Interactions hôte-virus à travers d'un gradient de pH du sol à l'échelle communautaire et individuelle

Sungeun Lee

ÉCOLE CENTRALE LYON

N° d'ordre NNT : 2020LYSEC20

THÈSE

Présentée devant :

ÉCOLE CENTRALE DE LYON

Pour obtenir le grade de :

DOCTEUR

De l'école doctorale :

Électronique, électrotechnique, automatique

UNIVERSITÉ DE LYON

Spécialité : Ingénierie pour le vivant

Par

Sungeun LEE

Virus-host interactions across a soil pH gradient at the community and individual scale

soutenue le 23 Septembre 2020 devant le jury composé de:

Graeme W. NICOL — Directeur de Thèse
Professeur, Ecole Centrale de Lyon, Ecully, France

Christina HAZARD — Co-directeur de Thèse
Docteur, Ecole Centrale de Lyon, Ecully, France

Michael DUBOW — Rapporteur
Professeur, Institut de Biologie Intégrative de la Cellule, Gif-sur-Yvette, France

Tim URICH — Rapporteur
Professeur, Universitat Greifswald, City, Allemagne

Sophie ABBY — Examinateur
Docteur, Laboratoire TIMC-IMAG, La Tronche, France

Joanne EMERSON — Examinateur
Assistant Professeur, University of California, Davis, CA, Etats-Unis

Timothy M. VOGEL — Examinateur
Professeur, Université Claude Bernard Lyon 1, Ecole Centrale de Lyon, Ecully, France

Acknowledgments

First of all, I would like to thank my great supervisors Dr. Christina HAZARD and Prof. Graeme W. NICOL for kindly answering my many questions with their scientific expertise and for their patience, support, valuable advice and endless encouragement through the years, as well as for the many fortuitous research opportunities.

I would also like to thank my adviser Prof. Timothy M. VOGEL who taught us how to be a critical scientist and for introducing me to Dr. Christina HAZARD.

I thank the members of my PhD jury, Joanne EMERSON, Michael DUBOW, Sophie ABBY, Tim URICH and Timothy VOGEL for evaluating my work.

I would like to thank Marie ROBIN and Rosaria FERRIGNO for the committee meetings.

I would also like to acknowledge all of my past and present colleagues at the Laboratoire Ampère, Ecole Centrale de Lyon: Romie, Concepcion, Rose, Benoît, Adrien, Arthur, Eva, Marion, all fellow PhD students, as well as Cécile, Laure, Christoph, Catherine, Pascal, Sébastien, Richard, Gilles, David, Alexiane and Agathe who have given me advice, support and friendship over the years. I would particularly like to thank Romie who always motivated me and for our many stimulating conversations, Concepcion for our co-working days with many discussions and Christoph for his sequencing help. A special thanks to Laure for her valuable feedback for manipulating soil viruses. I also thank Edith who supported me a lot with their smile. I thank Laurent POUILLOUX for his patience and help with solving server and computing problems.

I wish to thank the team of the Mary Firestone Lab, University of Berkeley for fruitful and fascinating collaborations, especially Alexa NICOLAS, Ella SIERADZKI and Mary K. FIRESTONE for their welcome during our visit. Also, thanks to Lucas P. P. BRAGA and Laurent PHILLIPOT from INRAE in Dijon for teaching me virus infection assay methods, and the crew of the Centre d'Imagerie Quantitative Lyon-Est (CIQLE) from the University of Lyon, Elisabeth ERRAZURIZ-CERDA who helped me with TEM. Thank you to Dr. Robin Walker at SRUC Craibstone Aberdeen for access to the pH plots at the Woodlands Field Experiment.

I am very grateful to the AXA Research Fund, the France Berkeley fund and the JGI CSP program for supporting this work.

Finally, I thank my family, my dad and uncle in heaven, my aunt and my husband Tommy who always stands by my side and my mom who has always encouraged me with her endless love.

Abstract

Soil viruses have potential to influence microbial community structure and subsequent ecosystem functioning by directly affecting the abundance of host cells by lysis and through their ability to transfer genes between hosts. Although our understanding of soil viral diversity and functioning has increased, the role of viruses and their interactions with prokaryotes in soil is limited. To gain a better understanding of virus-host interactions in soil, a long-term pH-manipulated soil gradient, which microbial community structure changes across, was investigated. The main objectives of this thesis were to 1) determine the influence of microbial community structure and soil pH on viruses using metagenomics and viromics (Chapter II), 2) determine the infectivity of soil viral populations from co-localized and foreign pH soil niches using a plaque assay approach combined with hybrid metagenomics sequencing (Chapter III) and 3) identify virus populations infecting specific soil microbial functional groups, specifically methanotrophs (Chapter IV) and nitrifiers (Chapter V), using DNA stable isotope probing combined with metagenomic deep sequencing. Viral community structure was found to change with soil pH, demonstrating that viral communities are tightly linked to host populations, but also may have narrow host ranges. Analysis of clustered regularly interspaced short palindromic repeats (CRISPR) arrays revealed dynamic virus-host interactions, with the number and size of CRISPR arrays distinct across contrasting pH soil. Profiling of the host-virus linkages between soil pH, suggests that viruses play a critical role in shaping the composition and function of the soil prokaryotic community. Surprisingly, greater infectivity of a host bacterium by virus populations was found when viruses and host bacterium were not co-localized in the same pH soil. Coevolutionary processes between the host and virus populations, such as restriction modification/virus-encoded methyltransferase and CRISPR-*Cas* system/spacer mutation, provide evidence for local adaptation, and that virus-bacterial host interactions play an integral part in the susceptibility of a host to infection and consequently in the regulation of soil bacterial populations. Targeting specific microbial functional groups via stable isotope probing allowed analysis of individual host-virus populations. Tracking carbon flow through prokaryotic and viral populations revealed active interactions between viruses and methanotroph and nitrifier hosts, and soil pH niche preferences. Evidence of horizontal gene transfer and virus-encoded auxiliary metabolic genes, such as glycoside hydrolase families, peptidases, particulate methane monooxygenase subunit C (*pmoC*), nitrogenase (*nifH*) and cytochrome cd1-nitrite reductase, supports that viruses are significant contributors to host functioning and carbon and nitrogen cycling in soil. Overall, this work demonstrated that soil viruses are important regulators of microbial communities through specific host lysis and dynamic virus-host interactions.

Table of Contents

List of Figures

Chapter III

Chapter IV

Chapter V

Chapter VI

List of Tables

Chapter V.

List of Supplementary Tables and Figures

Abbreviations

General abbreviations

^{13}C: carbon-13
ANI: Average nucleotide identity
AT: Adenine tymine
ATP: Adenosine triphosphate
BLAST: Basic local alignment search tool
C: Carbon
CH_4: Methane
CO_2: Carbon dioxide
CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
DB: Database
DNA: Deoxyribonucleic acid
SIP: Stable isotope probing
dNTP: Deoxyribonucleotide triphosphate
DR: Direct repeat
ds: double stranded
dTMP: deoxythymidine monophosphate
FAD: Flavin adenine dinucleotide
GC: Guanine cytosine
H: Hydrogen
HMM: Hidden Markov model
ICTV: International Committee on Taxonomy of Viruses
JGI: Joint Genome Institute
MAG: Metagenomic assembled genome
MC: Metagenomic contig
Mtase: Methyltransferase
mVC: metagenomic viral contigs
N: Nitrogen
N_2O: Nitrous oxide
NAD: Nicotinamide adenine dinucleotide
NC: Negative Control
NCBI: National Center for Biotechnology Information
NGS: Next generation sequencing
NH_4^+: Ammonium
NMDS: Non-metric multidimensional scaling
NO_2^-: Nitrite
NO_3^-: Nitrate
NR: Non-rebundant protein sequences
NT: Non-redundant nucleotide sequences

Organisms

AO: Ammonia oxidizer
AOA: Ammonia oxidizing archaea
AOB: Ammonia oxidizing bacteria
Ca.: *Candidatus*
Comammox: complete ammonia oxidizers
MOB: Methane oxidizing bacteria
NOB: Nitrite oxidizing bacteria
NSV: Nitrosopumilus spindle-shaped virus
sp.: species

Genes or proteins

16S rRNA: 16S ribosomal RNA
AMO: Ammonia monooxygenase
amoA: gene encoding alpha subunit of AMO
GH: Glycoside Hydrolase
MDH: Methanol dedydrogenase
MMO: Methane monooxygenase
NXR: Nitrite oxidoreductase
pMMO: Particular methane monooxygenase
pmoA: gene encoding alpha subunit of pMMO
sMMO: soluble methane monooxygenase

Chemical buffers

APC: amended potassium citrate
CsCl: Cesium chloride
CTAB: cetyltrimethylammonium bromide
PC: Potassium citrate
PEG: polyethylene glycerol
SDS: sodium dodecyl sulfate
SM: Storage medium
SP: Sodium pyrophosphage
TSA: Tryptic soy agar
TSB: Tryptic soy broth

Units

bp: base pair
CFU: Colony forming unit
CPM: Copies per million reads
Gb: Gigabyte

OD: Optical Density
ONF: Oligonucleotide frequency
ORF: Open reading frame
PCR: Polymerase chain reaction
PC: Positive Control
QC: Quality control
qPCR: Quantitative polymerase chain reaction
Rease: Restriction endonuclease
RM: Restriction modification
RNR: Ribonucleotide reductase
ss: single stranded
TEM: Transmission electron microscopy
VC: Viral contig
vVC: Viromic viral contig
kb: kilo-base pair
PFU: Plaque Forming Unit
RPK: Read per kilobase
Tb: Terabyte

CHAPTER I

General introduction:
Viruses, host interactions and the use of metagenomic approaches to understand their ecology

1.1. Overview

Viruses are an integral part of every environment and can infect all the primary domains of life (Clokie et al. 2011). Viruses in the marine environment are well recognized as important drivers of microbial community structure and ecosystem functioning through directly affecting the abundance of host cells by lysis and through their proficiency in transferring genes between hosts (Wommack and Colwell 2000; Weinbauer et al. 2003; Weinbauer and Rassoulzadegan 2004; Suttle 2005). In the oceans, viruses are responsible for killing approximately one-third of prokaryotic cells per day, and thereby altering carbon (C) and nutrient cycling at the global scale (Fuhrman 1999). In addition, marine viruses can impact productivity in a given ecosystem by manipulating host metabolic functions after viral infection via the expression of virus-encoded auxiliary metabolic genes (AMGs) (Breitbart et al. 2007). In comparison to the relatively homogeneous marine environment, there are more diverse habitats for microorganisms in soil due to the wide variation in their composition, spatial heterogeneity and physicochemical properties (Williamson et al. 2017). In one gram of soil up to 10 billion prokaryotes can be found (Raynaud and Nunan 2014). Similarly, viruses in soil may be as abundant and equally diverse, and therefore potentially play important roles in influencing microbial community structure and ecosystem functions (Williamson et al. 2017). Although our understanding of the scale of viral diversity and functioning in soil has increased, knowledge on virus-bacterial host interactions in soil remains limited (Han et al. 2017; Trubl et al. 2018; Graham et al. 2019; Emerson 2019). In this chapter, the biology and interactions of viruses and the current state of knowledge on soil viruses, metagenomic approaches used to study viral communities and the model soil pH gradient system that was studied, along with the specific research aims, is presented below.

1.1.1. Biology of viruses and their life cycle

Viruses are intracellular obligate parasites, meaning that they are completely dependent on their living host for replication, have no intrinsic metabolism and are dormant when not infecting a suitable host (Gelderblom 1996). All viruses consist of the same fundamental constituents, a protein capsid and a nucleic acid genome (King et al. 2011). The capsid describes the protein coat that surrounds the genome of the virus and is constructed of identical monomer proteins (capsomers) that self-assemble. Together, the capsid and the genome compose the virion, which is generally referred to as the infectious particle (King et al. 2011). The majority of characterized capsids have either a cubic or helical symmetry. However, some viruses, primarily those infecting bacteria or archaea, can contain more complex capsids (King et al. 2011). Virus genomes are either composed of DNA or RNA (Gelderblom 1996; King et al. 2011). DNA viruses can be either single stranded (ssDNA) or double stranded (dsDNA) and linear or circular. RNA viruses can also be single stranded (ssRNA) or double stranded (dsRNA) and unsegmented or segmented (Lodish et

al. 2000). Viruses of bacteria are typically referred to as bacteriophages (or phages), although these terms are not routinely used for the description of viruses that infect archaea. According to the International Committee on Taxonomy of Viruses (ICTV), viruses are taxonomically classified into orders, families, subfamilies, genera, species, isolates and variants of a host (Kuhn et al. 2010).

There are two main virus reproduction cycles, the lytic and lysogenic cycles (Figure 1.1). First, the virus attaches itself to the surface of the host cell through specific binding between viral surface glycoproteins and specific receptor molecules (Aswad and Katzourakis 2018). The virus then inserts its genetic material into the host cell via receptor-mediated endocytosis or other mechanisms. Viral or host enzymes degrade the viral capsid. In the lytic cycle, the viral genome induces the synthesis of viral constituents, assembly of new viral particles, and lysis of the infected cell, releasing new viruses that will spread and infect other host cells (Clokie et al. 2011). In the lysogenic cycle the viral genome is incorporated into the host cell's genome, or remains as a plasmid and thus reproduces as a prophage (Weinbauer et al. 2003; Weinbauer and Rassoulzadegan 2004). This is a latent form, in which the viral genes are present in the host without causing disruption of the cell (Clokie et al. 2011). Prophages are replicated and maintained in the following generations until an environmental stress of the host cell triggers a switch to the lytic cycle (Weinbauer et al. 2003). Lysogeny is an effective strategy for viral populations to persist when the abundance of host cells is low (Williamson et al. 2002; Weinbauer et al. 2003; Mann 2003; Chibani-Chennoufi et al. 2004; Kimura et al. 2008) or when host survival depends on periods of inactivity (Pantastico-Caldas et al. 1992; Kimura et al. 2008).

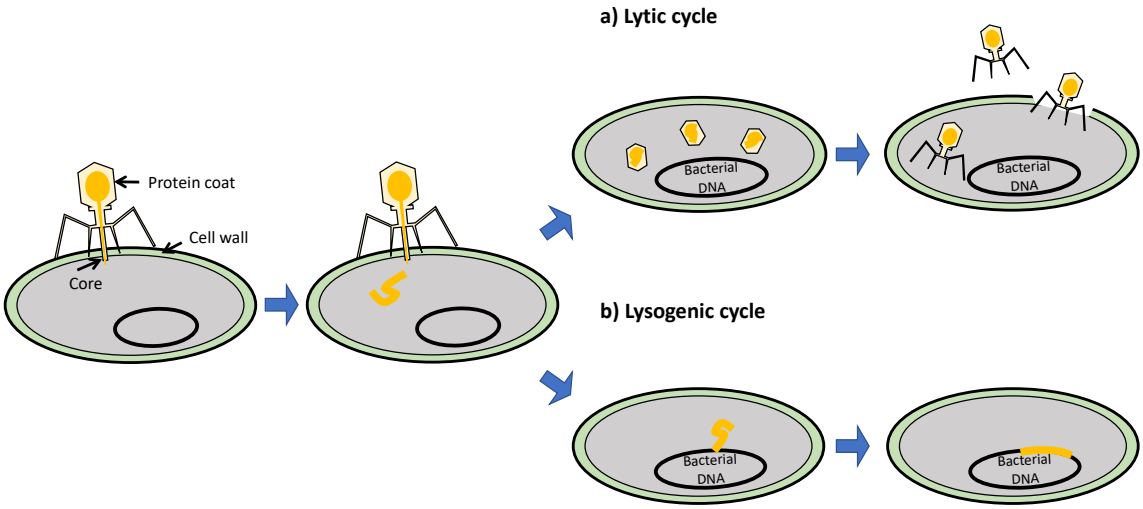


Figure 1.1. Schematic representation of a) the lytic and b) lysogenic cycles. The lytic cycle ends by releasing the mature viruses and the lysogenic cycle where viral DNA is incorporated into the host chromosome.

1.1.2. Abundance and diversity of prokaryotic viruses in soil

Prokaryotic viruses have been shown to be more abundant than prokaryotes, and are the most abundant and diverse biological entities in the biosphere (Fuhrman 1999; Williamson et al. 2013). It has been extrapolated from direct counts of virus-like particles in different environments that the global virosphere may contain up to $\sim 10^{31}$ viral particles (Edwards and Rohwer 2005; Breitbart and Rohwer 2005; Suttle 2005; Silveira and Rohwer 2016). Several studies measuring viral abundance in soils by transmission electron microscopy (TEM) or epifluorescent microscopy have shown large numbers of viral particles ranging from 10^7 to 10^{10} per gram of dry weight soil (Williamson et al. 2003, 2017; Han et al. 2017). In a recent study, viral particle abundance of four soil types was found to be similar but the different morphological groups of viruses had different relative abundances across the soil types (Reavy et al. 2015). Viral and bacterial abundance has been shown to increase with ecosystem productivity, generally lowest in dry, arid soils and greatest in moist soils rich in organic matter (Williamson et al. 2017). However, while viral abundance can vary between soils of different composition and geographic location (Williamson et al. 2005), soil viral abundance is comparatively stable relative to marine ecosystems. Viral abundance in marine environments has been shown to change over 2000-fold through the water column (Srinivasiah et al. 2008). However, comparison of viral abundances in soil is confounded by various parameters, including viral particle extraction and detection, and effects of soil properties (Trubl et al. 2016; Williamson et al. 2017). Counts may be greatly underestimated due to the difficulty of extracting all viruses present. The total number of free viruses in soil is probably 1 – 2 orders of magnitude greater than that of bacterial populations, and thus their relative abundance in comparison to prokaryotic cell numbers may be comparable to marine ecosystems (Watt et al. 2006; Trubl et al. 2018).

Viruses can be classified by morphology (van Regenmortel et al. 2000) with TEM the most extensively used technique for characterizing into morphotypes (Ackerman et al. 1978). Tailed phages constitute the order *Caudovirales*, accounting for 95% of all phages, and likely make up the majority of the viruses on the planet (Ackermann 1998; Maniloff and Ackermann 1998). The *Myoviridae* have a contractile tail, the *Siphoviridae* have a long non-contractile tail and the *Podoviridae* have a short non-contractile tail (Fauquet et al. 2005). Based on TEM and metagenomic analysis, viruses belonging to the order *Caudovirales* have been shown to dominate in soil (Zablocki et al. 2014; Ballaud et al. 2016; Han et al. 2017). Based on ICTV and the Virus-Host Database, there are currently 21 families, 57 subfamilies and 797 genera of bacterial viruses, and 19 families and 24 genera of archaeal viruses (Adriaenssens et al. 2020; Mihara et al. 2016). Based on cultivated viruses, archaeal viruses are morphologically more diverse than bacterial viruses, despite being underrepresented (Pietilä et al. 2014; Snyder et al. 2015). Viral metagenomics analyses suggest that global environmental viral diversity is vast and that the

majority of viral diversity has yet to be characterized (Angly et al. 2006; Mokili et al. 2012; Roux et al. 2015b). Between 60 to 99% of sequences within a virome typically have no significant similarity to sequences within a reference database, and is referred to as "viral dark matter" (Brum et al. 2015; Roux et al. 2015b). Although most research has focused on dsDNA viruses, recent work has also revealed distinct circular ssDNA viruses, and diverse and abundant RNA viruses in soil (Reavy et al. 2015; Starr et al. 2019).

1.1.3. Virus-host interactions

Viruses infect all types of cellular life that are present in soil, including eukaryotes and prokaryotes (Weinbauer and Rassoulzadegan 2004). Through both the lytic and lysogenetic cycle, viruses interact with their hosts resulting in impacts on microbial communities and nutrient cycling. As viruses cause cell lysis and the release of proteins and nucleic acids, they may have an important direct role in the cycling of carbon, nitrogen, sulfur and phosphorus in soil (Williamson et al. 2017). Soil viruses may increase the amount of available carbon which can influence microbial production and respiration (Williamson et al. 2017). The release of microbial cellular material via viral lysis, commonly referred to as the viral shunt, are reused by microorganisms and in part of the microbial loop (Suttle 2005). In marine systems, viruses were estimated to kill up to 40% of marine bacteria per day and to contribute to the viral shunt (Suttle, 2005). Although similar roles for soil viruses has been suggested, there is a lack of information about the contributions of the viral shunt in the soil food-web (Kuzyakov and Mason-Jones, 2018; Emerson et al. 2019).

The transfer of genetic material between viruses and hosts and then between microorganisms has important consequences. During the lysogenic cycle, prophages can alter host metabolism and the host phenotype, resulting in a change of fitness, and the expression of prophage genes can even protect their hosts from additional phage infection (Williamson et al. 2017). An additional role of viruses is transduction between prokaryotes mediated by viruses (Canchaya et al. 2003a). In the soil environment this is an important mechanism for transferring genes, resulting in host diversification and speciation (Wiedenbeck and Cohan 2011). In soil microcosms, transduction between introduced bacteria and phages has been observed although it has not yet been shown to occur *in situ* in indigenous soil bacteria, likely due to the technical difficulties of detecting rare transduction events in soil (Elsas et al. 2003).

