.phantome.org/PHACTS, last checked in June, 2019). In addition, the webserver CoreGenes3.5 helps determine core genes between genomes and is commonly used in ICTV species or genus proposals (e.g. Allison et al., 2016, Kropinski et al., 2015, Kropinski et al., 2018). Unfortunately, this server is currently available only for genomes archived at NCBI and it is not possible to analyze new genomes (http://binf.gmu.edu:8080/CoreGenes3.5/index.html, checked in June, 2019). Despite these limitations, other tools could be applied instead to characterize the phage genomes.

5.2. MARINE VIRUSES, PART OF THE MICRO-UNIVERSE FROM THE NORTH SEA

The application of a virus-specific pipeline provided the successful approach to answer of the questions addressed in the first research aim of this thesis:

"How distinct is the virus community composition between the coast and the open sea?"
"How is their geographic distribution in relation to environmental parameters?"
"What is the distribution of environmental relevant genes carried by phages?"

Overall, the North Sea is dominated by phages and it reveals biogeographic patterns between the coast and the open sea. Although almost a quarter of the sequences appear to be uncharacterized, the remaining identified virus sequences indicated clear differences between the stations (Garin-Fernandez et al., 2018), chapter 2). Although the diversity values were similar between each station, the higher richness and abundance values in the coastal area suggest that the environmental parameters can affect but not determine to shape the virus diversity. In this case, the biotic and abiotic factors might affect the abundance and diversity of viruses. As described in planktonic communities, the rare taxonomic groups (< 1 % per station) are highly diverse but in low abundance, while a small number of phage taxa dominate the community in the North Sea. These results suggest that the rare taxonomic groups might be activated under certain environmental and host conditions, as described by Breitbart and Rohwer (2005), and Jousset et al. (2017). On the other hand, the occurrence and geographic distribution varied between the identified virus families. For instance, algae viruses (Phycodnaviridae) were identified in higher abundance in the coastal areas, where their potential hosts are also highly abundant (Krause-Jensen et al., 2005, McQuatters-Gollop et al., 2009). In addition, fresh water-related cyanophages were detected at station 15, which represents the highest riverine influence. In the case of the identified Caudovirales, the Myoviridae phages (long contractile tail) are highly abundant and their occurrence decreased from the coast to the open sea. In contrast, the occurrence of Podoviridae (short noncontractile tailed phages) increased from the coast to the open sea and the occurrence of Siphoviridae was low throughout the North Sea (figure 2.3).

The coastal virus community was genetically more diverse than the open sea community (Garin-Fernandez et al., 2018). Apart from the genes related to hypothetical proteins and virion production, functional genes related to host virulence and metabolism were detected in low abundance (section 2.4.3). The comparison of virus-encoded genes analyzed in the North Sea virome shows differences between coastal and open sea regions. The indirect influence of the Atlantic inflow at station 24 (open sea) might cause these differences. Further studies are necessary to increase the knowledge about the ecological role of the virus-encoded genes, in terms of expression and host transference. The application of the newly developed viromic pipeline in this thesis yield new insights into the virus community composition which is the basis for future studies of virus community dynamics. However, continuous improvements of these tools will help increase our understanding in the future.

5.3. INCEPTION OF INFECTION: CHARACTERIZATION OF LYSOGENIC PHAGES FROM POTENTIALLY PATHOGENIC SPECIES

The application of a virus-specific pipeline to identify lysogenic *Vibrio* phages provided the information to answer the questions addressed in the second research aim of this thesis:

"How is the distribution of lysogen Vibrio in the North Sea?"
"What are these Vibrio phages?"
"Can these phages be related to pathogenicity in these strains?"

In general, potentially pathogenic *Vibrio* species were identified only near the coast in the North Sea (figure 5.3). Moreover, ca. 42 % of tested isolates carried prophages induced with mitomycin C (Garin-Fernandez and Wichels, submitted). These lysogen *Vibrio* strains were widely distributed along the coasts of the North Sea (figure 3.2) and they did not show a clear pattern in terms of sample site or isolation method (section 3.3.1).

Four lysogen *Vibrio* isolates were selected to sequence the genome of their inducible prophages. Each one of the selected isolates carried a single complete inducible prophage. The characterized *Vibrio* phages belong to the Caudovirales and Inoviridae groups. In detail, the phages correspond to the families Inoviridae (vB_VpaI_VP-3218 from *V. parahaemolyticus*), Podoviridae (vB_VpaP_VP-3220 from *V. parahaemolyticus*) and Myoviridae (vB_VpaM_VP-3212 from *V. parahaemolyticus*; vB_VchM_VP-3213 from *V. cholerae*). Among them, only the phage vB_VpaI_VP-3218 carried genes related to virulence (zonula occludens toxin gene, Garin-Fernandez et al., submitted). In addition, the resemblance of the studied tailed phages with other *Vibrio* phages with pathogenic related genes indicates a potential effect on the development of pathogenic *Vibrio* strains. For instance, the vB_VpaP_VP-3220 show similarity with Stx phages, which contribute to the production of Shiga toxin in nonpathogenic *E. coli* strains (Gamage et al., 2003, Colavecchio et al., 2017, Garin-Fernandez and Wichels, submitted).

If the increase of the surface sea water temperature continues, the abundance of potentially pathogenic *Vibrio* spp. and their phages may also increase. Other events, such as transport of alien species, may affect the rise of pathogenic or pandemic *Vibrio* strains, as discussed by Kirstein et al. (2016). Further studies are necessary to elucidate the effect of lysogenic phages in the development of pathogenic or pandemic *Vibrio* strains from the North Sea.

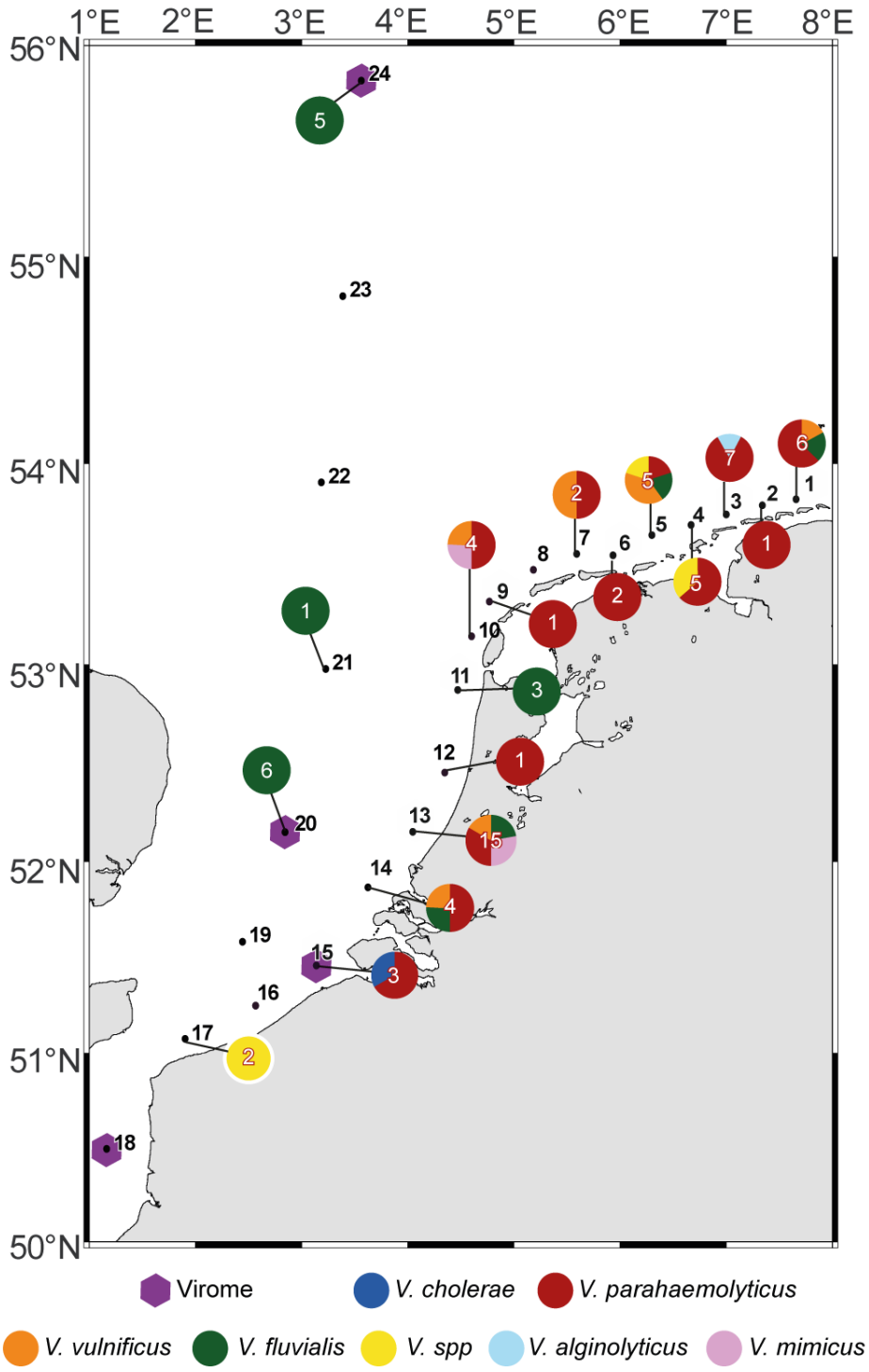


FIGURE 5. 3. Geographical occurrence of *Vibrio* species isolated from surface seawater and microplastic particles collected from the North Sea. The numbers in black indicates the station number from the research cruise HE430 on the RV Heincke in July/August 2014. Each pie chart represents the proportion of *Vibrio* isolates from each station, including the total isolates per station at the middle of each chart (white numbers). The violet hexagon indicates the sample stations for viromic analyses. Data collected from Garin-Fernandez and Wichels (submitted), Kirstein et al. (2016), Garin-Fernandez et al. (2018).

5.4. THE ROAD AHEAD FROM THE NORTH SEA: SUMMARY AND OUTLOOK

The application of NGS techniques has been highly effective for the exploration of virus diversity (Roux, 2019). Nevertheless, new challenges appear to face the limitations of these techniques. Despite the accelerated increase of marine virus studies, there is no statutory pipeline applied in marine phage ecology. The virus-specific pipeline settled in this thesis was applied to analyze an exemplary virus community in the North Sea, which can be applied to other samples. Overall, this pipeline offers a solid base that can be complemented with forthcoming new tools.

Chapter 2 was focused on the description of the virus community composition in the North Sea applying viromics. The knowledge gained in this study allowed for the detection of a biogeographical pattern of the identified virus groups, which may be influenced by the oceanographic features. In this study, four stations at the North Sea's coast and open sea were selected and analyzed. Further spatio-temporal monitoring using both viromics and metagenomics can bring new insights regarding the relationship of the viruses with their potential hosts in the North Sea. Due to their low DNA concentration, six additional stations from the research journey HE430 were sampled but the virus DNA was not sequenced (data not shown). Sample barcoding is an approach to label samples for multiplex sequencing and analysis (Meyer and Kircher, 2010), which can be applied to analyze these uncharacterized samples.

Despite the increase of virome and metagenome studies from several environments, the description of novel phage sequences is still a problem (Paez-Espino et al., 2016, Hurwitz et al., 2018). Since several *in silico* tools for (pro)phage sequence identification are based on reference-based searches, many novel phage cannot be identified with these methods (Hurwitz et al., 2018). For instance, non-tailed viruses such as Inoviridae phages are highly diverse, globally distributed and can often dominate the marine samples (Roux et al., 2019, Brum et al., 2013); however, due to their underrepresentation in curated genome databases, Inoviridae phages might be overlooked in the virus community descriptions by applying either a culture-based approach or NGS sequencing (Edwards, 2018, Roux et al., 2019). In addition, the VirSorter tool is widely used to identify virus sequences based on characteristic nucleotide patterns of DNA viruses, which may cause the omission of sequences from non-tailed phages (Edwards, 2018). The re-analysis of the North Sea virome using the machine learning program from Roux et al. (2019) might offer new insights about non-tailed phages in this region. Besides, Archaea virus contigs were identified only in short discarded contigs (< 3 Kbps, data not shown). Due to possible fragmentation and low abundance, viruses of Archaea might be overlooked from these analyses. Other approaches, such as RNA virus sampling, single-virus genomics (Martinez-Hernandez et al., 2017), linker amplification (Duhaime and Sullivan, 2012), cultivation, and virome sampling improvement

discussed in section 5.1.1, can be applied to analyze the virus community in the North Sea to obtain a broader scope.

As described in *V. cholerae*, new pathogenic strains may appear by horizontal gene transfer (HGT) in the environment (Banerjee et al., 2014, Hazen et al., 2010). The HGT by phages is an important factor of the genetic diversity in *Vibrio* spp. (Fortier and Sekulovic, 2013). In general, *Vibrio* phages were identified in low abundance in the North Sea virome (< 1 % per sampled virome, figure 2.4). The 10 virus biomass samples collected with iron chloride flocculation method were tested against the 41 isolates described in table A.4.1. However, no lytic activity was identified in double agar overlay assay (DAOA, data not shown). Considering the narrow host infection range of *Vibrio* phages, it is possible that the recovered virions would infect other *Vibrio* strains isolated in the same research cruise, such as *V. vulnificus*, *V. fluvialis*, *V. alginolyticus* and *V. mimicus* (figure 5.2). Potential host screening directly after virus biomass collection may improve the isolation of lytic *Vibrio* phages. Moreover, the constant monitoring of lytic *Vibrio* phages from marine coasts as described by Kauffman et al. (2018a) can be applied to improve the isolation approach. Due to a lack of infective host for lytic *Vibrio* phages from the North Sea, only lysogenic *Vibrio* phages were characterized in this thesis using a genomic approach.

Chapter 3 was focused on characterization of lysogenic phages from potentially pathogenic *Vibrio* species. The application of the virus-specific pipeline enables the successful identification of these phages at the family or at least genus level based on their genome sequence and without isolation. The tailed phages characterized in this chapter possess a mosaic genome structure. Phages with this mosaic genome structure are especially complex for taxonomic classification (Tolstoy et al., 2018, Ibrahim et al., 2018). Due to non-homologous recombination from viruses of different taxa, the features of these phages are highly dynamic and diverse (Tolstoy et al., 2018). Moreover, their general features differ from closer genus descriptions, which indicates that the diversity of *Vibrio* phages is far from being explored completely. The lack of closely related taxa might be related to the limited number of characterized phage genomes in the public databases. To enrich the characterization of these lysogenic phages, the prophage induction samples were additionally analyzed with electron microscopy (EM) to describe their virion morphology. However, only the induced sample from lysogen VN-3212 was successful (figure 5.4A). The Myoviridae phage vB_VpaM_VP-3212 identified with *in silico* approach (section 3.3.2) was confirmed by identification of characteristic virion morphology of Myoviridae phages with transmission EM pictures (figure 5.4A). The EM analysis with a larger sample volume can improve the characterization of these lysogenic phages. In addition, the characterized phage genomes can be used as template for labeled DNA probes for phageFISH analyses (Allers et al., 2013). The application of this technique enables the tracking of the proliferation cycle after induction, and

phage-host infection dynamics (Barrero-Canosa and Moraru, 2019). Moreover, other induction methods, such as UV radiation, can provide new insights of non-identified prophages.

Chapter 4 continues the characterization of *Vibrio* phage genomes, focused on the filamentous phage vB_VpaI_VP-3218. The results of this study bring new insights about *Vibrio* phages from the Inoviridae family and the possible role on pathogenicity in *V. parahaemolyticus*. Since both tailed and filamentous phages can produce single plaques (Nakasone et al., 2013), all induced prophage samples were tested against 154 potential *Vibrio* host strains, including clinical and environmental *V. parahaemolyticus*, *V. cholerae*, *V. vulnificus* and *V. mimicus* (data not shown). However, no strain was suitable for phage propagation using spot test or DAOA. The absence of plaque formation in host screening of lytic and lysogenic *Vibrio* phages might be due to the integration of prophages, which protects the lysogen against superinfection (Bondy-Denomy and Davidson, 2014, Hazen et al., 2010). Other resistance mechanisms is the acquisition of spacer sequences, named Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), which provides a rapid and efficient response against several possible phage infections (Marraffini, 2015).

Despite the importance of prophages in the development of pathogenic *Vibrio* strains, little is known about the relationship between lysogenic phages and *Vibrio* strains in the marine environment (Johnson et al., 2008, Banerjee et al., 2014). In this thesis, 13 lysogen *Vibrio* species from the North Sea were identified (section 3.3.1), and four genomes of inducible prophages were characterized (Garin-Fernandez et al., submitted, Garin-Fernandez and Wichels, submitted). Additionally, the DNA of five prophage inductions were sequenced (table 5.1, data not shown). The obtained reads were analyzed as described in section 3.2.4, and more than one virus contig was identified in each lysogen strain (table 5.1). Unfortunately, the characterization of the virus contigs from polylysogen strains was not possible. The assembled genome of the lysogen *Vibrio* strain can provide advantageous information for virus genome mapping and assembling for further data analyses. In addition, EM analysis can improve the characterization of these polylysogen strains and their phages. For instance, at least two different Myoviridae virions were identified in the VN-3247 isolate with EM (figure 5.3B), and 8 phage sequences from the assembled reads were identified with VirSorter (table 5.1). Besides, Erler (2015) assumed the influence of phages on the population development of potentially pathogenic *Vibrio* spp. from the North and Baltic Sea. These bacteria isolates are currently available in the culture collection from VibrioNet (Erler, 2015, Erler et al., 2015). The application of this approach to characterize these potential lysogen strains can bring new insights of the influence of horizontal gene transfer from *Vibrio* phages in the North Sea.

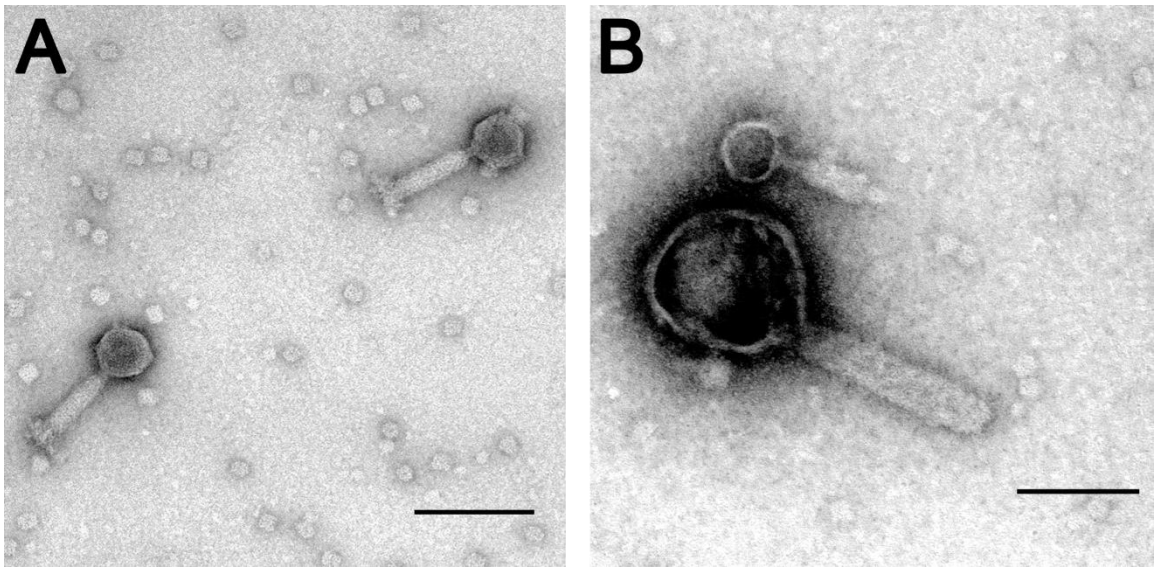


FIGURE 5. 4. Inducible tailed phages from potentially pathogenic *Vibrio* species. Transmission electron microscopic images from the Myoviridae phage vB_VpaM_VP-3212 induced from the monolysogen VN-3212 (A) and induced phages from the polylysogen VN-3247 (b). Scale bar represents 100 nm.

TABLE 5. 1. Species identification of polylysogen *Vibrio* strains and the viral contigs identified from the lysogenic induction. The tested isolates were firstly isolated by Kirstein et al. (2016). MALDI-TOF hit score results against Vibriobase (≥ 2.3 highly probable species; 2.0-2.29 secure genus and probable species identification), PCR results of species-specific (*toxR*), virulence-associated-gene (*trh*; *tdh*) , N-acetylneuraminate lyase of the sialic acid catabolism cluster (*nanA*), the mannitol fermentation operon (*manIIA*), flanking region of PRXII (*VVA1612*), and part of one putative chondroitinase AC lyase (*VVA1636*). The DNA from each lysogenic induction per isolate was sequenced and analyzed until VirSorter classification as described by Garin-Fernandez and Wichels (submitted)(chapter 3.2). The VirSorter identify possible viral sequences and classify them 'confidence' categories from 1 (highly probable phage sequence) to 6 (probable prophage sequence).

Identification of lysogen *V. parahaemolyticus.*

Isolate label (VibrioNet Number)	Station No.	Source	MALDI-TOF HIT-score	*toxR*	*trh*	*tdh*	Viral contigs predicted by Virsorter					
							# cat 1	# cat 2	# cat 3	# cat 4	# cat 5	# cat 6
VN-3231	3	Microplastic particle	2.38	+	-	-	1	3	8	0	0	1
VN-3234	1	Microplastic particle	2.47	+	-	-	2	12	7	0	1	0
VN-3247	14	Seawater	2.64	+	-	-	2	6	0	0	0	0
VN-3268	2	Seawater	2.41	+	-	-	0	0	0	0	3	0

Identification of lysogen *V. vulnificus*

Isolate label (VibrioNet Number)	Station No.	Source	MALDI-TOF HIT-score	*toxR*	*NanA*	*ManIIA*	*VVA1612*	VVA1636	Viral contigs predicted by Virsorter					
									# cat 1	# cat 2	# cat 3	# cat 4	# cat 5	# cat 6
VN-3264	14	Seawater	2.55	+	-	-	+	-	0	5	1	0	0	1

113

5.5. FINAL CONCLUSION

Marine phage genomics will continue improving with the progress of NGS technologies and development of more integrative virus-specific tools. As most microorganisms remain uncultivable, so do their viruses. The vast diversity of uncultivated viruses contains a tremendous potential to discover entirely new strategies tailored to specific hosts and environmental conditions, and the approaches to characterize them might become viable in the next years (Roux, 2019). In this manner, the accomplishments of this thesis included the development and initial application of novel tools to characterize marine phage genomes. The outcome of this research represents a snapshot of the virus community in in the North Sea, which reflect a high diversity of viruses distinct from published genomes. Hence, the biggest current challenge in this phage ecology is to characterize novel phage genomes.

This thesis represents an exemplary study of the virus community in the North Sea, with special emphasis on marine phages. The developed virus-specific pipeline as well as the obtained insights will contribute to expand the study of the virus diversity dynamics in other marine areas, including estimating the abundance and functional diversity of novel phages. For years to come, the discovery of novel virus groups will continue (Koonin and Dolja, 2018). The detailed characterization of novel phages under a comprehensive approach is highly necessary to improve the virus genomic databases. Hence, the application of (meta)genomic-based approach provides an impressive tool to face the new challenges for micro-universe exploration under an integrative scope.

Appendix

ADDITIONAL SCIENTIFIC PUBLICATIONS

A.1. DANGEROUS HITCHHIKERS? EVIDENCE FOR POTENTIALLY PATHOGENIC *VIBRIO SPP*. ON MICROPLASTIC PARTICLES

Authors: Inga V. Kirstein, Sidika Kirmizi, Antje Wichels, Alexa Garin-Fernandez, Rene Erler, Martin Löder, Gunnar Gerdts

Published in: (2016) Marine Environmental Research 120. 1-8.

DOI: 10.1016/j.marenvres.2016.07.004

Contribution: Experimental analyses for *Vibrio spp.* isolation and identification using MALDI-TOF (HE430, 2014). Contribution details in section 1.7.

A.2. SUPPLEMENT TO CHAPTER 1

TABLE A.2. 1. Exemplary tools and databases developed for virus genomics and viromics.

Tool/Database	Name	Description	Reference and software link
De novo Assembly	SPAdes*	Tool to assembly genomes and mini-metagenomes (metaSPAdes) from highly chimeric reads.	http://cab.spbu.ru/software/spades (Nurk et al., 2013)
	IDBA-UD*	Iterative De Bruijn Graph *De Novo* Assembler developed for short read sequencing data from single-cell and metagenomic reads.	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/index.html (Peng et al., 2012)
	Megahit*	Fast and memory-efficient NGS assembler via succinct de Bruijn graph. Optimized for metagenomic reads.	https://github.com/voutcn/megahit (Li et al., 2015)
	AV454	Assemble Viral 454. Designed for small and non-repetitive human virus genomes sequenced at high depth.	https://www.broadinstitute.org/viral-genomics/av454 (Henn et al., 2012)
	VrAP	Viral Assembly Pipeline. Tool based on SPAdes with additional read correction and filter steps. This tool classifies contigs to distinguish host from viral sequences.	https://www.rna.uni-jena.de/research/software/vrap-viral-assembly-pipeline (Fricke et al., in press)
	V-GAP	Viral Genome Assembly Pipeline. Developed for reconstruction of phage genome into single contig from shotgun reads.	(Nakamura et al., 2016)
Annotation	RAST*	Rapid Annotation using Subsystem Technology. Fully-automated service for annotation of prokaryotic genomes available as web tool or to download (as multi fasta).	http://rast.nmpdr.org (Aziz et al., 2008, Meyer et al., 2008, Overbeek et al., 2014)
	PROKKA	Rapid annotation tool developed for prokaryotic and viral genomes (as multi fasta).	https://github.com/tseemann/prokka (Seemann, 2014)
	Metavir	Original as web-based, currently is part of iVirus. Developed for annotation of viral metagenomic sequences as raw reads or assembled contigs (multi fasta).	(Bolduc et al., 2016, Roux et al., 2014)

TABLE A.2.1. Continuation.