1.1.3.1. Interactions between bacteria and viruses

The study of bacteriophage-host interactions remains a challenge because many host bacteria remain uncultured and techniques used for the isolation and characterization of phages is limited to those associated with cultured organisms (de Jonge et al. 2019). Based on infection assays, the

host range of bacteriophages is on a continuum, from broad to extremely narrow (Ross et al. 2016). For example, isolated bacteriophages from *Vibrio parahaemolyticus* were shown to not infect other strains of this species or other *Vibrio* species (Weinbauer and Rassoulzadegan 2004). In contrast, other bacteriophages have been shown to infect both multiple strains of the same species and multiple species (Greene and Goldberg 1985; Vinod et al. 2006; Uchiyama et al. 2008; Gupta and Prasad 2011; Khan Mirzaei and Nilsson 2015; Yu et al. 2016). For example, species of *Escherichia coli*, *Citrobacter freundii*, *Shigella sonnei*, *Enterobacter* and *Erwinia* were all found to be infected by the bacteriophage Mu (Ross et al. 2016). In soils, an understanding of infectivity rates and those of key functional microbial groups, such as methanotrophs and nitrifiers which directly regulate carbon and nitrogen cycling, is limited. However, viruses infecting methanotrophs have been isolated in several environments (Tiutikov et al. 1976; Tyutikov et al. 1980, 1983) and viruses from a recent soil metagenomic study have been predicted to associate with methanotroph hosts (Emerson et al. 2018).

Based on sequenced bacterial genomes, it has been estimated that 60 – 70% have prophages (Paul 2008). It has been proposed that lysogeny occurs during times when nutrients are scarce and host population sizes are small, and when environmental factors become favorable, the bacteriophage can extract itself from the host genome and enter into the lytic pathway (Williamson et al. 2002; Kimura et al. 2008). During lysogeny, a symbiotic relationship between the prophages and its bacterial host can promote the fitness of both prophages and host by expressing genes that increase the fitness of the host cell (Canchaya et al. 2003b). This process is known as lysogenic conversion (van Houte et al. 2016). For example, temperate bacteriophages may influence root nodule colonization, N-fixation efficiency and crop productivity by altering phenotypes of *rhizobium* through lysogenic conversion (Kimura et al. 2008). Also hosts and phages develop antiviral defense mechanisms and viral counterdefense mechanisms, respectively (Abedon 2012; Vasu and Nagaraja 2013). Many bacteria also have the clustered regularly interspaced short palindromic repeats (CRISPR)-*Cas* system as an adaptive defense against viruses (Bhaya et al. 2011). These mechanisms are further discussed in section 1.1.3.3.

1.1.3.2. Interactions between archaea and viruses

Most of the present knowledge on archaeal viruses is built on extremophiles, with characterized viruses infecting either hyperthermophilic *Crenarchaeota*, or halophilic or methanogenic *Euryarchaeota* (Snyder et al. 2015; Albers 2016; Quemin et al. 2016). However, the genomes of a number of marine archaeal viruses predicted to infect members of the *Euryarchaeota* and *Thaumarchaeota* have been assembled (Uchiyama et al. 2008; Philosof et al. 2017; Prangishvili et al. 2017; López-Pérez et al. 2019; Ahlgren et al. 2019). For example, the magrovirus group that infects the ubiquitous but uncultured marine group II *Euryarcheota* was recently discovered

(Philosof et al. 2017). In addition, viruses that infect ammonia oxidizing archaea (AOA) belonging to the phylum *Thaumarchaeota*, have also been assembled from metagenomic samples (López-Pérez et al. 2019; Ahlgren et al. 2019). These AOA viruses are most closely related to members of the order *Caudovirales* (Suttle 2005; Labonté et al. 2015; López-Pérez et al. 2019; Ahlgren et al. 2019). In some AOA viral genomes the presence of AMGs encoding for ammonia monooxygenase subunit C (*amoC*), a subunit of ammonia monooxygenase (AMO), was found, suggesting the potential role of *Thaumarchaeota* viruses in modulating nitrification in oceans (Roux et al. 2016; Ahlgren et al. 2019). Additionally, putative proviruses have been retrieved within the genome of the marine thaumarchaeon *Candidatus* Nitrosomariuns catalina SPOT01, and the genome of *Nitrososphera viennensis* EN76, the first isolated soil AOA (Krupovic et al. 2011; Ahlgren et al. 2019). Recently, three *Nitrosopumilus* spindle-shaped viruses (NSVs) were isolated from suspended particulate matter-rich seawater samples with a host AOA strain (Kim et al. 2019). These NSVs have a narrow host range and display high rates of adsorption on to their host cells, indicating efficient infectivity of the *Nitrosopumilus* host strain (Kim et al. 2019). Previous metagenomic studies have suggested that viruses infecting *Thaumarchaeota* have a major impact on thaumarchaeal functioning and mortality by cell lysis, and hence, modifying nitrogen and carbon cycles (Danovaro et al. 2016; López-Pérez et al. 2019; Ahlgren et al. 2019). Nevertheless, most marine AOA viruses have yet to be isolated due to difficulty in obtaining host cells in pure culture. Currently, no AOA virus has been isolated from soil.

1.1.3.3. The co-evolution 'arms race' between host cells and viruses

Host-virus interactions during virus replication involve virus adsorption to cell receptors and the entry of viral DNA into the host cell, resulting in a constant evolutionary arms race to change cell structures limiting viral infection and subsequent evolution of viruses to overcome these defensive adaptations (Golais et al. 2013). During virus replication the host cell can use antiviral defense mechanisms including the restriction-modification system, the CRISPR-*Cas* systems and abortive infections (Deveau et al. 2010; Labrie et al. 2010; Stern and Sorek 2011; tenOever 2016). The restriction-modification system is found in prokaryotes and provides a defense against foreign DNA through the action of restriction endonuclease and methyltransferases (Tock and Dryden 2005; Vasu and Nagaraja 2013). Endonucleases are proteins that recognize foreign DNA and cleave at specific palindromic DNA sequences (i.e. restriction sites) (Vasu and Nagaraja 2013). Unlike host bacterial DNA, viral DNA is not methylated by methyltransferase and therefore viral DNA is unprotected and cleaved by restriction endonucleases (Wilson and Murray 1991). Subsequently, viruses evolve mechanisms to counteract this form of protection. For example, the incorporation of unusual bases, such as 5-hydroxymethyluracil instead of thymine, into their genomes can modify restriction sites (Krüger and Bickle 1983). Some phages can code for their

own methyltransferase, incite the production of the host methyltransferase or possess genes encoding proteins that bind to restriction sites or mimic DNA proteins that can neutralize endonuclease action (Stern and Sorek 2011; Golais et al. 2013).

Nearly all of archaea and approximately half of all bacteria have been found to possess CRISPR arrays, which cooperate with CRISPR-associated proteins (encoded by *Cas* genes) to form the basis of the CRISPR-*Cas* adaptive immune systems in prokaryotes (Figure 1.2) (Jansen et al. 2002; Terns and Terns 2011). Host cells continuously acquire CRISPR spacer sequences from viruses to facilitate recognition and evasion of future viral infection. To evade newly acquired spacers, the viruses can mutate the targeted spacer sequence or phosphorylate the *Cas* proteins to evade the CRISPR-*Cas* system (Horvath and Barrangou 2010; Golais et al. 2013). Conversely, CRISPR repeats and their associated proteins evolve to escape a shut-down mechanism for the CRISPR system encoded by viruses (Wang et al. 2020). Thus, bacteria and viruses are locked in an arms race. The coevolution between host and viruses in the same habitat commonly occurs and is a key regulator of ecological and evolutionary processes in microbial communities (Koskella and Brockhurst 2014). The arms race may have long-term evolutionary consequences on the host population, and in an apparent compromised state, lysogeny may occur (Golais et al. 2013; Koskella and Brockhurst 2014).

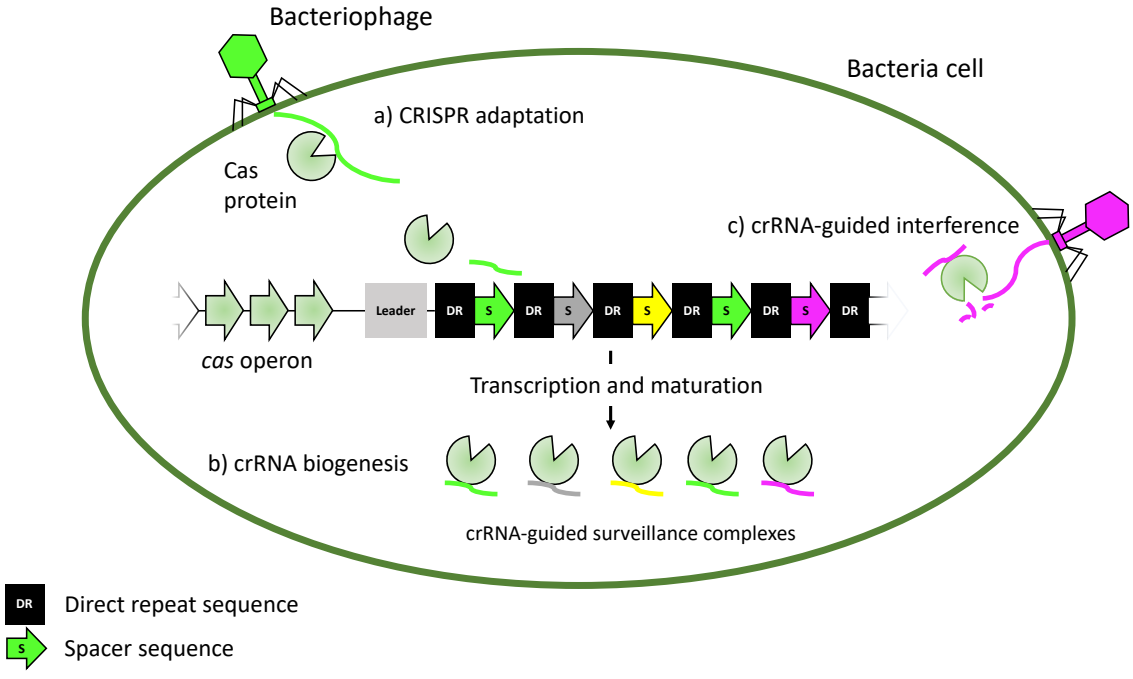


Figure 1.2. Schematic representation of the CRISPR-*Cas* mechanism. a) CRISPR adaptation where a new spacer sequence (S) from a phage genome is incorporated into the CRISPR system, adjacent to the leader sequence; b) CRISPR RNA (crRNA) biogenesis where spacer transcripts from the CRISPR array are transcribed into RNA and maturation into crRNA; and c) crRNA-guided interference where foreign DNA is recognized and silenced by the crRNA-guided surveillance complexes.

1.1.3.4. Auxiliary metabolic genes

Viruses often acquire host genes facilitating horizontal gene transfer between hosts (Hendrix et al. 2000; Miller et al. 2003; Lindell et al. 2004). Viral genomes include genes encoding proteins involved in the production of phage progeny, including DNA replication, nucleotide production, and RNA transcription (Lindell et al. 2004). Viral genes that do not contribute to viral replication, but have functions that alter host metabolism and may aid in the production of new viruses are termed auxiliary metabolic genes (AMGs) (Breitbart et al. 2007; Crummett et al. 2016; Jin et al. 2019). The AMGs of marine cyanophages have been the most studied (Mann 2003; Lindell et al. 2004; Sullivan et al. 2005; Millard et al. 2010). AMGs may be considered common or rare, with common AMGs among various lineages of hosts encoding metabolic functions that are essential under a range of conditions, whereas rare AMGs may be involved for only particular conditions (Crummett et al. 2016). For example, a recent analysis of virome data derived from the Pacific Ocean identified niche-specialized AMGs that contribute to depth-stratified host adaptation, such as those for high pressure deep-sea survival (Hurwitz et al. 2015). A large number of AMGs have also been identified that are associated with carbon metabolism in soil metagenomic analyses, including genes encoding glycoside hydrolases, endomannanases and chitosanases, suggesting the potential impact of viruses on carbon cycling in soil ecosystems (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019; Emerson 2019; Li et al. 2020). However, the adaptive significance of most soil viral AMGs is generally unclear.

1.2. Using metagenomics for studying viral communities in soil

1.2.1. Metagenomics

Research on the diversity and ecology of viruses in the ecosystems has been revolutionized by molecular based approaches (Mokili et al. 2012). Metagenomics can be described as the analysis of genomic DNA from environmental communities (Riesenfeld et al. 2004). The characterization of viral diversity in soil is challenging as most prokaryotic hosts are not cultivable and there are no universal marker genes common to all viral genomes, unlike for prokaryotic and eukaryotic communities, that allow for taxonomic discrimination (Edwards and Rohwer 2005). However, metagenomics can be used to describe the uncultured majority of organisms, and in combination with the development of bioinformatics tools, can be utilized to identify and provide genetic information about the viruses present in an environmental sample (Breitbart et al. 2002; Edwards and Rohwer 2005; Roux et al. 2015a; Ren et al. 2017). However, due to the complexity of soil which contains a vast microbial diversity, metagenomics typically only recovers complete or partial microbial genomes of the most abundant organisms given enough sequencing depth (Liolios et al. 2008; Mende et al. 2012). Therefore, generating samples enriched in viruses before performing metagenomic analysis (i.e. viromes) may facilitate more in-depth analysis of the viral

community composition. In order to produce soil viromes, viruses are typically extracted within a buffer, concentrated through filtering to remove large prokaryotic cells, and then precipitated (Trubl et al. 2016).

1.2.2. High throughput sequencing

High throughput sequencing (HTS) is massively parallel sequencing of millions of individual fragments of DNA (Tucker et al. 2009). Over the last two decades there have been many different HTS systems developed using a variety of different sequencing chemistries and approaches, with those of the various Illumina (e.g. MiSeq, HiSeq and NovaSeq) and Oxford Nanopore technologies commonly used today, and in this thesis.

Illumina sequencing is based on sequencing-by-synthesis, where single bases are detected via reversible dye terminators as they are incorporated into strands of DNA (Ambardar et al. 2016). In contrast, Oxford Nanopore sequencing directly detects the nucleotides without active DNA synthesis, and measures the change in electrical current of a nanopore as a single-stranded DNA strand is passed through (Branton et al. 2008; Feng et al. 2015). Sequencing systems vary in the number of sequences and the length of DNA fragments that can be processed. For example, the Illumina MiSeq and NovaSeq platforms can be used for sequencing amplicons or extracted genomic DNA, generating up to 15 GB or 6 TB data via 50 million or 20 billion paired-end reads, respectively. In contrast, the Oxford Nanopore platform is capable of producing long reads from a much-reduced number of DNA fragments. Nanopore sequencing has a relatively high error rate compared to short read sequencing, therefore the combination of both approaches where error correction of longer reads is performed using short read sequence data is commonly performed (Kono and Arakawa 2019).

1.2.3. Bioinformatic tools used in metagenomic analyses

The analysis of metagenomic sequencing data undergoes five basic steps: 1) quality trimming of output sequencing reads, 2) contig assembly, 3) binning, 4) functional and taxonomic annotation of the assembled contigs or bins and 5) quantification of bins or specific genes. In this section, each step is described, and the associated bioinformatic tools that were utilized throughout this thesis are presented.

1.2.3.1. Quality trimming

Sequenced reads contain adaptor sequences (added to DNA fragments to identify the sample) and the base sequence quality often decreases over the read. In addition, in paired-end sequencing, the quality of the second read can be low (Tan et al. 2018). As these errors in the sequence reads can compromise downstream analysis, it is necessary to filter the data to ensure any low-quality

reads and adaptor sequences are removed. The Phred scale is most commonly used to calculate quality scores (Ewing et al. 1998). A base call quality is assigned to each base call, which estimates the likelihood that the base call is incorrect. Generally, a median quality score greater than Q20, 1 in 100 chance of incorrect base call, is regarded as acceptable, and above Q30, 1 in 1000 chance of incorrect base call, is good. Two command line tools were used for quality trimming, the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and Trimmomatic (Bolger et al. 2014). The former has the advantage of providing QC metrics graphically but can result in a number of unpaired sequences after trimming unlike the latter which maintains read pairs (i.e. paired-end mode) and can efficiently find adapter sequences.

1.2.3.2. Contig assembly

Assembly involves the joining of short and long reads from the same genome into longer single contiguous sequences, called contigs. As *de-novo* assembly is not biased towards a reference genome, it is generally used to assemble novel genomes (Baker 2012). Three command line tools were used: MEGAHIT (Li et al. 2016), metaSPAdes (Nurk et al. 2017) and UniCycler (Wick et al. 2017). MEGAHIT is optimized for large and complex metagenomics reads, and therein the "meta-large" preset parameter was utilized. Like MEGAHIT, metaSPAdes constructs a De Bruijn graph (i.e. identification of sequence overlaps) of all reads, and transforms it into an assembly graph using several graph simplification procedures, and from the assembly graph reconstructs paths that correspond to long contigs within a metagenome (Bankevich et al. 2012; Nurk et al. 2013, 2017). MetaSPAdes is a commonly used assembler, for example, it is used in the Joint Genome Institute (JGI) bioinformatics pipeline. The efficiency of MEGAHIT and metaSPAdes assemblers was tested using virome data in Chapter II. UniCycler was utilized to produce a hybrid assembly using both Illumina reads and Nanopore long reads in Chapter III. Unicycler first produces an Illumina assembly graph using SPAdes, and then uses Nanopore long reads to build bridges that can resolve repeats in the genome, yielding a complete genome assembly.

1.2.3.3. Binning

Binning involves the clustering of contigs with similar sequence attributes, such as *k*-mer composition, codon usage and similar read coverages into the same bin (Uritskiy et al. 2018). In this thesis, the modules implemented in the MetaWRAP tool were largely used, as this tool performs all of the main procedures of metagenomic analysis (read-quality control, assembly, taxonomic profiling, binning, functional annotation, visualization and quantification). In addition, MetaWRAP uses three metagenomic binning software: MaxBin2, metaBAT2 and CONCOCT, and then performs a hybrid approach considering the three bin sets to produce a consolidated improved bin set. The CheckM tool implemented in MetaWRAP was used to assess the quality of

genomes recovered from the metagenomes (Parks et al. 2015). It provides robust estimates of the completion and contamination of each genome by using collected sets of genes that are ubiquitous and single-copy genes within a phylogenetic lineage. Bins can be defined as a high-quality metagenomic assembled genome (MAG), completeness > 90% and contamination < 5%, medium-quality MAG, completeness > 70% and contamination < 10%, and partial genome, contamination > 50% and contamination < 4%. However, it is important to note that using MAGs can lead to the loss of information, as only a relatively small proportion of reads are successfully assembled and binned in complex metagenome datasets (Maguire et al. 2020). Bins were visualize using the tool Anvi'o (Eren et al. 2015).

1.2.3.4. Functional and taxonomic annotation of assembled contigs or bins

For analyzing contigs, the basic local alignment search tool (BLAST) can be used to perform local alignments to find regions of local similarity between sequences. BLASTn and BLASTp can be used to compare query nucleotide or protein sequences to sequence databases, respectively, and provide the statistical significance of matches. A disadvantage of this method is in long processing times, and even with high computing servers, it can take numerous days to weeks. Alternatively, Diamond BLASTp can quickly align query protein sequences against a non-redundant (*nr*) sequence database (Madden et al. 1996; Buchfink et al. 2015). The National Center for Biotechnology Information (NCBI) hosts a non-redundant nucleotide (nt) and protein (nr) sequence collection that is commonly used as the reference database in BLAST searches. Additionally, the curated protein sequence database Swiss-Prot, which is part of the UniProt database collection, was utilized (Table 1.1).

Table 1.1. Information about the reference databases used within this thesis.

Database	Host	Type	Number of sequences	Length of sequences (bp)
nt	NCBI	Nucleotide	53,777,267	237,410,501,766
nr	NCBI	Protein	115,570,790	42,364,384,627
Refseq	NCBI	Protein	49,770,189	15,556,247,796
Swiss-Prot	UniProt	Protein	560,823	201,585,439
viruses	NCBI	Nucleotide	12,156	317,115,877
viruses	NCBI	Protein	315,213	79,514,978

To annotate the taxonomy and function of contigs, the tools Kaiju and InterProScan 5 are commonly used, respectively (Jones et al. 2014; Menzel et al. 2016). Kaiju is a program for taxonomic classification sequence reads or contigs from whole metagenomic sequencing (Menzel et al. 2016). Each read (or contig) is translated into an amino acid sequence, which is searched for in the reference database. The reference sequence (RefSeq) collection, hosted by NCBI, provides an inclusive, non-redundant well-annotated set of sequences, including proteins (Table 1.1). InterProScan 5 utilizes nucleotide and amino acid sequences for matches against a collection of

protein signature databases, including CATH-Gene3D, CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, SFLD and SMART (Jones et al. 2014). This tool provides protein families and domains, and the gene ontology and pathways (KEGG, MetaCyc and Reactome).

The taxonomic and functional analysis of MAGs is different from that of contigs. Taxonomic identification can be carried out using the genome taxonomy database tool kit (GTDB-Tk) (Chaumeil et al. 2020). This tool uses a set of 120 bacterial marker genes and 122 archaeal marker genes and the FastANI tool to estimate the average nucleotide identity (ANI) between a MAG and reference genome. If the ANI between the MAG and genome is > 95% and the alignment fraction is > 0.65, then the MAG is classified as belonging to a species. Gene prediction of MAGs can be performed with the prokaryotic dynamic programming gene finding algorithm (Prodigal) (Hyatt et al. 2010) and functional analysis with Diamond BLASTp, Kyoto encyclopedia of genes and genomes (KEGG) Mapper (Kanehisa and Goto 2000) and InterProScan 5 (Jones et al. 2014). The contig annotation tool (CAT) and bin annotation tool (BAT) are also commonly used for classifying the taxonomy of long DNA sequences and MAGs (Meijenfeldt et al. 2019).

1.2.3.5. Quantification of bins or specific genes

The process of aligning short reads to a reference sequence of a complete genome or *de-novo* assembly is called read mapping (Schatz et al. 2010). Many programs have been developed to map reads to a reference sequence, and which vary in their algorithm. Read mapping allows to define the mean coverage of each contig and provides input for binning through clustering algorithms that use contig coverage patterns across samples (Eren et al. 2015; Uritskiy et al. 2018). The resulting abundances from read mapping can be normalized by the coverage and the genome size. Here, the Salmon mapper implemented in the metaWRAP tool (Patro et al. 2017) and the Bowties2 tool implemented in Anvi'o were utilized (Langdon 2015).