Tool/Data base	Name	Description	Reference and software link
Database (db)	ICTV	International Committee on Taxonomy of Viruses. Focused on organize and authorize the taxonomy and nomenclature of viruses. Contain a detailed database with the approved virus species.	https://talk.ictvonline.org/taxonomy (Lefkowitz et al., 2018)
	ViralZone	Web resource with general molecular and epidemiological information of each virus family, including access to UniProtKB/Swiss-Prot viral protein entries (EMBL-EBI).	http://viralzone.expasy.org (Hulo et al., 2010)
	GenBank NCBI*	NIH genetic sequence database (U.S.). Associated with the International Nucleotide Sequence Database Collaboration (INSDC), together with the DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA). Also provides the viral genome sequence data resource.	https://www.ncbi.nlm.nih.gov/genbank; http://www.ncbi.nlm.nih.gov/genome/viruses (Brister et al., 2015)
	ENA EMBL-EBI*	European Nucleotide Archive, part of the European Bioinformatic Institute. Associated with the (INSDC). This database provides nucleotide sequences with their input information (e.g. sample, experimental setup, software configuration), output machine data (e.g. reads and quality scores), and other related information (e.g. assembly, mapping, gene annotation).	https://www.ebi.ac.uk/ena (Toribio et al., 2017)
	JGI IMG/VR	Curated virus db associated with the Genomes On-Line Database (GOLD), that contains both UViGs ans iVGs. The virus data is obtained from the Microbiomes system IMG/M database (https://img.jgi.doe.gov), that contains reference virus genomes from NCBI, and predicted virus sequences from metagenomes and scaffolds.	https://img.jgi.doe.gov/cgi-bin/vr/main.cgi (Paez-Espino et al., 2016, Pati et al., 2013)
	The Actinobacteriophage database	Database focused in Actinobacteriophages, genomes from PHS.	https://phagesdb.org (Russell and Hatfull, 2016)
	ACLAME	A Classification of Mobile genetic Elements. This database collects information from various resources about mobile genetic elements from plasmids, viruses and prophages.	http://aclame.ulb.ac.be (Toussaint et al., 2009)

TABLE A.2.1. Continuation.

Tool/Data base	Name	Description	Reference and software link
Viral sequence screening	VirFinder	k-mer based tool for identification of viral sequences in metagenomic data.	https://github.com/jessieren/VirFinder (Ren et al., 2017)
	virMine	An *in silico* detection tool of virus sequences from metagenomic reads	https://github.com/thatzopoulos/virMine (Garretto et al., 2019)
	ViraMiner	Deep learning-based method to identify and quantify virus reads from raw metagenomic contigs of human biospecimens.	https://github.com/rega-cev/virulign (Tampuu et al., 2019)
	virMiner	Web-based phage contig prediction tool using on random forest algorithm.	http://sbb.hku.hk/VirMiner (Zheng et al., 2019)
Prophage screening	PHAST	Phage Search Tool. Web-based tool for identification of prophage sequences.	http://phast.wishartlab.com (Zhou et al., 2011)
	PHASTER	Upgraded web-based tool PHAST for fast identification and annotation of prophage sequences in bacteria genome, metagenomes and viromes.	http://phaster.ca (Arndt et al., 2016)
	Phage_Finder	Heuristic computer program written in PERL to identify prophage regions within bacterial genomes.	http://phage-finder.sourceforge.net (Fouts, 2006)
	Prophinder	Web-based program for prophages detection, based on BLASTP against the ACLAME db.	http://aclame.ulb.ac.be/Tools/Prophinder (Lima-Mendez et al., 2008)
	PhiSpy	Computer program written in C++, Python and R to identify prophages sequences in complete bacterial genomes.	https://github.com/linsalrob/PhiSpy (Akhter et al., 2012)
Others virus-based tools	VIRALpro	Part of Scratch suite. Identification of capsid and tail protein sequences in single query protein sequence, based on the amino acid composition and secondary structure prediction.	http://scratch.proteomics.ics.uci.edu (Galiez et al., 2016)
	Virfam	Web-based tool for classification of tailed phages based on head-neck-tail modules in their genomes.	http://biodev.extra.cea.fr/virfam (Lopes et al., 2014)

TABLE A.2.1. Continuation.

Tool/Data base	Name	Description	Reference and software link
	ClassiPhages	This tool uses a reference phage-derived-HMMs scoring matrix to develop and train an Artificial Neural Network (ANN) to find patterns for phage genome classification.	(Chibani et al., 2019)
	VICTOR	Virus Classification and Tree Building Online Resource. Developed by DSMZ, this tool permits to classify phages based on the amino acid or nucleotide sequences as multi or single fasta. The resulting trees (mostly monophyletic) revealed the classification of whole genome comparison (ICTV db).	https://ggdc.dsmz.de/victor.php (Meier-Kolthoff and Göker, 2017)
	PhageTerm	Fast tool to determine DNA termini and phage packaging mechanisms using NGS data. Also available on the Galaxy-based server (https://galaxy.pasteur.fr).	https://sourceforge.net/projects/phageterm (Garneau et al., 2017)
	HostPhinder	Based on the query similarity within the reference complete phage genomes (NCBI, EMBL-EBI and other phagedb from latest 2017), it can predict the bacterial host species based on a query phage genome (as single fasta).	https://cge.cbs.dtu.dk/services/HostPhinder (Villarroel et al., 2016)
	PHACTS	Phage Classification Tool Set. PERL-based tool to predict proliferation cycle (as temperate or virulent) based on a local reference database, similarity algorithm of predicted proteins and a supervised Random Forest classifier.	http://www.phantome.org/PHACTS (Bailey et al., 2012)
	WIsH	Tool written in C++ code for prediction of prokaryotic hosts based on the phage genome sequence.	https://github.com/soedinglab/wish (Galiez et al., 2017)
	iMicrobe	Integrated and federated system that interconnects diverse microbiome data sets, tools, and community resources from own data users. This system is associated with CyVerse to analyze data in the web-based friendly platform Discovery Environment (https://de.cyverse.org). This system contains its own database, which is connected to iVirus.	https://imicrobe.us (Hurwitz, 2014)
	VirSorter (iVirus)	Original as web-based tool, currently is part of iVirus (http://ivirus.us) and available in Cyverse (https://de.iplantcollaborative.org/de/). Detection of virus sequence in multi fasta sequences, sorted 6 confidence levels. It can be based on reference-dependent or independent method.	(Bolduc et al., 2016)

* = Not developed for virus analysis. However, this tool contains important information in this field.

A.3. SUPPLEMENT TO CHAPTER 2

Rarefaction analysis, general information.

To obtain the figure A.3.1, all virus genes predicted by VIRSorter and identified by BLAST (Material and Methods, section 2.5) were analyzed using *rarecurve* function from Community Ecology Package (Vegan package for R version 3.3.2) (Oksanen et al., 2017).

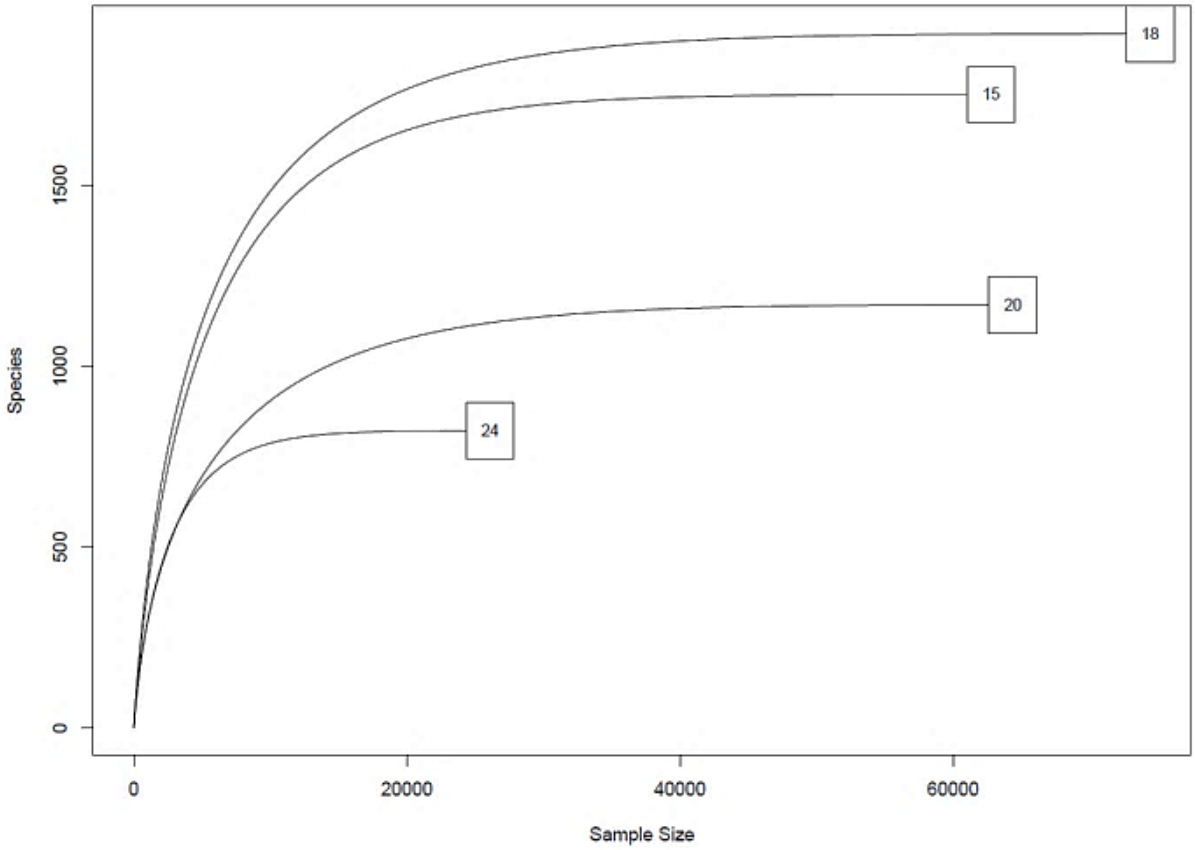
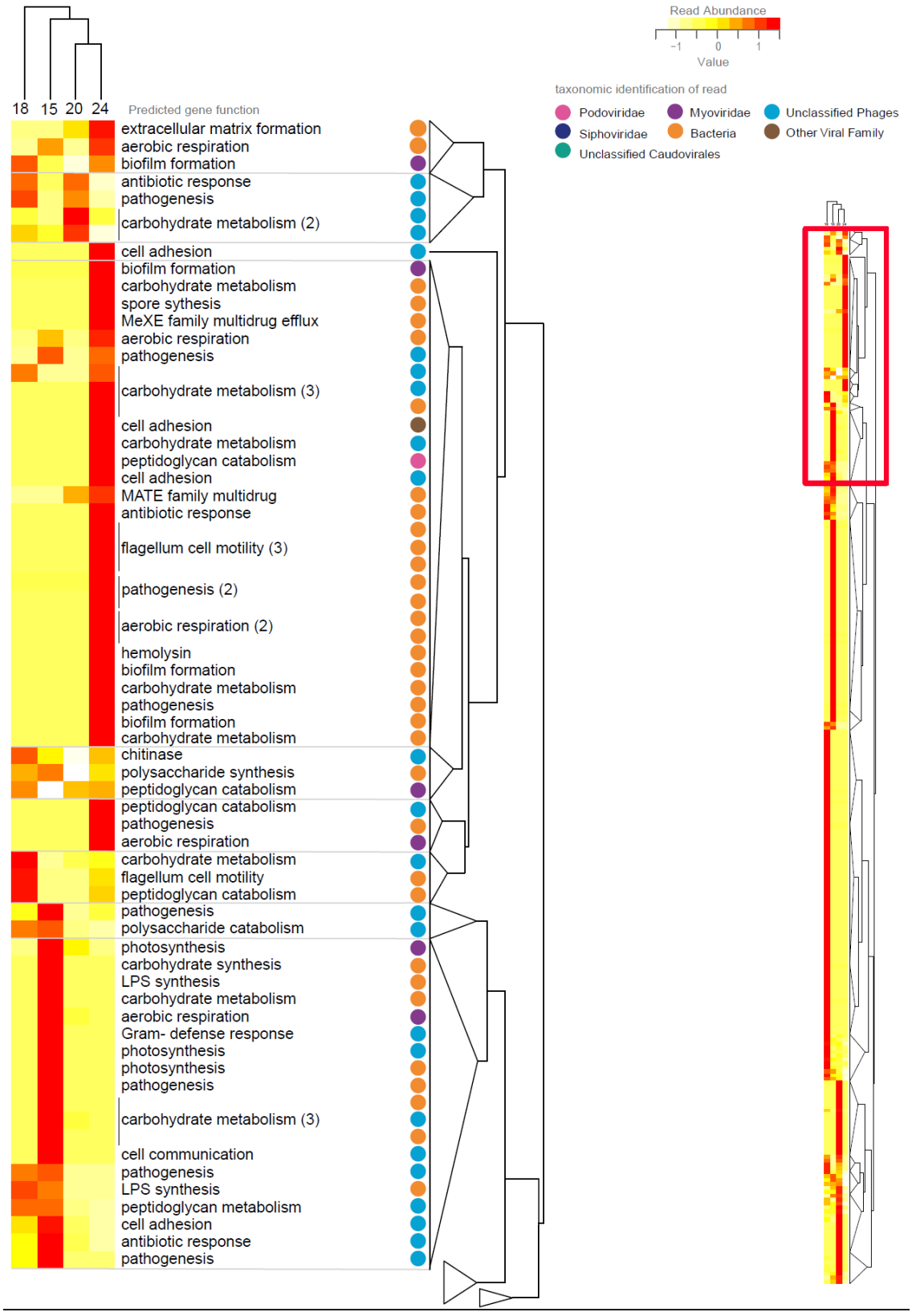


FIGURE A.3.1. Rarefaction analysis of hits to identified proteins from all North Sea viromes.

Heatmap subset of host-related genes, general information.

The total number of the classified virus proteins from each North Sea virome was compared as described in experimental procedures from chapter 2. The dataset of this heatmap was not reduced. Identified genes of contigs larger than 3 kbps were first classified in base of their molecular or biological function of the encoded protein and then separated by the contig identified family. The abundance of the classified functions was z-scored and clustered by Bray Curtis dissimilarity index. The dendrogram represents the Average Hierarchical analysis of the classified gene function (y axis) of each sampling station (x axis).

FIGURE A.3.2 (NEXT PAGE). Heatmap comparison of the filtered protein functions from virus-encoded genes. Identified genes from contigs larger than 3 kbps were classified in base of their taxonomical family and molecular or biological function of the encoded protein. The gene abundance was transformed to z-score and clustered by Bray Curtis dissimilarity index. The dendrogram represents the Average Hierarchical analysis of the classified gene function (y axis) of each sampling station (x axis).


Read Abundance
−1
0
1
Value
taxonomic identification of read
Podoviridae
Myoviridae
Unclassified Phages
Siphoviridae
Bacteria
Other Viral Family
Unclassified Caudovirales
18
15
20
24
Predicted gene function
extracellular matrix formation
aerobic respiration
biofilm formation
antibiotic response
pathogenesis
carbohydrate metabolism (2)
cell adhesion
biofilm formation
carbohydrate metabolism
spore sythesis
MeXE family multidrug efflux
aerobic respiration
pathogenesis
carbohydrate metabolism (3)
cell adhesion
carbohydrate metabolism
peptidoglycan catabolism
cell adhesion
MATE family multidrug
antibiotic response
flagellum cell motility (3)
pathogenesis (2)
aerobic respiration (2)
hemolysin
biofilm formation
carbohydrate metabolism
pathogenesis
biofilm formation
carbohydrate metabolism
chitinase
polysaccharide synthesis
peptidoglycan catabolism
peptidoglycan catabolism
pathogenesis
aerobic respiration
carbohydrate metabolism
flagellum cell motility
peptidoglycan catabolism
pathogenesis
polysaccharide catabolism
photosynthesis
carbohydrate synthesis
LPS synthesis
carbohydrate metabolism
aerobic respiration
Gram- defense response
photosynthesis
photosynthesis
pathogenesis
carbohydrate metabolism (3)
cell communication
pathogenesis
LPS synthesis
peptidoglycan metabolism
cell adhesion
antibiotic response
pathogenesis

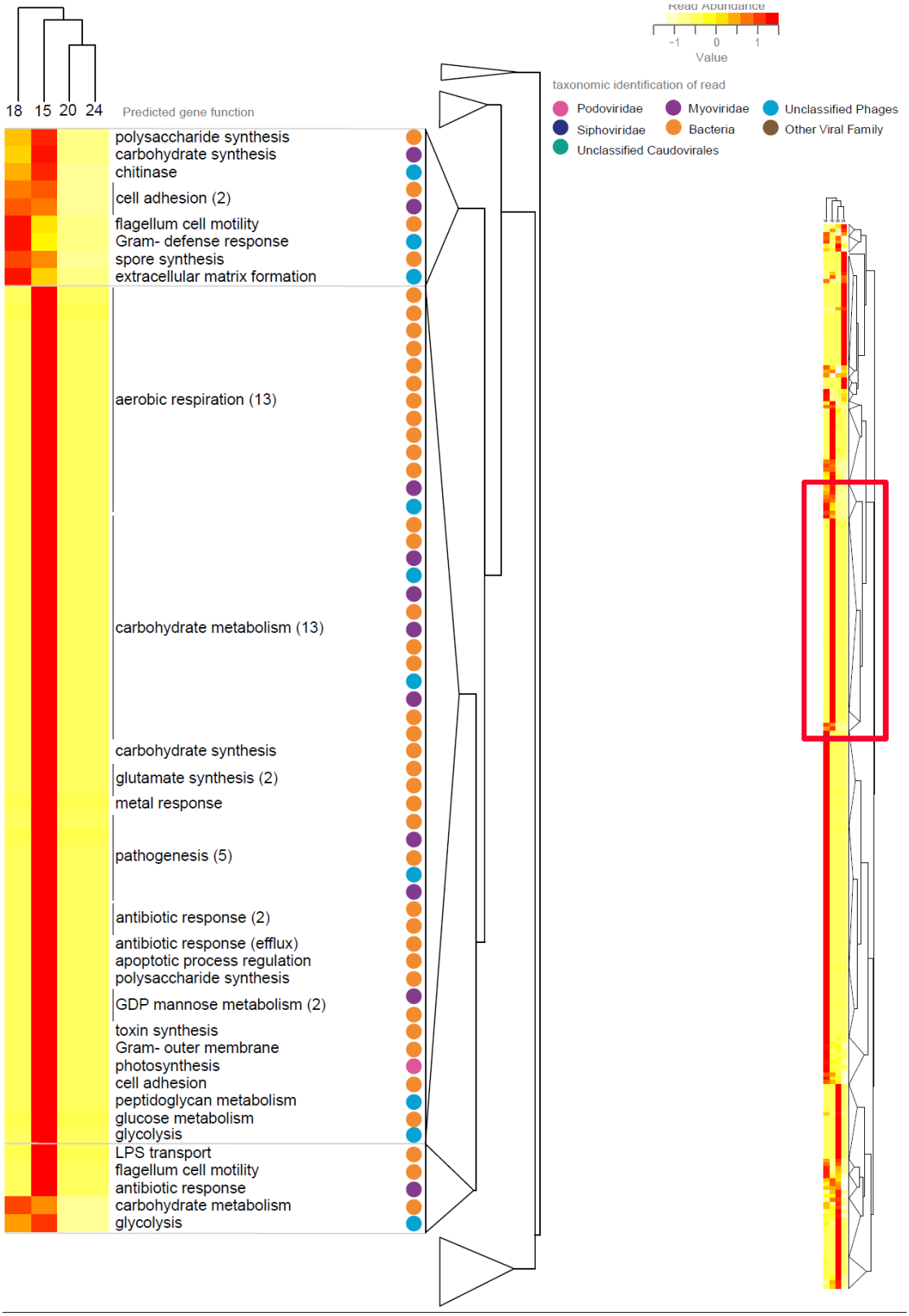


FIGURE A.3.2. Continuation.

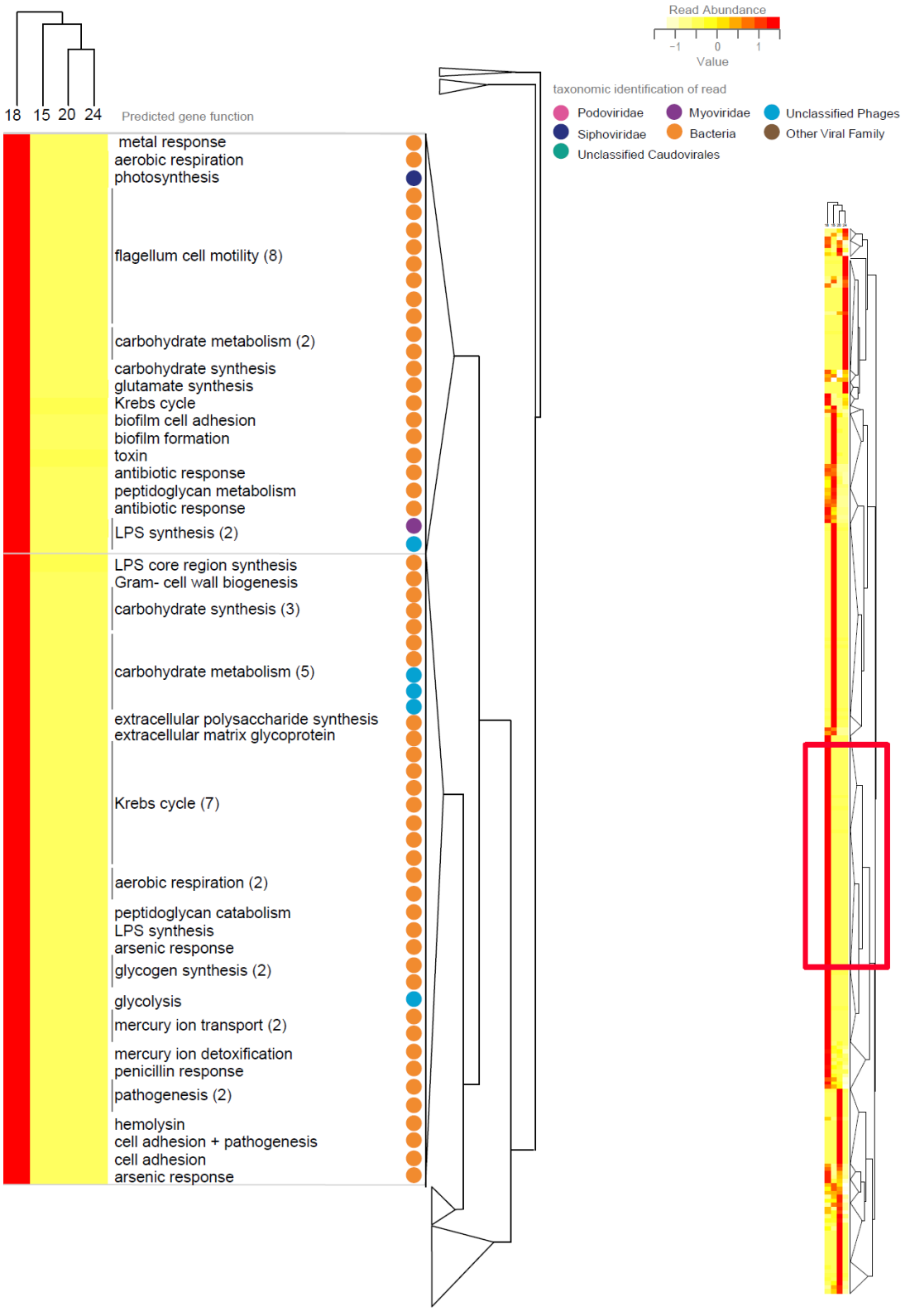


FIGURE A.3.2. Continuation.

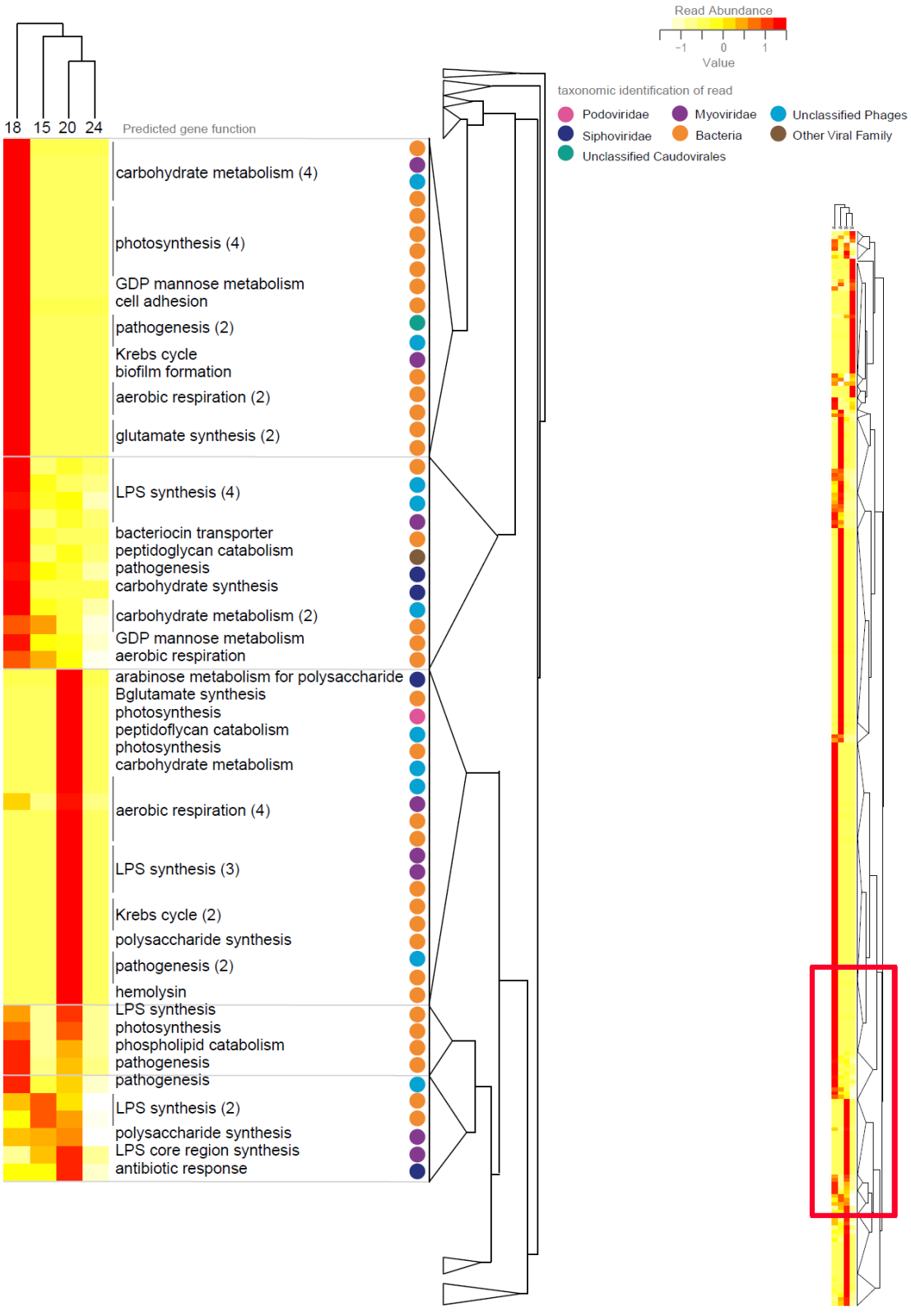


FIGURE A.3.2. Continuation.

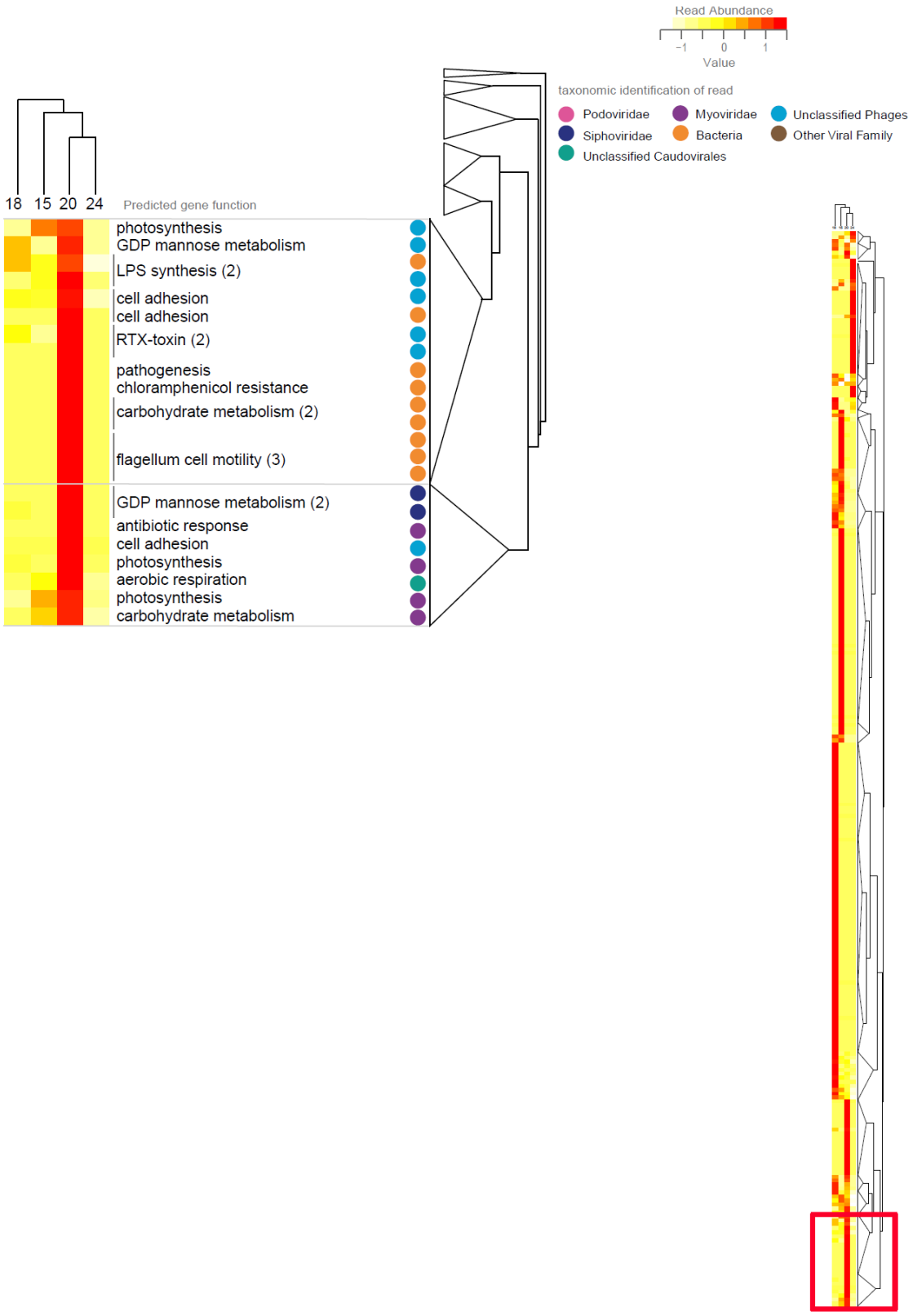


FIGURE A.3.2. Continuation.

TABLE A.3.1. Number of reads, total assembled contigs and total predicted viral contigs.

The total of assembled Megahit contigs were analyzed using the VirSorter tool (default settings, Virome Decontamination mode) (section 2.3) prior a 3 Kbps cutoff for improved recall of the virus community classification.

Station No.	# reads	Mbps	Avg. read length	Stdev read length	# contigs	Avg. coverage per contig	Avg. contig length	**Total viral contigs predicted by VirSorter**							
								# viral contigs	N50 (Kbps)	# cat 1	# cat 2	# cat 3	# cat 4	# cat 5	# cat 6
15	8,904,876	1,905.12	210.38	84.82	227,675	4.66	675.04	2,577	627	228	2,127	222	0	0	0
18	16,898,592	1,227.36	212.15	82.24	368,223	6.50	694.44	2,706	665	226	2,240	239	0	1	0
20	7,235,850	1,718.34	233.57	78.77	210,601	6.85	594.82	1,719	423	133	1,458	127	0	1	0
24	3,224,142	659.93	201.26	84.93	123,167	3.98	626.30	1,193	302	120	1,033	39	0	1	0

TABLE A.3.2. Percentage of virus contig's BLAST hit per station and overall the North Sea virome.

The subset of identified virus contigs from table 2.2 were additionally classified as Podoviridae, Siphoviridae, Myoviridae, Unclassified Caudovirales, Unclassified phage, Phycodnaviridae and Other Virus Family**.**

BLAST homolog		**St. 15** Contigs (%)	**St. 18** Contigs (%)	**St. 20** Contigs (%)	**St. 24** Contigs (%)	**Total Station** Contigs (%)
Unclassified phage		59.67	68.57	64.80	56.76	63.46
Caudovirales	Podoviridae	2.95	3.44	6.03	13.34	4.76
	Siphoviridae	3.32	2.57	4.03	1.56	3.00
	Myoviridae	27.22	18.21	17.76	19.78	21.53
	Unclassified-Caudovirales	1.81	1.73	3.17	2.19	2.07
Phycodnaviridae		2.53	0.90	0.71	0.62	1.42
Other-Virus-Family		2.50	4.58	3.50	5.75	3.76

Diversity and species richness indices, general information.

To determinate the values on table A.3.3, the percentage of identified virus species per sampling station were evaluated with following formula using PRIMER 6 software version 6.1.11 (PRIMER-E Ltd., Plymouth, United Kingdom).

Species richness

Margalef $d = (S-1)/\text{Log}(N)$

Diversity index

Shannon $H' = -\text{SUM}(Pi \times \text{Log}(Pi))$

S: total species

N: total individuals

H': Log base 2

TABLE A.3.3. Diversity and species richness indices for identified taxonomic virus groups in North Sea. Total species (S), species richness (d) and Shannon (H') values were determinate based on the percentage of identified virus species per sampling station.

Station No.	S	N	d	H'
15	61	4,674	7.101	3.796
18	55	4,616	6.4	3.205
20	33	2,367	4.119	3.036
24	29	1,347	3.886	3.177

A.4. SUPPLEMENT TO CHAPTER 3

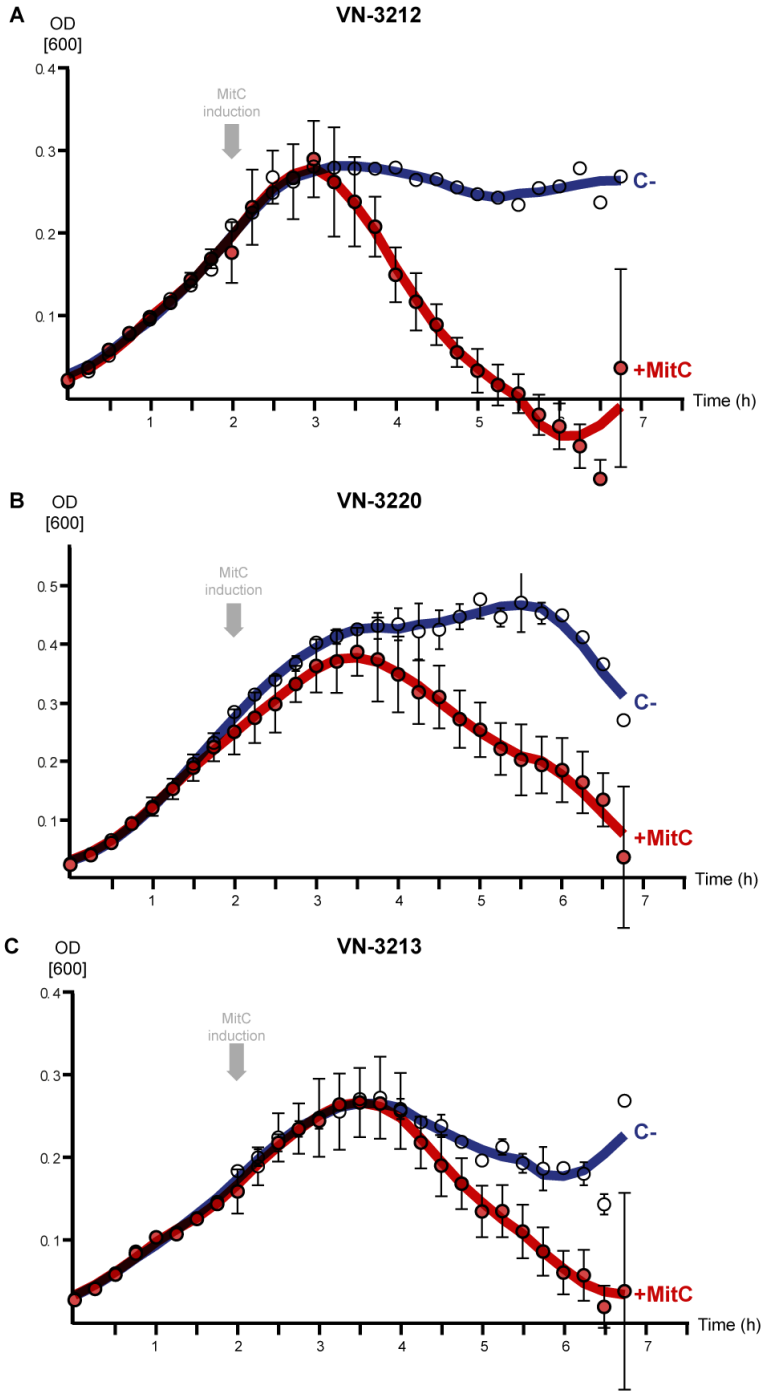


FIGURE A.4. 1 Prophage induction of *V. parahaemolyticus* VN-3212 (A), VN-3220 (B), and *V. cholerae* VN-3213 (C). Mitomycin C (+MitC, red) and MB-50% broth (control C-, blue) was added to the culture at optical density 0.1-0.2 (600 nm) (MitC induction, grey arrow) and incubated at 37 °C and shaking. The induction (+MitC) was performed in eight replicates, and the negative control (C-) was performed in duplicate. The error bar indicates the standard error of the mean from replicate experiments.

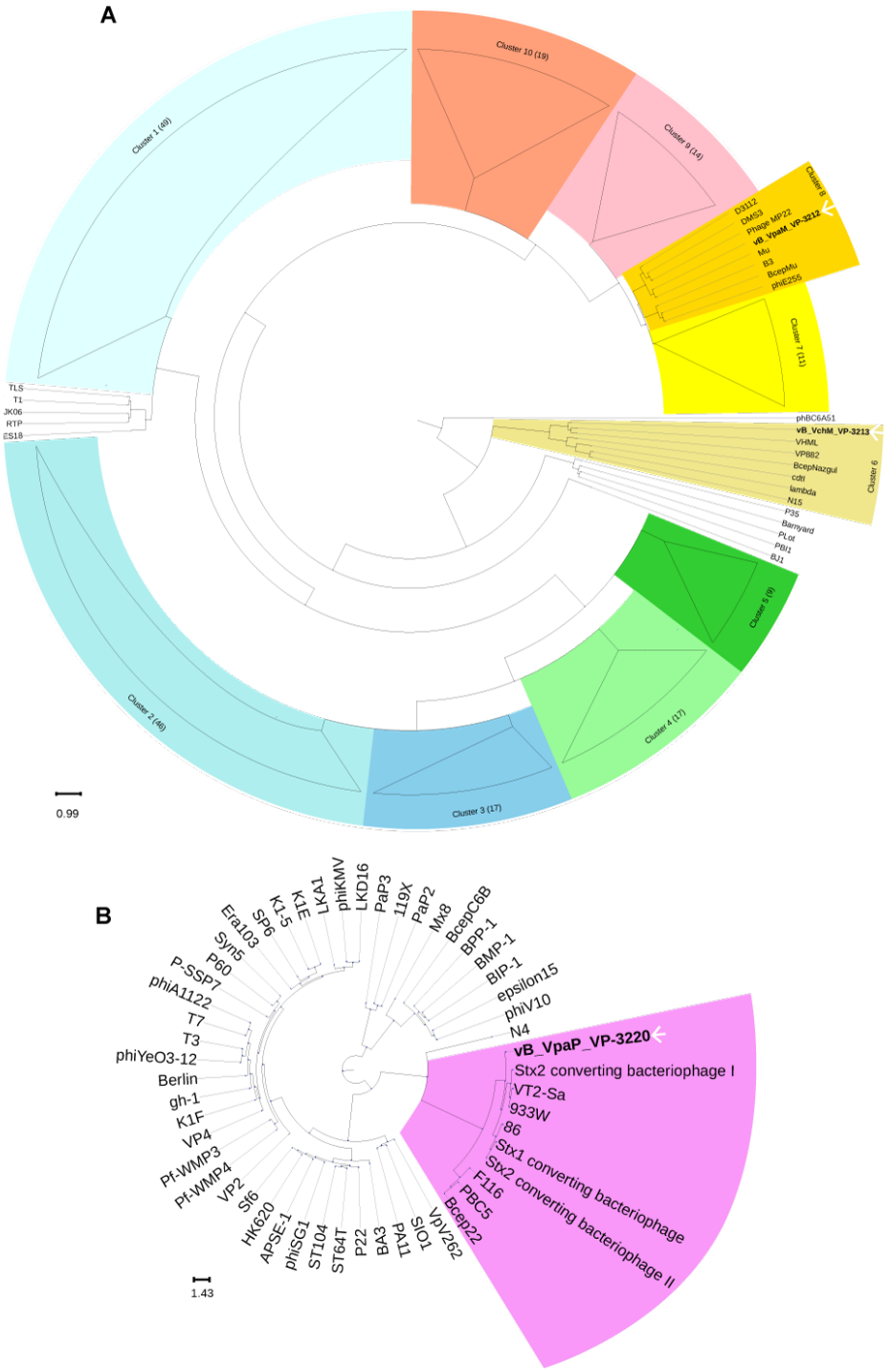


FIGURE A.4. 2. Tree representation of Myoviridae Type 1 phage clustering, including the phages vB_VpaM_VP-3212 and vB_VchM_VP-3213 (A); and Podoviridae Type 3 clustering with phage vB_VpaP_VP-3220 (B). The phage genomes were analyzed with Virfam Caudovirales classifier, the resulting trees were built automatically in Virfam with Aclame (Lopes et al., 2014).

TABLE A.4. 1. Species identification of *V. parahaemolyticus* tested for inducible prophages. (Vp = *V. parahaemolyticus*; Vc = *V.cholerae*). MALDI-TOF hit score results against Vibriobase (≥ 2.3 highly probable species; 2.0-2.29 secure genus and probable species identification), PCR results of species-specific (*toxR*), virulence-associated-gene (*trh*; *tdh*) cholera toxin gene (*ctxA*), and serotypes (O139; O1) screening, and prophage induction results using mitomycin C (+ = positive, - = negative).

Identification of *V. parahaemolyticus*

Isolate label (VibrioNet Number)	Station No.	Source	MALDI-TOF HIT-score	*VptoxR*	*trh*	*tdh*	Inducible prophage
VN-3216[a]	1	Seawater	2.57	+	-	-	-
VN-3218[a]	1	Seawater	2.48	+	-	-	+
VN-3257[b]	1	Seawater	2.49	+	-	-	-
VN-3234[b]	1	Microplastic particle	2.47	+	-	-	+
VN-3268[b]	2	Seawater	2.41	+	-	-	+
VN-3222[a]	3	Seawater	2.48	+	-	-	+
VN-3265[b]	3	Seawater	2.6	+	-	-	-
VN-3223[a]	3	Microplastic particle	2.64	+	-	-	-
VN-3225[b]	3	Microplastic particle	2.42	+	-	-	-
VN-3228[b]	3	Microplastic particle	2.22	+	-	-	-
VN-3231[b]	3	Microplastic particle	2.38	+	-	-	+
VN-3255[b]	4	Seawater	2.34	+	-	-	-
VN-3266[b]	4	Seawater	2.24	+	-	-	-
VN-3275[b]	4	Seawater	2.6	+	-	-	-
VN-3251[b]	5	Seawater	2.49	+	-	-	-
VN-3212*	**6**	**Seawater**	**2.37**	**+**	**-**	**-**	**+**
VN-3261[b]	9	Seawater	2.41	+	-	-	+
VN-3273[b]	10	Seawater	2.6	+	-	-	+
VN-3278[b]	10	Seawater	2.53	+	-	-	-
VN-3221[a]	12	Seawater	2.56	+	-	-	-
VN-3254[b]	13	Seawater	2.33	+	-	-	-
VN-3249[b]	13	Seawater	2.38	+	-	-	+
VN-3263[b]	13	Seawater	2.32	+	-	-	-
VN-3272[b]	13	Seawater	2.5	+	-	-	-
VN-3274[b]	13	Seawater	2.45	+	-	-	-
VN-3285[b]	13	Seawater	2.6	+	-	-	-
VN-3240[b]	14	Seawater	2.44	+	-	-	-
VN-3247[b]	14	Seawater	2.64	+	-	-	+
VN-3220[a]	**15**	**Seawater**	**2.55**	**+**	**-**	**-**	**+**
VN-3258[b]	15	Seawater	2.69	+	-	-	+

Identification of *V. cholerae*

Isolate label (VibrioNet Number)	Station No.	Source	MALDI-TOF HIT-score	*VctoxR*	*ctxA*	O139 /O1	Inducible prophage
VN-3213[a]	**15**	**Seawater**	**2.39**	**+**	**-**	**-**	**+**

References

[a] This study.
[b] Kirstein Kirstein et al. (2016)

TABLE A.4. 2. Coding sequences (CDS) identified in the vB_VpaM_VP-3212 phage, including the position of transcription and sequence length of each CDS. The product was defined after comparison of automated annotation and genome context comparison. Description details include additional information based on annotation software used (Rast, Prokka, Metavir) or genome synteny comparison of the tBLASTx similarity and genome context from the most similar phage genomes (synteny), as well as the protein superfamily, defined with Virfam Caudovirales classifier. Synteny details and Virfam results were not used for final annotated flat file.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS01	735 – 1	735	Transcriptional regulator, C	• Phage repressor protein C from *Vibrio* phage martha 12B12 (synteny). • Transcriptional regulator prophage MuSo2, and Cro/CI family from *Shewanella oneidensis* MR-1 (Rast). • Product similar to phage_cluster_895_PFAM-Peptidase_S24 (Metavir).
CDS02	909 – 1133	225	Transcriptional regulator, Ner family	• Negative regulator of early transcription from Enterobacteria phage Mu (synteny). • Ner-like regulatory protein from *Photorhabdus asymbiotica* subsp. *asymbiotica* and *Salmonella enterica* subsp. *enterica* serovar *Choleraesuis* str. SC-B67 (Rast). • Product similar to phage_cluster_9975_PFAM-HTH_35 (Metavir).
CDS03	1140 – 3134	1995	Transposase	• Transposase from Enterobacteria phage Mu and *Vibrio* phage martha 12B12 (synteny). • Putative mobile element protein, similar to transposase from *S. oneidensis* MR-1, and DNA transposition protein A from Enterobacteria phage (Rast). • DDE-recombinase A (Prokka). • Product similar to phage_cluster_6696_PFAM-DDE_2 (Metavir).
CDS04	3170 – 4117	948	Mobile element	• DNA transposition protein from Enterobacteria phage Mu and *Vibrio* phage martha 12B12 (synteny). • Mobile element protein, similar to putative prophage MuSo2, DNA transposition protein from *S. oneidensis* MR-1; and DNA transposition protein B from Enterobacteria phage Mu (Rast).
CDS05	4126 – 4365	240	Hypothetical protein	• CDS05-06 sequence similar to phage_cluster_6695_PFAM-Phage-MuB_C (Metavir).
CDS06	4369 – 4521	153	Hypothetical protein	
CDS07	4530 – 4904	375	Hypothetical protein	
CDS08	4897 – 5220	324	kil	• kil protein from Enterobacteria phage Mu (Synteny). • Hypothetical protein (Rast).
CDS09	5230 – 5424	195	Hypothetical protein	
CDS10	5414 – 6031	618	Hypothetical protein	• hypothetical protein from *Vibrio* phage martha 12B12 (synteny). • hypothetical protein from *S. oneidensis* MR-1 (Rast).
CDS11	6044 – 6316	273	Hypothetical protein	• hypothetical protein from *Vibrio* phage martha 12B12 (synteny).

TABLE A.4. 2. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS12	6309 - 6527	219	Hypothetical protein	• hypothetical protein from *Vibrio* phage martha 12B12 (synteny).
CDS13	6520 - 6747	228	Hypothetical protein	• hypothetical protein from *Vibrio* phage martha 12B12 (synteny).
CDS14	6744 - 7310	567	DNA methyl transferase	• phage-associated DNA methyl transferase from Bacteriophage K139 and Phage protein from K139-like from *Vibrio* phage kappa (Rast).
CDS15	7307 - 7579	273	Hypothetical protein	• hypothetical protein from *S. oneidensis* MR-1 (Rast). • Product similar to gi_481019186_ref_YP_007877541.1_ (Metavir).
CDS16	7576 - 8109	534	Hypothetical protein	• hypothetical protein from *Haemophilus ducreyi* 35000HP (Rast). • Product similar to phage cluster 1901 (Metavir).
CDS17	8112 - 8309	198	Hypothetical protein	• Product similar to phage cluster 682 (Metavir).
CDS18	8398 - 8634	237	Hypothetical protein	
CDS19	8643 - 8822	180	Hypothetical protein	
CDS20	8826 - 9341	516	Hypothetical protein	• hypothetical protein from *S. oneidensis* MR-1 and *Acidovorax avenae* subsp. *citrulli* AAC00-1 (Rast).
CDS21	9334 - 9810	477	Hypothetical protein	
CDS22	9879 - 10280	402	Mor	• Middle operon regulator (Mor) from Enterobacteria phage Mu (synteny). • Putative prophage MuSo2, positive regulator of late transcription from *S. oneidensis* MR-1 (Rast).
CDS23	10367 - 10933	567	Secretion activator protein	• Secretion activator protein from *S. oneidensis* MR-1 and *Delftia acidovorans* SPH-1 (Rast).
CDS24	10933 - 11205	273	Putative Mor	• Hypothetical protein from *S. oneidensis* MR-1 (Rast).
CDS25	11190 - 11786	597	Hypothetical protein	• Hypothetical protein from *S. oneidensis* MR-1 (Rast). • Product similar to Phage_cluster_253_PFAM-Glyco_hydro_108 (Metavir).
CDS26	11786 - 12001	216	C4-type zinc finger protein	• C4-type zinc finger protein from *Vibrio* phage martha 12B12 (synteny). • Product similar to gi_481019176_ref_YP_007877531.1_ (Metavir).
CDS27	12001 - 12306	306	Hypothetical protein	• Hypothetical protein from *S. oneidensis* MR-1 (Rast). • Product similar to gi_481019176_ref_YP_007877531.1_ (Metavir).
CDS28	12315 - 12608	294	Mu-like prophage protein gp26	• Mu-like phage gp26 from *S. oneidensis* MR-1; and phage protein from Enterobacteria phage Mu (Rast).
CDS29	12618 - 13184	567	Terminase, small subunit gp27	• Phage terminase, small subunit from *S. oneidensis* MR-1; and Mu-like phage gp27 from *H. ducreyi* 35000HP (Rast).
CDS30	13184 - 13375	192	Phage protein	

TABLE A.4.2. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS31	13496 - 15058	1563	Terminase, large subunit gp28	• Portal protein (gp28) from Enterobacteria phage Mu and portal protein (VPCG_00023) from *Vibrio* phage martha 12B12 (synteny). • Putative portal protein prophage MuSo2 from *S. oneidensis* MR-1; Mu-like phage gp28 protein from *H. ducreyi* 35000HP (Rast). • Terminase (TermL) from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS32	15055 - 16617	1563	Portal protein gp29	• Hypothetical protein (VPCG_00022) from *Vibrio* phage martha 12B12 (synteny). • Mu-like prophage FluMu protein gp29 from *S. oneidensis* MR-1 and *Azorhizobium caulinodans* ORS 571 (Rast). • Portal protein from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS33	16610 - 17389	780	Putative capsid assembly protein F	• Virion morphogenesis late F ORF from Enterobacteria phage Mu and F protein (VPCG_00021) from *Vibrio* phage martha 12B12 (synteny). • Phage (Mu-like) virion morphogenesis protein from *S. oneidensis* MR-1 (Rast). • Product sequence part of Phage_cluster_1285_PFAM-DUF935 (Metavir).
CDS34	17610 - 17470	141	Hypothetical protein	
CDS35	17691 - 18668	978	Protease I, gp32	• I protein (VPCG_00020) from *Vibrio* phage martha 12B12 (synteny). • Putative prophage MuSo2, protein Gp32 from *S. oneidensis* MR-1; and Mu-like prophage I protein from *A. caulinodans* ORS 571 (Rast).
CDS36	18668 - 19567	900	Major capsid protein	• Phage major capsid protein from Enterobacteria phage Mu and *Vibrio* phage martha 12B12 (synteny). • Phage major capsid protein #Fam0046 from *S. oneidensis* MR-1, and *Desulfovibrio vulgaris* str. Miyazaki F (Rast). • Major capsid protein (MCP) from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS37	19646 - 19903	258	Hypothetical protein	• Hypothetical protein from *Vibrio* phage martha 12B12 (synteny).
CDS38	19906 - 20445	540	Hypothetical protein	• Hypothetical protein from *Vibrio* phage martha 12B12 (synteny). • Hypothetical protein from *S. oneidensis* MR-1 (Rast).
CDS39	20473 - 21057	585	Hypothetical protein	• Hypothetical protein from *Vibrio* phage martha 12B12 (synteny). • Hypothetical protein from *S. oneidensis* MR-1; and phage protein from unclassified SPO1-like viruses *Listeria* phage A511 (Rast).
CDS40	21076 - 21510	435	Adaptor protein gp36	• Gene product J. • Mu-like prophageprotein GP36 from *S. oneidensis* MR-1 (Rast). • Adaptor protein (Ad1) from Myoviridae of Type1 (Cluster 8) (Virfam).