1.2.4. Virus sequence analyses

As viral sequences can have higher homology to prokaryotic or eukaryotic genes (Breitbart et al. 2002, 2003; Angly et al. 2006; Schoenfeld et al. 2008; Blomström et al. 2010), when processing viromes it is important to remove microbial contaminated sequences by mapping to prokaryotic or eukaryotic reference genomes. After assembly of metagenomic or viromic data, viral contigs can be predicted through either a reference-based or reference-free approach. Predicted viral contigs can then be classified into viral proteomic trees based on genome-wide similarity, and different approaches can be used to try and link viruses to their hosts, such as CRISPR arrays, genomic similarity and gene homology (Sanguino et al. 2015; Ahlgren et al. 2017; Galiez et al. 2017).

1.2.4.1. Virus prediction

Viral genomic sequences can be predicted with gene-based similarity (e.g. VirSorter) and *k*-mer frequency-based approaches (e.g. VirFinder and DeepVirFinder). The gene-based similarity approach uses hidden Markov models (HMM) based on gene annotation to recognize coding regions that are homologous to viral origin hallmark genes. The *k*-mer frequency-based approach is useful for short contigs with few predicted proteins or when proteins do not have similarity to known viruses. Within this thesis, both approaches were utilized, implementing VirSorter and DeepVirFinder.

VirSorter is the most recent tool developed to predict viral signals, including prophages and lytic viruses (Roux et al. 2015a). VirSorter identifies viral sequences enriched in genes with similarity to virome databases, and are predicted into one of three categories depending on the presence or absence of viral origin hallmark genes annotated as major capsid protein, portal, terminase large subunit, spike, tail, virion formation or coat, and viral-like genes based on the viromes databases and deletion of protein family (Pfam)-affiliated genes. Category 1, "most confident" predictions, are sequences that contain either enrichment in viral-like or non-*Caudovirales* genes and at least one hallmark viral gene detected. Category 2, "likely" predictions, are sequences that contain either enrichment in viral-like or non-*Caudovirales* genes or a viral hallmark gene and have at least one of the following metrics: depletion in Pfam affiliation, enrichment in uncharacterized genes, enrichment in short genes or depletions in strand switch. Category 3, "possible" predictions, are sequences that have neither a viral hallmark gene nor enrichment in viral-like or non-*Caudovirales* genes but have at least two of the aforementioned metrics, but at least one requires having a considerable significance score. Also, a contig is determined viral if a predicted sequence has more than 80% of the predicted genes on a contig, (Roux et al. 2015a). Identified prophage are classified in same way as viruses, into category 4 ("most confident"), 5 ("likely") and 6 ("possible") prophages (Roux et al. 2015a).

Conversely, DeepVirFinder is a reference-free and alignment-free approach for detecting viral sequences in metagenomic data based on deep machine learning methods. DeepVirFinder was trained based on a large number of viral reference genomes, and learned a convolutional neural network that accurately identifies viral sequences at all contig lengths (Ren et al. 2017, 2018b). The use of these tools has allowed for the discovery of a large number of new viruses from prokaryotic metagenomes leading to marked advances in our knowledge of virus-host interactions.

1.2.4.2. Viral populations

Classifying viruses is challenging due to the absence of universal marker genes (Edwards and Rohwer 2005). Although, recent developments in sequencing technologies have enabled the

increase of large viral genome fragments making viral classification easier (Paez-Espino et al. 2016; Roux et al. 2016). Several approaches for classifying viruses have been proposed, based on genomic, proteomic and specific gene-based comparative strategies (Quan et al. 2016). Generally, viral genomes or sequences (> 10 kb) with a similarity threshold of 95% nucleotide sequence identity are classified at the species rank (Quan et al. 2016; Roux et al., 2019). These populations have been suggested to be considered as viral population units (Hurwitz et al. 2015). Similarity of protein-encoding genes within viral genomes, and which are used to generate viral proteomic trees, has demonstrated that this method can be used as the foundation for a genome-based taxonomical system (Nishimura et al. 2017). This method allows for monitoring virus biodiversity by grouping viruses into taxa that predicts several aspects of virus biology (Rohwer and Edwards 2002). However, this approach is limited when determination of BLASTp based similarity is faced by distantly related protein sequences (Chibani et al. 2019).

In this thesis, the viral proteomic tree server (VipTree) for classification of viral contigs based on genome-wide similarity was used (Nishimura et al. 2017). The genomes of host-associated viruses are compared with reference viral genomes stored in the Virus-Host database (DB), which contains a total of 2687 prokaryote-associated dsDNA viruses and 1119 eukaryote-associated dsDNA viruses (Mihara et al. 2016). All-against-all similarity scores are computed from the results of tBLASTx. Additionally, the viral contig automatic clustering and taxonomy (vConTACT) tool was used with NCBI's RefSeq database (Bolduc et al. 2017; Bin Jang et al. 2019). This is a genome-based network analysis of the shared viral protein content with a reference database. Prokaryotic virus taxonomy is inferred by clustering viral protein sequences through all-to-all BLASTp searches (E value < 10^{-4} and bit score > 50). Based on the number of shared protein clusters between viral genomes and reference genomes, a similarity score for each pair is calculated based on a generated *p*-value by the total number of pairwise genome comparisons. The resulting network with genome pairs (similarity score > 1) was visualized with Cytoscape software (source: http://cytoscape.org/).

1.2.4.3. Virus-host linkage

The most confident approach for linking viruses to their hosts is by CRISPR array analysis. CRISPR arrays are composed of direct repeats (DRs), which are a succession of 24 – 47 bp of microbial origin, interspaced by invader-derived spacers (S) (Figure 1.2) (Mojica et al. 2009). After insertion of the spacer in the AT rich leader end of the CRISPR, spacer sequences are transcribed and maturated into small interfering RNAs (crRNAs) providing immunity to the phage (Jansen et al. 2002; Terns and Terns 2011). The analysis of CRISPR arrays present in host genomes and spacer sequences to identify viral genome fragments allows for hosts and their associated viruses to be linked (Andersson and Banfield 2008; Mojica et al. 2009). There are several software programs

developed that use different algorithms in order to search for CRISPR sequences in metagenomic data. In this thesis, three different tools were used: the CRISPR recognition tool (CRT) that searches for a series of short exact repeats (Bland et al. 2007), the CRISPR assembler (Crass) that uses k-mer searches (Skennerton et al. 2013) and CRISPRCasFinder that detects *Cas* genes (Couvin et al. 2018).

As CRISPR arrays are often not found within host contigs, alternative approaches have been developed to link viruses with hosts, such as oligonucleotide frequency (ONF), shared genomic regions, host-virus abundance correlation and tRNA matches. In this thesis, an ONF approach was used with 'who is the host' (WIsH) tool (Galiez et al. 2017). The potential host of a query virus can be predicted by identifying the host to which it has the greatest similarity of *k*-mer frequencies. WIsH was developed by adopting a suited probabilistic approach (Galiez et al. 2017). A homogeneous Markov model of order *k* (*k* = 8, as the accuracy is maximal for order 8) can be trained for each potential host contig. The likelihood of a viral contig under each of the trained Markov models are then computed, and the host whose model yields the highest likelihood is *de novo* predicted. In addition, *p*-values are computed using the parameters of the Gaussian null-distributions of each Markov model. Unlike other *k*-mer based approaches this tool can be used with short contigs (< 3 kb). In combination with ONF analysis, gene homology between predicted hosts and viruses was assessed using the various tools described in section 1.2.3.4.

1.2.5. High performance computing

Bioinformatics often require high hardware requirements, such as multicore servers and also high levels of expertise and 'know-how'. Using a graphical window program for interacting with the shell (i.e. a command-line interface with the operating system), Terminal, for MacOS, or PuTTY, for Windows or Linux, a computer can be connected to a server through the secure shell (SSH) client program. In the studies within this thesis, the bioinformatics was performed using the MacOS operating system and two high performance computing clusters (i.e. servers), NEWTON and MUSCLOR. The NEWTON server has a high number of central processing units (CPUs) and was used for analyzing the larger datasets, and when the tools required for analyses required high hardware equipment (e.g. *de-novo* assembly and > 200 GB RAM). NEWTON is composed of 84 nodes of 1,704 CPUs and 12 TB of RAM. Nodes are separated by several networks depending on CPU and RAM needs. MUSCLOR was used to analyze the smaller datasets and is composed of 24 CPUs and 40 GB of RAM. Using Terminal, shell command lines were used to execute various tasks, such as the installation of tools. The tools were installed within environments created using the Python command-line based program Anaconda (source: https://www.anaconda.com/). For the

efficient handling of FASTA and FASTQ sequence files (converting, searching, filtering, deduplicating, splitting, etc.), the Seqkit tool was installed and utilized (Shen et al. 2016).

1.3. The model soil pH gradient

The use of environmental gradients, such as temperature, nutrient, salinity, moisture, oxygen and pH for addressing various questions about the distribution, interactions and assembly of microbial communities is long standing in the field of microbial ecology and has led to a greater understanding of microbial dynamics. Soil pH has repeatedly been shown to influence microbial diversity and composition, with bacterial diversity greater in neutral than in acidic soils (Lauber et al. 2009; Rousk et al. 2010; Zhalnina et al. 2014; Ren et al. 2018a). As such, a well-established and studied soil pH gradient was utilized in this thesis for investigating the impact of both prokaryotic community structure and soil pH on viruses and their interactions with different hosts. The Scottish Rural College (SRUC) Craibstone campus, located in Aberdeen, Scotland, UK (57°11, 2°12) has a number of long-established agricultural experimental systems, one of which is a series of eight parallel plots in which each plot has pH gradient ranging from 4.5 to 7.5 at 0.5 pH unit intervals. The plots are under an 8-year crop rotation (winter wheat, potatoes, spring barley, swedes, spring oat, 3 years of grass with no re-sowing), and all receive moderate NPK fertilization before sowing, except for years 2 and 3 of the grass rotation, and has been maintained for over 50 years (Figure 1.3). The pH of each sub-plot is controlled by the addition of either lime or aluminum sulfate (Bartram et al. 2014). As the crop plant is rotated on an annual basis, the effect of one plant species on the microbial community is not a major driver of community structure. Total carbon, total nitrogen and organic matter do not change significantly across the gradient (Nicol et al. 2008). As crop yield, health and quality are underpinned by the maintenance of soil agricultural ecosystem function driven by soil microorganisms, studying the viruses that control microbial host abundances is of agronomic interest. For example, several studies have been conducted to identify the viruses associated with *Rhizobium* sp. that can enhance crop productivity (Kimura et al. 2008; Santamaría et al. 2014).

Previous studies conducted across the soil pH gradient at Craibstone have demonstrated that soil pH influenced microbial diversity and community composition (Nicol et al. 2008; Bartram et al. 2014). For example, based on 16S rRNA and *amoA* genes the structure of ammonia oxidizing bacteria and archaea have been demonstrated to change with soil pH, with distinct nitrifier populations in the acidic and neutral soil (Stephen et al. 1998; Nicol et al. 2008; Gubry-Rangin et al. 2011). In particular, *Acidobacteria* increased in the high pH soil, whereas *Actinobacteria* were proportionately greater in the low-pH soil. The genera *Burkholderia* and *Paucibacter* (*Betaproteobacteria*) were only found within the low-pH soils, whereas the following genera were high-pH associated: *Nitrosospira*, *Denitratisoma*, *Paucimonas*, *Herbaspirillum*, *Tepidimonas* and

Polaromonas (*Betaproteobacteria*). Within the *Alphaproteobacteria*, the genus *Phenylobacterium* was associated with low-pH soil, whereas *Devosia*, *Roseomonas*, *Labrys*, *Methylosinus*, *Fulvimarina*, *Filomicrobium*, *Rhodobacter*, *Hyphomicrobium*, *Bartonella* and *Mesorhizobium* were only within the high-pH soil. Within *Gammaproteobacteria*, the genera *Dyella* and *Rhodanobacter* were only found within the low-pH soil, whereas the genus *Lysobacter* was only observed within the high and medium-pH soils (Bartram et al. 2014). As such, this soil pH gradient is an excellent model environment for investigating virus-host interactions in different soil pH niches.

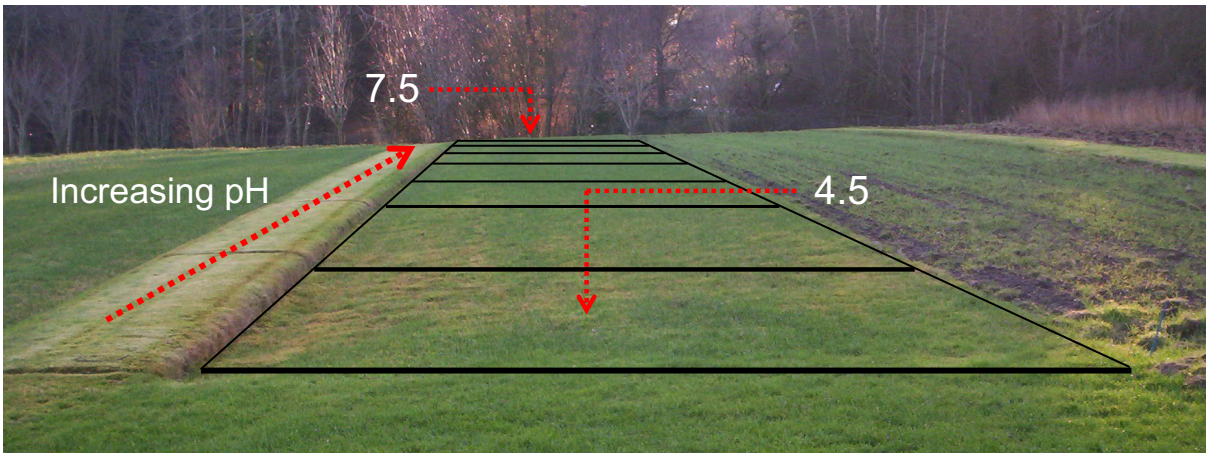


Figure 1.3. The pH-controlled plots sampled at the Scottish Agricultural College, Craibstone, Scotland.

1.4. Overview of research aims

In this thesis, four research studies that utilized the soils of the pH gradient (Chapter II – V) are presented and concludes with an overall discussion on soil viruses in the context of the findings of the thesis (Chapter VI). Viral diversity and host-virus interactions across soil pH have not yet been sufficiently described. Therein, the research study presented in Chapter II aims to determine the influence of microbial community structure and soil pH on viruses using metagenomics and viromics. In contrast to marine environments, a general understanding of the extent to which virus-prokaryotic host interactions regulate prokaryotic populations in soil is lacking. While some viruses will have the ability to infect a range of hosts in highly diverse prokaryotic soil communities, coevolutionary processes within ecological niches may tightly control the susceptibility of hosts. To gain a better understanding of the extent to which virus-prokaryotic host interactions regulate soil prokaryotic populations, Chapter III presents a study that assesses the infectivity of a host bacterium to co-localized and increasingly allochthonous sources of virus populations isolated from across the soil pH gradient using a culture based plaque assay approach, followed by electron transmission microscopy and hybrid metagenomic sequencing to determine viral diversity and virus-bacterial host interactions.

Lastly, two studies are presented that aim to identify virus populations infecting specific soil microbial functional groups in contrasting pH soil, specifically methanotrophs (Chapter IV) and nitrifiers (Chapter V), using DNA stable isotope probing (SIP) combined with metagenomic deep sequencing. Currently, there are limited studies that have determined active relationships between soil microbial communities and their associated viruses *in situ*. SIP involves the assimilation of substrates enriched in a heavy isotope into cellular microbial biomass of environmental samples (Radajewski et al. 2000). Molecular analysis of isotopically labeled DNA provides phylogenetic and functional information about the active microorganisms responsible for the metabolism of a substrate (Radajewski et al. 2000; Chen et al. 2008; Prosser and Nicol 2012). As viruses are obligate intracellular parasites using genetic material of their host cell, consequently, the associated viruses of the active microorganisms will also be isotopically labeled (Gelderblom 1996; Lee et al. 2012). By combining metagenomic deep sequencing with DNA-SIP and determining host-virus linkages through CRISPR array and ONF analysis, the active virus-host interactions in a complex soil system were investigated and the impact of viruses on microbial communities assessed by identifying the host genes that were potentially transferred among their associated viruses (Chapter IV and V).

Investigating the impact of viruses on soil methanotroph and nitrifier communities also has important ecological ramifications. Soil nitrifiers and methanotrophs play a crucial role in the production and consumption of greenhouse gasses by oxidizing ammonium and producing N_2O (Prosser et al. 2020) and consuming methane (Dedysh and Knief 2018), respectively. Carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) are relevant greenhouse gasses that contribute to climate change (IPCC 2013). Human activities, such as the management of agricultural soil, have stimulated the production of these greenhouse gasses by soil microbes (Canfield et al. 2010). Methane is the second most abundant greenhouse gas after CO_2, accounting for an estimated 20% of global warming (Tate 2015; Nisbet et al. 2016). Aerobic methane-oxidizing bacteria are ubiquitous in soil, and utilize CH_4 as their sole energy and carbon source (Dedysh and Knief 2018). Ammonia oxidation by nitrifiers is coupled with autotrophic CO_2 fixation. Ammonia oxidation itself generates N_2O and produces the substrate nitrate that is subsequently used to produce N_2O through denitrification (Nicol et al. 2008; Prosser and Nicol 2012). Therefore, an understanding of the controls and influences on soil methanotroph and nitrifier communities is crucial for contributing to efforts in the mitigation of climate change.

CHAPTER II

Prokaryotic and viral community structure, functional diversity and host-virus interactions in contrasting pH soils

2.1. Abstract

Microorganisms, together with their associated viruses, are involved in a range of processes that contribute to biogeochemical cycling and the altering of the evolution and ecology of the soil microbiome. While microbes and viruses are key members of soil communities, viral-host dynamics in soil are not well understood. Abiotic factors, particularly soil pH, can strongly affect microbial community structure in soil, but how this impacts viral distribution, diversity, function and host-virus interactions is unknown. This work aimed to test the overarching hypothesis that, compared to prokaryotic hosts, viral communities are less distinct between two samples of contrasting pH due to viral host ranges being greater than the constraints of pH on prokaryotic community structure. Soil collected from the extreme ends of a well-studied pH-manipulated soil system was utilized to: 1) compare metagenomes (i.e. the untargeted whole community) and viromes for the analysis of soil viruses, 2) determine the influence of soil pH and microbial host distribution on viral diversity and community structure, 3) characterize the potential microbial and viral functional differences between soil pH and 4) identify virus-host relationships and the potential impact of these viruses on their hosts' functioning. DNA extracted from pH 4.5 and 7.5 soils and soil viral filtrates were sequenced on an Illumina NovaSeq platform. Sequencing yielded in 8,061 assembled metagenomic contigs (MCs) and 8,867 assembled viromic contigs (VCs) greater than 10 kb. Virus prediction between metagenomes and viromes revealed size selection methodology biases. Both microbial and viral community structure was influenced by soil pH, but contrary to our hypothesis, viral community structure changed more than that of prokaryotes between the two pH. Widespread auxiliary metabolic genes (AMGs) and specific "core" AMGs to the pH soils were found including those encoding for glycoside hydrolases and peptidases, which are involved in carbon cycling. A greater number and a broader size range of clustered regularly interspaced short palindromic repeats (CRISPR) arrays were detected in the pH 4.5 soil, suggesting higher viral infection frequencies in the pH 4.5 soil. Results support that viruses play an important role in shaping the composition and function of soil prokaryotic communities and that they are significant contributors to the carbon cycle. Viral core AMGs may serve as an adaptive mechanism for microbes in different pH soil niches.

2.2. Introduction

Soils contain a vast diversity of inhabitants. For example, in one gram of soil, there are typically up to 10^9 – 10^{10} microorganisms, including eukaryotes, prokaryotes and viruses. Soil microorganisms are important for maintaining soil quality and fertility in both natural and agricultural ecosystems and have key roles in biogeochemical cycles (Hill et al. 2000; Prosser and Nicol 2008; Chaparro et al. 2012; Aislabie and Deslippe 2013). Microbial diversity and community structure regulate ecological functions, and as such, it is essential to understand the impact of

abiotic and biotic factors on soil microbial communities (Baumann et al. 2013; Vivant et al. 2013; Philippot et al. 2013). Soil pH has repeatedly been shown to influence microbial diversity and community structure. Amongst the abiotic factors measured within studies, soil pH is often found to have had the greatest effect on microbial diversity and composition, with bacterial diversity greater in neutral than in acidic soils (Lauber et al. 2009; Rousk et al. 2010; Zhalnina et al. 2014; Ren et al. 2018a). Soil pH can directly drive prokaryotic community structure through affecting the taxa that have a narrow range in pH tolerance, and indirectly through a change in mineral nutrient availability and ion toxicity (Moser and Weisse 2011; Zhalnina et al. 2014; Lammel et al. 2018). Biotic soil interactions, such as predation and competition, have also been shown to influence microbial community structure (Feng et al. 2017; Fernández et al. 2018; Karakoç et al. 2018). For example, predation pressure on host cells from viruses can contribute to maintaining microbial communities (Buckling and Rainey 2002; Rodriguez-Valera et al. 2009; Koskella and Brockhurst 2014).