TABLE A.4.2. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS41	21510 - 22052	543	Neck protein Ne1	• Phage virion morphogenesis protein (VPCG_00014) from *Vibrio* phage martha 12B12 (synteny). • Putative prophage MuSo2, virion morphogenesis protein from *S. oneidensis* MR-1 (Rast). • Neck protein (Ne1) from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS42	22049 - 22639	591	Tail competition protein gp37	• Hypothetical protein from *Vibrio* phage martha 12B12 (synteny). • Mu-like prophage FluMu protein gp37 from *S. oneidensis* MR-1 (Rast). • Tail competition protein (Tc) from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS43	22653 - 22889	237	Hypothetical protein	• Hypothetical protein from Enterobacteria phage Mu and *Vibrio* phage martha 12B12 (synteny).
CDS44	22892 - 24370	1479	Tail sheath protein gpL	• Major tail subunit from Enterobacteria phage Mu; and tail sheath protein GpL (VPCG_00011) from *Vibrio* phage martha 12B12 (synteny). • Bacteriophage tail sheath protein from *S. oneidensis* MR-1; and tail sheath protein from *Escherichia* phage D108 (Rast). • Sheath protein from Myoviridae of Type1 (Cluster 8) (Virfam).
CDS45	24386 - 24739	354	Phage tail tube protein	• Hypothetical protein with possible phage tail tube protein function from Enterobacteria phage Mu; and tail tube protein (VPCG_00010) from *Vibrio* phage martha 12B12 (synteny). • Phage tail tube protein from *S. oneidensis* MR-1 (Rast).
CDS46	24739 - 25110	372	Tail assembly protein gp41	• Hypothetical protein with possible phage tail assembly chaperone proteins, E, or 41 or 14 function from Enterobacteria phage Mu; and extracellular solute-binding protein family 7 protein (VPCG_0009) from *Vibrio* phage martha 12B12 (synteny). • Hypothetical protein from *V. cholerae* O1 biovar el tor str. N16961; and Mu-like prophage FluMu protein gp41 from *S. oneidensis* MR-1 (Rast).
CDS47	25227 - 26903	1677	Tape measure protein gp42	• Phage tail measure protein (VPCG_0008) from *Vibrio* phage martha 12B12 (synteny). • Similar to Mu-like prophage protein from *V. cholerae* O1 biovar el tor str. N16961; and Mu-like prophage FluMu protein gp42 from *S. oneidensis* MR-1 (Rast).
CDS48	26915 - 28234	1320	Phage tail/DNA circulation protein N	• DNA circulation protein (VPCG_00007) from *Vibrio* phage martha 12B12 (synteny). • Phage tail/DNA circulation protein from *S. oneidensis* MR-1 and *Shigella flexneri* bacteriophage V (Rast).

TABLE A.4.2. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS49	28227 - 29321	1095	Baseplate hub protein	• Low similarity with putative tail protein (phage late control gene D protein, GPD) from Enterobacteria phage Mu, and highly similar to P protein (VPCG_00006) from *Vibrio* phage martha 12B12 (synteny). • Prophage tail protein (Phage late control gene D protein) from *S. oneidensis* MR-1 and *Chromobacterium violaceum* ATCC 12472 (Rast).
CDS50	29322 - 29927	606	Baseplate puncturing device gp45	• Baseplate assembly protein gp45 from Enterobacteria phage Mu, and hypothetical protein (VPCG_00005) from *Vibrio* phage martha 12B12 (synteny). • Prophage baseplate assembly protein V from *S. oneidensis* MR-1; and phage baseplate from Enterobacteria phage Mu (Rast).
CDS51	29930 - 30382	453	Baseplate protein gp46	• Hypothetical protein with phage protein gp46 function from Enterobacteria phage Mu, and hypothetical protein (VPCG_00004) from *Vibrio* phage martha 12B12 (synteny). • Bacteriophage protein gp46 from *S. oneidensis* MR-1 (Rast).
CDS52	30372 - 31439	1068	Baseplate protein gp47	• Hypothetical protein with phage protein gp47 function from Enterobacteria phage Mu, and baseplate J-like protein (VPCG_00003) from *Vibrio* phage martha 12B12 (synteny). • Phage FluMu protein gp47 from *S. oneidensis* MR-1; and phage protein Enterobacteria phage Mu (Rast).
CDS53	31424 - 32011	588	Tail protein	• Hypothetical protein from Enterobacteria phage Mu and *Vibrio* phage martha 12B12 (synteny). • Prophage tail protein from *S. oneidensis* MR-1 (Rast).
CDS54	32032 - 33918	1887	Long tail fiber protein	• Prophage long tail protein from *S. oneidensis* MR-1 (Rast).
CDS55	33929 - 35935	2007	5-methylcytosine-specific restriction related	• Conserved domain protein from *Flavobacterium psychrophilum* JIP02/86, and hypothetical protein from *Desulfotalea psychrophila* LSv54 (Rast).

TABLE A.4. 3. Coding sequences (CDS) identified in the vB_VpaP_VP-3220 phage, including the position of transcription and sequence length of each CDS. The product was defined after comparison of automated annotation and genome context comparison. Description details include additional information based on annotation software used (Rast, Prokka, Metavir) or genome synteny comparison of the tBLASTx similarity and genome context from the most similar phage genomes (synteny), as well as the protein superfamily, defined with Virfam Caudovirales classifier. Synteny details and Virfam results were not used for final annotated flat file.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS01	303 - 1463	1161	Integrase	• Similar to site-specific recombinase, phage integrase from unclassified Lambda-like viruses Stx2-converting phage 1717, and Enterobacteria phage 2851 (Rast). • Hypothetical protein (Prokka) • Product function similar to phage_cluster_3275_PFAM-Phage_integrase (Metavir).
CDS02	1870 - 1445	426	Transcriptional regulator	• XRE family transcriptional regulator (WP_020840939) from *V. parahaemolyticus* (tBLASTx). • Similar to hypothetical protein (Rast, Prokka).
CDS03	2088 – 1879	210	Hypothetical protein	
CDS04	2699 - 2073	627	Adenine DNA methyltransferase	• Adenine methylase (VPAG_00059) from *Vibrio* phage douglas 12A4 (synteny). • Similar to adenine DNA methyltransferase, phage-associate from unclassified Lambda-like viruses Stx2-converting phage 86; and adenine methylase from *Streptococcus pyogenes* MGAS2096 (Rast). • Product function similar to phage_cluster_395_PFAM-MT-A70 (Metavir).
CDS05	3886 - 2696	1191	Hypothetical protein	• Product function similar to hypothetical protein VPAG_00058 from *Vibrio* phage douglas 12A4, part of phage cluster gi_481019127_ref_YP_007877483.1_ (Metavir).
CDS06	4710 - 3886	825	Phage protein	• Similar to phage protein from unclassified Lambda-like viruses Enterobacteria phage cdtI, and phage protein from *Shigella flexneri* bacteriophage V (Rast). • Product function similar to phage_cluster_4539_PFAM-DUF2303 (Metavir).
CDS07	5152 - 4724	429	Hypothetical protein	
CDS08	6111 - 5476	636	Hypothetical protein	• Similar to hypothetical protein from *Vibrio fischeri* ES114 (Rast). • Product function similar to phage_cluster_667 (Metavir).
CDS09	6403 - 6717	315	Hypothetical protein	
CDS10	6792 - 7751	960	Replication initiation protein	• Product function similar to phage_cluster_1092 (Metavir).
CDS11	7696 - 8181	486	Hypothetical protein	• Product function similar to hypothetical protein VPGG_00001 from *Vibrio* phage VBM1, part of phage cluster gi_472340789_ref_YP_007674310.1_ (Metavir).
CDS12	8178 - 8543	366	Hypothetical protein	
CDS13	8606 - 9007	402	Hypothetical protein	• Product function similar to hypothetical protein, part of phage_cluster_5210 (Metavir).

TABLE A.4.3. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS14	9004 - 9309	306	Phage protein	Product function similar to phage_cluster_2433_PFAM-DUF4406 (Rast).
CDS15	9317 - 9466	150	Hypothetical protein	
CDS16	9469 - 9660	192	Hypothetical protein	
CDS17	9704 - 10513	810	Phage antirepressor protein	• Phage antirepressor protein from *Acidithiobacillus ferrooxidans* ATCC 23270 (Rast). • Product function similar to Phage_cluster_12_PFAM-Bro-N (Metavir).
CDS18	10527 - 10799	273	Hypothetical protein	
CDS19	10877 - 11335	459	Hypothetical protein	• Product function similar to hypothetical protein VPAG_00034 from *Vibrio* phage douglas 12A4, part of phage cluster gi_481019103_ref_YP_007877459.1_ (Metavir).
CDS20	11405 - 11701	297	Hypothetical protein	
CDS21	11944 - 12372	429	Hypothetical protein	
CDS22	12296 - 12556	261	Hypothetical protein	
CDS23	12477 - 12869	393	Hypothetical protein	
CDS24	12880 - 13386	507	Hypothetical protein	• Product function similar to hypothetical protein VPAG_00017 from *Vibrio* phage douglas 12A4, part of phage cluster gi_481019086_ref_YP_007877442.1_ (Metavir).
CDS25	13373 - 13651	279	Hypothetical protein	
CDS26	13656 - 13865	210	Hypothetical protein	
CDS27	14658 - 14203	456	Hypothetical protein	
CDS28	15850 - 14873	978	Hypothetical protein	• Product function similar to ORF54 from *Lactococcus* phage TP901-1, part of phage cluster gi_13786585_ref_NP_112717.1__PFAM-FRG (Metavir).
CDS29	16529 - 15996	534	Putative DnaJ-related protein	
CDS30	17153 - 16728	426	Hypothetical protein	
CDS31	17673 - 17314	360	Hypothetical protein	
CDS32	18181 - 17825	357	Hypothetical protein	
CDS33	19065 - 18334	732	Hypothetical protein	• Hypothetical protein from *Dechloromonas aromatica* RCB (Rast).
CDS34	19997 - 20239	243	Hypothetical protein	
CDS35	20252 - 20587	336	Hypothetical protein	• Hypothetical protein from *Vibrio sp*. MED222 (Rast).
CDS36	20628 - 21356	729	Hypothetical protein	
CDS37	21381 - 21506	126	Hypothetical protein	
CDS38	21599 - 22021	423	Phage-related hypothetical protein	• Similar to phage-related hypothetical protein from *Aeromonas hydrophila* subsp. hydrophila ATCC 7966 (Rast). • Product similar to Phage_cluster_2683 (Metavir).

TABLE A.4.3. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS39	22060 - 22977	918	Hypothetical protein	• Product function similar to hypothetical protein VPAG_00011 from *Vibrio phage* douglas 12A4, part of phage cluster gi_481019080_ref_YP_007877436.1_ (Metavir).
CDS40	22967 - 24760	1794	Terminase large subunit	• Terminase large subunit (pMIN27_52) from Enterobacteria phage Min27, and large subunit terminase (VPAG_00010) from *Vibrio* phage douglas 12A4 (synteny). • Phage protein from Podoviridae Enterobacteria phage Min27; DNA packaging from Bacteriophage 933W; and phage protein from Stx1 converting bacteriophage (Rast). • Product function similar to phage_cluster_1744_PFAM-Pfam-B_3639 (Metavir). • Terminase (TermL) from Podoviridae of Type3 (Virfam).
CDS41	24772 - 26877	2106	Portal protein	• Putative portal proteins (pMIN27_53–54) from Enterobacteria phage Min27, and portal protein (VPAG_00009) from *Vibrio* phage douglas 12A4 (synteny). • DNA packaging from Bacteriophage 933W; and phage protein from Bacteriophage VT2-Sa (Rast). • Product function similar to phage_cluster_2090_PFAM-Pfam-B_7381 (Metavir). • Portal protein from Podoviridae of Type3 (Virfam).
CDS42	26897 - 27409	513	Phage protein	• Product function similar to probable DNA polymerase, part of phage_cluster_933 (Metavir).
CDS43	27414 - 27731	318	Hypothetical protein	• Product function similar to phage_cluster_1131_PFAM-Phage_gp49_66 (Metavir).
CDS44	27758 - 28243	486	Phage protein	• Product function similar to Phage_cluster_4449 (Metavir).
CDS45	28451 - 29563	1113	Phage protein	• Hypothetical protein (pMIN27_55) from Enterobacteria phage Min27 (synteny). • Phage protein from unclassified Lambda-like virus Stx2-converting phage 86, and phage protein from Bacteriophage VT2-Sa (Rast). • Product function similar to phage_cluster_13759 (Metavir).
CDS46	29578 - 30801	1224	Major capsid protein	• Hypothetical protein (pMIN27_56) from Enterobacteria phage Min27 (synteny). • Phage protein from unclassified Lambda-like virus Stx2-converting phage 86, and phage protein from Bacteriophage 933W (Rast). • Product function similar to phage_cluster_7279_PFAM-DUF4043 (Metavir). • Major capsid protein (MCP) from Podoviridae of Type3 (Virfam).
CDS47	30861 - 31307	447	Hypothetical protein	
CDS48	31421 - 31885	465	Phage protein	• Product function similar to hypothetical protein VPAG_00004 from *Vibrio* phage douglas 12A4, part of phage cluster gi_481019073_ref_YP_007877429.1_ (Metavir).
CDS49	31885 - 32463	579	Adaptor protein	• Product part of gi_481019072_ref_YP_007877428.1_ (Metavir). • Adaptor protein (Ad3) from Podoviridae of Type3 (Virfam).

TABLE A.4.3. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS50	32472 - 33143	672	Phage protein	• Hypothetical protein (pMIN27_60) from Enterobacteria phage Min27 (synteny). • Phage protein from unclassified Lambda-like viruses Stx2-converting phage 86, and phage protein from Stx2 converting bacteriophage I (Rast). • Product function similar to hypothetical protein, part of phage_cluster_1736 (Metavir).
CDS51	33163 - 37374	4212	Putative phage tail fiber protein	• Conserved protein of unknown function from *Pseudoalteromonas haloplanktis* TAC125 (Rast). • Product function similar to hypothetical protein VPDG_00121 from *Vibrio phage* henriette 12B8, part of phage cluster gi_481019742_ref_YP_007878093.1_ (Metavir).
CDS52	37371 - 38231	861	Hypothetical protein	• Product function similar to hypothetical protein VPDG_00120 from *Vibrio phage* henriette 12B8, part of phage cluster gi_481019741_ref_YP_007878092.1_ (Metavir).
CDS53	38353 - 38577	225	Hypothetical protein	
CDS54	38577 - 40187	1611	Putative head-closure phage protein	• Hypothetical protein (pMIN27_64) from Enterobacteria phage Min27 (synteny). • Phage protein from Bacteriophage VT2-Sa, and phage protein from Stx2 converting bacteriophage I (Rast). • Product function similar to probable scaffolding protein, part of phage_cluster_2027 (Metavir). • Head-closure protein (Hc3) from Podoviridae of Type3 (Virfam).
CDS55	40188 - 42605	2418	Phage tail fiber protein, phage host specificity protein J	• Similar to phage tail fiber protein #Phage host specificity protein J from Stx1 converting bacteriophage, and phage tail fiber protein #Phage host specificity protein J from Stx2 converting bacteriophage I (Rast). • Product function similar to phage_cluster_9484_PFAM-DUF3672 (Metavir).
CDS56	42668 - 43084	417	Hypothetical protein	
CDS57	43077 - 43754	678	Phage protein	• Similar to phage protein from Stx2 converting bacteriophage II, and phage protein from Bacteriophage VT2-Sa (Rast). • Product function similar to phage_cluster_9048_PFAM-Pfam-B_9451 (Metavir).
CDS58	43774 - 44280	507	Phage protein	• Product function similar to hypothetical protein, part of phage_cluster_7280 (Metavir).
CDS59	44290 - 44550	261	Phage protein	
CDS60	44562 - 46313	1752	Hypothetical protein	• Product function similar to hypothetical protein VPAG_00063 from *Vibrio* phage douglas 12A4, part of phage cluster gi_481019132_ref_YP_007877488.1 (Metavir).

TABLE A.4.3. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS61	46405 - 56862	10458	Hypothetical protein	• Hypothetical protein (VPAG_00062) from *Vibrio* phage douglas 12A4 (synteny). • Conserved phage mega protein from *Stenotrophomonas maltophilia* K279a, and conserved phage mega protein from Podoviridae *Thalassomonas* phage BA3 (Rast). • Product function similar to phage_cluster_3084 (Metavir).
CDS62	56905 - 57456	552	Hypothetical protein	
CDS63	57733 - 57894	162	Hypothetical protein	

TABLE A.4. 4. Coding sequences (CDS) identified in the vB_VchM_VN-3213 phage, including the position of transcription and sequence length of each CDS. The product was defined after comparison of automated annotation and genome context comparison. Description details include additional information based on annotation software used (Rast, Prokka, Metavir) or genome synteny comparison of the tBLASTx similarity and genome context from the most similar phage genomes (synteny), as well as the protein superfamily, defined with Virfam Caudovirales classifier. Synteny details and Virfam results were not used for final annotated flat file.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS01	230 - 460	231	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to phage cluster PFAM-DUF3423 (Metavir).
CDS02	435 - 764	330	Hypothetical protein	• Hypothetical protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS03	866 - 3442	2577	Major capsid protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Major capsid protein (MCP) from Myoviridae of Type1 (Cluster 6) (Virfam).
CDS04	3522 - 4115	594	ATPase component of ABC transporter	• ATPase component of ABC transporter from *Vibrio vulnificus* CMCP6 (Rast).
CDS05	4167 - 4694	528	Hypothetical protein	• Hypothetical protein from *Vibrio vulnificus* CMCP6; and phage protein from Podoviridae *Thalassomonas* phage BA3 (Rast). • Product function similar to phage_cluster_933 (Metavir).
CDS06	4684 - 4980	297	YaeH protein	• UPF0325 function (Rast). • Chromosome segregation ATPase from *Vibrio vulnificus* CMCP6 (Rast).
CDS07	4990 - 5595	606	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS08	5588 - 6082	495	Neck protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to hypothetical protein, part of phage_cluster_3539 (Metavir). • Neck protein (Ne1) from Myoviridae of Type1 (Cluster 6) (Virfam).
CDS09	6095 - 6640	546	Tail completion protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Tail completion protein (Tc) from Myoviridae of Type1 (Cluster 6) (Virfam)
CDS10	6633 - 7139	507	Phage P2 baseplate assembly protein gpV	• Similar to phage P2 baseplate assembly protein gpV from *Vibrio vulnificus* CMCP6 (Rast).
CDS11	7142 - 7435	294	10.2 kDa protein in segC-Gp6 intergenic	
CDS12	7436 - 7783	348	dTDP-glucose pyrophosphorylase	• Similar to dTDP-glucose pyrophosphorylase from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to probable replicase of phage_cluster_1829 (Metavir).

TABLE A.4.4. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS13	7780 - 8646	867	Phage-related baseplate assembly protein	• Baseplate assembly protein gene J (VHMLp32) from *Vibrio* phage VHML, and gp34, phage baseplate assembly protein (BTHphiE202_0034) from *Burkholderia* phage phiE202 (synteny). • Phage-related baseplate assembly protein from *Vibrio vulnificus* CMCP6, phage-related baseplate assembly protein from *Stenotrophomonas maltophilia* K279a, and phage-related baseplate assembly protein from P2-like viruses *Burkholderia* phage phiE202 (Rast). • Product function similar to phage_cluster_1429_PFAM-Baseplate_J (Metavir).
CDS14	8646 - 9296	651	Phage P2-related tail formation	• gp33, phage tail protein I (BTHphiE202_0033) from *Burkholderia* phage phiE202 (synteny). • Bacteriophage P2-related tail formation protein from *Vibrio vulnificus* CMCP6, phage tail fibers from *Pseudomonas aeruginosa* PAO1, and tail protein I from P2-like viruses *Mannheimia* phage phiMHaA1 (Rast). • Product function similar to phage_cluster_1334_PFAM-Tail_P2_I (Metavir).
CDS15	9293 - 10402	1110	Phage tail fiber	• Phage-related tail fibers protein *Vibrio vulnificus* CMCP6, phage tail fibers from K139-like phages Vibrio phage K139, and phage tail fibers from bacteriophage K139 (Rast). • Product function similar to putative tail fiber protein from *Vibrio* phage VPUSM 8, part of phage cluster gi_557307513_ref_YP_008766855.1 (Metavir).
CDS16	10402 - 10920	519	Phage tail fiber	• Phage tail fibers from K139-like phages *Vibrio* phage kappa, phage tail fibers from bacteriophage K139, and phage tail fibers from K139-like phages *Vibrio* phage K139 (Rast). • Product function similar to putative tail fiber assembly protein from *Vibrio* phage VPUSM 8, • part of phage cluster gi_557307514_ref_YP_008766856.1_ (Metavir).
CDS17	11008 - 12174	1167	Tail sheath protein FI	• Major tail sheath protein gene FI (VHMLp38) from *Vibrio* phage VHML, and gp30, phage tail sheath protein (BTHphiE202_0030) from *Burkholderia* phage phiE202 (synteny). • Phage tail sheath protein FI from *Vibrio vulnificus* CMCP6, phage tail sheath monomer from *Roseobacter* sp. GAI101, and ORF39 from *Thioalkalivibrio* sp. HL-EbGR7 (Rast). • Product function similar to phage_cluster_1045_PFAM-Phage_sheath_1 (Metavir). • Sheath protein (Sheath) from Myoviridae of Type1 (Cluster 6) (Virfam).
CDS18	12187 - 12699	513	Major tail tube protein FII	• Phage tail tube protein FII from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to phage_cluster_1431_PFAM-Phage_tube (Metavir) • Major tail protein (MTP) from Myoviridae of Type1 (Cluster 6) (Virfam).
CDS19	12708 - 13064	357	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to PFAM-FluMu_gp41 (Metavir).

TABLE A.4.4. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS20	13137 - 15278	2142	Tail protein	• Putative phage tail tape measure protein, TP901 family from *Vibrio vulnificus* CMCP6, and methyl-accepting chemotaxis protein from *Pseudomonas aeruginosa* PAO1 (Rast). • Product function similar to phage tail tape measure protein from *Vibrio* phage martha 12B12, part of phage cluster gi_481019154_ref_YP_007877509.1__PFAM-PhageMin_T (Metavir).
CDS21	15288 - 15680	393	Phage tail protein U	• Phage protein U from *Vibrio vulnificus* CMCP6, and pyocin R2_PP, tail formation protein from *Synechococcus elongatus* PCC 6301 (Rast). • Product function similar to phage_cluster_1599_PFAM-Phage_P2_GpU (Metavir).
CDS22	15683 - 15889	207	Baseplate protein X	• P2-like prophage tail protein X from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to phage_cluster_1598_PFAM-Phage_tail_X (Metavir).
CDS23	15880 - 16887	1008	Phage protein D	• Phage protein D from *Vibrio vulnificus* CMCP6, phage protein D from Rhodobacter sphaeroides 2.4.1, and phage protein D *Vibrio harveyi* bacteriophage VHML (Rast). • Product function similar to phage_cluster_1048_PFAM-Phage_GPD (Metavir).
CDS24	17123 - 16884	240	Hypothetical protein	
CDS25	18116 - 17160	957	Hypothetical protein	
CDS26	18418 - 19293	876	Hypothetical protein	
CDS27	20697 - 19891	807	Transcriptional regulator	• Predicted transcriptional regulator from *Vibrio vulnificus* CMCP6 (Rast).
CDS28	20832 - 21020	189	Hypothetical protein	
CDS29	21210 - 21689	480	Phage phi gp55-like protein	• Bacteriophage phi gp55-like protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to phage_cluster_14655 (Metavir).
CDS30	21719 - 22420	702	Tyrosine recombinase XerC	• Site-specific recombinase XerC from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to phage_cluster_415_PFAM-Phage_integrase (Metavir).
CDS31	22408 - 22620	213	Hypothetical protein	
CDS32	22635 - 22946	312	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS33	22939 - 23238	300	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS34	23222 - 23527	306	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS35	23493 - 26252	2760	DNA segregation ATPase FtsK/SpoIIIE and related	• conserved hypothetical protein gp11 (BTHphiE202_0011) from *Burkholderia* phage phiE202 (synteny). • Hypothetical protein from *Vibrio vulnificus* CMCP6, and phage protein from P2-like viruses *Ralstonia* phage phiRSA1 (Rast). • Product function similar to phage_cluster_7128_PFAM-Pfam-B_9177 (Metavir).