The influence of abiotic and biotic factors on soil viral communities is less understood in comparison to that of host communities. Soil pH has been found to be a strong predictor of viral community structure, with viral abundance negatively correlated with soil pH (Narr et al. 2017; Williamson et al. 2017; Adriaenssens et al. 2017). The attachment of viruses to the soil surface can be influenced by soil pH, and this may affect the persistence of viruses in soils (Lance and Gerba 1984; Loveland et al. 1996; Williamson et al. 2017). However, as viruses are strictly dependent on their hosts for reproduction, the distribution of host populations is more likely the dominant factor predicting viral community structure. Based on phages that have been tested by infection assays, the host range of phages spans a wide continuum, from extremely narrow to broad (Moebus and Nattkemper 1981). However, it has been argued that infection assays are biased and that most of the isolated bacteriophages used in previous experiments have been from a single host strain (Hyman and Abedon 2010; Ross et al. 2016). Though, metagenomic based studies have demonstrated that viruses can exhibit broad-host ranges, and could possibly be prevalent in soils (Adriaenssens et al. 2017). In marine environments, it is thought that viruses with a narrow host-range are prevalent when their hosts are abundant, whereas broad host-range viruses are prevalent when the abundance of different host cells are low or variable (Woolhouse et al. 2001; Sullivan et al. 2003; Elena et al. 2009; Dekel-Bird et al. 2015; Doron et al. 2016). Comparatively, studies on the host-range of viruses in soils are lacking.

Microbial abundance, diversity and activity in environments are tempered by virus-host interactions (Fuhrman 1999; Suttle 2005; Rodriguez-Valera et al. 2009; Clokie et al. 2011). In nature, it has been estimated that there are ten viruses for every microbial host cell (Watt et al. 2006; Emerson et al. 2018). Through the lytic cycle, hosts are infected, allowing for viral reproduction followed by the lysis of host cells to release the new viral particles. Due to the death

of host cells and the modulation of host cell metabolisms during infection, viruses greatly influence global biogeochemical cycles (Fuhrman 1999; Weinbauer and Rassoulzadegan 2004; Suttle 2005; Breitbart et al. 2007). Following viral lysis, the bioavailability of various compounds, such as carbon, nitrogen, phosphorus, sulfur and iron can enhance prokaryotic heterotrophic metabolism and nutrient turnover (Middelboe et al. 1996; Fuhrman 1999; Wilhelm and Suttle 1999; Danovaro et al. 2016). Host-derived genes carried by viruses, referred to as Auxiliary metabolic genes (AMGs), can be acquired from their immediate host or more distantly-related hosts (Sharon et al. 2009; Sullivan et al. 2010; Kelly et al. 2013; Crummett et al. 2016). Recent soil virome studies have reported a large number of AMGs, and have identified AMGs that encode for enzymes involved in carbon metabolism, such as glycoside hydrolases (GH), endomannanase and chitosanase, suggesting viruses can impact soil carbon cycling (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019).

To capture viral communities in environmental samples, metagenomics has been widely applied (Riesenfeld et al. 2004; Edwards and Rohwer 2005). Due to the vast microbial diversity in soil ecosystems, only relatively few viral genomes (mostly partial), in comparison to prokaryotic genomes, have been recovered from soil metagenomes (Mavromatis et al. 2007; Mende et al. 2012; Schulz et al. 2018). However, the lack in viral genomes may be overcome with viromics, the sequencing of virus enriched genomic DNA extracted after a filtering step to reduce microbial complexity prior to DNA extraction (Breitbart et al. 2002; Edwards and Rohwer 2005). Although viromes may better reflect the actual viral community composition, some viruses may be neglected, such as large viruses (García-López et al.; Halary et al. 2016). Additionally, viruses cannot be linked to their hosts using viromics. Therefore, combining metagenomics and viromics with deep sequencing is likely the most advantageous for investigations of soil viruses and hosts, and their interactions.

To gain a better understanding of the influence of soil pH and the microbial community in structuring the viral community, and virus-host interactions in contrasting soil pH, metagenomes and viromes derived from soil collected at either ends of a continuous soil pH gradient (pH 4.5 and 7.5) were analyzed. A soil pH gradient model system in which both bacterial and archaeal community structures have been previously demonstrated to be strongly influenced by soil pH was utilized (Nicol et al. 2008; Bartram et al. 2014). It was hypothesized that soil viral community structure would be influenced by soil pH and microbial community structure, but that viral community structure would change less than compared to prokaryotic hosts, potentially due to viruses having host ranges greater than the constraint of host community structure imposed by soil pH. The overall objectives of this study were to: 1) compare between metagenomics and viromics for the recovery of soil viruses, 2) determine the influence of soil pH and microbial host distribution on viral diversity and community structure, 3) characterize the potential microbial

and viral functional differences between soil pH and 4) identify virus-host relationships and the potential impact of these viruses on their hosts' functioning.

2.3. Materials and Methods

2.3.1. Soil sampling and physicochemical analyses

Soil was collected from the Craibstone Research Station, SRUC, Aberdeen, Scotland (57°11, 2°12) (see Chapter I, section 1.3.). Three replicate surface soil samples (top 10 cm) from the pH 4.5 and 7.5 plot were collected randomly on 11 January 2019. Soil samples were sieved (2 mm) and stored at 4°C. Soil pH was measured by mixing 5 g of wet soil and 10 ml of deionized water with a rotator, centrifuged at 25°C at 1,000 x *g* for 10 min, and measured with a pH meter (Multi-parameter analyser C532, CONSORT, Turnhout, Belgium). Mean soil pH for the pH 4.5 and 7.5 plots were 4.22 (0.06 ±SD) and 7.34 (0.04 ±SD) respectively. To determine soil moisture content, 3 g of wet soil was dried at 120°C and the difference between wet and dry weight was calculated. Mean soil moisture content for pH 4.5 and 7.5 plots were 17.6 (0.38 ±SD) and 9.5 (0.33 ±SD).

2.3.2. Virus isolation from soil samples

Isolation of viruses was performed within one week of soil collection. Prior to use, buffers for viral extraction were filtered at 0.02 μm to remove viruses and prokaryotic and eukaryotic cells. For each pH 4.5 and 7.5 replicate soil sample, 10 g of soil was weighed in a 50 ml sterile centrifuge tube, and 20 ml of amended 1% potassium citrate buffer (APC) was added (1 g of potassium citrate, 0.144 g of sodium phosphate dibasic heptahydrate, 0.024 g of potassium phosphate, 10% of phosphate buffered-saline, 5 mM of ethylenediaminetetraacetic acid (EDTA), 150 mM of magnesium sulfate). To release the viral particles from the soil, samples were vigorously mixed for 15 min using a vortex (Vortex-Genie 2 mixer, Sigma Aldrich, St. Louis, MO, USA), followed by 4 min of sonication at 20% amplitude with 1 min of sonication followed by 30 sec pause (Vibra-cell, Sonics, Newtown, CT, USA) . The samples were centrifuged for 10 min at 4°C at 10,000 x *g* to pellet the debris, and the supernatant was transferred into a new 15 ml sterile centrifuge tube. Based on previous evidence that sequential re-extraction of the initial soil maximizes virus recovery (Trubl et al. 2016), the pelleted debris was re-suspended in 20 ml of APC buffer, followed by vortexing and sonication. The two sequential filtrates from each sample were respectively combined into a 50 ml tube. The samples were centrifuged for 20 min at 4°C at 10,000 x *g*. To remove prokaryotic and eukaryotic cells from the viral particles, the supernatant was passed through 0.45 and 0.2 μm syringe filters. Viral particles were concentrated using the polyethylene glycerol (PEG) precipitation method. Viral particles were precipitated by adding 10% (w/v) solid PEG 8000 (Alfa Aesar, Carlsbad, CA, USA) and 0.6% (w/v) sodium chloride (NaCl), and incubated overnight at 4°C. Precipitates were centrifuged for 45 min at 4°C at 10,000 x *g*. The pellet was resuspended with

400 µl of Tris-EDTA (TE) buffer (Sigma Aldrich, St. Louis, MO, USA) in 2 ml tubes, and treated with DNase I (RQ1 DNase; Promega, Madison, WI, USA) to remove any free DNA.

2.3.3. DNA extraction and sequencing

Viral particles from the filtrates were lysed by adding sodium dodecyl sulfate (SDS) (0.5% final concentration) and proteinase K (100 µg ml^{-1} final concentration) followed by incubation for 1 h at 37°C. To each sample, 100 µl of 5% cetyltrimethylammonium bromide (CTAB) in 120 mM phosphate buffer (2.58 g K_2HPO_4 $3H_2O$, 0.1 g KH_2PO_4, 5 g CTAB, 2.05 g NaCl, water to 100 ml) was added and samples were incubated for 10 min at 65°C (Dry bath system, STAR Lab, Orsay, France). An equal volume of 25:24:1 phenol/chloroform/isoamyl alcohol (PCI) was added to the lysate and vortexed. All samples were centrifuged for 2 min at 4°C at 16,000 x *g*. The supernatants were transferred to 1.5 ml tubes, and an equal volume of 24:1 chloroform/isoamyl alcohol (CI) was added. Samples were vortexed and centrifuged for 2 min at 4°C at 16,000 x *g*. The supernatant was transferred to a 1.5 ml tube, and 2 volumes of PEG/NaCl solution (30% of PEG, 1.4 mM of NaCl) was added. The samples were incubated for 2 h at 4°C and then centrifuged for 10 min at 4°C at 16,000 x *g*. Pellets were resuspended in 100 µl of TE buffer in 1.5 ml tube and RNase A treated (Thermo Fisher, Carlsbad, CA, USA). The steps involving the addition of PCI, CI, and PEG precipitation were repeated twice to reduce humic contaminants from the soil. Afterwards, samples were centrifuged for 10 min at 4°C at 16,000 x *g*, and the PEG solution was removed, and pellets washed by adding 1 ml of 70% ice-cold ethanol and dried at room temperature. Pellets were resuspended in 100 µl of TE buffer in 1.5 ml tube, and then bead purified (Promega, Madison, WI, USA) and eluted with 50 µl of TE. The quality of viral DNA was assessed with 1.5% agarose gel electrophoresis and quantified using Qubit dsDNA HS Assay kit (Thermo Fisher, Carlsbad, CA, USA). The presence of bacterial DNA in the samples was checked by amplifying the 16S rRNA gene using the PA/PH primer pair (Edward et al., 1989). Each 50 µl PCR reaction contained 10X buffer, 0.4 µM of each primer, 10 µl of 10 mM dNTP, 0.2 µl of Taq polymerase and 1 µl of viral DNA and PCR grade water for the negative control (Thermo Fisher, Carlsbad, CA, USA). The thermal cycling program consisted of an initial denaturation step for 3 min at 95°C, followed by 30 cycles of amplification (30 s at 95°C, 30 s at 55°C, 1 min at 72°C) on a PCR cycler (Biometra, Göttingen, Germany).

For extracting total microbial DNA, soil (0.50 g) was added to a 2-ml microcentrifuge tube containing 0.5 g Lysing Matrix B (MP Bio, Illkirch-Graffenstaden, France) with 0.5 ml phenol-chloroform-isoamyl alcohol (25:24:1) (v/v/v) and 0.5 ml 5% CTAB in 120 mM phosphate buffer (Nicol et al., 2008). Cells were lysed by bead beating for 30 s at 4 m s^{-1}. After centrifugation at 16,000 x *g* for 5 min at 5°C, the nucleic acid containing aqueous supernatant was transferred into a new 1.7 ml microcentrifuge tube, and further extracted with an equal volume of chloroform-

isoamyl alcohol (24:1 v/v). After further centrifugation at 16,000 x *g* for 5 min at 5°C, the supernatant was removed and added to two volumes of 30% PEG8000 (w/v) in 1.6 mM NaCl and incubated overnight at 5°C. Samples were centrifuged for 16,000 x *g* for 20 min at 5°C before removing the PEG NaCl solution and washing pellet with 1 ml 70% (v/v) ice-cold ethanol. After centrifugation at 16,000 x *g* for 5 min at 5°C, the pellet was dried at room temperature. DNA was eluted in 50 µl of sterile molecular grade water. DNA was extracted from three 0.5 g soil samples for each microcosm and pooled in order to obtain a total of ~10 µg of total genomic DNA. The quality and quantity of DNA were determined by agarose gel electrophoresis (1.5%) and Qubit dsDNA BR Assay kit, respectively.

2.3.4. Bioinformatic analyses of metagenomes and viromes

An overview of the bioinformatics workflow is presented in Figure 2.1. Briefly, the quality-controlled reads from each metagenome and virome were assembled into contigs and taxonomically and functionally annotated via the JGI pipeline. These assemblies were used in community analyses. The assembled metagenomic contigs (> 300 bp) were used for virus prediction and co-assembled for downstream analyses. Contigs from the viromes were co-assembled and subjected to virus prediction (> 300 bp) and compared to the metagenomic predicted viruses. To match hosts with viruses, co-assembled contigs (> 10 kb) and virus contigs (> 10 kb) were used for clustered regularly interspaced short palindromic repeats (CRISPR) array and oligo nucleotide frequency (ONF) analysis. Gene homology between host-linked viruses and their associated hosts was analyzed through protein alignment. To confirm the potential host-virus linkages, gene homology of the host-linked virus was assessed. Functional and genome-based network analyses were also performed.

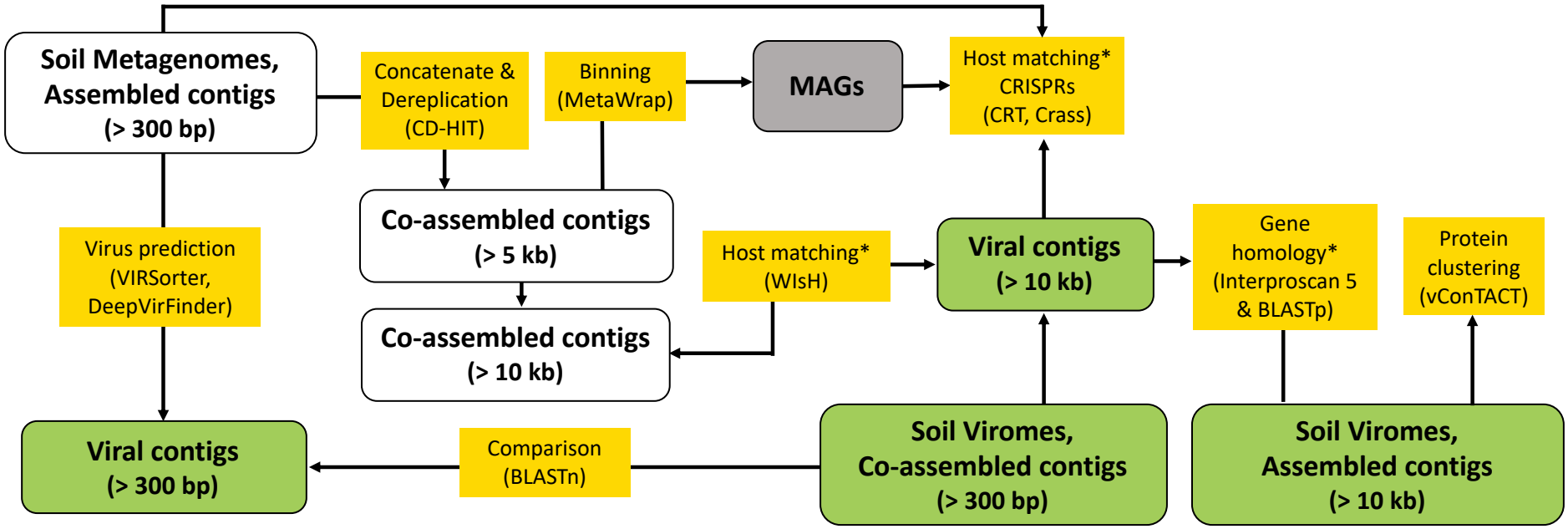


Figure 2.1. Schematic overview of the bioinformatics workflow. Square boxes in white, gray and green represent the host assembled contigs and metagenome-assembled genomes and viral contigs, respectively. Square boxes in yellow represent the bioinformatics performed and the key tools utilized are in parentheses. *Gene homology analyses between host MAGs or contigs and host-associated metagenomic viral contigs (mVCs) were realized (AMGs, axillary metabolic genes).

2.3.4.1. Sequence quality filtering, contig assembly and annotation

The JGI bioinformatics pipeline was utilized for sequence quality filtering and contig assembly for metagenomes and viromes. The tool BBDuk v38.51 (JGI's BBTools Team) was used to remove known sequencing artifacts, contaminants and low-quality sequences (quality score of 20 and minimum read length of 51 bp). To remove any contaminants (human, cat, dog and mouse, and prokaryotes and eukaryotes from viromes), reads were subjected to mapping with BBMap v38.34.1 (JGI's BBTools Team). The quality-controlled reads of each sample were *de novo* assembled into contigs using MetaSPAdes v3.13.0.2 (Nurk et al. 2017). The input read set was mapped to the final assembly, and coverage information was generated with BBmap (JGI's BBTools Team). Taxonomic profiling, using the nr protein database, and functional profiling, using the KEGG, fPham and COG databases, were carried out on the assembled metagenomic and viromic contigs through the DOE-JGI Metagenome Annotation Pipeline v5.0.3 (Huntemann et al. 2016).

Contigs of the metagenomic assemblies were concatenated as one co-assembled sample, and sequence redundancy was reduced using the psi-cd-hit-DNA tool with 95% identity (Fu et al. 2012). Contig names were simplified using the anvi-script-reformat-fasta from anvio 5 (Eren et al. 2015). Binning of metagenomic contigs (MCs) into metagenomic assembled genomes (MAGs) was performed (see Chapter IV, section 4.3.5.2.), but this only resulted in eight MAGs with < 68% completeness and one MAG with 97% completeness (Supplementary Table 2.1), and thus were not used in further analyses.

For the viromes, assembly and co-assembly using MEGAHIT (Li et al. 2016) was compared to the assemblies produced by MetaSPAdes (Li et al. 2016; Nurk et al. 2017). As co-assembly with MEGAHIT resulted in the greatest number of unique viral contigs (VCs) (see Table 2.2), these co-assemblies were used in down stream analyses (i.e. diversity, community structure and AMG analyses). The MEGAHIT contigs (VCs > 10 kb) were taxonomically annotated using the Kaiju tool with the NCBI viral proteins database and visualized using Krona (Ondov et al. 2011), and functionally annotated using InterProScan 5 with E value < 10^{-5} (Jones et al. 2014; Menzel et al. 2016).

2.3.4.2. Comparison of viral recovery between metagenomes and viromes

The VirSorter tool was used to predict viruses from the MCs and VCs that were > 10 kb (Roux et al. 2015). The tool DeepVirFinder was used to predicted viruses from MCs and VCs that were > 300 bp (Ren et al. 2018b). No metagenomic viral contigs (mVCs) were predicted using VirSorter. As such, to compare the recovery of soil viruses between the metagenomes and viromes the DeepVirFinder predicted metagenomic viral contigs (mVCs) and virome viral contigs (VCs) were used. Two virus databases were created from the viromes, VCs ≥ 10 kb and VCs ≥ 300 bp, using

the makeblastdb function in BLAST. DeepVirFinder predicted mVCs were aligned to these two databases using BLASTn with 100% identity. Taxonomy of the mVCs and VCs were assigned using Kaiju with the NCBI viral proteins database (Menzel et al. 2016), and the taxonomic compositions were visualized using the Krona tool (Ondov et al. 2011).

2.3.4.3. Analysis of microbial and viral diversity and community structure

The MCs and VCs (>10 kb) were annotated using Kaiju with the NCBI nr database (Menzel et al. 2016). To visualize the relative abundance of annotated MCs a taxa barplot was produced using the ggplot2 R package. Annotated VCs were visualized with Krona (Ondov et al. 2011). To infer VC taxonomy, a genome-based network analysis of the shared protein content from VCs was performed using the vConTACT 2.0 tool with a reference database containing 2,102 bacterial and archaeal viruses from NCBI RefSeq (v94) (Bolduc et al. 2017; Bin Jang et al. 2019). A total of 711,479 protein sequences, derived from the viromes (173,433 sequences), VCs (269,901 sequences) and reference database (268,145 sequences), were subjected to all-to-all BLASTp searches (E value < 10^{-4} and bit score > 50) and grouped into protein clusters. The resulting network with genome pairs (similarity score > 1) was visualized with Cytoscape software (version 3.8.0, http://cytoscape.org/).

Using the MCs and VCs, diversity indices, including Shannon-Weaver's index, Simpson's index and richness were calculated using the vegan package, and diversity indices were plotted using the boxplot function in R (R core team, 2019). To test for significant difference in alpha diversity metrics between the pH 4.5 and 7.5 metagenomes and viromes, respectively, the Student's t-test function was used (R core team, 2019). To determine the relative abundance of both the MCs and VCs, the Salmon v0.9.1 in the Quant_bins module was used to index the contigs and align reads from each metagenome or virome back to the contigs (Patro et al. 2017; Uritskiy et al. 2018). To visualize the variation in the relative abundance of the MCs and VCs across the metagenomes and viromes, respectively, a heatmap was produced using the heatmaply R package (R core team, 2019). To compare the number of unique MCs and VCs for each soil pH, non-reproducible MCs and VCs (i.e. contigs not present in all three replicates) were removed, and the unique MCs and VCs were extracted. The unique MCs and VCs were annotated using the Kaiju tool with the NCBI nr and viral proteins database, respectively (Menzel et al. 2016). A taxa bar plot was produced using the ggplot2 R package to visualize the relative abundance of the unique MCs (R core team, 2019). A taxa bar plot was not produced for the unique VCs as most of the VCs were uncharacterized. Significance was tested for each taxa between soil pH using the Student's t-test or Wilcoxon-Mann-Whitney's test, when Bartlett's test for homogeneity of variances was not significant, using the ggpubr R package (R core team, 2019).