TABLE A.4.4. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS36	26263 - 26556	294	Phage protein	• ORF51 (VHMLp49) from *Vibrio* phage VHML (synteny). • Phage protein from Myoviridae *Vibrio* phage VP882, and phage protein from *Vibrio harveyi* bacteriophage VHML (Rast). • Product function similar to phage_cluster_14375 (Metavir).
CDS37	26553 - 26810	258	Hypothetical protein	• Hypothetical protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to hypothetical protein VPUSM8P11 from *Vibrio* phage VPUSM 8, part of phage cluster gi_557307485_ref_YP_008766827.1 (Metavir).
CDS38	26797 - 27048	252	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS39	27079 - 27594	516	Transcriptional regulator, PadR family	• Transcriptional regulator, PadR family from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to putative transcriptional regulator from *Vibrio* phage vB_VpaS_MAR10, part of phage cluster gi_428782101_ref_YP_007111862.1 PFAM-PadR (Metavir).
CDS40	27791 - 27985	195	Hypothetical protein	•
CDS41	28003 - 29169	1167	Integrase	• Unknown protein and DNA-directed RNA polymerase, beta subunit/140 kD subunit, both from *Vibrio vulnificus* CMCP6, and mobile element protein from *Serratia marcescens* Db11 (Rast). • Product function similar to hypothetical protein, part of phage_cluster_6427 (Metavir)
CDS42	30016 - 29453	564	Secretion activating protein	• ORF19 (VHMLp19) from *Vibrio* phage VHML (synteny). • Putative secretion activating protein *Vibrio vulnificus* CMCP6, secretion protein from *Vibrio harveyi* bacteriophage VHML, and phage protein from Myoviridae *Vibrio* phage VP882 (Rast). • Product function similar to phage_cluster_1376_PFAM-Glyco_hydro_108 (Metavir).
CDS43	30657 - 31703	1047	Tyrosine recombinase XerD	• Site-specific recombinase XerD from *Vibrio vulnificus* CMCP6, and integrase/recombinase (XerC/CodV family) from *Acidobacteria bacterium* Ellin345 (Rast). • Product function similar to phage_cluster_415_PFAM-Phage_integrase (Metavir).
CDS44	31700 - 31963	264	10 kDa chaperonin groS3	• Heat shock protein 60 family co-chaperone GroEs from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to PFAM-Cpn10 (Metavir).
CDS45	32106 - 33251	1146	Hypothetical protein	
CDS46	33656 - 34351	696	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).
CDS47	34348 - 35313	966	Hypothetical protein	
CDS48	35306 - 36904	1599	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast).

TABLE A.4.4. Continuation.

CDS	Position of transcription	Nucleotides (bp)	Product	Description details
CDS49	36901 - 38280	1380	Portal protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Portal protein (Portal) from Myoviridae of Type1 (Cluster 6) (Virfam).
CDS50	38286 - 39374	1089	Hypothetical protein	• Unknown protein from *Vibrio vulnificus* CMCP6 (Rast). • Product function similar to hypothetical protein, part of phage_cluster_4063 (Metavir).
CDS51	39465 - 39656	192	Hypothetical protein	
CDS52	39748 - 40416	669	hypothetical protein	
CDS53	40825 - 40451	375	hypothetical protein	

149

A.5. SUPPLEMENT TO CHAPTER 4

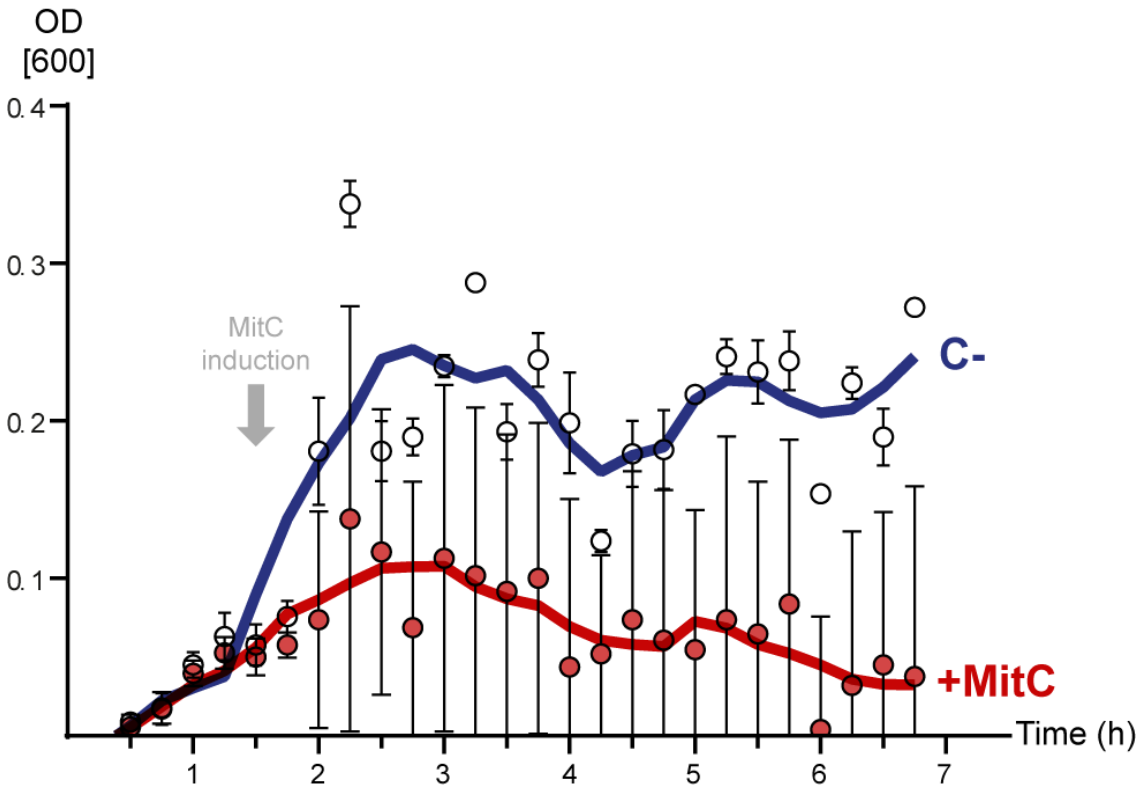


FIGURE A.5.1. Prophage induction screening of *V. parahaemolyticus* VN-3218. Mitomycin C (+MitC, red) and MB-50% broth (control C-, blue) was added to the culture at optical density 0.1-0.2 (600 nm) (MitC induction, grey arrow) and incubated at 37 °C and shaking. The induction test (+MitC) was performed in eight replicates, and the negative control (C-) was performed in duplicate. The error bar indicates the standard error of the mean from replicate experiments.

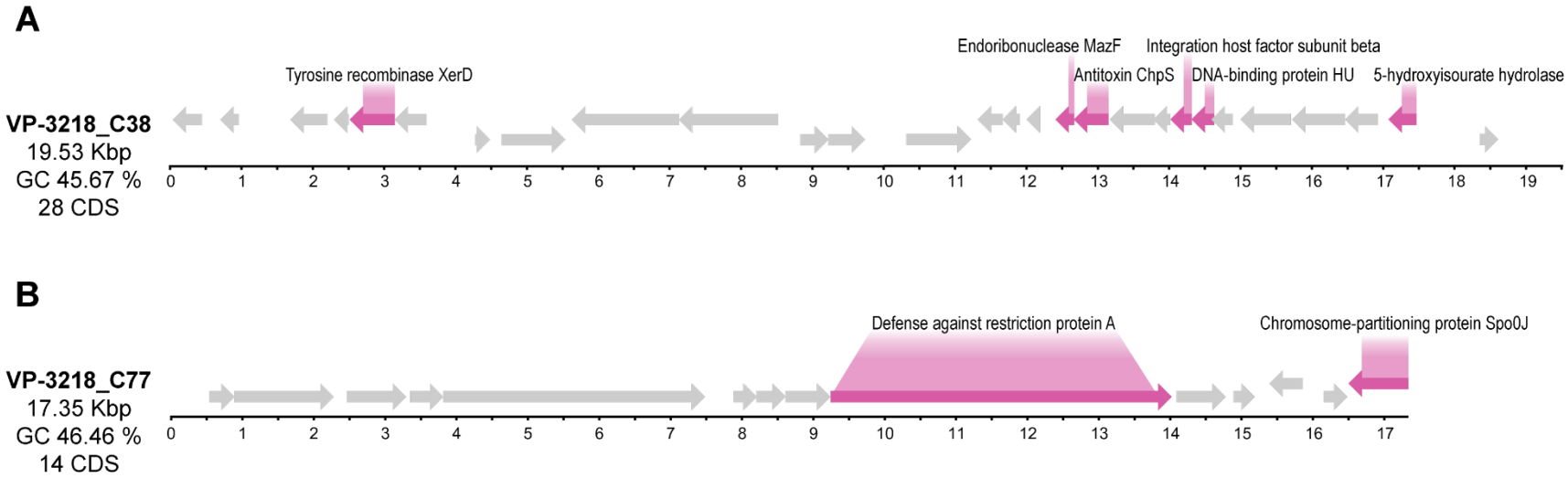


FIGURE A.5.2. Incomplete prophages sequences from *V. parahaemolyticus* VN-3218. The incomplete prophages VP-3218_C38 (A) and VP-3218_C77 (B) were identified from the to the assembled contig 38 and 77, respectively. Sequences were assembled with SPAdes, identified with PHASTER and annotated with Prokka (section 4.2.3). Predicted hypothetical proteins are marked in grey, automatically annotated proteins are marked in pink with their corresponding function.

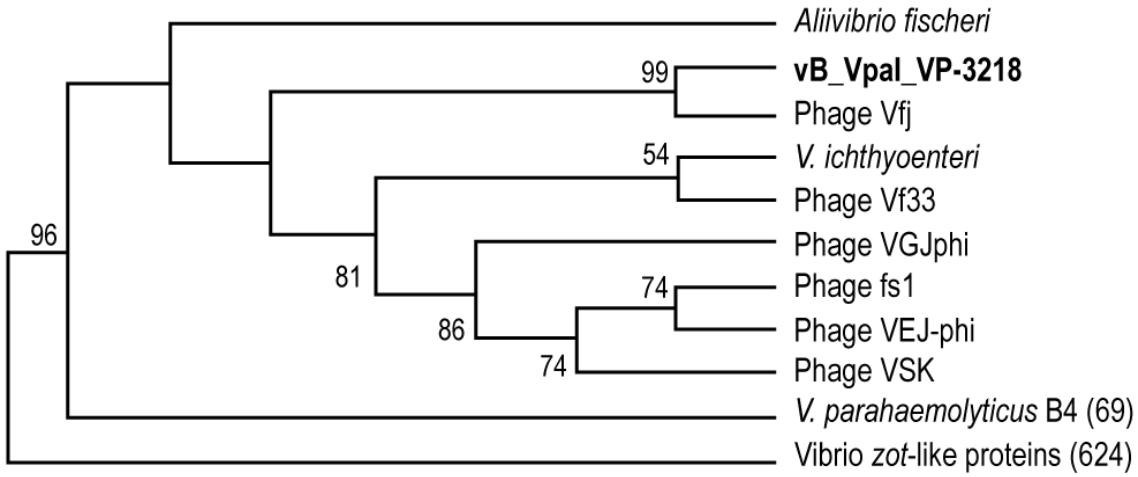


FIGURE A.5.3. Maximum Likelihood tree of zot-encoding prophages. Unrooted phylogenetic tree constructed from 702 zot-like toxin amino acid sequences from and the CDS08 encoding sequence of vB_VpaI_VP-3218, using the maximum likelihood algorithm with 1,000 bootstrap replicates. The tree shows the most similar sequences corresponding to *V. parahaemolitycus* B4 cluster and *Vibrio* phages. Numbers at the nodes represent percent bootstrap support where unlabeled nodes had bootstraps of > 50 %.

TABLE A.5.1. Coding sequences (CDS) identified in the vB_VpaI_VP-3218 phage and their description details.

CDS	Position of transcription	Nucleotides (bp)	Product (gene abbr., Figure 4.2A)	Description details		
				Analogous protein (based on genome position)	Rast/Seed	VirSorter cluster (Metavir+MAG)
CDS01	1 - 2031	2031	Replication gene A protein (*repA*)	Analogous to replication protein (orf698) from VFJ.	Similar to RstA-like protein from *V. cholerae*, RstA-like protein *Vibrio* virus CTXphi; and putative replication protein from *Photobacterium profundum* 3TCK.	
CDS02	2109 - 2420	312	Putative replication gene B protein (*repB*)	Analogous to ssDNA-binding V protein (orf104) from VFJ.		Product function homologous to DNA replication protein DnaC, part of phage_cluster_ 1359_PFAM-Phage_GPA.
CDS03	2431 - 2550	120	Hypothetical protein			
CDS04	2551 - 2667	117	Putative minor capsid protein (*gIX*)	Analogous to minor coat IX protein (orf38-2) from VFJ.		
CDS05	2691 - 2954	264	Putative major coat protein (*gVIII*)		Similar to phage major capsid protein from Fs-2; and phage major capsid protein of *P. profundum* 3TCK.	
CDS06	3032 - 4657	1626	Putative receptor-binding protein (*gIII*)	Analogous to attachment III protein (orf499) from VFJ.	Similar to phage protein from phage VfO3K6.	

TABLE A.5.1. Continuation.

CDS07	4657 - 5016	360	Putative minor coat protein (*gVI*)	Analogous to minor coat VI protein (orf124) from VFJ.	Similar to putative capsid protein of *P. profundum* 3TCK; and phage minor capsid protein from Fs-2.	Product function homologous to minor coat VI protein from VFJ, part of phage cluster gi_514361039_ref_ YP_008130280.1
CDS08	5016 - 6089	1074	Phage morphogenesis protein (*gI*)	Analogous to I protein (orf361) from VFJ, with possibly zot toxin function.	Similar to phage maturase Inoviridae G1p from Fs-2; and gene I protein from *P. profundum* 3TCK	Product function homologous to Phage_cluster_4966 _PFAM-Zot.
CDS09	6107 - 7459	1353	Phage morphogenesis protein (*gIV*)	Analogous to IV protein (orf442) from VFJ.	Type II secretion pathway protein D, virion export protein. Similar to Gene IV protein from *P. profundum* 3TCK; and phage assembly protein from Fs-2.	
CDS10	8135 - 7497	639	Hypothetical protein	Contains restriction endonuclease-like superfamily domain.	Similar to hypothetical protein from *Nitratiruptor sp.* SB155-2.	
CDS11	8325 - 9428	1104	putative ATP/GTP phosphatase (*ATP/GTP phos*)			
CDS12	9431 - 10081	651	Hypothetical protein			

TABLE A.5.1. Continuation**.**

CDS13	10381 - 10109	273	Transcriptional regulator (*regA*)	Bacteriophage f237 ORF9. Similar to phage protein from phage VfO3K6, and transcriptional regulator of phage VEJphi	Product homologous to hypothetical protein of Phage_cluster_7477
CDS14	10611 - 10450	162	Transcriptional regulator (*regB*)	Similar to bacteriophage f237 ORF10 from *P. profundum* 3TCK; and transcriptional regulator of phage VGJphi	Product homologous to Phage_cluster_5834

ACKNOWLEDGEMENTS

The opportunity to do my PhD in Germany and accomplish my dream to study abroad could be a reality thanks to the scholarship of German Academic Exchange Service (DAAD). I would like to thank the opportunities given by the International Max Planck Research School of Marine Microbiology (MarMic). The research of this thesis was funded by the Alfred Wegener Institute for Polar and Marine Research (AWI), Micro B3 (Marine Microbial Biodiversity, Bioinformatics, Biotechnology) and Ocean Sampling Day. I would like to thank the whole team of the RV Heincke (AWI) for this support of the samples taken for this thesis. I am grateful for the scientific support and the laboratory spaces to perform my thesis in the best possible way at the AWI Biologische Anstalt Helgoland (BAH), as well as the bioinformatic support and technology access at the Max Planck Institute for Marine Microbiology Bremen.

I would like to thank my PhD thesis committee, namely Prof. Dr. Frank Oliver Glöckner, Prof. Dr. Matthias Ullrich, Dr. Antje Wichels, and Dr. Gunnar Gerdts for the great supervision during this process. Thanks to all of you for your valuable insights and support to solve the obstacles throughout the course of my research project. I am deeply thankful to Frank Oliver Glöckner, to trust on my research and give me all the support since the first application for DAAD scholarship. Thanks for all your feedback, to give me the time for fruitful scientific discussion, and accept me as part of your working group.

I would like to thank to Ivaylo Kostadinov for his support for the sequence and genome submissions, Emiliano Pereira-Flores for all the support and advices for the bioinformatic analyses, and Georg Krohne for his support for the TEM pictures. I would like to thank to Julian Schubert, Leoni Mack, Nina Heinzmann and Candice Thorstenson for your valuable support during this thesis and help me to be a better scientist.

I want to express my gratitude to my direct supervisor Antje Wichels. Thank you to answer my very first *shy* e-mail asking for a place in your research group back in 2011. Thank you for your patience in this process and to help me in every step into the *phage-world*. Thanks to support the diversity in research and give me the chance to find my own path. I am deeply thankful to make me feel home in a tiny island so far away of my hometown… and specially thanks to believe on myself when I was not able to. I want to express my gratitude to Gunnar Gerdts, to support me, especially during the first stage of my PhD thesis. Thank you for all your clever ideas to improve my research, and to push me to talk German, *obwohl mein deutsch nich so gut war… oder ist es endlich besser?* Also, I would like to thank Hilke Döpke, and the whole Microbial Ecology working group for your scientific and technical support, as well as Renate and Kathrin to help me with every tiny doubt I had while living on the island.

I also would like to thank the unconditional support of Cristina Dorador and Martha Hengst. Thanks for keeping the contact during these years and push me to write my first contact e-mail for a PhD position. Thank you to believe on me and support my ideas since my undergraduate studies.

All of this would not be possible without the little help of my friends, many now colleagues. Thanks to Álvaro, Eduardo, Diego, Chalito (and Mic Robio!), Pauli, Chiry, *gracias por estar allí.* Also, thanks to my friends from the German course at the Goethe Institute in Dresden, especially Vale, Sami and Karlitos.

I would like to thank my friends on Helgoland, specially to my pizza pizza girls! *Dankeschön* Sidi, Claudita, Inguita, Birtecita, Steffen, and Ced to be my friends, support, colleagues and family. Thanks for your scientific discussions, to (over)think the scientist life together, the coffee and the beers, as well as your hugs, to cry and laugh together…without all you this journey would be completely different. Thanks to Chris, Meri, Svenja, Rebi, Judith, and Lisa on Helgoland, and Thomas in Bremen (*Danke für die Biere und die Kaffee! Du hast mein 'neue' Zuhause in Bremen ein besserer Ort*). *Danke Uli und meine neue Familie hier in Deutschland.* I would like to thank Alejandra Mera for all the grammar feed-back during this process, but especially for our friendship.

Muchas Gracias Ivo. I still do not learn a language with the correct words to write you my gratitude and admiration. Thank you to be together and to do not let me down, without you I would not be able to stand up with all this strength! So now: *Hey Ho, Let's Go!*

Gracias a mis amigues en Chile, gracias a infinitas a Karlita, Giannina, Rafita, Marjorie, Afro, Marta, Ivania, la gente del teatro, la pintura, el cómic y la vida. Ustedes son parte importante de mi historia y de quién soy hoy.

Esta aventura ha sido gracias a la paciencia y el apoyo incondicional de mi mamá Corina Fernández Peña, mi hermana Milena Varela Fernández, mi abuelo Gabriel Fernández Canque, también a Juan Verdejo Orsola, Vicente, yayita y a tooodo el gran clan Fernández que me hacen sentir su amor y recordarme que mis raíces son la base para florecer. Gracias hermana por inspirarme, y a mis ancestres por darme la confianza. Gracias por creer en mí, *Yuspagarpa*. Espérenme con guatia que hay mucho para celebrar, ¡les amo mucho!

¡Jallalla!

I am thankful for all the awesome people I met in this PhD adventure. Thanks for everyone who believe in myself, my drawings, *my viruses*, my ideas and dreams.

REFERENCES

Abedon, S. T. **2008**. Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses, *Cambridge University Press.*

Ackermann, H. W. **2003**. Bacteriophage observations and evolution. *Research in Microbiology,* 154**,** 245-251.

Ackermann, H. W. **2007**. 5500 Phages examined in the electron microscope. *Archives of Virology,* 152**,** 227-243.

Ackermann, H. W. **2011**. Bacteriophage taxonomy. *Microbiology Australia,* 32**,** 90-94.

Ackermann, H. W. & DuBow, M. S. **1987**. Viruses of Prokaryotes: General Properties of Bacteriophages, *CRC Press Inc*.

Ackermann, H. W., Kasatiya, S. S., Kawata, T., Koga, T., Lee, J. V., Mbiguino, A., Newman, F. S., Vieu, J. F. & Zachary, A. **1984**. Classification of *Vibrio* bacteriophages. *Intervirology,* 22**,** 61-71.

Adams, M. **1959.** Bacteriophages. *New York: Interscience Publishers. Inc*.

Adams, M. J., Lefkowitz, E. J., King, A. M. Q., Harrach, B., Harrison, R. L., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Mushegian, A. R., Nibert, M., Sabanadzovic, S., Sanfaçon, H., Siddell, S. G., Simmonds, P., Varsani, A., Zerbini, F. M., Gorbalenya, A. E. & Davison, A. J. **2017**. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). *Archives of Virology,* 162**,** 2505-2538.

Adriaenssens, E. M., Krupovic, M., Knezevic, P., Ackermann, H.-W., Barylski, J., Brister, J. R., Clokie, M. R. C., Duffy, S., Dutilh, B. E., Edwards, R. A., Enault, F., Jang, H. B., Klumpp, J., Kropinski, A. M., Lavigne, R., Poranen, M. M., Prangishvili, D., Rumnieks, J., Sullivan, M. B., Wittmann, J., Oksanen, H. M., Gillis, A. & Kuhn, J. H. **2017**. Taxonomy of prokaryotic viruses: 2016 update from the ICTV bacterial and archaeal viruses subcommittee. *Archives of Virology,* 162**,** 1153-1157.

Adriaenssens, E. M., Wittmann, J., Kuhn, J. H., Turner, D., Sullivan, M. B., Dutilh, B. E., Jang, H. B., van Zyl, L. J., Klumpp, J., Lobocka, M., Moreno Switt, A. I., Rumnieks, J., Edwards, R. A., Uchiyama, J., Alfenas-Zerbini, P., Petty, N. K., Kropinski, A. M., Barylski, J., Gillis, A., Clokie, M. R. C., Prangishvili, D., Lavigne, R., Aziz, R. K., Duffy, S., Krupovic, M., Poranen, M. M., Knezevic, P., Enault, F., Tong, Y., Oksanen, H. M. & Rodney Brister, J. **2018**. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Archives of Virology,* 163**,** 1125-1129.

Akhter, S., Aziz, R. K. & Edwards, R. A. **2012**. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Research,* 40**,** e126-e126.

Alanis Villa, A., Kropinski, A. M., Abbasifar, R., Abbasifar, A. & Griffiths, M. W. **2012**. Genome sequence of temperate *Vibrio parahaemolyticus* bacteriophage vB_VpaS_MAR10. *Journal of Virology,* 86**,** 13851-13852.

Alavandi, S. V. & Poornima, M. **2012**. Viral Metagenomics: A tool for virus discovery and diversity in Aquaculture. *Indian journal of virology : an official organ of Indian Virological Society,* 23**,** 88-98.

Allers, E., Moraru, C., Duhaime, M. B., Beneze, E., Solonenko, N., Barrero-Canosa, J., Amann, R. & Sullivan, M. B. **2013**. Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology,* 15**,** 2306-2318.

Allison, H. E., Brister, J. R., Kuhn, J. H., Kropinski, A. M., Adriaenssens, E. M., Plunkett, G. & Węgrzyn, G. **2016**. To create one (1) new subfamily, *Sepvirinae* containing three (3) new genera, in the family Podoviridae. *ICTV approved proposal* 2016.045a-oB.

Alonso, M. C., Rodríguez, J. & Borrego, J. J. **1999**. Enumeration and isolation of viral particles from oligotrophic marine environments by tangential flow filtration. *International Microbiology,* 2**,** 227-232.

Altschul, S. F., Madden, T. L., Schaffer, 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**,** 3389-3402.

Anand, T., Bera, B. C., Vaid, R. K., Barua, S., Riyesh, T., Virmani, N., Hussain, M., Singh, R. K. & Tripathi, B. N. **2016**. Abundance of antibiotic resistance genes in environmental bacteriophages. *Journal of General Virology,* 97**,** 3458-3466.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A. & Rohwer, F. **2006**. The marine viromes of four oceanic regions. *PLoS Biology,* 4**,** e368.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D. S. **2016**. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research,* 44**,** W16-W21.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. **2008**. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics,* 9**,** 75.

Bailey, B. A., McNair, K. & Edwards, R. A. **2012**. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics,* 28**,** 614-618.

Baker-Austin, C., Trinanes, J., Gonzalez-Escalona, N. & Martinez-Urtaza, J. **2017**. Non-Cholera Vibrios: The microbial barometer of climate change. *Trends in Microbiology,* 25**,** 76-84.

Baker-Austin, C., Trinanes, J., Salmenlinna, S., Löfdahl, M., Siitonen, A., Taylor, N. G. H. & Martinez-Urtaza, J. **2016**. Heat wave–associated Vibriosis, Sweden and Finland, 2014. *Emerging Infectious Diseases,* 22**,** 1216-1220.

Baker-Austin, C., Trinanes, J. A., Taylor, N. G. H., Hartnell, R., Siitonen, A. & Martinez-Urtaza, J. **2013**. Emerging *Vibrio* risk at high latitudes in response to ocean warming. *Nature Climate Change,* 3**,** 73-77.

Banerjee, R., Das, B., Balakrish Nair, G. & Basak, S. **2014**. Dynamics in genome evolution of *Vibrio cholerae*. *Infection, Genetics and Evolution,* 23**,** 32-41.

Barrero-Canosa, J. & Moraru, C. **2019**. PhageFISH for monitoring phage infections at single cell level. *In:* Clokie, M. R. J., Kropinski, A. & Lavigne, R. (eds.) Bacteriophages: methods and protocols, Volume IV. *New York, NY: Springer New York*.

Barylski, J., Enault, F., Dutilh, B. E., Schuller, M. B. P., Edwards, R. A., Gillis, A., Klumpp, J., Knezevic, P., Krupovic, M., Kuhn, J. H., Lavigne, R., Oksanen, H. M., Sullivan, M. B., Wittmann, J., Tolstoy, I., Rodney Brister, J., Kropinski, A. M. & Adriaenssens, E. M. **2018**. Analysis of spounaviruses as a case study for the overdue reclassification of tailed bacteriophages. *bioRxiv***,** 220434.

Bastías, R., Higuera, G., Sierralta, W. & Espejo, R. T. **2010**. A new group of cosmopolitan bacteriophages induce a carrier state in the pandemic strain of *Vibrio parahaemolyticus*. *Environmental Microbiology,* 12**,** 990-1000.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S. & Sonnhammer, E. L. **2004**. The Pfam protein families database. *Nucleic Acids Research,* 32**,** D138-D141.

Baudoux, A. C., Hendrix, R. W., Lander, G. C., Bailly, X., Podell, S., Paillard, C., Johnson, J. E., Potter, C. S., Carragher, B. & Azam, F. **2012**. Genomic and functional analysis of *Vibrio* phage SIO-2 reveals novel insights into ecology and evolution of marine siphoviruses. *Environmental Microbiology,* 14**,** 2071-86.

Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K. & Wommack, K. E. **2007**. Metagenomic characterization of Chesapeake Bay virioplankton. *Applied and Environmental Microbiology,* 73**,** 7629-7641.

Berard, S., Chateau, A., Pompidor, N., Guertin, P., Bergeron, A. & Swenson, K. M. **2016**. Aligning the unalignable: bacteriophage whole genome alignments. *BMC Bioinformatics,* 17**,** 30.

Bergh, O., Børsheim, K. Y., Bratbak, G. & Heldal, M. **1989**. High abundance of viruses found in aquatic environments. *Nature,* 340**,** 467-468.

Beumer, A. & Robinson, J. B. **2005**. A Broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Applied and Environmental Microbiology,* 71**,** 8301.

Bier, N., Jäckel, C., Dieckmann, R., Brennholt, N., Böer, I. S. & Strauch, E. **2015**. Virulence profiles of *Vibrio vulnificus* in german coastal waters, a comparison of North Sea and Baltic Sea isolates. *International Journal of Environmental Research and Public Health,* 12, 15943-15959.

Bietz, M. J. & Lee, C. P. Collaboration in metagenomics: sequence databases and the organization of scientific work. *In:* Wagner, I., Tellioğlu, H., Balka, E., Simone, C. & Ciolfi, L., eds. ECSCW 2009, 2009// 2009 *London. Springer London,* 243-262.

Bils, F., Moyano, M., Aberle, N., van Damme, C. J. G., Nash, R. D. M., Kloppmann, M., Loots, C. & Peck, M. A. **2019**. Broad-scale distribution of the winter protozooplankton community in the North Sea. *Journal of Sea Research,* 144**,** 112-121.

Böer, S. I., Heinemeyer, E. A., Luden, K., Erler, R., Gerdts, G., Janssen, F. & Brennholt, N. **2013**. Temporal and spatial distribution patterns of potentially pathogenic *Vibrio* spp. at recreational beaches of the german North Sea. *Microbial Ecology,* 65**,** 1052-1067.

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

Bondy-Denomy, J. & Davidson, A. R. **2014**. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *Journal of Microbiology,* 52**,** 235-242.

Brandsma, J., Martínez, J. M., Slagter, H. A., Evans, C. & Brussaard, C. P. D. **2013**. Microbial biogeography of the North Sea during summer. *Biogeochemistry,* 113**,** 119-136.

Breitbart, M. **2012**. Marine viruses: truth or dare. *Annual Review of Marine Science,* 4**,** 425-448.

Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. **2018**. Phage puppet masters of the marine microbial realm. *Nature Microbiology,* 3**,** 754-766.

Breitbart, M. & Rohwer, F. **2005**. Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology,* 13**,** 278-84.

Breitbart, M., Rohwer, F. & Abedon, S. T. **2005**. Phage ecology and bacterial pathogenesis. *In:* Waldor, M. K. and Friedman, D. I., Adhya, S. L.(eds.), Phages. *American Society of Microbiology,* 66-91.

Breitwieser, F. P., Lu, J. & Salzberg, S. L. **2017**. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics,* bbx120.

Brister, J. R., Ako-adjei, D., Bao, Y. & Blinkova, O. **2015**. NCBI Viral Genomes Resource. *Nucleic Acids Research,* 43**,** D571-D577.

Brown, M. V., Lauro, F. M., DeMaere, M. Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M. J., Fuhrman, J. A., Andrews-Pfannkoch, C., Hoffman, J. M., McQuaid, J. B., Allen, A., Rintoul, S. R. & Cavicchioli, R. **2012**. Global biogeography of SAR11 marine bacteria. *Molecular Systems Biology,* 8**,** 595.

Brum, J., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., De Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., Schwenck, S., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Coordinators, T. O., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E. & Sullivan, M. B. **2015**. Patterns and ecological drivers of ocean viral communities. *Science,* 348**,** 1261498.

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

Brum, J. R. & Sullivan, M. B. **2015**. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nature Reviews Microbiology,* 13**,** 147.

Brussow, H., Canchaya, C. & Hardt, W. D. **2004**. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews,* 68**,** 560-602.

Brüssow, H. & Hendrix, R. W. **2002**. Phage genomics: small is beautiful. *Cell,* 108**,** 13-16.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. **2003**. Prophage genomics. *Microbiology and Molecular Biology Reviews,* 67**,** 238-276.

Casas, V. & Maloy, S. **2011**. Role of bacteriophage-encoded exotoxins in the evolution of bacterial pathogens. *Future Microbiology,* 6**,** 1461-1473.

Casas, V., Miyake, J., Balsley, H., Roark, J., Telles, S., Leeds, S., Zurita, I., Breitbart, M., Bartlett, D., Azam, F. & Rohwer, F. **2006**. Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. *FEMS Microbiology Letters,* 261**,** 141-149.

Casjens, S. **2003**. Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology,* 49**,** 277-300.

Castillo, D., Kauffman, K., Hussain, F., Kalatzis, P., Rørbo, N., Polz, M. F. & Middelboe, M. **2018**. Widespread distribution of prophage-encoded virulence factors in marine *Vibrio* communities. *Scientific Reports,* 8**,** 9973.

Cerveny, K. E., DePaola, A., Duckworth, D. H. & Gulig, P. A. **2002**. Phage therapy of local and systemic disease caused by *Vibrio vulnificus* in iron-dextran-treated mice. *Infection and Immunity,* 70**,** 6251.

Chakraborty, S., Mukhopadhyay, A. K., Bhadra, R. K., Ghosh, A. N., Mitra, R., Shimada, T., Yamasaki, S., Faruque, S. M., Takeda, Y., Colwell, R. R. & Nair, G. B. **2000**. Virulence genes in environmental strains of *Vibrio cholerae*. *Applied and Environmental Microbiology,* 66**,** 4022-4028.

Chang, B., Miyamoto, H., Taniguchi, H. & Yoshida, S.-i. **2002**. Isolation and genetic characterization of a novel filamentous bacteriophage, a deleted form of phage f237, from a pandemic *Vibrio parahaemolyticus* O4:K68 strain. *Microbiology and Immunology*, 46**,** 565-569.

Chang, B., Taniguchi, H., Miyamoto, H. & Yoshida, S.-i. **1998**. Filamentous bacteriophages of *Vibrio parahaemolyticus* as a possible clue to genetic transmission. *Journal of Bacteriology*, 180**,** 5094-5101.

Chen, H. **2013.** Generate high-resolution Venn and Euler plots. *In:* http://cran.r-project.org/web/packages/VennDiagram/index.html

Chen, J., Ram, G., Penadés, J. R., Brown, S. & Novick, R. P. **2015**. Pathogenicity island-directed transfer of unlinked chromosomal virulence genes. *Molecular cell*, 57**,** 138-149.

Chénard, C., Chan, A. M., Vincent, W. F. & Suttle, C. A. **2015**. Polar freshwater cyanophage S-EIV1 represents a new widespread evolutionary lineage of phages. *The ISME journal*, 9**,** 2046-2058.

Cheng, W.-C., Jan, I. S., Chen, J.-M., Teng, S.-H., Teng, L.-J., Sheng, W.-H., Ko, W.-C. & Hsueh, P.-R. **2015**. Evaluation of the bruker biotyper Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry System for identification of blood isolates of *Vibrio* species. . *Journal of Clinical Microbiology*, 53**,** 1741-1744.

Chibani, C. M., Meinecke, F., Farr, A., Dietrich, S. & Liesegang, H. **2019**. ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks. *bioRxiv***,** 558171.

Chow, C.-E. T. & Suttle, C. A. **2015**. Biogeography of Viruses in the Sea. *Annual Review of Virology*, 2**,** 41-66.

Clokie, M. R., Kropinski, A. M. & Lavigne, R. **2018**. Bacteriophages, Methods and Protocols Volume III, *Springer Protocols, Humana Press.*

Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. **2011**. Phages in nature. *Bacteriophage*, 1**,** 31-45.

Clokie, M. R. J. & Kropinski, A. M. **2009**. Bacteriophages: methods and protocols, *New York, Humana Press.*

Colavecchio, A., Cadieux, B., Lo, A. & Goodridge, L. D. **2017**. Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the Enterobacteriaceae family - a review. *Frontiers in Microbiology*, 8**,** 1108.

Comeau, A. M., Tremblay, D., Moineau, S., Rattei, T., Kushkina, A. I., Tovkach, F. I., Krisch, H. M. & Ackermann, H.-W. **2012**. Phage morphology recapitulates phylogeny: the comparative genomics of a new group of myoviruses. *PLoS ONE*, 7**,** e40102.

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.

Covacci, A., Falkow, S., Berg, D. E. & Rappuoli, R. **1997**. Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori*? *Trends in Microbiology*, 5**,** 205-208.

Croci, L., Suffredini, E., Cozzi, L., Toti, L., Ottaviani, D., Pruzzo, C., Serratore, P., Fischetti, R., Goffredo, E., Loffredo, G. & Mioni, R. **2007**. Comparison of different biochemical and molecular methods for the identification of *Vibrio parahaemolyticus. Journal of Applied Microbiology*, 102**,** 229-237.

Cunningham, B. R., Brum, J. R., Schwenck, S. M., Sullivan, M. B., John, S. G. & Wommack, K. E. **2015**. An inexpensive, accurate, and precise wet-mount method for enumerating aquatic viruses. *Applied and Environmental Microbiology*, 81**,** 2995-3000.

Das, B. **2014**. Mechanistic insights into filamentous phage integration in *Vibrio cholerae. Frontiers in Microbiology*, 5**,** 650.

Davis, B. M., Moyer, K. E., Boyd, E. F. & Waldor, M. K. **2000**. CTX Prophages in Classical Biotype *Vibrio cholerae*: functional phage genes but dysfunctional phage genomes. *Journal of Bacteriology*, 182**,** 6992-6998.

Davis, B. M. & Waldor, M. K. **2003**. Filamentous phages linked to virulence of *Vibrio cholerae*. *Current Opinion in Microbiology*, 6**,** 35-42.

Day, L. A. **2011**. Family Inoviruses. *In:* King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (eds.) Virus Taxonomy: Classification and Nomenclature of Viruses. Ninth Report of the International Committee on Taxonomy of Viruses. *San Diego: Elseiver Avademic Press.*

Demerec, M. & Fano, U. **1945**. Bacteriophage-resistant mutants in *Escherichia coli*. *Genetics*, 30**,** 119-136.

Dennehy, J. J. **2014**. What ecologists can tell virologists. *Annual Review of Microbiology*, 68**,** 117-35.

Desselberger, U. **2018**. Virus taxonomy—a taxing task. *Archives of Virology*, 163**,** 2019-2020.

Di Pinto, A., Terio, V., Novello, L. & Tantillo, G. **2011**. Comparison between thiosulphate-citrate-bile salt sucrose (TCBS) agar and CHROMagar *Vibrio* for isolating *Vibrio parahaemolyticus*. *Food Control*, 22**,** 124-127.

Dieckmann, R., Strauch, E. & Alter, T. **2010**. Rapid identification and characterization of *Vibrio* species using whole-cell MALDI-TOF mass spectrometry. *Journal of Applied Microbiology*, 109**,** 199-211.

Diepenbroek, M., Glöckner, F. O., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., Kostadinov, I., Nieschulze, J., Seeger, B., Tolksdorf, R. & Triebel, D. **2014**. Towards an integrated biodiversity and ecological research data management and archiving platform: the German federation for the curation of biological data (GFBio). E. Plödereder, L. Grunske, E. Schneider, D. Ull (Eds.), Informatik 2014 – Big Data Komplexität meistern (GI-Edition), Lecture Notes in Informatics (LNI) – Köllen Verlag, Bonn. *Proceedings* 232, 1711-1724.

Dou, C., Xiong, J., Gu, Y., Yin, K., Wang, J., Hu, Y., Zhou, D., Fu, X., Qi, S., Zhu, X., Yao, S., Xu, H., Nie, C., Liang, Z., Yang, S., Wei, Y. & Cheng, W. **2018**. Structural and functional insights into the regulation of the lysis–lysogeny decision in viral communities. *Nature Microbiology*, 3**,** 1285-1294.

Dudhagara, P., Bhavsar, S., Bhagat, C., Ghelani, A., Bhatt, S. & Patel, R. **2015**. Web resources for metagenomics studies. *Genomics, Proteomics & Bioinformatics*, 13**,** 296-303.

Duhaime, M. B. **2010**. Exploring the Marine Virosphere from Genome Context to Content. Doctor of Philosophy thesis, *Jacobs University.*

Duhaime, M. B., Solonenko, N., Roux, S., Verberkmoes, N. C., Wichels, A. & Sullivan, M. B. **2017**. comparative omics and trait analyses of marine *Pseudoalteromonas* phages advance the phage OTU concept. *Frontiers in Microbiology*, 8**,** 1241.

Duhaime, M. B. & Sullivan, M. B. **2012**. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, 434**,** 181-186.

Duhaime, M. B., Wichels, A. & Sullivan, M. B. **2016**. Six *Pseudoalteromonas* strains isolated from surface waters of Kabeltonne, offshore Helgoland, North Sea. *Genome Announcements*, 4.

Duhaime, M. B., Wichels, A., Waldmann, J., Teeling, H. & Glöckner, F. O. **2011a**. Ecogenomics and genome landscapes of marine Pseudoalteromonas phage H105/1. *The ISME journal*, 5**,** 107-21.

Duhaime, M. B., Wichels, A., Waldmann, J., Teeling, H. & Glöckner, F. O. **2011b**. Ecogenomics and genome landscapes of marine Pseudoalteromonas phage H105/1. *The ISME journal*, 5**,** 107-121.

Dutilh, B. E. **2014**. Metagenomic ventures into outer sequence space. *Bacteriophage*, 4**,** e979664.

Eddy, S. R. **1998**. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14**,** 755-763.

Edgar, R. C. **2010**. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26**,** 2460-2461.

Edwards, C. **2018**. Learning to See, Machine learning turns the spotlight on elusive viruses. *Communications of the ACM,* 61**,** 13-15.

Edwards, R. A. & Rohwer, F. **2005**. Viral metagenomics. *Nat Rev Micro,* 3**,** 504-510.

Eisma, D., Johnston, R., Cadogan John Ivan, G., Cadogan John Ivan, G., Clark, R. B. & Hartley, J. P. **1987**. The North Sea: an overview. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences,* 316**,** 461-485.

Emeis, K.-C., van Beusekom, J., Callies, U., Ebinghaus, R., Kannen, A., Kraus, G., Kröncke, I., Lenhart, H., Lorkowski, I., Matthias, V., Möllmann, C., Pätsch, J., Scharfe, M., Thomas, H., Weisse, R. & Zorita, E. **2015**. The North Sea — A shelf sea in the Anthropocene. *Journal of Marine Systems,* 141**,** 18-33.

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B. & Petit, M. A. **2017**. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME journal,* 11**,** 237-247.

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G. & Sorek, R. **2017**. Communication between viruses guides lysis–lysogeny decisions. *Nature,* 541**,** 488.

Erler, R. **2015**. Identification and Characterization of *Vibrio* species. Doctor of Philosophy thesis, *Jacobs University*.

Erler, R., Wichels, A., Heinemeyer, E.-A., Hauk, G., Hippelein, M., Reyes, N. T. & Gerdts, G. **2015**. VibrioBase: A MALDI-TOF MS database for fast identification of *Vibrio* spp. that are potentially pathogenic in humans. *Systematic and Applied Microbiology,* 38**,** 16-25.

Farrah, S. R. **1987**. Ecology of phage in freshwater environments. *In:* Goyal, S. M., Gerba, C. P. & Bitton, G. (eds.) Phage Ecology (Goyal, SM, Gerba, CP and Bitton, G., Eds.). *John Wiley and Sons.*

Faruque, S. M. & Mekalanos, J. J. **2003**. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends in Microbiology,* 11**,** 505-510.

Faruque, S. M. & Mekalanos, J. J. **2012**. Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. *Virulence,* 3**,** 556-65.

Fauquet, C. M. & Martelli, G. P. **2013**. Viral Classification and Nomenclature. *eLS.*

Fernández, L., Rodríguez, A. & García, P. **2018**. Phage or foe: an insight into the impact of viral predation on microbial communities. *The ISME journal,* 12**,** 1171-1179.

Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. **2010**. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences***,** 201007615.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. & al, e. **1995**. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269**,** 496.

Fortier, L. C. & Sekulovic, O. **2013**. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence,* 4**,** 354-365.

Fouts, D. E. **2006**. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research,* 34**,** 5839-5851.

Frank, H. & Moebus, K. **1987**. An electron microscopic study of bacteriophages from marine waters. *Helgoländer Meeresuntersuchungen,* 41**,** 385-414.

Fuhrman, J. A. **1999**. Marine viruses and their biogeochemical and ecological effects. *Nature,* 399**,** 541-548.

Fuhrman, J. A., Liang, X. & Noble, R. T. **2005**. Rapid detection of enteroviruses in small volumes of natural waters by real-time quantitative reverse transcriptase PCR. *Applied and Environmental Microbiology,* 71**,** 4523-4530.

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L. & Brown, J. H. **2008**. A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences,* 105**,** 7774-7778.

Galiez, C., Coste, F., Magnan, C. N. & Baldi, P. **2016**. VIRALpro: a tool to identify viral capsid and tail sequences. *Bioinformatics,* 32**,** 1405-1407.

Galiez, C., Siebert, M., Enault, F., Vincent, J. & Soding, J. **2017**. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics,* 33**,** 3113-3114.

Gamage, S. D., Strasser, J. E., Chalk, C. L. & Weiss, A. A. **2003**. Nonpathogenic *Escherichia coli* can contribute to the production of Shiga toxin. *Infection and Immunity,* 71**,** 3107-3115.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F. & Conesa, A. **2012**. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics,* 28**,** 2678-2679.

García-López, R., Vázquez-Castellanos, J. F. & Moya, A. **2015**. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Frontiers in Bioengineering and Biotechnology,* 3**,** 141.

Garin-Fernandez, A., Glöckner, F. O. & Wichels, A. **submitted**. Genomic characterization of filamentous phage vB_VpaI_VP-3218, an inducible prophage of *Vibrio parahaemolyticus*.

Garin-Fernandez, A., Pereira-Flores, E., Glöckner, F. O. & Wichels, A. **2018**. The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages. *Marine Genomics,* 41**,** 31-41.

Garin-Fernandez, A. & Wichels, A. **submitted**. Looking for the hidden: Characterization of lysogenic phages in potential pathogenic *Vibrio* species from the North Sea.

Garneau, J. R., Depardieu, F., Fortier, L.-C., Bikard, D. & Monot, M. **2017**. PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Scientific Reports,* 7**,** 8292.

Garretto, A., Hatzopoulos, T. & Putonti, C. **2019**. VirMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ,* 7**,** e6695.

Gerdts, G. & Rohardt, G. **2016.** Physical oceanography during HEINCKE cruise HE430. PANGAEA.

Gerdts, G., Wichels, A., Döpke, H., Klings, K. W., Gunkel, W. & Schütt, C. **2004**. 40-year long-term study of microbial parameters near Helgoland (German Bight, North Sea): historical view and future perspectives. *Helgoland Marine Research,* 58**,** 230-242.

Giebel, H.-A., Kalhoefer, D., Lemke, A., Thole, S., Gahl-Janssen, R., Simon, M. & Brinkhoff, T. **2010**. Distribution of *Roseobacter* RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. *The ISME journal,* 5**,** 8-19.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. **1990**. Genetic diversity in Sargasso Sea bacterioplankton. *Nature,* 345**,** 60.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W. H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S.-J., Kvilekval, K., Manjunath, B. S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S. M., Cranston, K. A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D. J., White, J. W., Leebens-Mack, J., Donoghue, M. J., Spalding, E. P., Vision, T. J., Myers, C. R., Lowenthal,

D., Enquist, B. J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L. & Stanzione, D. **2011**. The iPlant Collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science,* 2**,** 34-34.

Grose, J. H. & Casjens, S. R. **2014**. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. *Virology,* 468-470**,** 421-443.

Groth, A. C. & Calos, M. P. **2004**. Phage integrases: biology and applications. *Journal of Molecular Biology,* 335**,** 667-678.

Guidolin, A., Morelli, G., Kamke, M. & Manning, P. A. **1984**. *Vibrio cholerae* bacteriophage CP-T1: characterization of bacteriophage DNA and restriction analysis. *Journal of Virology,* 51**,** 163-169.

Guillou, N., Chapalain, G. & Thais, L. **2009**. Three-dimensional modeling of tide-induced suspended transport of seabed multicomponent sediments in the eastern English Channel. *Journal of Geophysical Research: Oceans (1978–2012),* 114.

Handling editor: Ruth, T., McQuatters-Gollop, A., Bedford, J. & Johns, D. **2018**. A century of change in North Sea plankton communities explored through integrating historical datasets. *ICES Journal of Marine Science,* 76**,** 104-112.

Harrington, C., Del Casale, A., Kennedy, J., Neve, H., Picton, B. E., Mooij, M. J., O'Gara, F., Kulakov, L. A., Larkin, M. J. & Dobson, A. D. W. **2012**. Evidence of bacteriophage-mediated horizontal transfer of bacterial 16S rRNA genes in the viral metagenome of the marine sponge Hymeniacidon perlevis. *Microbiology,* 158**,** 2789-2795.

Hartnell, R. E., Stockley, L., Keay, W., Rosec, J. P., Hervio-Heath, D., Van den Berg, H., Leoni, F., Ottaviani, D., Henigman, U., Denayer, S., Serbruyns, B., Georgsson, F., Krumova-Valcheva, G., Gyurova, E., Blanco, C., Copin, S., Strauch, E., Wieczorek, K., Lopatek, M., Britova, A., Hardouin, G., Lombard, B., in't Veld, P., Leclercq, A. & Baker-Austin, C. **2019**. A pan-European ring trial to validate an International Standard for detection of *Vibrio cholerae*, *Vibrio parahaemolyticus* and *Vibrio vulnificus* in seafoods. *International Journal of Food Microbiology,* 288**,** 58-65.

Hassan, F., Kamruzzaman, M., Mekalanos, J. J. & Faruque, S. M. **2010**. Satellite phage TLCϕ enables toxigenic conversion by CTX phage through dif site alteration. *Nature,* 467**,** 982-985.

Hazen, T. H., Martinez, R. J., Chen, Y., Lafon, P. C., Garrett, N. M., Parsons, M. B., Bopp, C. A., Sullards, M. C. & Sobecky, P. A. **2009**. Rapid identification of *Vibrio parahaemolyticus* by Whole-Cell Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. *Applied and Environmental Microbiology,* 75**,** 6745-6756.

Hazen, T. H., Pan, L., Gu, J. D. & Sobecky, P. A. **2010**. The contribution of mobile genetic elements to the evolution and ecology of Vibrios. *FEMS Microbiology Ecology,* 74**,** 485-99.

Heidelberg, J. F., Paulsen, I. T., Nelson, K. E., Gaidos, E. J., Nelson, W. C., Read, T. D., Eisen, J. A., Seshadri, R., Ward, N., Methe, B., Clayton, R. A., Meyer, T., Tsapin, A., Scott, J., Beanan, M., Brinkac, L., Daugherty, S., DeBoy, R. T., Dodson, R. J., Durkin, A. S., Haft, D. H., Kolonay, J. F., Madupu, R., Peterson, J. D., Umayam, L. A., White, O., Wolf, A. M., Vamathevan, J., Weidman, J., Impraim, M., Lee, K., Berry, K., Lee, C., Mueller, J., Khouri, H., Gill, J., Utterback, T. R., McDonald, L. A., Feldblyum, T. V., Smith, H. O., Venter, J. C., Nealson, K. H. & Fraser, C. M. **2002**. Genome sequence of the dissimilatory metal ion–reducing bacterium *Shewanella oneidensis*. *Nature Biotechnology,* 20**,** 1118-1123.

Hendrix, R. W. **2003**. Bacteriophage genomics. *Current Opinion in Microbiology,* 6**,** 506-511.

Hengstl, T., Weßler, S., Thelesklaf, P., Oberdanner, C. & Verna, S. **2012.** Automated anerobic cell growth curve monitoring. *Tecan Trading AG*.