To compare between prokaryotic and viral community structures, non-metric multidimensional scaling (NMDS) and %GC coverage plots were produced. NMDS plots were produced from Bray-Curtis distance matrixes using the relative abundance of the MCs and VCs that were > 10 kb. The metaMDS function in the vegan R package was used for NMDS analysis, and the plot was produced using the ggplot2 R package (R core team, 2019). Significant difference in the Bray-Curtis dissimilarity between the pH 4.5 and 7.5 soil was tested using the Mann-Whitney *U* test (R core team, 2019). To generate %GC coverage plots, the MCs and VCs were indexed and the reads from the metagenomes and viromes were re-mapped with the bowtie2 mapper v2.3.0, respectively (Langmead and Salzberg 2012). The GC_cov_annotate.pl function from MetaWRAP-Bloblogy module was used to generate a blobplot file with the %GC content, coverage and taxonomy of each contig (Uritskiy et al. 2018). Blobplots were made using the makeblobplot.R function from the MetaWRAP-Bloblogy module (Uritskiy et al. 2018) within R (R core team, 2019).

2.3.4.4. Functional comparative analysis of metagenomes and viromes

From the soil metagenomes and viromes, the relative abundance of the COG functional gene categories was calculated and visualized using the ggbarplot function in the ggplot2 R package (R core team, 2019). Significance difference between the metagenomes and viromes for each functional gene category was tested using either the Student's t-test or Wilcoxon-Mann-Whitney's test, according to Bartlett's test for the homogeneity of variances, using the ggpubr R package (R core team, 2019).

2.3.4.5. Linking viruses to hosts using CRISPR array and ONF analysis

The CRISPR Recognition Tool (CRT) was used to identify CRISPR arrays from the assembled MCs (Bland et al. 2007). The CRISPR arrays were assembled from quality-passed reads using the Crass assembler (Skennerton et al. 2013). The direct repeats (DRs) and spacers were extracted using Linux commands and located using the Seqkit locate function with an exact match, and positive and negative strand search (Shen et al. 2016). Spacer sequences were searched for in mVCs and VCs, and DRs were searched for in the MCs. The DRs that matched individual contigs were annotated using the Kaiju tool with the NCBI nr database (Menzel et al. 2016).

For ONF analysis between the MCs and the VCs that were > 10 kb, the WIsH tool was used (Galiez et al. 2017). Predicted host-virus linkages with a *p*-value > 0.05 were considered potential matches (Galiez et al. 2017). Potential host MCs were taxonomically annotated with Kaiju using the NCBI nr database (Menzel et al. 2016).

2.3.4.6. Analysis of gene homology

To identify potential AMGs, gene prediction of VCs was completed using Prodigal (Hyatt et al. 2010) and homology searches were conducted using InterProScan 5 (E value < 10^{-5}) and using Diamond Blastp with the NCBI-nr database (E value < 10^{-5})(Jones et al. 2014; Buchfink et al. 2015).

Gene homology between host-linked VCs and their associated hosts was investigated. Gene prediction of the host-linked VCs and MCs was performed using Prodigal (Hyatt et al. 2010). Protein alignment between viral and host origin proteins was conducted with BLASTp (identity > 30%, E value < 10^{-5} and query cover > 70%) (Madden et al. 1996). Shared protein sequences were annotated using InterProScan 5 (E value < 10^{-5})(Jones et al. 2014). Additionally, gene homology was assessed between host-linked viruses and prokaryotes using Diamond BLASTp with the NCBI-nr database (E value < 10^{-5}) (Buchfink et al. 2015). From the Diamond BLASTp output file of each host-linked VC, the number of viral genes homologous to the same taxa as the associated host was calculated using Linux commands.

2.4. Results

2.4.1. Sequence summary of metagenomes and viromes

The six soil metagenomes consisted of 43 - 63 GB of sequencing data (Table 2.1). A total of 0.9 billion sequences were retained after quality filtering, with between 125 – 182 million reads per metagenome. Contig assembly of the sequences yielded approximately 10 million contigs, ranging from 1 – 2 million contigs per metagenome, with an average contig length of 529 bp. The six soil viromes consist of 44 - 99 GB of sequence data (Table 2.1). Approximately 1 billion sequences were retained after quality filtering, with between 141 – 321 million reads per virome. Contig assembly of the sequences yielded approximately 3.8 million contigs, ranging from 0.4 – 1 million contigs per virome.

Each virome was assembled through the JGI pipeline using MetaSPAdes. Assembly resulted in 7,114 VCs that were ≥ 10 kb, with between 497 - 1,993 VCs per virome (Table 2.2). Of the 7,114 VCs, 1,116 VCs were redundant, thus resulting in a total of 5,998 unique VCs (ANI < 95%). In comparison, assembly with MEGAHIT resulted in 13,533 VCs, representing 8,708 unique VCs (ANI < 95%). Co-assembly of the viromes using MEGAHIT yielded a total of 8,867 contigs, and the greatest number of unique VCs (8,867 VCs) (ANI < 95%). The MEGAHIT co-assembled VCs were used in downstream analyses.

Table 2.1. Sequencing summary data of pH 4.5 and 7.5 soil metagenomes (M) and viromes (V).

Sample ID	Pre-QC number of reads	Post-QC number of reads	Contig count (> 300 bp)	Contig count (> 10 kb)	Average contig length (bp)
M-pH 4.5-1	173,027,852	172,144,878	2,411,931	3,193	605.2
M-pH 4.5-2	126,396,146	125,492,704	1,597,507	2,081	625.7
M-pH 4.5-3	183,901,288	182,316,528	2,592,543	3,450	609.3
M-pH 7.5-1	146,357,310	145,190,380	1,223,894	439	449.7
M-pH 7.5-2	136,105,224	134,467,810	1,119,374	399	436
M-pH 7.5-3	151,830,478	149,253,584	1,304,821	366	449.3
V-pH 4.5-1	159,611,088	157,122,298	431,483	955	674.2
V-pH 4.5-2	145,039,036	141,380,766	365,720	544	629.2
V-pH 4.5-3	164,348,572	162,908,784	382,368	497	629.1
V-pH 7.5-1	154,699,216	153,740,568	630,012	1,933	727.8
V-pH 7.5-2	149,269,120	148,165,304	601,128	1,488	696.2
V-pH 7.5-3	323,696,904	321,563,540	898,764	1,697	673.6

QC, quality control

Table 2.2. Assembly summary data (contigs > 1 kb) of pH 4.5 and 7.5 viromes (V) using MetaSPAdes and MEGAHIT.

Sample ID	MetaSPAdes	MEGAHIT
V-pH 4.5-1	955	1,448
V-pH 4.5-2	544	1,130
V-pH 4.5-3	497	1,166
V-pH 7.5-1	1,933	3,212
V-pH 7.5-2	1,488	2,789
V-pH 7.5-3	1,697	3,788
Total VCs	7,114	13,533
Unique VCs	5,998	8,708

VC, viral contig

2.4.2. Comparison of viral recovery between metagenomes and viromes

Virus prediction from the metagenomes resulted in 296,679 metagenomic viral contigs (mVCs), with between 34,855 – 72,061 mVCs per metagenome. Predicted mVCs were only found when contigs ≥ 300 bp were analyzed using DeepVirFinder. The pH 4.5 soil metagenomes yielded approximately 2-fold more mVCs compared to the pH 7.5 soil metagenomes (Table 2.3). From the viromes, VirSorter predicted 259 virome viral contigs (vVCs), with between 17 - 77 vVCs per virome. In comparison, DeepVirFinder predicted 1,082,122 vVCs, with between 109,477 - 308,719 vVCs per virome (Table 2.3). Of these, 13,539 vVCs were greater than 10 kb, with between 1,131 - 3,789 vVCs per virome (data not shown).

Of the 296,679 DeepVirFinder predicted mVCs (> 300 bp), only 352 mVCs from the pH 4.5 metagenomes were found in the > 10 kb VCs database. None of the pH 7.5 mVCs matched to the > 10 kb VCs database (Figure 2.2a). A greater number of the mVCs were found in the > 300 bp VCs database, 14,948 and 9,035 mVCs from the pH 4.5 and 7.5 metagenomes, respectively. (Figure 2.2). The mVCs that did not match to any of the VCs in the two databases were taxonomically annotated. The majority of these mVCs were *Mimiviridae* (40%), followed by *Phycodnaviridae* (21%) (Figure 2.3).

Table 2.3. The number of predicted viruses from each pH 4.5 and 7.5 soil metagenome (M) and virome (V) using VirSorter and DeepVirFinder.

Sample ID	Predicted number of viruses	
	VirSorter (contigs > 10 kb)	DeepVirFinder (contigs > 300 bp)
M-pH 4.5-1	0	67,177
M-pH 4.5-2	0	43,806
M-pH 4.5-3	0	72,061
M-pH 7.5-1	0	38,350
M-pH 7.5-2	0	34,855
M-pH 7.5-3	0	40,430
V-pH 4.5-1	29	124,870
V-pH 4.5-2	20	109,477
V-pH 4.5-3	17	113,706
V-pH 7.5-1	54	218,664
V-pH 7.5-2	62	206,686
V-pH 7.5-3	77	308,719

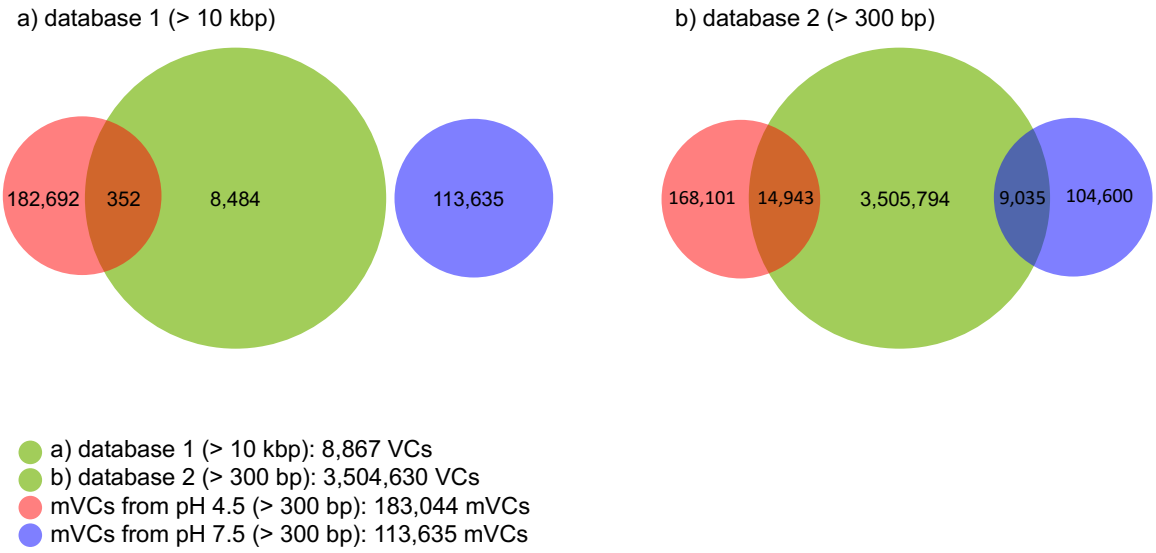


Figure 2.2. Venn diagram showing the number of metagenomic viral contigs (mVCs) found in the a) > 10 kb viral contigs (VCs) database (database 1; 8,867 VCs) and b) > 300 bp VCs database (database 2: 3,504,630 VCs) from the pH 4.5 (red) and 7.5 (blue) soil metagenomes.

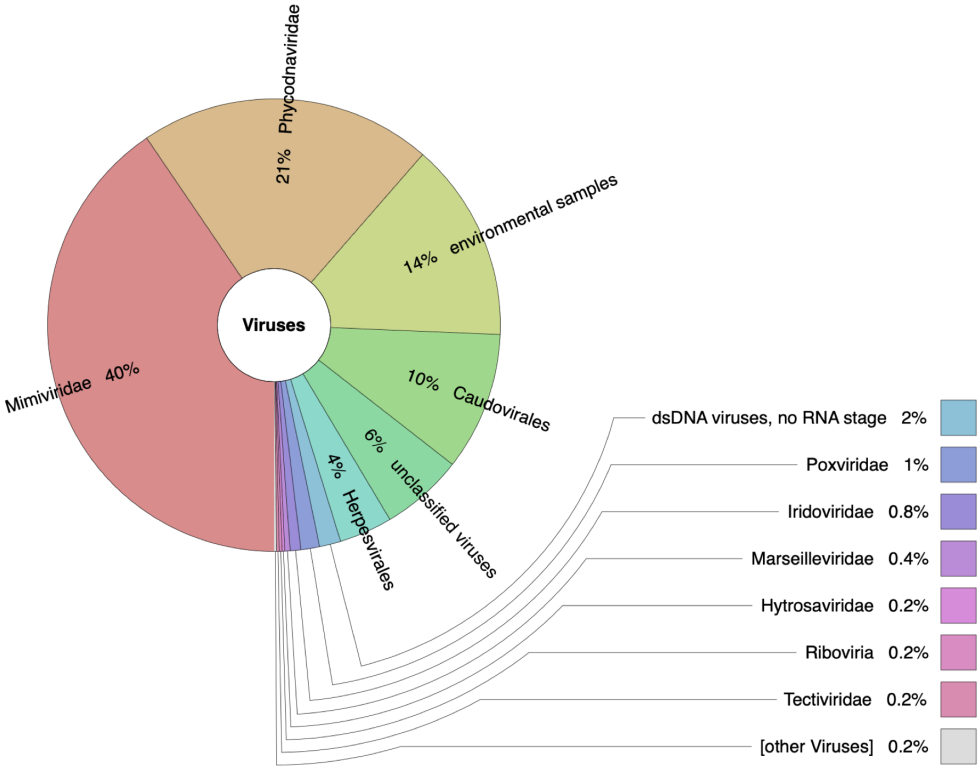


Figure 2.3. Taxonomic annotation of the 272,702 metagenomic viral contigs (mVCs) that were not found in the viral contigs (VCs) databases from the viromes.

2.4.3. Microbial and viral diversity and community structure

Taxonomic annotation of the contigs from the pH 4.5 and 7.5 soil metagenomes resulted in 34% and 26% of the contigs annotated, respectively. For all metagenomes, the contigs were predominantly assigned to the bacterial domain (98%), with less than 2% of the contigs assigned to the archaeal domain, and less than 1% to Eukarya and viruses (Figure 2.4). At the phylum level, pH 4.5 and 7.5 soils were both dominated by *Actinobacteria* and *Proteobacteria*, consisting of 68% of the contigs (Figure 2.5). The pH 4.5 soil metagenomes had a significantly greater abundance of *Actinobacteria* (48%) than the pH 7.5 soils (37%), whereas the pH 7.5 soil metagenomes had significantly higher abundances of *Proteobacteria* than the pH 4.5 soils (pH 4.5, 20%; pH 7.5, 31%) (Figure 2.5). There was a trend of greater diversity (Shannon-Weaver's and Simpson's index) in the pH 7.5 soils, but richness was slightly greater in the pH 4.5 soils, however, there was no significant difference between these two soils (Figure 2.6). Comparison of the normalized relative abundance of MCs (> 10 kb, ANI > 95%, 8,061 MCs) revealed distinct differences between the pH 4.5 and 7.5 soil (Figure 2.7). The removal of the non-reproducible MCs (i.e. contigs not present in all three replicates), resulted in 7,571 MCs. Of these, 453 MCs were unique to pH 4.5 soil, and 46 MCs were unique to pH 7.5 soil. Of the 7,072 MCs that were common to both pH soils, 73 MCs (0.9%) demonstrated a 10-fold greater abundance in the pH 4.5 soil, whereas 6,440 MCs (93%) had a 90-fold greater abundance in the pH 7.5 soil. Overall, the majority of the unique MCs were *Actinobacteria* (32%) and *Proteobacteria* (29%) (Figure 2.8). All 15 of the identified phyla had significantly different relative abundances between soil pH (Student's t-test, p-value < 0.05) (Figure 2.8).

The majority of the VCs (92%) were taxonomically unknown (Figure 2.9). Of the classified VCs, 94% were *Caudovirales* and 5% were unclassified bacterial viruses (Figure 2.9). The genome-based network analysis resulted in 3,617 viral clusters (Figure 2.10). Only 138 VCs (pH 4.5, 18 VCs; pH 7.5, 22 VCs) were clustered into 31 viral clusters with RefSeq viruses, including *Arthrobacter*, *Bacillus*, *Burkholderia*, *Flavobacterium*, *Pseudomonas*, *Ralstonia*, *Rhodococcus* and *Vibrio* phage (data not shown). Most of the VCs and those specific to either pH 4.5 and 7.5 soil clustered together, although there were some specific pH viral clusters (Figure 2.10). Viral diversity (Shannon-Weaver's and Simpson's index) and richness were significantly greater in the pH 7.5 soil than the pH 4.5 soil (Student's t-test, p-value < 0.05) (Figure 2.11), and viral community structure was distinctly different between the pH 4.5 and 7.5 soil (Figure 2.12). Removal of the non-reproducible VCs resulted in 5,057 VCs. Of these, 2,402 and 1,408 VCs were unique to pH 4.5 and pH 7.5 soil, respectively, with 1,247 shared VCs. The total number of unique VCs was 2-fold less in the pH 7.5 soil.

In the NMDS plots, microbial and viral community structure were distinct between soil pH, however, differences in contig abundances between soil pH lead to a greater dispersion in the

microbial than viral community (Figure 2.13). %GC coverage plots of the microbial and viral populations showed a consistent overlap of MCs and VCs for both soil pH (Figure 2.14a – d). In the %GC coverage plots, the microbial populations exhibited different community structure between soil pH (Figure 2.14c), whereas the viral populations were similar (Figure 2.14c and d).

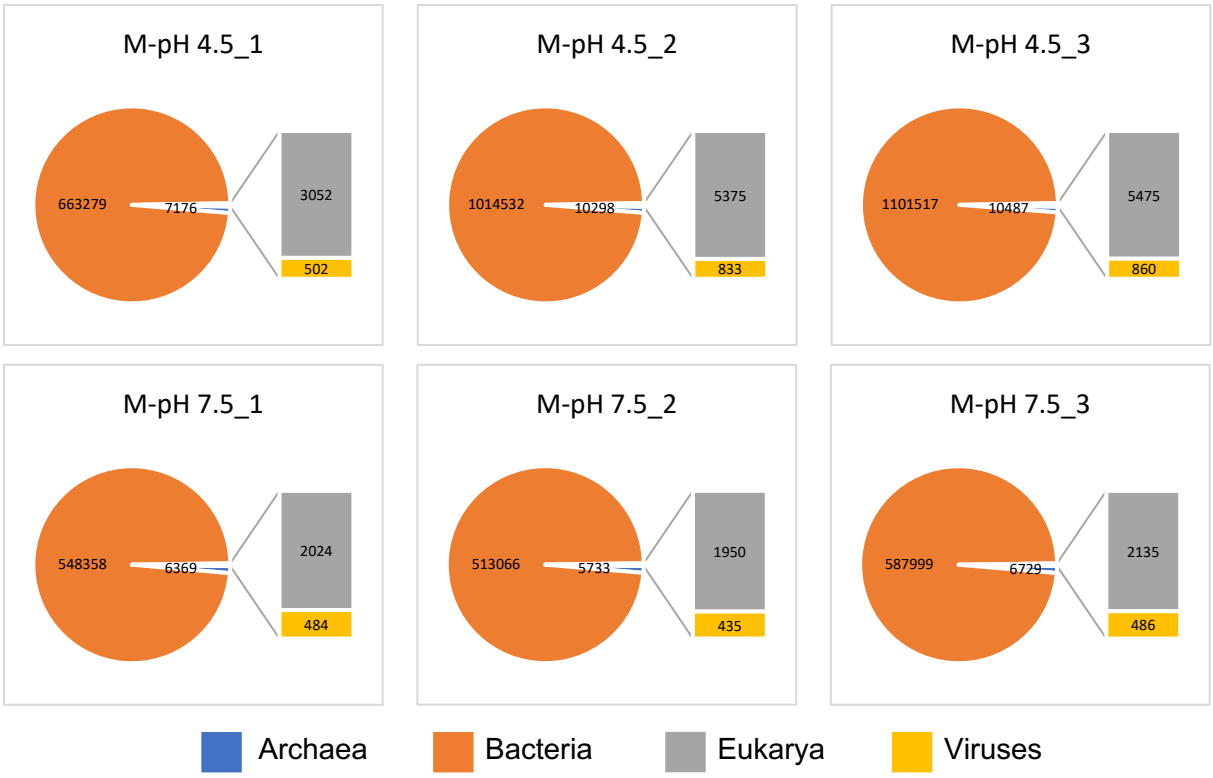


Figure 2.4. Taxonomic annotation of the pH 4.5 and 7.5 soil metagenomes at the domain level.

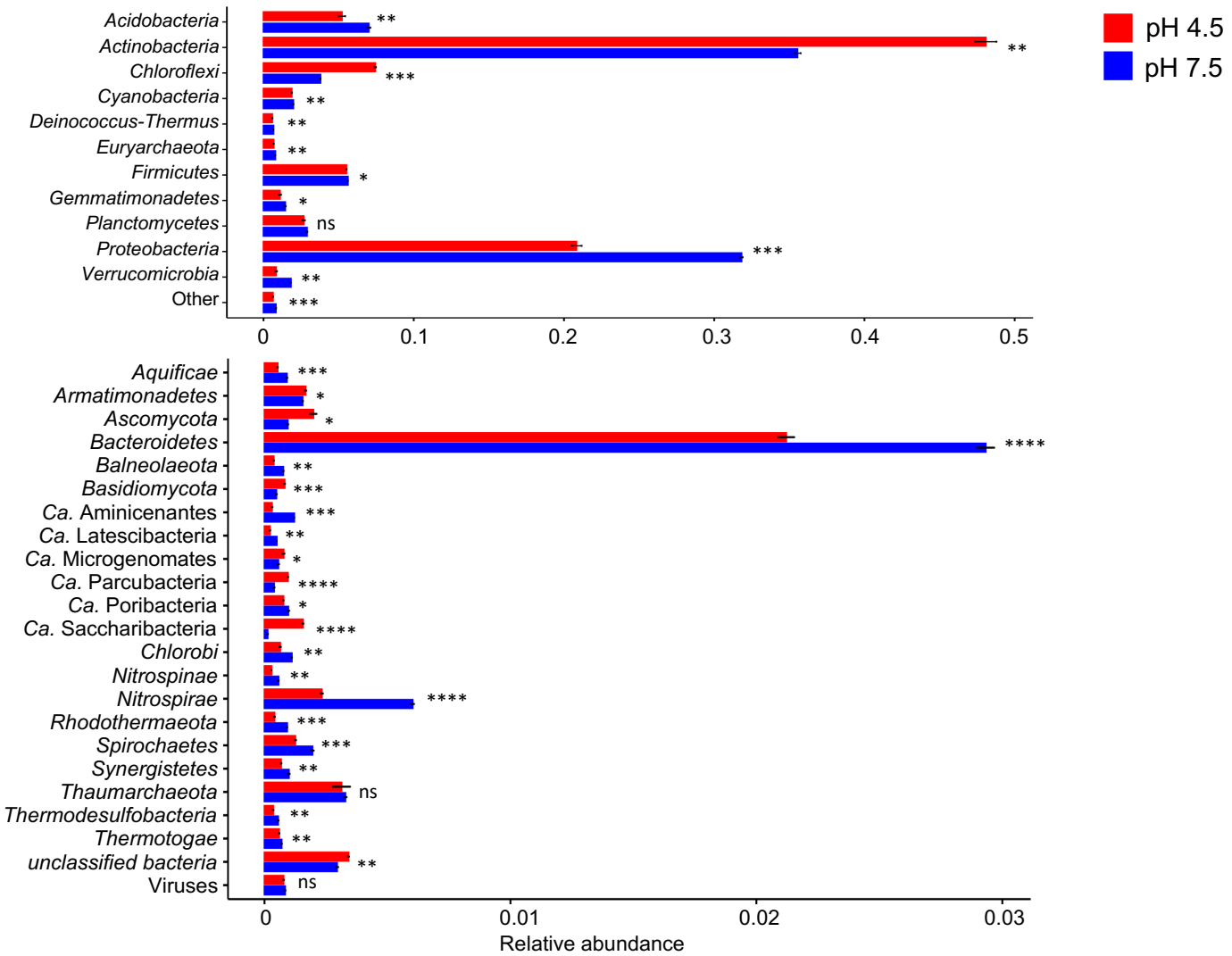


Figure 2.5. Relative abundance of the pH 4.5 (red) and 7.5 (blue) soil microbial communities. Significance differences in relative abundance for each taxon between the pH soils was tested using the Student's t-test; $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.001$ (***); $p <= 0.0001$ (****). The error bars are derived from three replicates (± SE).

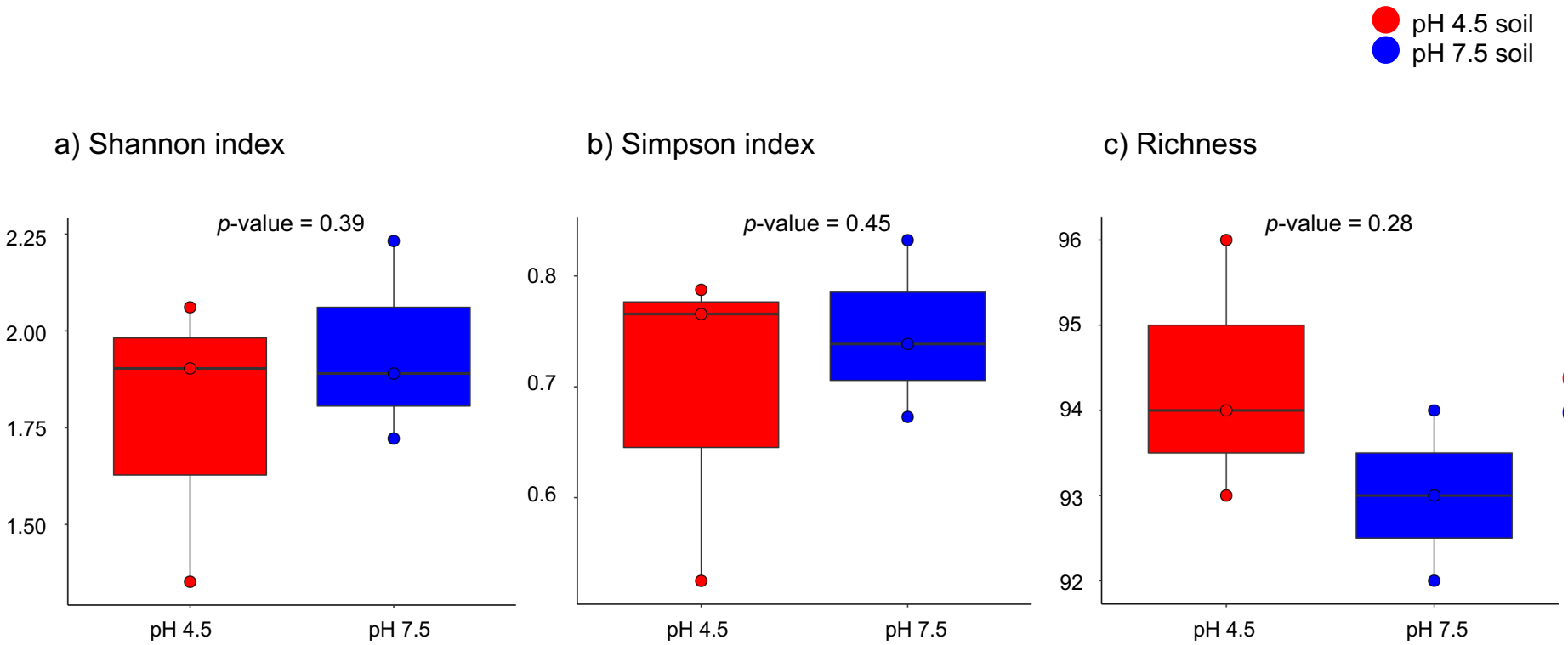


Figure 2.6. Microbial diversity, a) Shannon's index and b) Simpson's index, and c) richness of the pH 4.5 and 7.5 soil. Significant differences in indices between the pH 4.5 and 7.5 soils was tested using the Student's t-test. The error bars are derived from three replicates (± SE).

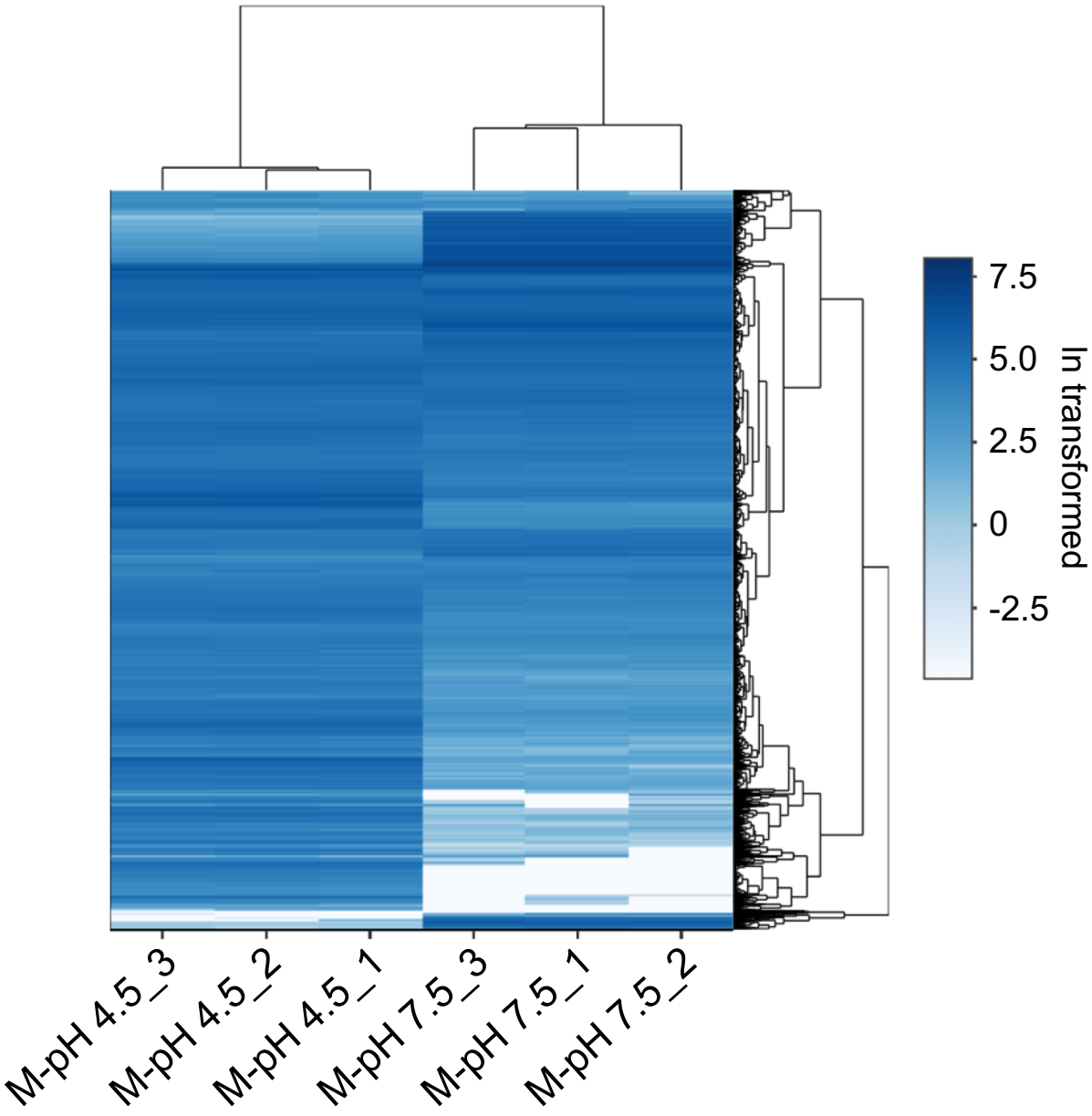


Figure 2.7. The normalized relative abundance of metagenomic contigs (MCs) in the pH 4.5 and 7.5 soil metagenomes (M).

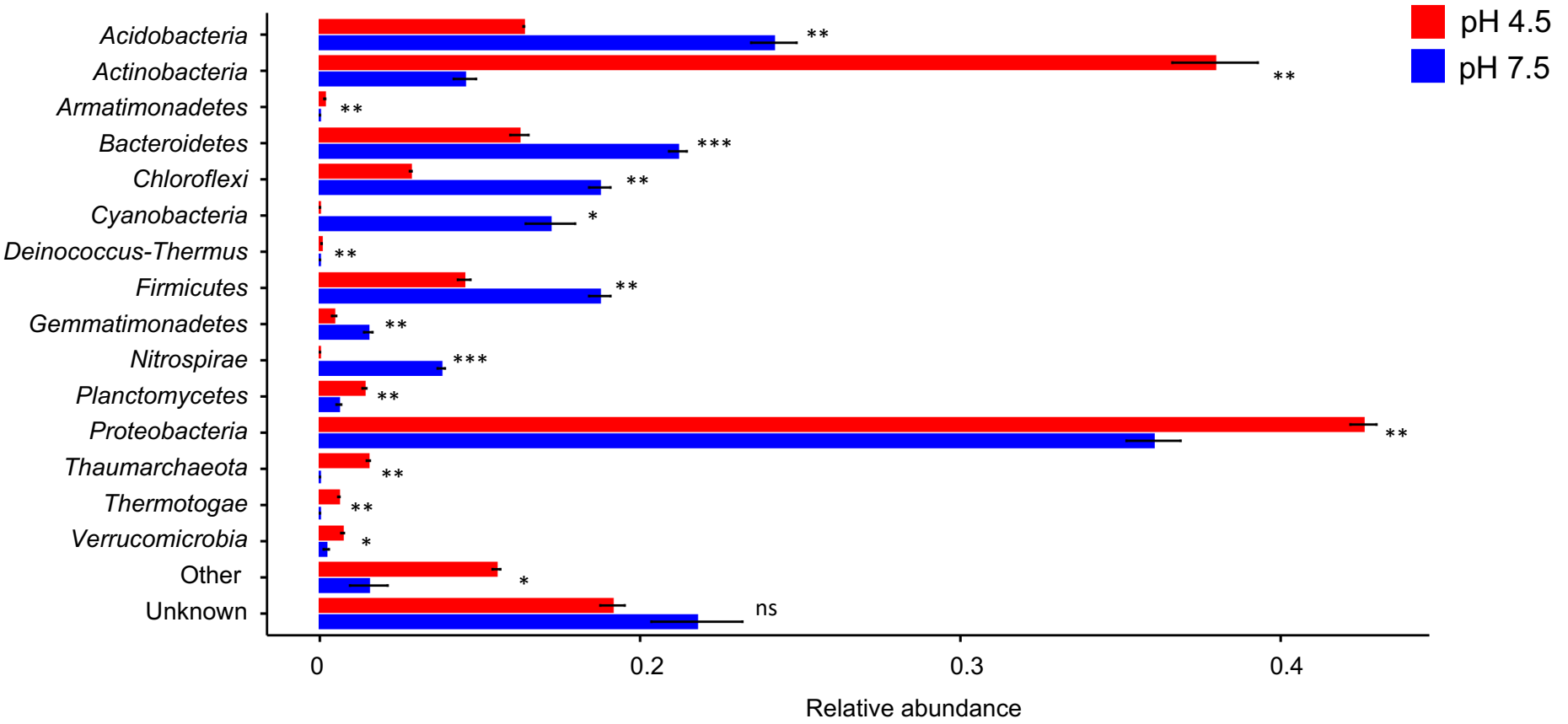


Figure 2.8. Relative abundance of the unique metagenomic contigs (MCs) from the pH 4.5 (red) and 7.5 soil (blue). Significance tested using the Student's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.001$ (***). Error bars represent the standard error of three replicates.

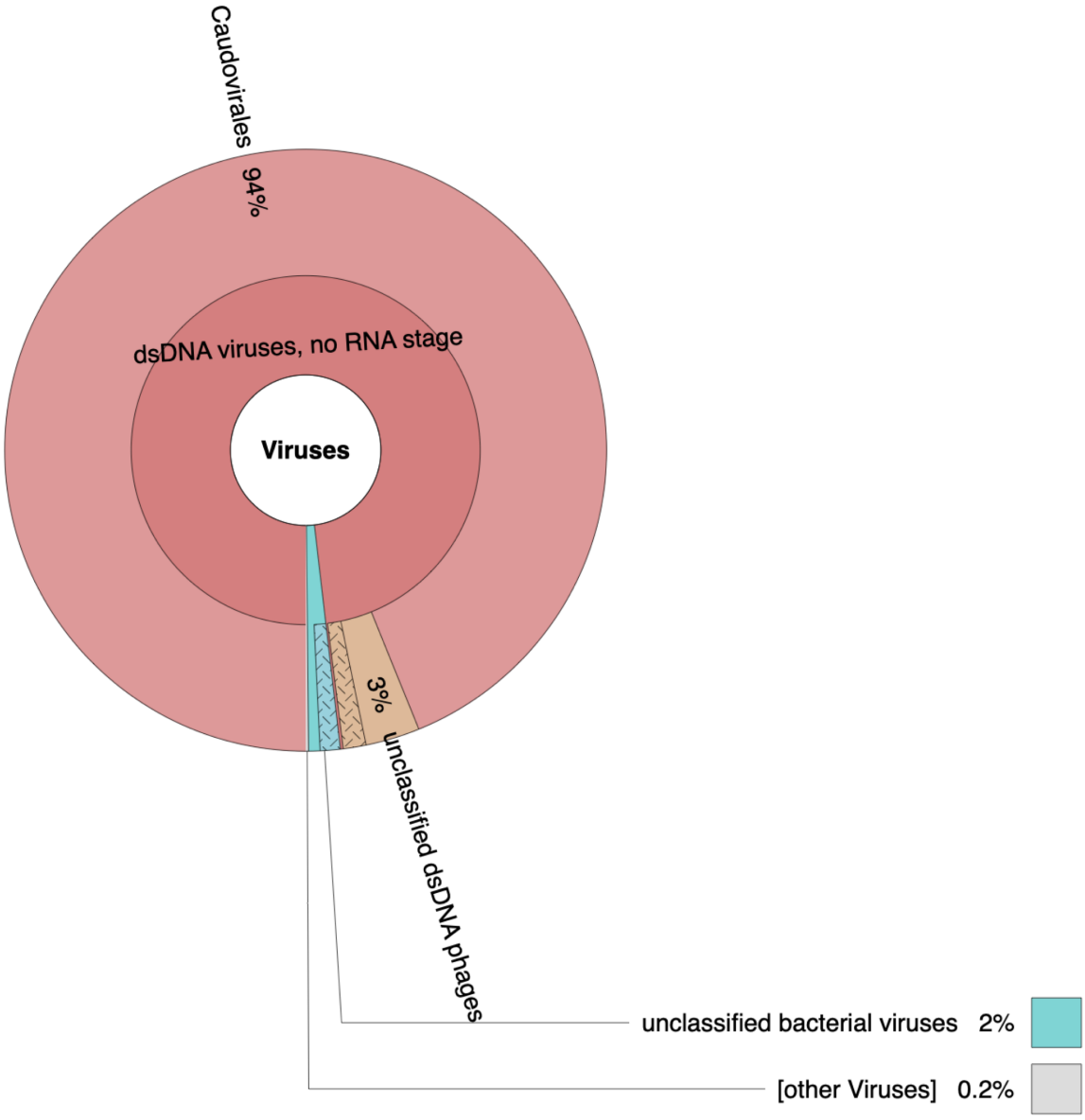


Figure 2.9. Taxonomic annotation of the virome contigs (VCs) from the pH 4.5 and 7.5 soil.

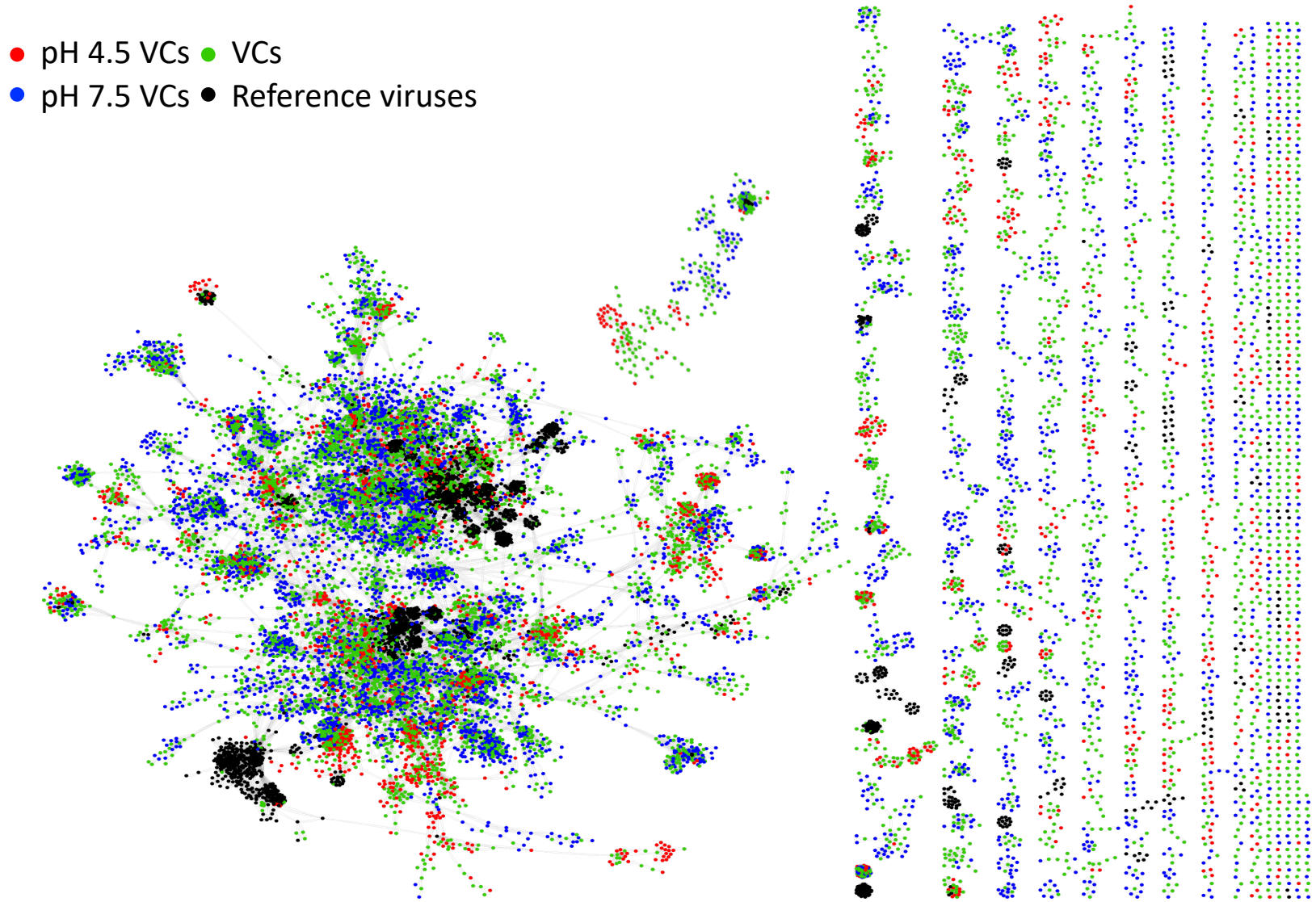


Figure 2.10. Network of shared predicted protein content among the specific viral contigs (VCs) of pH 4.5 (700 VCs) and 7.5 (1679 VCs), non-pH specific VCs (2248 VCs) and reference viruses (464 viral genomes). Nodes (circles) represent viral genomes or VCs and the shared edges (lines) indicate shared protein content.