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D. & Allen, T. M. **2012**. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathogens,* 8**,** e1002529.

Henn, M. R., Sullivan, M. B., Stange-Thomann, N., Osburne, M. S., Berlin, A. M., Kelly, L., Yandava, C., Kodira, C., Zeng, Q., Weiand, M., Sparrow, T., Saif, S., Giannoukos, G., Young, S. K., Nusbaum, C., Birren, B. W. & Chisholm, S. W. **2010**. Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE,* 5.

Hobbie, J. E., Daley, R. J. & Jasper, S. **1977**. Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Applied and Environmental Microbiology,* 33**,** 1225-1228.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., Verberkmoes, N. C. & Sullivan, M. B. **2013**. Twelve previously unknown phage genera are ubiquitous in global oceans. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 12798-12803.

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. **2017**. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME journal,* 11**,** 1511.

Huang, S., Zhang, S., Jiao, N., Chen, F. & Wommack, K. E. **2015**. Marine cyanophages demonstrate biogeographic patterns throughout the global ocean. *Applied and Environmental Microbiology,* 81**,** 441-452.

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. & Le Mercier, P. **2010**. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research,* 39**,** D576-D582.

Hulo, C., Masson, P., Le Mercier, P. & Toussaint, A. **2015**. A structured annotation frame for the transposable phages: A new proposed family "Saltoviridae" within the Caudovirales. *Virology,* 477**,** 155-163.

Hurwitz, B. iMicrobe: advancing clinical and environmental microbial research using the iPlant cyberinfrastructure. *Proceedings of the International Plant and Animal Genome XXII Conference, San Diego, CA, 2014.*

Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. **2013a**. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology,* 15**,** 1428-1440.

hurwitz, b. l., hallam, s. & sullivan, m. b. **2013b**. metabolic reprogramming by viruses in the sunlit and dark ocean. *genome biology,* 14.

Hurwitz, B. L., Ponsero, A., Thornton, J. & U'Ren, J. M. **2018**. Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets. *Virus Research,* 244**,** 110-115.

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**,** e57355.

Hurwitz, B. L., U'Ren, J. M. & Youens-Clark, K. **2016**. Computational prospecting the great viral unknown. *FEMS Microbiology Letters,* 363.

Hurwitz, B. L. & U'Ren, J. M. **2016**. Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology,* 31**,** 161-168.

Hurwitz, B. L., Westveld, A. H., Brum, J. & Sullivan, M. B. **2014**. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 10714-10719.

Hwang, J., Park, S. Y., Park, M., Lee, S., Jo, Y., Cho, W. K. & Lee, T.-K. **2016**. Metagenomic characterization of viral communities in Goseong Bay, Korea. *Ocean Science Journal,* 51**,** 599-612.

Ibrahim, B., McMahon, D. P., Hufsky, F., Beer, M., Deng, L., Mercier, P. L., Palmarini, M., Thiel, V. & Marz, M. **2018**. A new era of virus bioinformatics. *Virus Research,* 251**,** 86-90.

Ignacio-Espinosa, J. C. & Fuhrman, J. A. **2018**. A non-tailed twist in the viral tale. *Nature,* 554**,** 38-39.

Iida, T., Hattori, A., Tagomori, K., Nasu, H., Naim, R. & Honda, T. **2001**. Filamentous phage associated with recent pandemic strains of *Vibrio parahaemolyticus*. *Emerging Infectious Diseases,* 7**,** 477-478.

Ikema, M. & Honma, Y. **1998**. A novel filamentous phage, FS-2, of *Vibrio cholerae* 0139. *Microbiology,* 144**,** 1901-1906.

Jannasch, H. W. & Jones, G. E. **1959**. Bacterial populations in sea water as determined by different methods of enumeration. *Limnology and Oceanography,* 4**,** 128-139.

Jensen, M. A., Faruque, S. M., Mekalanos, J. J. & Levin, B. R. **2006**. Modeling the role of bacteriophage in the control of cholera outbreaks. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 4652-4657.

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F. & Zhang, X. **2012**. Comparison of metagenomic samples using sequence signatures. *BMC Genomics,* 13**,** 730-730.

Jiang, S. C. & Paul, J. H. **1996**. Occurrence of lysogenic bacteria in marine microbial communities as determined by prophage induction. *Marine Ecology Progress Series,* 142**,** 27-38.

Jiang, S. C. & Paul, J. H. **1998**. Significance of Lysogeny in the marine environment: studies with isolates and a model of lysogenic phage production. *Microbial Ecology,* 35**,** 235-243.

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

Johnson, C. N., Flowers, A. R., Young, V. C., Gonzalez-Escalona, N., DePaola, A., Noriea, N. F. & Grimes, D. J. **2008**. Genetic Relatedness Among tdh+ and trh+*Vibrio parahaemolyticus* Cultured from Gulf of Mexico Oysters (Crassostrea virginica) and Surrounding Water and Sediment. *Microbial Ecology,* 57**,** 437.

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Kusel, K., Rillig, M. C., Rivett, D. W., Salles, J. F., van der Heijden, M. G., Youssef, N. H., Zhang, X., Wei, Z. & Hol, W. H. **2017**. Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME journal,* 11**,** 853-862.

Kamp, D. **1987**. The evolution of Mu. *Phage Mu. In* Phage Mu (Symonds, N., Toussaint, A., van de Putte, P. & Howe, M. M., eds), *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY,* 259-269.

Kang, I., Oh, H. M., Kang, D. & Cho, J. C. **2013**. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proceedings of the National Academy of Sciences,* 110**,** 12343-12348.

Kauffman, K. M., Brown, J. M., Sharma, R. S., VanInsberghe, D., Elsherbini, J., Polz, M. & Kelly, L. **2018a**. Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data,* 5**,** 180114.

Kauffman, K. M., Hussain, F. A., Yang, J., Arevalo, P., Brown, J. M., Chang, W. K., VanInsberghe, D., Elsherbini, J., Sharma, R. S., Cutler, M. B., Kelly, L. & Polz, M. F. **2018b**. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature,* 554**,** 118.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. **2012**. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28**,** 1647-1649.

Keen, E. C., Bliskovsky, V. V., Malagon, F., Baker, J. D., Prince, J. S., Klaus, J. S. & Adhya, S. L. **2017**. Novel "Superspreader" bacteriophages promote horizontal gene transfer by transformation. *mBio,* 8**,** e02115-16.

Kim, H. U., Kim, S. Y., Jeong, H., Kim, T. Y., Kim, J. J., Choy, H. E., Yi, K. Y., Rhee, J. H. & Lee, S. Y. **2011**. Integrative genome-scale metabolic analysis of *Vibrio vulnificus* for drug targeting and discovery. *Molecular Systems Biology,* 7**,** 460.

Kim, Y. R., Lee, S. E., Kim, C. M., Kim, S. Y., Shin, E. K., Shin, D. H., Chung, S. S., Choy, H. E., Progulske-Fox, A., Hillman, J. D., Handfield, M. & Rhee, J. H. **2003**. Characterization and pathogenic significance of *Vibrio vulnificus* antigens preferentially expressed in septicemic patients. *Infection and Immunity,* 71**,** 5461-5471.

King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. **2012**. Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses, *Academic Press, London;Waltham.*

Kirstein, I. V., Kirmizi, S., Wichels, A., Garin-Fernandez, A., Erler, R., Loder, M. & Gerdts, G. **2016**. Dangerous hitchhikers? Evidence for potentially pathogenic Vibrio spp. on microplastic particles. *Marine Environmental Research,* 120**,** 1-8.

Klumpp, J., Fouts, D. E. & Sozhamannan, S. **2013**. Bacteriophage functional genomics and its role in bacterial pathogen detection. *Briefings in Functional Genomics,* 12**,** 354-365.

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., Little, M., Luque, A., McDole-Somera, T., McNair, K., de Oliveira, L. S., Quistad, S. D., Robinett, N. L., Sala, E., Salamon, P., Sanchez, S. E., Sandin, S., Silva, G. G. Z., Smith, J., Sullivan, C., Thompson, C., Vermeij, M. J. A., Youle, M., Young, C., Zgliczynski, B., Brainard, R., Edwards, R. A., Nulton, J., Thompson, F. & Rohwer, F. **2016**. Lytic to temperate switching of viral communities. *Nature,* 531**,** 466-470.

Koboldt, Daniel C., Steinberg, Karyn M., Larson, David E., Wilson, Richard K. & Mardis, E. R. **2013**. The next-generation sequencing revolution and its impact on genomics. *Cell,* 155**,** 27-38.

Koonin, E. V. & Dolja, V. V. **2018**. Metaviromics: a tectonic shift in understanding virus evolution. *Virus Research,* 246**,** A1-A3.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R., Abdallah, R. Z., Sonnenschein, E. C., Cariou, T., O'Gara, F., Jackson, S., Orlic, S., Steinke, M., Busch, J., Duarte, B., Cacador, I., Canning-Clode, J., Bobrova, O., Marteinsson, V., Reynisson, E., Loureiro, C. M., Luna, G. M., Quero, G. M., Loscher, C. R., Kremp, A., DeLorenzo, M. E., Ovreas, L., Tolman, J., LaRoche, J., Penna, A., Frischer, M., Davis, T., Katherine, B., Meyer, C. P., Ramos, S., Magalhaes, C., Jude-Lemeilleur, F., Aguirre-Macedo, M. L., Wang, S., Poulton, N., Jones, S., Collin, R., Fuhrman, J. A., Conan, P., Alonso, C., Stambler, N., Goodwin, K., Yakimov, M. M., Baltar, F., Bodrossy, L., Van De Kamp, J., Frampton, D. M., Ostrowski, M., Van Ruth, P., Malthouse, P., Claus, S., Deneudt, K., Mortelmans, J., Pitois, S., Wallom, D., Salter, I., Costa, R., Schroeder, D. C., Kandil, M. M., Amaral, V., Biancalana, F., Santana, R., Pedrotti, M. L., Yoshida, T., Ogata, H., Ingleton, T., Munnik, K., Rodriguez-Ezpeleta, N., Berteaux-Lecellier, V., Wecker, P., Cancio, I., Vaulot, D., Bienhold, C., Ghazal, H., Chaouni, B., Essayeh, S., Ettamimi, S., Zaid el, H., Boukhatem, N., Bouali, A., Chahboune, R., Barrijal, S., Timinouni, M., El Otmani, F., Bennani, M., Mea, M., et al. **2015a**. The ocean sampling day consortium. *Gigascience,* 4**,** 27.

Kopf, A., Kostadinov, I., Wichels, A., Quast, C. & Glöckner, F. O. **2015b**. Metatranscriptome of marine bacterioplankton during winter time in the North Sea assessed by total RNA sequencing. *Mar Genomics,* 19**,** 45-46.

Krause-Jensen, D., Greve, T. M. & Nielsen, K. **2005**. Eelgrass as a Bioindicator Under the European Water Framework Directive. *Water Resources Management,* 19**,** 63-75.

Krishnamurthy, S. R. & Wang, D. **2017**. Origins and challenges of viral dark matter. *Virus Research,* 239**,** 136-142.

Kropinski, A. M., Adriaenssens, E. M., Alanis Villa, A., Hertwig, S. & Kuhn, J. H. **2015**. To create one (1) new genus, *Vhmlvirus*, including three (3) new species in the family Myoviridae. *ICTV approved proposal* 2015.024a-dB.

Kropinski, A. M., Adriaenssens, E. M., Pourcel, C. & Toussaint, A. **2018**. To create one (1) new family, Saltoviridae, containing three existing genera. *ICTV proposal* 2018.139B.

Kropinski, A. M., Prangishvili, D. & Lavigne, R. **2009**. Position paper: the creation of a rational scheme for the nomenclature of viruses of Bacteria and Archaea. *Environmental Microbiology,* 11**,** 2775-7.

Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. **2011**. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiology and Molecular Biology Reviews,* 75**,** 610-35.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. **2018**. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology Evolution,* 35**,** 1547-1549.

Kutter, E. & Sulakvelidze, A. **2004**. Bacteriophages: Biology and Applications, *Taylor & Francis press.*

Laber, C. P., Hunter, J. E., Carvalho, F., Collins, J. R., Hunter, E. J., Schieler, B. M., Boss, E., More, K., Frada, M., Thamatrakoln, K., Brown, C. M., Haramaty, L., Ossolinski, J., Fredricks, H., Nissimov, J. I., Vandzura, R., Sheyn, U., Lehahn, Y., Chant, R. J., Martins, A. M., Coolen, M. J. L., Vardi, A., DiTullio, G. R., Van Mooy, B. A. S. & Bidle, K. D. **2018**. Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology,* 3**,** 537-547.

Lan, S.-F., Huang, C.-H., Chang, C.-H., Liao, W.-C., Lin, I. H., Jian, W.-N., Wu, Y.-G., Chen, S.-Y. & Wong, H.-c. **2009**. Characterization of a new plasmid-like prophage in a pandemic *Vibrio parahaemolyticus* O3:K6 Strain. *Applied and Environmental Microbiology,* 75**,** 2659-2667.

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G. & Smith, D. B. **2018**. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research,* 46**,** D708-D717.

Lekunberri, I., Subirats, J., Borrego, C. M. & Balcázar, J. L. **2017**. Exploring the contribution of bacteriophages to antibiotic resistance. *Environmental Pollution,* 220**,** 981-984.

Lema, K. A., Willis, B. L. & Bourne, D. G. **2014**. Amplicon pyrosequencing reveals spatial and temporal consistency in diazotroph assemblages of the *Acropora millepora* microbiome. *Environmental Microbiology,* 16**,** 3345-3359.

Leterme, S. C., Pingree, R. D., Skogen, M. D., Seuront, L., Reid, P. C. & Attrill, M. J. **2008**. Decadal fluctuations in North Atlantic water inflow in the North Sea between 1958–2003: impacts on temperature and phytoplankton populations. *Oceanologia,* 50**,** 59-72.

Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. **2015**. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

Li, H. & Durbin, R. **2009**. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. **2008**. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics,* 24**,** 863-865.

Lima, F. P. & Wethey, D. S. **2012**. Three decades of high-resolution coastal sea surface temperatures reveal more than warming. *Nature Communications,* 3**,** 704.

Lin, H., Yu, M., Wang, X. & Zhang, X.-H. **2018**. Comparative genomic analysis reveals the evolution and environmental adaptation strategies of Vibrios. *BMC Genomics,* 19**,** 135.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R. & Pallen, M. J. **2012**. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology,* 10**,** 599.

Lopes, A., Tavares, P., Petit, M.-A., Guérois, R. & Zinn-Justin, S. **2014**. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics,* 15**,** 1027.

Lorenz, N., Reiger, M., Toro-Nahuelpan, M., Brachmann, A., Poettinger, L., Plener, L., Lassak, J. & Jung, K. **2016**. Identification and Initial Characterization of Prophages in *Vibrio campbellii*. *PLoS ONE,* 11**,** e0156010.

Lozupone, C. A. & Knight, R. **2007**. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences,* 104**,** 11436-11440.

Lucas, J., Wichels, A. & Gerdts, G. **2016**. Spatiotemporal variation of the bacterioplankton community in the German Bight: from estuarine to offshore regions. *Helgoland Marine Research,* 70.

Lucas, J., Wichels, A., Teeling, H., Chafee, M., Scharfe, M. & Gerdts, G. **2015**. Annual dynamics of North Sea bacterioplankton: seasonal variability superimposes short-term variation. *FEMS Microbiology Ecology,* 91**,** fiv099-fiv099.

Luo, H. & Moran, M. A. **2014**. Evolutionary Ecology of the Marine Roseobacter Clade. *Microbiology and Molecular Biology Reviews,* 78**,** 573-587.

Magrane, M. & Consortium, U. **2011**. UniProt Knowledgebase: a hub of integrated protein data. *Database: The Journal of Biological Databases and Curation,* 2011**,** bar009.

Makino, K., Oshima, K., Kurokawa, K., Yokoyama, K., Uda, T., Tagomori, K., Iijima, Y., Najima, M., Nakano, M., Yamashita, A., Kubota, Y., Kimura, S., Yasunaga, T., Honda, T., Shinagawa, H., Hattori, M. & Iida, T. **2003**. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. *The Lancet,* 361**,** 743-749.

Malainine, S. M., Moussaoui, W., Prévost, G., Scheftel, J. M. & Mimouni, R. **2013**. Rapid identification of *Vibrio parahaemolyticus* isolated from shellfish, sea water and sediments of the Khnifiss lagoon, Morocco, by MALDI-TOF mass spectrometry. *Letters in Applied Microbiology,* 56**,** 379-386.

Maniloff, J. & Ackermann, H. W. **1998**. Taxonomy of bacterial viruses: establishment of tailed virus genera and the other Caudovirales. *Archives of Virology,* 143**,** 2051-2063.

Marchant, H., Davidson, A., Wright, S. & Glazebrook, J. **2000**. The distribution and abundance of viruses in the Southern Ocean during spring. *Antarctic Science,* 12**,** 414-417.

Marraffini, L. A. **2015**. CRISPR-Cas immunity in prokaryotes. *Nature,* 526**,** 55-61.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Peña, M. J., Martínez, J. M., Anton, J., Gasol, J. M., Rosselli, R., Rodriguez-Valera, F., Sullivan, M. B., Acinas, S. G. & Martinez-Garcia, M. **2017**. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature Communications,* 8**,** 15892.

Martinez-Urtaza, J., Bowers, J. C., Trinanes, J. & DePaola, A. **2010**. Climate anomalies and the increasing risk of *Vibrio parahaemolyticus* and *Vibrio vulnificus* illnesses. *Food Research International,* 43**,** 1780-1790.

Martinez-Urtaza, J., Simental, L., Velasco, D., DePaola, A., Ishibashi, M., Nakaguchi, Y., Nishibuchi, M., Carrera-Flores, D., Rey.Alvarez, C. & Pousa, A. **2005**. Pandemic *Vibrio parahaemolyticus* O3:K6, Europe. *Emerging Infectious Diseases,* 11**,** 1319-1320.

Martinez-Urtaza, J., Trinanes, J., Abanto, M., Lozano-Leon, A., Llovo-Taboada, J., Garcia-Campello, M., Pousa, A., Powell, A., Baker-Austin, C. & Gonzalez-Escalona, N. **2018**. Epidemic dynamics of *Vibrio parahaemolyticus* illness in a hotspot of disease emergence, Galicia, Spain. *Emerging Infectious Diseases,* 24**,** 852-859.

Martínez, J. M., Swan, B. K. & Wilson, W. H. **2014**. Marine viruses, a genetic reservoir revealed by targeted viromics. *The ISME journal,* 8**,** 1079-1088.

Maslov, S. & Sneppen, K. **2017**. Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. *Scientific Reports,* 7**,** 39642.

Mathis, M., Elizalde, A., Mikolajewicz, U. & Pohlmann, T. **2015**. Variability patterns of the general circulation and sea water temperature in the North Sea. *Progress in Oceanography,* 135**,** 91-112.

Matsuzaki, S., Inoue, T., Tanaka, S., Koga, T., Kuroda, M., Kimura, S. & Imai, S. **2000**. Characterization of a Novel *Vibrio parahaemolyticus* Phage, KVP241, and its relatives frequently isolated from seawater. *Microbiology and Immunology,* 44**,** 953-956.

Matsuzaki, S., Tanaka, S., Koga, T. & Kawata, T. **1992**. A broad-host-range vibriophage, KVP40, isolated from sea water. *Microbiology and Immunology,* 36**,** 93-97.

Mcdaniel, L., Paul, J. H. & De la Rosa, M. **2006**. Temperate and lytic cyanophages from the Gulf of Mexico. *Marine Biological Association of the United Kingdom,* 86**,** 517-527.

McLeod, S. M., Kimsey, H. H., Davis, B. M. & Waldor, M. K. **2005**. CTXφ and Vibrio cholerae: exploring a newly recognized type of phage–host cell relationship. *Molecular Microbiology,* 57**,** 347-356.

McLeod, S. M. & Waldor, M. K. **2004**. Characterization of *XerC*- and *XerD*-dependent CTX phage integration in *Vibrio cholerae*. *Molecular Microbiology,* 54**,** 935-947.

McQuatters-Gollop, A., Gilbert, A. J., Mee, L. D., Vermaat, J. E., Artioli, Y., Humborg, C. & Wulff, F. **2009**. How well do ecosystem indicators communicate the effects of anthropogenic eutrophication? *Estuarine, Coastal and Shelf Science,* 82**,** 583-596.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P. & Lopez, R. **2013**. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Research,* 41**,** W597-600.

Meier-Kolthoff, J. P. & Göker, M. **2017**. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics,* 33**,** 3396-3404.

Metzker, M. L. **2009**. Sequencing technologies — the next generation. *Nature Reviews Genetics,* 11**,** 31.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. & Kubal, M. **2008**. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf,* 9.

Meyer, M. & Kircher, M. **2010**. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols,* 2010**,** pdb.prot5448.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S. & Ogata, H. **2016**. Linking virus genomes with host taxonomy. *Viruses,* 8.

Mikheenko, A., Saveliev, V. & Gurevich, A. **2016**. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics,* 32**,** 1088-1090.

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., Lee, J., Szczypinski, B. & Fraser, C. M. **2003**. Complete genome sequence of the broad-host-range vibriophage KVP40: Comparative genomics of a T4-related bacteriophage. *Journal of Bacteriology,* 185**,** 5220-5233.

Miyoshi, S. **2013**. Extracellular proteolytic enzymes produced by human pathogenic *Vibrio* species. *Front Microbiol,* 4**,** 339.

Mizuno, C., Ghai, R. & Rodriguez-Valera, F. **2014**. Evidence for metaviromic islands in marine phages. *Frontiers in Microbiology,* 5**,** 27.

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. **2013**. Expanding the marine virosphere using metagenomics. *PLoS Genetics,* 9**,** e1003987.

Model, P. & Russel, M. **1988**. Filamentous bacteriophages. In: Calendar, R. (ed.) The bacteriophages, vol 2. *Plenum Press, New York,* 375–456.

Moebus, K. **1980**. A method for the detection of bacteriophages from ocean water. *Helgoländer Meeresuntersuchungen,* 34**,** 1-14.

Moebus, K. **1983**. Lytic and inhibition responses to bacteriophages among marine bacteria, with special reference to the origin of phage-host systems. *Helgoländer Meeresuntersuchungen,* 36**,** 375-391.

Moebus, K. **1987**. Ecology of marine bacteriophages. *In:* Goyal, S. M., Gerba, C. P. & Bitton, G. (eds.) Phage Ecology (Goyal, SM, Gerba, CP and Bitton, G., Eds.). *John Wiley and Sons*.

Moebus, K. **1991**. Preliminary observations on the concentration of marine bacteriophages in the water around Helgoland. *Helgoländer Meeresunters,* 45**,** 411-422.

Moebus, K. **1992**. Further investigations on the concentration of marine bacteriophages in the water around Helgoland, with reference to the phage-host systems encountered. *Helgoländer Meeresunters,* 46**,** 275-292.

Moebus, K. **1997a**. Investigations of the marine lysogenic bacterium H24. I. General description of the phage-host system. *Marine Ecology Progress Series,* 148**,** 217-228.

Moebus, K. **1997b**. Investigations of the marine lysogenic bacterium H24. II. Development of pseudolysogeny in nutrient-rich broth culture. *Marine Ecology Progress Series,* 148**,** 229-240.

Morgan, G. J. **2016**. What is a virus species? Radical pluralism in viral taxonomy. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 59**,** 64-70.

Morgan, G. J., Hatfull, G. F., Casjens, S. & Hendrix, R. W. **2002**. Bacteriophage *Mu* genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. *Journal of Molecular Biology,* 317**,** 337-359.

Mork, M., Swallow John, C., Currie, R. I., Gill Adrian, E. & Simpson, J. H. **1981**. Circulation phenomena and frontal dynamics of the Norwegian coastal current. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences,* 302**,** 635-647.

Motro, Y. & Moran-Gilad, J. **2017**. Next-generation sequencing applications in clinical bacteriology. *Biomolecular Detection and Quantification,* 14**,** 1-6.

Muhire, B. M., Varsani, A. & Martin, D. P. **2014**. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE,* 9**,** e108277.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. **2015**. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences,* 10**,** 18.