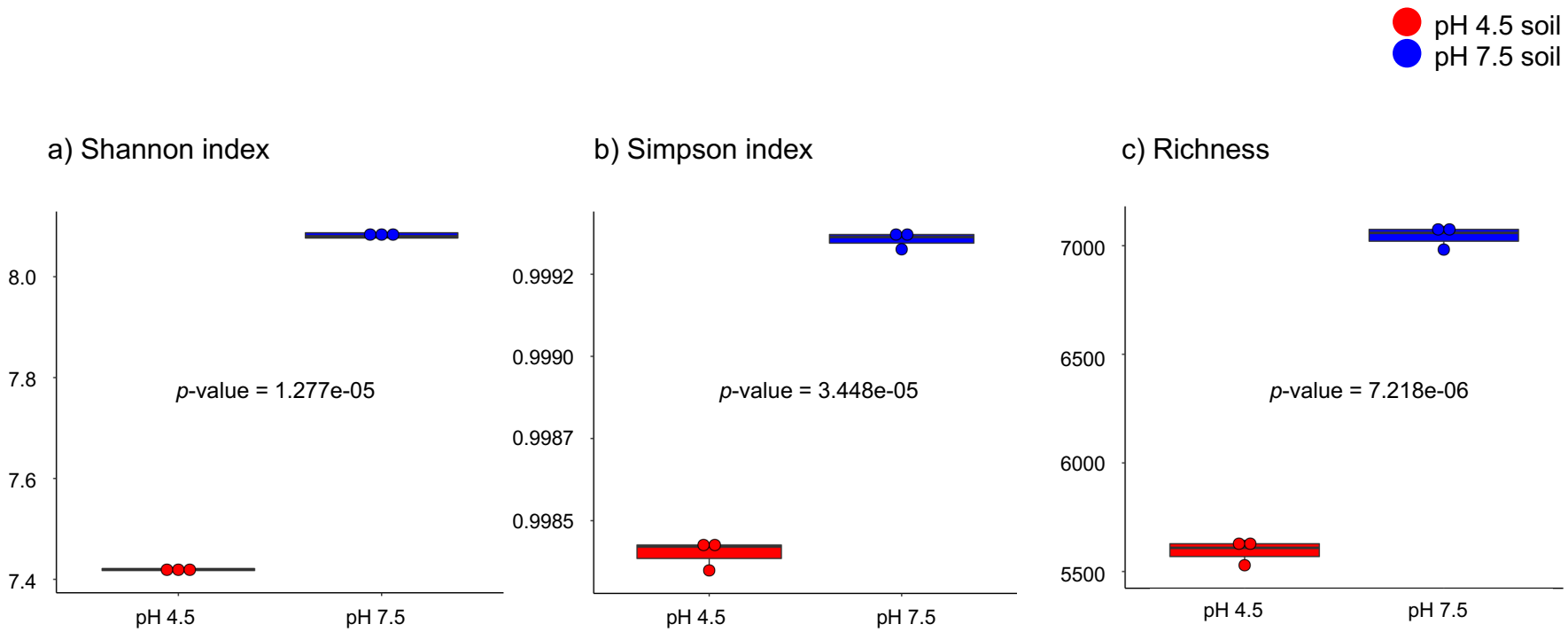


Figure 2.11. Viral diversity as determined by a) Shannon's index and b) Simpson's index, and c) richness of the pH 4.5 and 7.5 soil. Significant differences in indices between the pH 4.5 and 7.5 soils was tested using the Student's t-test. Error bars represent the standard error of the mean three replicates.

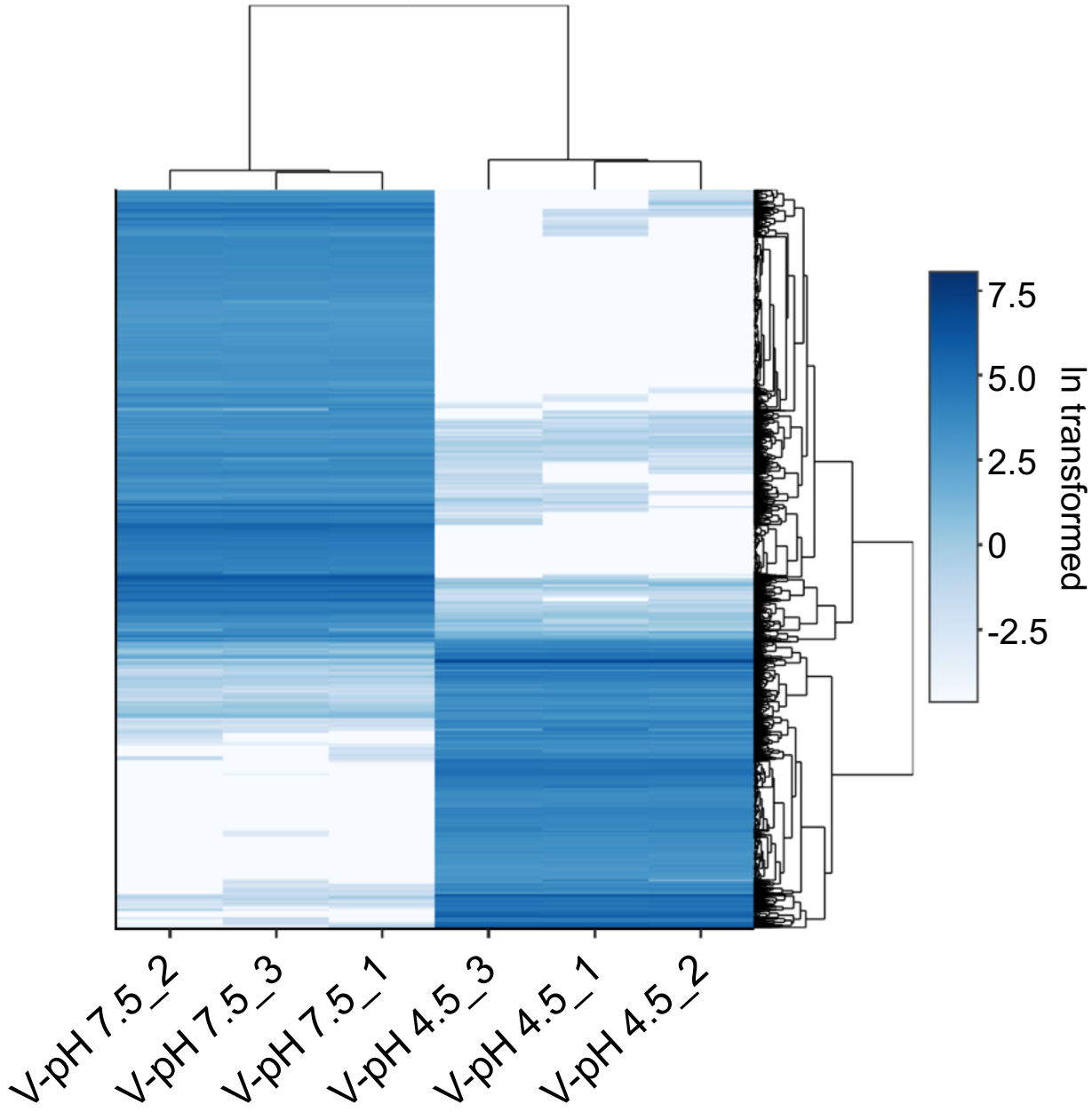


Figure 2.12. The relative abundance of virome contigs (VCs) in the pH 4.5 and 7.5 soil viromes (V).

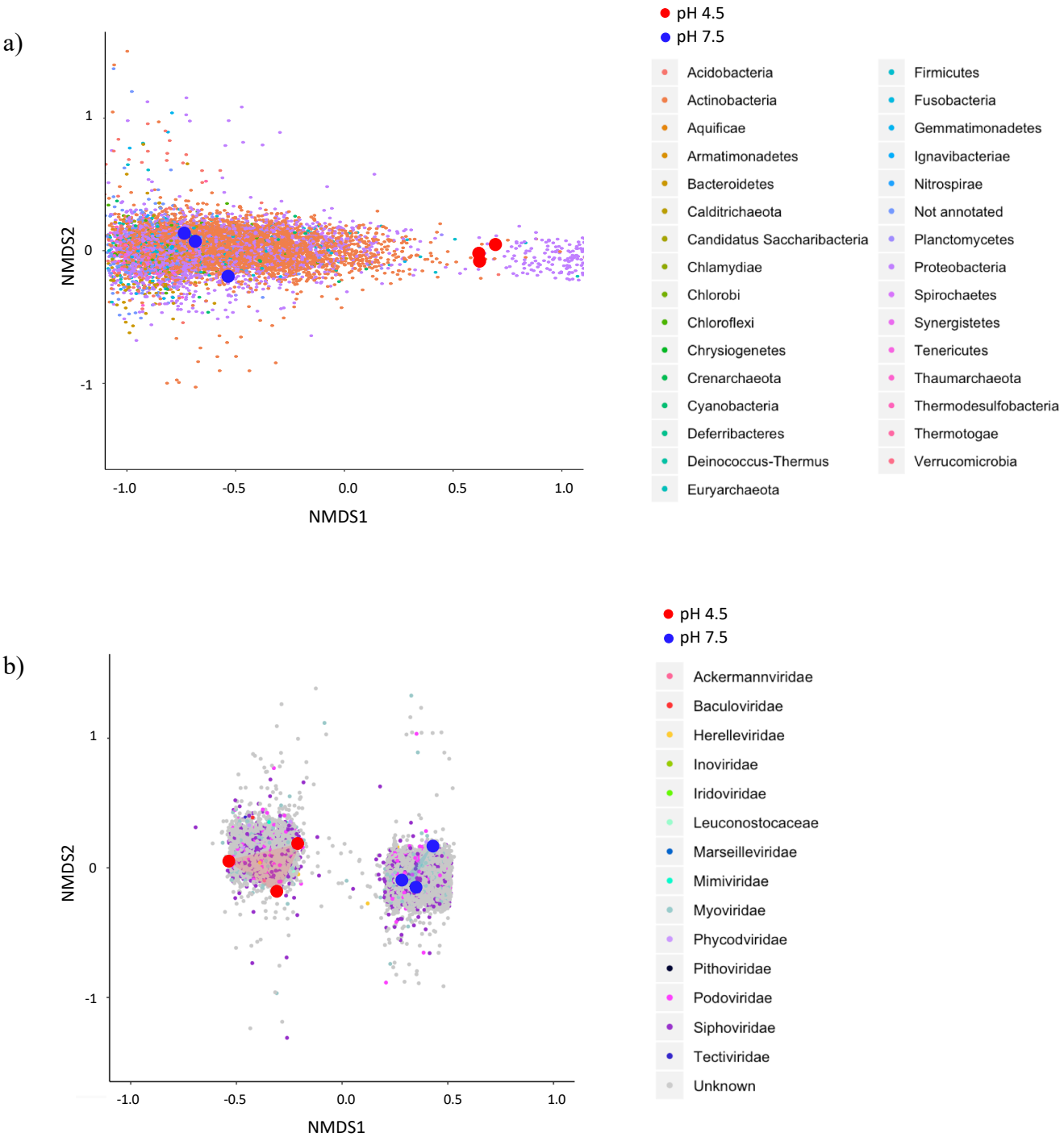


Figure 2.13. Non-metric multidimensional scaling (NMDS) plot of the a) prokaryotic and b) viral communities of the pH 4.5 (red) and 7.5 (blue) soil.

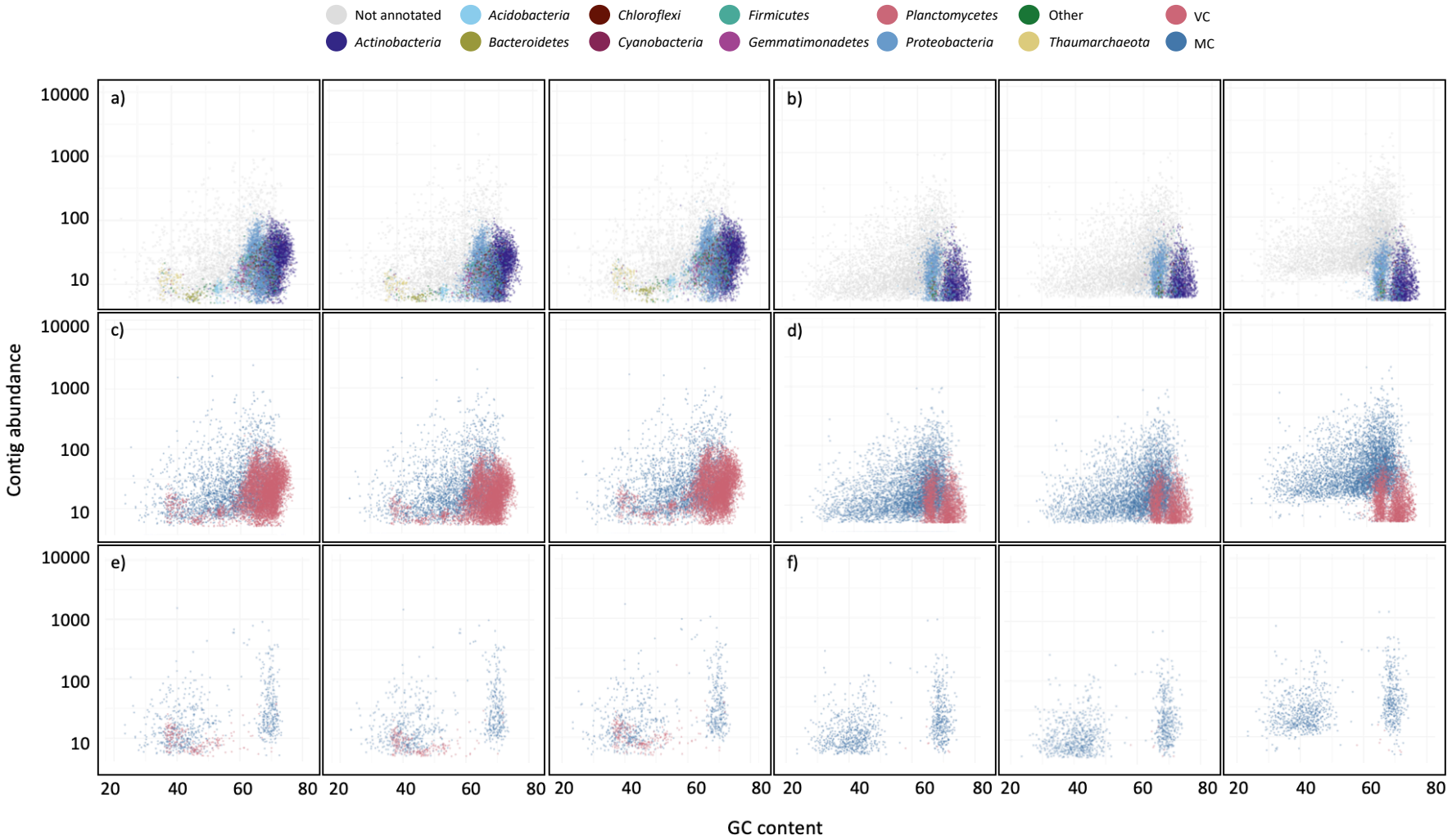


Figure 2.14. %GC coverage plots of the prokaryotic and viral populations (contigs > 10 kb) in the a) pH 4.5 and b) 7.5 soil replicates, the overlap between the metagenomic contigs (MCs) and virome contigs (VCs) in the c) pH 4.5 and d) 7.5 soil replicates, and the host-virus linked populations (contigs > 10 kb) in the e) pH 4.5 and f) 7.5 soil replicates.

2.4.4. Comparison of functional diversity between metagenomes and viromes

A total of 7,582,318 COG functional genes were annotated from the metagenomes (contigs > 300 bp), with between 913,077 - 1,973,918 genes per metagenome. Of these, 2,533 were unique COG genes. A total of 599,242 COG functional genes were annotated from the viromes (contigs > 300 bp), with between 71,197 - 145,989 genes per virome. Of these, 1,263 were unique COG genes. Soil pH did not significantly influence the functional profiles of either the viromes or metagenomes (Student's t-test, *p*-value > 0.05), however, the viromes and metagenomes had distinct functional profiles (Figure 2.15). Among the 25 predicated COG gene categories, most of the annotated genes from the metagenomes were predicted as amino acid transport and metabolism (pH 4.5, 9.6%; pH 7.5, 10%) and general function prediction only (pH 4.5, 11.9%; pH 7.5, 11.4%) (Figure 2.15). In comparison, most of the annotated genes from the viromes were predicted as replication, recombination and repair function (pH 4.5, 17.6%; pH 7.5, 18.4%) and mobilome: prophage, transposons function (pH 4.5, 13.2%; pH 7.5, 16%). The metagenomes of both pH soils in comparison to the viromes had significantly greater abundance in 10 out of the 25 COG categories. Conversely, 8 of the COG categories were significantly greater in abundance in the viromes than the metagenomes. The cell mortality, nucleotide transport and metabolism categories had significantly greater abundances in the pH 4.5 soil viromes than metagenomes.

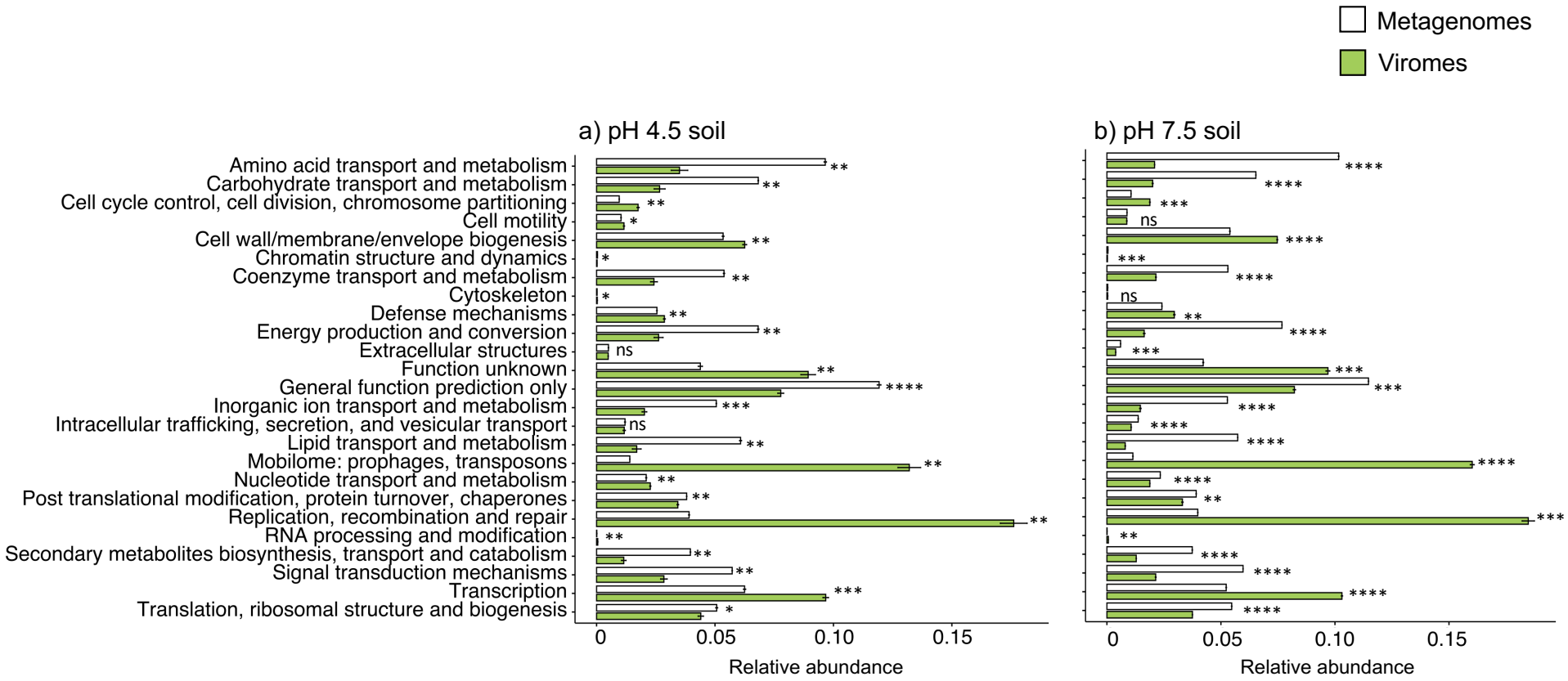


Figure 2.15. Relative abundance of the COG categories of metagenomes (white) and viromes (green) from a) pH 4.5 and b) 7.5 soil. Significant difference in the relative abundance of the COG categories between the metagenomes and viromes were tested using the Student's t-test; $p > 0.05$ (ns), $p <= 0.05$ (*), $p <= 0.01$ (**), $p <= 0.001$ (***), $p <= 0.0001$ (****). Error bars represent the standard error of the mean of triplicate samples.