Muramatsu, K. & Matsumoto, H. **1991**. Two generalized transducing phages in *Vibrio parahaemolyticus* and *Vibrio alginolyticus*. *Microbiology and Immunology*, 35**,** 1073-1084.
Nakamura, Y., Yasuike, M., Nishiki, I., Iwasaki, Y., Fujiwara, A., Kawato, Y., Nakai, T., Nagai, S., Kobayashi, T., Gojobori, T. & Ototake, M. **2016**. V-GAP: Viral genome assembly pipeline. *Gene*, 576**,** 676-80.
Nakasone, N., Ikema, M., Higa, N., Yamashiro, T. & Iwanaga, M. **2013**. A filamentous phage of *Vibrio parahaemolyticus* O3:K6 isolated in Laos. *Microbiology and Immunology*, 43**,** 385-388.
Nasu, H., Iida, T., Sugahara, T., Yamaichi, Y., Park, K.-S., Yokoyama, K., Makino, K., Shinagawa, H. & Honda, T. **2000**. A filamentous phage associated with recent pandemic *Vibrio parahaemolyticus* O3:K6 strains. *Journal of Clinical Microbiology*, 38**,** 2156-2161.
Noble, R. T. & Fuhrman, J. A. **1998**. Use of SYBR Green I for wapid epifliorescence counts of marine viruses and bacteria. *Aquat Microb Ecol*, 14**,** 113-118.
Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., McLean, J., Lasken, R., Clingenpeel, S. R., Woyke, T., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. Assembling genomes and mini-metagenomes from highly chimeric reads. *In:* Deng, M., Jiang, R., Sun, F. & Zhang, X., eds. Research in Computational Molecular Biology RECOMB 2013. *Lecture Notes in Computer Science*, vol 7821. *Springer, Berlin*, Heidelberg, 2013// 2013. 158-170.
O'Shea, Y. A. & Boyd, E. F. **2002**. Mobilization of the *Vibrio* pathogenicity island between *Vibrio cholerae* isolates mediated by CP-T1 generalized transduction. *FEMS Microbiology Letters*, 214**,** 153-157.
Oakey, H. J., Cullen, B. R. & Owens, L. **2002**. The complete nucleotide sequence of the *Vibrio harveyi* bacteriophage VHML. *Journal of Applied Microbiology*, 93**,** 1089-1098.
Oakey, H. J. & Owens, L. **2000**. A new bacteriophage, VHML, isolated from a toxin-producing strain of *Vibrio harveyi* in tropical Australia. *Journal of Applied Microbiology*, 89**,** 702-709.
Oberbeckmann, S., Fuchs, B. M., Meiners, M., Wichels, A., Wiltshire, K. H. & Gerdts, G. **2012**. Seasonal dynamics and modeling of a *Vibrio* community in coastal waters of the North Sea. *Microbial Ecology*, 63**,** 543-51.
Oberbeckmann, S., Wichels, A., Maier, T., Kostrzewa, M., Raffelberg, S. & Gerdts, G. **2011a**. A polyphasic approach for the differentiation of environmental *Vibrio* isolates from temperate waters. *FEMS Microbiology Ecology*, 75**,** 145-62.
Oberbeckmann, S., Wichels, A., Wiltshire, K. H. & Gerdts, G. **2011b**. Occurrence of *Vibrio parahaemolyticus* and *Vibrio alginolyticus* in the German Bight over a seasonal cycle. *Antonie Van Leeuwenhoek*, 100**,** 291-307.
Okonechnikov, K., Conesa, A. & García-Alcalde, F. **2016**. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32**,** 292-294.
Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, H. H., Szoecs, E. & Wagner, H. **2017.** Community Ecology Package. *In:* https://cran.r-project.org/package=vegan, R. f. (ed.).
Oppenheimer, C. H. & Zobell, C. E. **1952**. The growth and viability of sixty three species of marine bacteria influenced by hydrostatic pressure. *Journal of Marine Research*, 11**,** 10-18.
Otto, L., Zimmerman, J. T. F., Furnes, G. K., Mork, M., Saetre, R. & Becker, G. **1990**. Review of the physical oceanography of the North Sea. *Netherlands Journal of Sea Research*, 26**,** 161-238.
Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F. & Stevens, R. **2014**. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42**,** D206-D214.
Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N. & Kyrpides, N. C. **2016**. Uncovering Earth's virome. *Nature*, 536**,** 425.
Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. **2017**. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nature Protocols*, 12**,** 1673.
Park, K.-S., Ono, T., Rokuda, M., Jang, M. H., Okada, K., Iida, T. & Honda, T. **2004**. Functional characterization of two type III secretion systems of *Vibrio parahaemolyticus*. *Infection and Immunity*, 72**,** 6659-65.
Patel, A., Noble, R. T., Steele, J. A., Schwalbach, M. S., Hewson, I. & Fuhrman, J. A. **2007**. Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols*, 2**,** 269-276.
Pati, A., Ratner, A., Szeto, E., Chen, I. M. A., Pagani, I., Huang, J., Chu, K., Billis, K., Mavromatis, K., Palaniappan, K., Tennessen, K., Pillay, M., Huntemann, M., Ivanova, N. N., Varghese, N., Tringe, S., Kyrpides, N. C. & Markowitz, V. M. **2013**. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42**,** D568-D573.
Paul, J. H. **2008**. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME journal*, 2**,** 579-589.
Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. **2012**. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28**,** 1420-1428.
Percival, S., Chalmers, R., Embrey, M., Hunter, P., Sellwood, J. & Wyn-Jones, P. **2004**. Microbiology of Waterborne Diseases: Microbiological Aspects and Risks, *Elsevier Science.*
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.
Plunkett, G., 3rd, Rose, D. J., Durfee, T. J. & Blattner, F. R. **1999**. Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *Journal of Bacteriology*, 181**,** 1767-1778.
Proctor, L. M. **1997**. Advances in the study of marine viruses. *Microscopy Research and Technique*, 37**,** 136-161.
Pryshliak, M., Hammerl, J. A., Reetz, J., Strauch, E. & Hertwig, S. **2014**. *Vibrio vulnificus* phage PV94 is closely related to temperate phages of *V. cholerae* and other *Vibrio* species. *PLoS ONE*, 9**,** e94707.
Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. **2014**. Shedding new light on viral photosynthesis. *Photosynthesis Research*, 126**,** 71-97.
Quilici, M.-L., Robert-Pillot, A., Picart, J. & Fournier, J.-M. **2005**. Pandemic *Vibrio parahaemolyticus* O3:K6 spread, France. *Emerging Infectious Diseases*, 11**,** 1148-1149.
Quinlan, A. R. & Hall, I. M. **2010**. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26**,** 841-842.
Rachel, T. N. & Jed, A. F. **1998**. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology*, 14**,** 113-118.
Rajeev, L., Malanowska, K. & Gardner, J. F. **2009**. Challenging a Paradigm: the Role of DNA Homology in Tyrosine Recombinase Reactions. *Microbiology and Molecular Biology Reviews*, 73**,** 300-309.
Rakonjac, J. **2012**. Filamentous Bacteriophages: Biology and Applications. *eLS*.

Rappé, M. S. & Giovannoni, S. J. **2003**. The Uncultured Microbial Majority. *Annual Review of Microbiology*, 57**,** 369-394.
Rastrojo, A. & Alcamí, A. **2017**. Aquatic viral metagenomics: Lights and shadows. *Virus Research*, 239**,** 87-96.
Raya, R. l. R. & H'bert, E. M. **2009**. Isolation of Phage via Induction of Lysogens. *In:* Clokie, M. R. J. & Kropinski, A. M. (eds.) Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions. *Totowa, NJ: Humana Press.*
Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. **2017**. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5**,** 69.
Rihtman, B., Meaden, S., Clokie, M. R., Koskella, B. & Millard, A. D. **2016**. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. *PeerJ*, 4**,** e2055.
Rink, B., Grüner, N., Brinkhoff, T., Ziegelmüller, K. & Simon, M. **2011**. Regional patterns of bacterial community composition and biogeochemical properties in the southern North Sea. *Aquatic Microbial Ecology*, 63**,** 207-222.
Rishovd, S., Marvik, O. J., Jacobsen, E. & Lindqvist, B. H. **1994**. Bacteriophage P2 and P4 Morphogenesis: Identification and Characterization of the Portal Protein. *Virology*, 200**,** 744-751.
Roessner, C. A. & Ihler, G. M. **1984**. Proteinase sensitivity of bacteriophage lambda tail proteins *gpJ* and *pH* in complexes with the lambda receptor. *Journal of Bacteriology*, 157**,** 165-170.
Rohwer, F., Youle, M., Maughan, H. & Hisakawa **2014**. Life In Our Phage World. *Wholon, San Diego CA.*
Rosario, K. & Breitbart, M. **2011**. Exploring the viral world through metagenomics. *Current Opinion in Virology*, 1**,** 289-297.
Roux, F. L., Wegner, K. M., Baker-Austin, C., Vezzulli, L., Osorio, C. R., Amaro, C., Ritchie, J. M., Defoirdt, T., Destoumieux-Garzón, D., Blokesch, M., Mazel, D., Jacq, A., Cava, F., Gram, L., Wendling, C. C., Strauch, E., Kirschner, A. & Huehn, S. **2015a**. The emergence of *Vibrio* pathogens in Europe: ecology, evolution, and pathogenesis (Paris, 11–12th March 2015). *Frontiers in Microbiology*, 6**,** 830.
Roux, S. **2019**. A Viral Ecogenomics Framework To Uncover the Secrets of Nature's "Microbe Whisperers". *mSystems*, 4**,** e00111-19.
Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., Emerson, J. B., Enault, F., Fuhrman, J. A., Hingamp, P., Hugenholtz, P., Hurwitz, B. L., Ivanova, N. N., Labonté, J. M., Lee, K.-B., Malmstrom, R. R., Martinez-Garcia, M., Mizrachi, I. K., Ogata, H., Páez-Espino, D., Petit, M.-A., Putonti, C., Rattei, T., Reyes, A., Rodriguez-Valera, F., Rosario, K., Schriml, L., Schulz, F., Steward, G. F., Sullivan, M. B., Sunagawa, S., Suttle, C. A., Temperton, B., Tringe, S. G., Thurber, R. V., Webster, N. S., Whiteson, K. L., Wilhelm, S. W., Wommack, K. E., Woyke, T., Wrighton, K. C., Yilmaz, P., Yoshida, T., Young, M. J., Yutin, N., Allen, L. Z., Kyrpides, N. C. & Eloe-Fadrosh, E. A. **2018**. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology*, 37**,** 29.
Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., Vaqué, D., Tara Oceans, C., Bork, P., Acinas, S. G., Wincker, P. & Sullivan, M. B. **2016**. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537**,** 689-693.
Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. **2017**. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, 5**,** e3817.
Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. **2015b**. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3**,** e985.
Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. **2015c**. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife*, 4**,** e08490.
Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Cheng, J.-F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. & Eloe-Fadrosh, E. A. **2019**. Cryptic inoviruses are pervasive in bacteria and archaea across Earth's biomes. *bioRxiv***,** 548222.
Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. **2014**. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics*, 15.
Russell, D. A. & Hatfull, G. F. **2016**. PhagesDB: the actinobacteriophage database. *Bioinformatics*, 33**,** 784-786.
Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. **2000**. Artemis: sequence visualization and annotation. *Bioinformatics*, 16**,** 944-945.
Sabehi, G., Shaulov, L., Silver, D. H., Yanai, I., Harel, A. & Lindell, D. **2012**. A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. *Proceedings of the National Academy of Sciences*, 109**,** 2037-2042.
Sakib, S. N., Reddi, G. & Almagro-Moreno, S. **2018**. Environmental role of pathogenic traits in *Vibrio cholerae*. *Journal of Bacteriology*.
Salmond, G. P. & Fineran, P. C. **2015**. A century of the phage: past, present and future. *Nature Reviews Microbiology* 13**,** 777-86.
Sato, T., Shimizu, T., Watarai, M., Kobayashi, M., Kano, S., Hamabata, T., Takeda, Y. & Yamasaki, S. **2003**. Genome analysis of a novel Shiga toxin 1 (Stx1)-converting phage which is closely related to Stx2-converting phages but not to other Stx1-converting phages. *Journal of Bacteriology*, 185**,** 3966-3971.
Seed, K. D., Bodi, K. L., Kropinski, A. M., Ackermann, H. W., Calderwood, S. B., Qadri, F. & Camilli, A. **2011**. Evidence of a dominant lineage of *Vibrio cholerae*-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. *mBio*, 2**,** e00334-10.
Seemann, T. **2014**. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30**,** 2068-2069.
Sen, A. & Ghosh, A. N. **2005**. New *Vibrio cholerae* O1 biotype ElTor bacteriophages. *Virol J*, 2**,** 28.
Sime-Ngando, T. **2014**. Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Frontiers in Microbiology*, 5**,** 355.
Simmonds, P., Adams, M. J., Benko, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., Sabanadzovic, S., Sullivan, M. B., Suttle, C. A., Tesh, R. B., van der Vlugt, R. A., Varsani, A. & Zerbini, F. M. **2017**. Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15**,** 161-168.
Skopina, M. Y., Vasileva, A. A., Pershina, E. V. & Pinevich, A. V. **2016**. Diversity at low abundance: The phenomenon of the rare bacterial biosphere. *Microbiology*, 85**,** 272-282.
Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. **2006**. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103**,** 12115-20.
Solonenko, S. A. & Sullivan, M. B. **2013**. Preparation of metagenomic libraries from naturally occurring marine viruses. *In:* DeLong, E. F. (ed.) Methods in Enzymology. *Academic Press.*

Southward, A. J., Langmead, O., Hardman-Mountford, N. J., Aiken, J., Boalch, G. T., Dando, P. R., Genner, M. J., Joint, I., Kendall, M. A., Halliday, N. C., Harris, R. P., Leaper, R., Mieszkowska, N., Pingree, R. D., Richardson, A. J., Sims, D. W., Smith, T., Walne, A. W. & Hawkins, S. J. **2004**. Long-term oceanographic and ecological research in the Western English Channel. *Advances in Marine Biology. Academic Press.*

Srinivasan, P. & Ramasamy, P. **2017**. Morphological characterization and biocontrol effects of *Vibrio vulnificus* phages against Vibriosis in the shrimp aquaculture environment. *Microbial Pathogenesis,* 111**,** 472-480.

Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M. & Poisson, G. **2012**. Are we missing half of the viruses in the ocean? *The ISME journal,* 7**,** 672.

Su, L. K., Lu, C. P., Wang, Y., Cao, D. M., Sun, J. H. & Yan, Y. X. **2010**. Lysogenic infection of a Shiga toxin 2-converting bacteriophage changes host gene expression, enhances host acid resistance and motility. *Molecular Biology,* 44**,** 54-66.

Subirats, J., Sanchez-Melsio, A., Borrego, C. M., Balcazar, J. L. & Simonet, P. **2016**. Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes. *International Journal Antimicrobial Agents,* 48**,** 163-7.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. & Chisholm, S. W. **2005**. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol,* 3**,** e144.

Sullivan, M. B., Weitz, J. S. & Wilhelm, S. **2016**. Viral Ecology Comes of Age. *Environmental Microbiology Rep*.

Sullivan, M. J., Petty, N. K. & Beatson, S. A. **2011**. Easyfig: a genome comparison visualizer. *Bioinformatics,* 27**,** 1009-1010.

Summer, E. J., Gonzalez, C. F., Carlisle, T., Mebane, L. M., Cass, A. M., Savva, C. G., LiPuma, J. & Young, R. **2004**. *Burkholderia cenocepacia* phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. *Journal of Molecular Biology,* 340**,** 49-65.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G. & Bork, P. **2015**. Structure and function of the global ocean microbiome. *Science,* 348**,** 1261359.

Sündermann, J. & Pohlmann, T. **2011**. A brief analysis of North Sea physics. *Oceanologia,* 53**,** 663-689.

Suttle, C. A. **2005**. Viruses in the sea. *Nature,* 437**,** 356-61.

Suttle, C. A. **2007**. Marine viruses-major players in the global ecosystem. *Nature Reviews Microbiology* 5**,** 801-12.

Suttle, C. A. **2016**. Environmental microbiology: Viral diversity on the global stage. *Nature Microbiology,* 1**,** 16205.

Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. **2019**. Choice of assembly software has a critical impact on virome characterisation. *Microbiome,* 7**,** 12.

Systems, C. B. D. M., Commission on Geosciences, E. R., Studies, D. E. L. & Council, N. R. **1995**. Understanding Marine Biodiversity, *National Academies Press*.

Tampuu, A., Bzhalava, Z., Dillner, J. & Vicente, R. **2019**. ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *bioRxiv***,** 602656.

Tangherlini, M., Dell'Anno, A., Zeigler Allen, L., Riccioni, G. & Corinaldesi, C. **2016**. Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Scientific Reports,* 6**,** 28428.

Taniguchi, H., Sato, K., Ogawa, M., Udou, T. & Mizuguchi, Y. **1984**. Isolation and characterization of a filamentous phage, Vf33, specific for *Vibrio parahaemolyticus*. *Microbiology and Immunology,* 28**,** 327-337.

Thingstad, T. F. **2000**. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography,* 45**,** 1320-1328.

Thingstad, T. F. & Lignell, R. **1997**. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology,* 13**,** 19-27.

Thompson, F. L., Iida, T. & Swings, J. **2004**. Biodiversity of Vibrios. *Microbiology and Molecular Biology Reviews,* 68**,** 403-431.

Thurber, R. V. **2009**. Current insights into phage biodiversity and biogeography. *Current Opinion in Microbiology,* 12**,** 582-7.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. **2009**. Laboratory procedures to generate viral metagenomes. *Nature Protocols,* 4**,** 470-483.

Tolstoy, I., Kropinski, A. M. & Brister, J. R. **2018**. Bacteriophage taxonomy: an evolving discipline. *In:* Azeredo, J. & Sillankorva, S. (eds.) Bacteriophage Therapy: From Lab to Clinical Practice. *New York, NY: Springer New York.*

Toribio, A. L., Alako, B., Amid, C., Cerdeno-Tarraga, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R., Goodgame, N., Ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R., Liu, X., Martinez-Villacorta, J., Pakseresht, N., Rajan, J., Reddy, K., Rosello, M., Silvester, N., Smirnov, D., Vaughan, D., Zalunin, V. & Cochrane, G. **2017**. European Nucleotide Archive in 2016. *Nucleic Acids Res,* 45**,** D32-D36.

Torrella, F. & Morita, R. Y. **1979**. Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications. *Applied and Environmental Microbiology,* 37**,** 774.

Toussaint, A., Lima-Mendez, G. & Leplae, R. **2009**. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research,* 38**,** D57-D61.

Tragin, M. & Vaulot, D. **2018**. Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. *Scientific Reports,* 8**,** 14020.

Tucker, K. P., Parsons, R., Symonds, E. M. & Breitbart, M. **2011**. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *The ISME journal,* 5**,** 822-30.

Tye, B.-K., Huberman, J. A. & Botstein, D. **1974**. Non-random circular permutation of phage P22 DNA. *Journal of Molecular Biology,* 85**,** 501-527.

Van Regenmortel, M. H. V., Ackermann, H.-W., Calisher, C. H., Dietzgen, R. G., Horzinek, M. C., Keil, G. M., Mahy, B. W. J., Martelli, G. P., Murphy, F. A., Pringle, C., Rima, B. K., Skern, T., Vetten, H. J. & Weaver, S. C. **2013**. Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. *Archives of Virology,* 158**,** 1115-1119.

Vezzulli, L., Brettar, I., Pezzati, E., Reid, P. C., Colwell, R. R., Höfle, M. G. & Pruzzo, C. **2012**. Long-term effects of ocean warming on the prokaryotic community: evidence from the vibrios. *The ISME journal,* 6**,** 21-30.

Vezzulli, L., Colwell, R. R. & Pruzzo, C. **2013**. Ocean warming and spread of pathogenic Vibrios in the aquatic environment. *Microbial Ecology,* 65**,** 817-825.

Vezzulli, L., Grande, C., Reid, P. C., Hélaouët, P., Edwards, M., Höfle, M. G., Brettar, I., Colwell, R. R. & Pruzzo, C. **2016**. Climate influence on *Vibrio* and associated human diseases during the past half-century in the coastal North Atlantic. *Proceedings of the National Academy of Sciences,* 113**,** E5062-E5071.

Villarroel, J., Kleinheinz, K., Jurtz, V., Zschach, H., Lund, O., Nielsen, M. & Larsen, M. **2016**. HostPhinder: a phage host prediction tool. *Viruses*, 8**,** 116.

Vincent, A. T., Derome, N., Boyle, B., Culley, A. I. & Charette, S. J. **2017**. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *Journal of Microbiological Methods*, 138**,** 60-71.

Waldor, M. K. & Mekalanos, J. J. **1996**. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, 272**,** 1910.

Wang, Q., Kan, B. & Wang, R. **2013**. Isolation and characterization of the new mosaic filamentous phage VFJΦ of *Vibrio cholerae*. *PLoS ONE*, 8**,** e70934.

Warnes, G. R., Bolker, B. & Lumley, T. **2005.** Package 'gtools', various R Programming Tools. in: https://CRAN.R-project.org/package=gtools.

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A. C., Allen, M. J., Sullivan, M. B. & Temperton, B. **2019**. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ*, 7**,** e6800.

Weinbauer, M. G. **2004**. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, 28**,** 127-181.

Weinbauer, M. G., Brettar, I. & Höfle, M. G. **2003**. Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. *Limnology and Oceanography*, 48**,** 1457-1465.

Weinbauer, M. G. & Rassoulzadegan, F. **2004**. Are viruses driving microbial diversification and diversity? *Environmental Microbiology*, 6**,** 1-11.

Weynberg, K. D. **2018**. Viruses in marine ecosystems: from open waters to coral reefs. *In:* Malmstrom, C. M. (ed.) *Advances in Virus Research*. Academic Press.

Wichels, A., Biel, S. S., Gelderblom, H. R., Brinkhoff, T., Muyzer, G. & Schütt, C. **1998**. Bacteriophage diversity in the North Sea. *Applied and Environmental Microbiology*, 64**,** 4128-4133.

Wichels, A., Gerdts, G. & Schütt, C. **2002**. *Pseudoalteromonas* spp. phages, a significant group of marine bacteriophages in the North Sea. *Aquatic Microbial Ecology*, 27**,** 233-239.

Wickham, H. **2009**. ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York*.

Wilhelm, S. W. & Suttle, C. A. **1999**. Viruses and nutrient cycles in the sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, 49**,** 781-788.

Williamson, K. E., Helton, R. R. & Wommack, K. E. **2012**. Bias in bacteriophage morphological classification by transmission electron microscopy due to breakage or loss of tail structures. *Microsc Res Tech*, 75**,** 452-7.

Williamson, S. J. **2011**. Viral Metagenomics. Handbook of Molecular Microbial Ecology II. *John Wiley & Sons, Inc.*

Williamson, S. J., Houchin, L. A., McDaniel, L. & Paul, J. H. **2002**. Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Applied and Environmental Microbiology*, 68**,** 4307.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M. & Venter, J. C. **2008**. The sorcerer II global ocean sampling expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, 3**,** e1456.

Wiltshire, K. H., Kraberg, A., Bartsch, I., Boersma, M., Franke, H.-D., Freund, J., Gebühr, C., Gerdts, G., Stockmann, K. & Wichels, A. **2010**. Helgoland Roads, North Sea: 45 years of change. *Estuaries and Coasts*, 33**,** 295-310.

Wiltshire, K. H., Malzahn, A. M., Wirtz, K., Greve, W., Janisch, S., Mangelsdorf, P., Manly, B. F. J. & Boersma, M. **2008**. Resilience of North Sea phytoplankton spring bloom dynamics: An analysis of long-term data at Helgoland Roads. *Limnology and Oceanography*, 53**,** 1294-1302.

Winter, C., Garcia, J. A., Weinbauer, M. G., DuBow, M. S. & Herndl, G. J. **2014**. Comparison of deep-water viromes from the atlantic ocean and the mediterranean sea. *PLoS ONE*, 9**,** e100600.

Wommack, K. E., Bhavsar, J. & Ravel, J. **2008**. Metagenomics: Read Length Matters. *Applied and Environmental Microbiology*, 74**,** 1453-1463.

Wommack, K. E. & Colwell, R. K. **2000**. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology Molecular Biology Reviews*, 64**,** 69-114.

Xue, H., Xu, Y., Boucher, Y. & Polz, M. F. **2012**. High frequency of a novel filamentous phage, VCYΦ, within an environmental *Vibrio cholerae* population. *Applied and Environmental Microbiology*, 78**,** 28-33.

Yen, M., Cairns, L. S. & Camilli, A. **2017**. A cocktail of three virulent bacteriophages prevents *Vibrio cholerae* infection in animal models. *Nature Communications*, 8**,** 14187.

Yen, M. & Camilli, A. **2017**. Mechanisms of the evolutionary arms race between *Vibrio cholerae* and Vibriophage clinical isolates. *Int Microbiol*, 20**,** 116-120.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B. W., Blaser, M. J., Bonazzi, V., Booth, T., Bork, P., Bushman, F. D., Buttigieg, P. L., Chain, P. S. G., Charlson, E., Costello, E. K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J. A., Gallery, R. E., Gevers, D., Gibbs, R. A., Gil, I. S., Gonzalez, A., Gordon, J. I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A. L., Kelley, S. T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C. L., Legg, T., Ley, R. E., Lozupone, C. A., Ludwig, W., Lyons, D., Maguire, E., Methe, B. A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K. E., Nemergut, D., Neufeld, J. D., Newbold, L. K., Oliver, A. E., Pace, N. R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D. A., Assunta-Sansone, S., Schloss, P. D., Schriml, L., Sinha, R., Smith, M. I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J. M., Ward, D. V., Weinstock, G. M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J. R., Yatsunenko, T. & Glöckner, F. O. **2011**. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29**,** 415-420.

Zafar, N., Mazumder, R. & Seto, D. **2002**. CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes. *BMC Bioinformatics*, 3**,** 12.

Zehr, J. P., Mellon, M., Braun, S., Litaker, W., Steppe, T. & Paerl, H. W. **1995**. Diversity of heterotrophic nitrogen fixation genes in a marine cyanobacterial mat. *Applied and Environmental Microbiology*, 61**,** 2527-2532.

Zeidner, G., Bielawski, J. P., Shmoish, M., Scanlan, D. J., Sabehi, G. & Béjà, O. **2005**. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environmental Microbiology*, 7**,** 1505-1513.

Zeigler Allen, L., McCrow, J. P., Ininbergs, K., Dupont, C. L., Badger, J. H., Hoffman, J. M., Ekman, M., Allen, A. E., Bergman, B. & Venter, J. C. **2017**. The Baltic Sea virome: Diversity and transcriptional activity of DNA and RNA viruses. *mSystems*, 2**,** e00125-16.

Zhang, R. & Zhang, C. T. **2004**. A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics*, 20**,** 612-22.

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., Ellisman, M., Deerinck, T., Sullivan, M. B. & Giovannoni, S. J. **2013**. Abundant SAR11 viruses in the ocean. *Nature*, 494**,** 357-60.

Zhao, Y., Wang, K., Ackermann, H. W., Halden, R. U., Jiao, N. & Chen, F. **2010**. Searching for a "hidden" prophage in a marine bacterium. *Applied Environmental Microbiology*, 76**,** 589-95.

Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M.-A., Imamovic, L., Chow, B. K. C., Rode, A. A., Bytzer, P., Sommer, M. & Panagiotou, G. **2019**. Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome*, 7**,** 42.

Zhou, J., Zhang, W., Yan, S., Xiao, J., Zhang, Y., Li, B., Pan, Y. & Wang, Y. **2013**. Diversity of virophages in metagenomic data sets. *Journal of Virology*, 87**,** 4225.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. **2011**. PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, 39**,** W347-W352.

Zobell, C. **1946**. Marine microbiology, a monograph on hydrobacteriology. *Waltham Mass, Chronica Botanica Company*.