2.4.5. Host-virus linkage

2.4.5.1. CRISPR array analysis

Using the CRT tool, a total of 444 CRISPR arrays were screened from the soil metagenomes (pH 4.5, 383 CRISPR arrays with 1,885 spacers; pH 7.5, 62 CRISPR arrays with 232 spacers) (Table 2.4). After removing redundancy, 1,483 and 205 unique spacers were found in the pH 4.5 and 7.5 soil metagenomes, respectively. Only one spacer was common across the soil metagenomes. The pH 4.5 soil metagenomes had 6-fold more CRISPR arrays than the pH 7.5 soil metagenomes (Table 2.5). Also, the size of the CRISPR array (i.e. number of spacers) was greater in the pH 4.5 than 7.5 soil metagenomes (Table 2.4). Out of the 383 CRISPR arrays from the pH 4.5 soil, 72 MCs were annotated, and belonged to the *Actinobacteria* (39 MCs), *Proteobacteria* (9 MCs), *Chloroflexi* (5 MCs), *Acidobacteria* (4 MCs), *Planctomycetes* (4 MCs), *Firmicutes* (4 MCs), *Cyanobacteria* (2 MCs), *Gemmatimonadetes* (1 MC) and *Thermotogae* (1 MC). Out of the 62 CRISPR arrays from the pH 7.5 soils, 16 MCs were annotated, and belonged to *Actinobacteria* (11 MCs), *Proteobacteria* (2 MCs), *Acidobacteria* (2 MC), *Chloroflexi* (1 MC) and *Planctomycetes* (1 MC). In total, 14 spacers from five CRISPRs in pH 4.5 soils were matched to 5 VCs (> 10 kb) whereas only one spacer from one CRISPR in pH 7.5 soils was matched to one VC (> 10 kb) (Table 2.5). The MCs that contained DRs were searched to investigate host-virus linkages. However, only DRs from one CRISPR in pH 7.5 soil were matched to MCs (> 10 kb), resulting in nine host-virus linkages (Table 2.6). Out of the nine host-virus linkages from pH 7.5 soils, eight hosts belonged to *Actinobacteria* (*Streptomyces*, *Blastococcus*, *Leptothrix*, *Kribbella* and two unknown genera) and one to *Proteobacteria* (*Leptothrix*) (Table 2.6).

Using the Crass tool, a total of 2,997 CRISPR arrays were assembled from the pH 4.5 and 7.5 soil metagenomes (pH 4.5, 1,896 CRISPR arrays with 34,738 spacers; pH 7.5, 1,101 CRISPR arrays with 8,349 spacers). After removing redundancy, 34,433 and 8,285 unique spacers were found in the pH 4.5 and 7.5 soil metagenomes, respectively. In total, 41 spacers were common across the soil metagenomes. Only three spacers from pH 4.5 soils were matched to three VCs (> 10 kb), while their DRs did not match to any MCs (> 10 kb) (Table 2.5).

2.4.5.2. ONF analysis

ONF analysis using the WIsH tool with 8,061 MCs (> 10 kb) and 8,867 VCs (> 10 kb), resulted in the prediction of 1,923 host-virus linkages (p-value of 0.05). A total of 303 MCs were predicted as hosts of the 1,923 VCs (Table 2.7). Most of the VCs were linked to *Actinobacteria* (661 VCs), including *Mycobacterium* (466 VCs) and *Actinoplanes* (179 VCs), and *Thaumarchaeota* (489 VCs), particularly *Nitrososphaera* (464 VCs), and to unknown contigs (307 VCs) (Table 2.7). Of the 1,923 host-linked virus populations, 209 and 330 VCs were uniquely present in the pH 4.5 and 7.5 viromes, respectively (Table 2.7). A similar proportion of unique VCs was predicted between the

pH soils (Table 2.7). The unique VCs from both pH soils were mostly predicted to infect the *Thaumarchaeota* (pH 4.5, 65 VCs; pH 7.5, 104 VCs) (Table 2.7). Both pH soil viral populations were predominantly predicted to infect several phyla, including *Actinobacteria*, *Bacteroidetes* and *Proteobacteria* (Table 2.7). Similar %GC coverage plots of host-linked virus populations were observed between pH soils (Figure 2.13e and f).

Table 2.4. Summary data of the CRISPR array analysis.

Metagenome ID	Number of CRISPR arrays	Number of spacers	Length of CRISPR array sequences		
			sum	average	maximum
M-pH 4.5_1	142	733	50,054	344.8	2,275
M-pH 4.5_2	77	419	27,544	353.1	1,543
M-pH 4.5_3	163	843	38,906	346.3	2,334
M-pH 7.5_1	20	87	5,523	263	467
M-pH 7.5_2	21	90	5,502	250.1	487
M-pH 7.5_3	21	77	4,758	216.3	392

Table 2.5. Spacer sequences matching to viral contigs (VC) > 10 kb.

CRISPR ID	Virus ID	Matched Sequence	Strand	Start	End
CRT_pH 4.5_3_CRISPR_12_Spacer_3	VC_01252	CGTTTTCCTGAGTCCGGGAAAACGACTGCTCAG	+	3967	3999
CRT_pH 4.5_3_CRISPR_34_Spacer_1	VC_02336	ACGTTTTCAATGACTTAGCTGGATAATTTTTCCTTCC	-	32693	32729
CRT_pH 4.5_3_CRISPR_34_Spacer_2	VC_02336	GCGTTTCCCCTCCGTGTCTTTAAAAAGCCACAATTCA	-	32628	32664
CRT_pH 4.5_3_CRISPR_38_Spacer_1	VC_06823	AACTGGCCCTGGCCGGCGGGCTGCTGGTAGGACGGCTG	+	3469	3506
CRT_pH 4.5_3_CRISPR_38_Spacer_2	VC_06823	TACTGCTGGGCTGCCTGGACCCCGGCCTGCGTGGCGTCCGCCAC	+	3529	3572
CRT_pH 4.5_3_CRISPR_38_Spacer_3	VC_06823	CCGAGCGGCTGGCCGTTGAAGCCGGTCGCGTCAGGCTGCCACTG	+	3595	3638
CRT_pH 4.5_3_CRISPR_64_Spacer_2	VC_08093	ATGCCGACGGTGGTGAATCCCACGAAGCCGAC	+	1903	1934
CRT_pH 4.5_3_CRISPR_77_Spacer_7	VC_07400	GCGCGATGTTCACTCGGCTGCCCATGAGGT	-	75886	75915
CRT_pH 4.5_3_CRISPR_77_Spacer_6	VC_07400	TCGGATACTTCGCGACCTGCGGGCGAGCAGCACCACGGGG	-	75938	75977
CRT_pH 4.5_3_CRISPR_77_Spacer_3	VC_07400	TCGGATGCGAGCCGTGGCGAGGAGGTGATACGTGCCTTG	-	76122	76160
CRT_pH 4.5_3_CRISPR_77_Spacer_2	VC_07400	TCGGATTCGATCTGGTGCGTGGTCGCCTTGCCTACCATC	-	76183	76221
CRT_pH 4.5_3_CRISPR_77_Spacer_1	VC_07400	CCATATAGCGACGACGCGCTGGGTCATGCCCGCATGGAG	-	76244	76282
CRT_pH 4.5_3_CRISPR_77_Spacer_5	VC_07400	TCGGATGGTGACCGGCGGGCTGACCAACGGCACGCTCCC	-	76000	76038
CRT_pH 4.5_3_CRISPR_77_Spacer_4	VC_07400	TCGGATCGGACTGACGCTTCTGTGTCAATCCCGTAGTAC	-	76061	76099
CRT_pH 7.5_1_CRISPR_02_Spacer_6	VC_02123	TCCGGCCAAGAAGGCCGCAGC	+	7425	7445
Crass_CRISPR_pH 4.5_Spacer_330	VC_08205	AGTGGACGGGCGCTCACCGGGAGGCGCTAGCAG	+	25924	25956
Crass_CRISPR_pH 4.5_Spacer_361	VC_00892	TTATCTCGGTTGCTGCCGCGTCATTGGCGCTCGC	-	3406	3439
Crass_CRISPR_pH 4.5_Spacer_472	VC_04507	GGGTGTGGCTCCGGTCCAGTTGATGCGGACGATG	+	6022	6055

Table 2.6. Direct repeat (DR) sequences matching to metagenomic contigs (MC) > 10 kb.

CRISPR ID	Host ID	Matched sequence	Strand	Annotation of MC (phylum; genera)
pH 7.5_1_CRISPR_02_DR	MC_00008	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Proteobacteria; Leptothrix*
pH 7.5_1_CRISPR_02_DR	MC_00698	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Actinobacteria; Streptomyces*
pH 7.5_1_CRISPR_02_DR	MC_02023	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Actinobacteria; Blastococcus*
pH 7.5_1_CRISPR_02_DR	MC_11448	GAAGAAGGCTCCGGCCAAGAAGGC	-	*Actinobacteria;* NA
pH 7.5_1_CRISPR_02_DR	MC_11502	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Actinobacteria; Kribbella*
pH 7.5_1_CRISPR_02_DR	MC_19271	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Actinobacteria; Streptomyces*
pH 7.5_1_CRISPR_02_DR	MC_19277	GAAGAAGGCTCCGGCCAAGAAGGC	-	*Actinobacteria;* NA
pH 7.5_1_CRISPR_02_DR	MC_19278	GAAGAAGGCTCCGGCCAAGAAGGC	-	*Actinobacteria; Microlunatus*
pH 7.5_1_CRISPR_02_DR	MC_20600	GAAGAAGGCTCCGGCCAAGAAGGC	+	*Actinobacteria; Streptomyces*

NA, not annotated

Table 2.7. Virus- host linkage between co-assembled viral contigs (VCs) and prokaryotic contigs of the soil metagenomes using the WIsH tool.

Host phylum	Number of host-linked VCs	Number of unique VCs for pH 4.5	Number of unique VCs for pH 7.5
Acidobacteria	20	5	6
Actinobacteria	661	30	55
Bacteroidetes	227	29	57
Candidatus *Saccharibacteria*	20	3	3
Chlamydiae	18	1	4
Cyanobacteria	4	1	1
Euryarchaeota	10	4	1
Firmicutes	9	4	2
Gemmatimonadetes	1	0	1
Planctomycetes	1	0	0
Proteobacteria	146	23	27
Spirochaetes	2	0	0
Thaumarchaeota	496	65	107
Unknown	307	43	66
Verrucomicrobia	1	1	0
Total	1923	209	330

2.4.6. Gene homology

2.4.6.1. Auxiliary metabolic genes

The VCs (> 10 kb) of the co-assembled viromes yielded a greater number of AMGs (2,017 AMGs) compared to the VCs of the assembled pH 4.5 and 7.5 viromes (1184 AMGs). From the 8,867 co-assembled VCs, a total of 269,901 viral genes were predicted, of which 113,770 genes were annotated and 2,481 of these genes had a unique function. The majority of the annotated genes (5.9%, 6,799 genes) were viral proteins, such as major capsid protein, tail protein, integrase, portal, tail protein and terminase, and conventional bacterial proteins (3.3%, 3,824 genes), including nucleic acid synthesis and metabolism, that are critical for the reproduction and survival of viruses. After the removal of known viral-associated protein families, the total number of unique protein families was 2,071 (90%, 103,241 genes). AMGs encoding for proteins in the glycoside hydrolase (GH) family (GH family 1, 5, 8, 10, 12, 16, 19, 22, 24, 25 26, 37, 42, 45, 71, 372 genes) and peptidase family (A24, A8, C1A, C26, C39, G1, G2, M2, M4, M10, M11, M14, M15, M16, M17, M20, M23, M24, M41, M48, M50, M54, S1, S8, S11, S24, S26, S49, S53, U9, 309 genes) were identified (Supplementary Table 2.2). Of these, GH 25 (56 genes), GH 24 (38 genes), GH 16 (14 genes), GH 19 (14 genes), GH 26 (13 genes) GH 5 (12 genes genes), M15 (80 genes), M23 (61 genes) and C1 (29 genes) were the most abundant families (Supplementary Table 2.2). In addition, AMGs encoding proteins for membrane transport functions were also identified, such as ATPases (127 genes), ABC transporters (5017 genes), and other membrane transporters, including mechanosensitive channel (2 genes), potassium channels (1 gene), aquaporin transporter (1 gene)

and sodium/phosphate symporters (1 gene). AMGs encoding protein cofactors, such as FAD- (31 genes), NAD-binding domain (108 genes), [2Fe-2S] ferredoxin domain (16 genes), copper-binding site (3 genes) and cupredoxin like domain (6 genes) were identified (Supplementary Table 2.3).

From the VCs of the assembled pH 4.5 and 7.5 viromes, a total of 173,433 viral genes (pH 4.5 VCs, 50,908; pH 7.5 VCs, 122,525) were predicted, of which, 73,078 genes (pH 4.5 VCs, 37,215; pH 7.5 VCs, 51,849) were annotated, and 1,486 of these genes had a unique function. The comparison of AMGs from the pH 4.5 and 7.5 VCs showed 2-fold abundant viral proteins including major capsid protein, tail protein, integrase, portal, tail protein and terminase in pH 7.5 VCs (3156 genes) compared to pH 4.5 VCs (1449 genes). Similarly, conventional bacterial proteins that are critical for the reproduction and survival of viruses were 2-fold more abundant in the pH 7.5 VCs (1,708 genes) than pH 4.5 VCs (737 genes). After the removal of known viral-associated protein families, the total number of unique protein families was 704 and 966 in pH 4.5 and 7.5 VCs, respectively. Of these, 484 protein families were present in both pH 4.5 and 7.5 VCs, while 482 and 219 protein families were specific to pH 4.5 and 7.5 VCs, respectively. There was a greater abundance of GH and peptidases in the pH 7.5 than 4.5 soil (Figure 2.16). The GH families 5, 19, 25 and 26, and the peptidases C1A, C26, C39, M15A, M23, M41, S1, S1C and S49 were present in both pH VCs (GH families, 169 genes; peptidases, 126 genes). The GH family 37 and peptidase G1 and U4 were only identified in pH 4.5 VCs (GH families, 3 genes; peptidases, 5 genes), whereas the pH 7.5 VCs had AMGs encoding for the GH families 6, 8, 9, 16, 18, 26, 46, 71 and 81, and peptidases A8, M10, M13, M15B, M15C, M2, M4, M48 and SB1 (GH families genes 26 genes; peptidases, 39 genes) (Supplementary Table 2.2). While the AMGs coding for ATPases (pH 4.5 VCs, 26 genes; pH 7.5 VCs, 41 genes) and FAD- and NAD-binding domain (pH 4.5 VCs, 23 genes; pH 7.5 VCs, 52 genes) were identified in both soil pH, the ABC transporter type 1 (5 genes) and ferrodoxin-like domain (13 genes) were only present in pH 7.5 VCs (Supplementary Table 2.3).

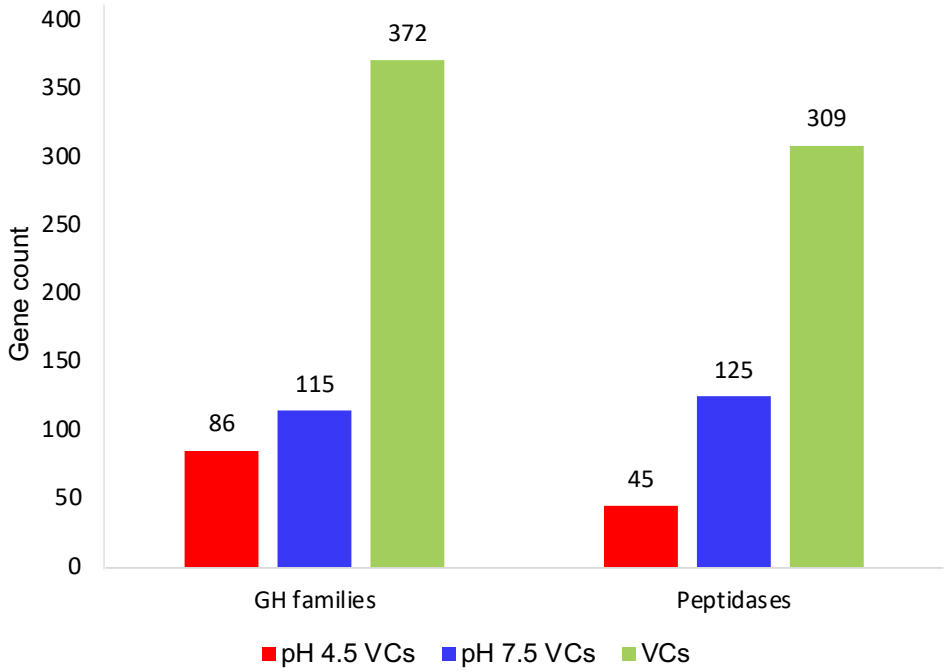


Figure 2.16. Number of genes coding for glycoside hydrolase (GH) families and peptidases in the viral contigs (VCs) of the assembled pH 4.5 (red) and 7.5 (blue) viromes, and the VCs of the co-assembled viromes (green).

2.4.6.2. Gene homology between viruses and their associated hosts

The single virus, VC_02123, that was linked to 9 MCs through CRISPR array analysis (Table 2.6), had a homologous gene involved in DNA replication (alpha subunit of DNA polymerase III) of an *Actinobacteria* of an unknown genus (BLASTp, identity = 31.7%, E value = $8.06e^{-139}$ and bit score = 448). Gene homology analysis between WIsH linked viruses and hosts revealed that of the 1,923 predicted host-virus linkages only 139 had at least one homologous gene (Table 2.8). Of the 1,540 homologous genes, 837 genes were annotated (E value < 10^{-3}). Most of the genes were unknown and uncharacterized hypothetical proteins. The VCs infecting *Mycobacterium* had genes coding for WhiB-like iron-sulfur binding domain (9 genes) and transcription factor WhiB (4 genes). Interestingly, the VCs infecting *Actinobacteria*, including *Bifidobacterium*, *Nocardioides*, *Streptomyces* and unclassified genera, had genes coding for a glycoside hydrolase family (5 genes). Several viral hallmark genes, such as phage tail (3 genes), portal (3 genes) and terminase (4 genes), and bacterial genes involved in DNA replication (e.g. transcription factor (14 genes) and DNA-binding domain (19 genes)), were also homologous between VCs and host (data not shown). A total of 43 MCs had the same series of genes to a VC (identity = 100%), indicating viral presence in these MCs.

Table 2.8. Gene homology between the WIsH predicted viral contigs (VC) and host metagenomic contigs (MC) for determining the number of shared genes (BLASTp; identity > 30%, E value > 0.001, bit score > 30 and query cover > 70%).

Virus ID	Virus annotation	Host ID	Host annotation Phylum	Genus	WisH (*p*-value)	Number of homologs with predicted host
VC_0000000009	*Siphoviridae*	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	2
VC_0000000016	Unknown	MC_0000032464	Unknown	Unknown	0	26
VC_0000000032	Unknown	MC_0000019697	Unknown	Unknown	0	22
VC_0000000149	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	5
VC_0000000155	*Siphoviridae*	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	2
VC_0000000178	Unknown	MC_0000022142	Unknown	Unknown	1.90e-06	7
VC_0000000215	Unknown	MC_0000019275	*Actinobacteria*	*Nocardioides*	0	32
VC_0000000396	Unknown	MC_0000000341	Unknown	Unknown	0	22
VC_0000000432	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.03	1
VC_0000000433	Unknown	MC_0000021247	Unknown	Unknown	3.40e-4	1
VC_0000000492	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000000536	Unknown	MC_0000020058	Unknown	Unknown	0	24
VC_0000000609	Unknown	MC_0000020710	*Candidatus* Saccharibacteria	NA	0	20
VC_0000000696	Unknown	MC_0000019267	Actinobacteria	*Mycobacterium*	0.03	1
VC_0000000715	*Siphoviridae*	MC_0000022142	Unknown	Unknown	0	5
VC_0000000877	*Siphoviridae*	MC_0000000013	Unknown	Unknown	0	119
VC_0000000914	Unknown	MC_0000000418	Unknown	Unknown	0	25
VC_0000000951	Unknown	MC_0000002403	Unknown	Unknown	0	19
VC_0000000986	Unknown	MC_0000000580	Unknown	Unknown	1.90e-3	5
VC_0000001065	Unknown	MC_0000019272	*Actinobacteria*	*Streptosporangium*	0	14
VC_0000001066	*Siphoviridae*	MC_0000021247	Unknown	Unknown	1.11e-16	2
VC_0000001254	*Siphoviridae*	MC_0000019279	*Actinobacteria*	*Streptomyces*	0	45
VC_0000001313	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1

VC_0000001327	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.04	1
VC_0000001415	Myoviridae	MC_0000019977	Unknown	Unknown	0	10
VC_0000001445	Unknown	MC_0000019344	*Candidatus* Saccharibacteria	NA	0	35
VC_0000001779	Unknown	MC_0000002608	Unknown	Unknown	0	7
VC_0000001785	Siphoviridae	MC_0000032613	Unknown	Unknown	0.01	2
VC_0000001796	Siphoviridae	MC_0000019286	Unknown	Unknown	0	71
VC_0000001813	Siphoviridae	MC_0000001409	*Candidatus* Saccharibacteria	NA	0	15
VC_0000001823	Mimiviridae	MC_0000002436	*Candidatus* Saccharibacteria	NA	0	6
VC_0000001937	Unknown	MC_0000022142	Unknown	Unknown	0	5
VC_0000001985	Unknown	MC_0000020483	*Proteobacteria*	*Yersinia*	0	11
VC_0000002006	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.01	2
VC_0000002014	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.02	1
VC_0000002048	Unknown	MC_0000019746	*Actinobacteria*	*Bifidobacterium*	5.86e-4	67
VC_0000002054	Unknown	MC_0000019746	Unknown	Unknown	0	27
VC_0000002081	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000002133	Unknown	MC_0000033612	Unknown	Unknown	0	9
VC_0000002138	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000002155	Unknown	MC_0000020170	Unknown	Unknown	2.97e-3	4
VC_0000002186	Unknown	MC_0000019275	*Actinobacteria*	*Nocardioides*	0	32
VC_0000002247	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.01	1
VC_0000002262	Unknown	MC_0000022142	Unknown	Unknown	0	4
VC_0000002373	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	3.26 e-3	1
VC_0000002376	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	6.75 e-3	1
VC_0000002431	Unknown	MC_0000000003	*Actinobacteria*	*Streptomyces*	4.37e-11	9
VC_0000002523	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.02	1
VC_0000002529	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.02	1
VC_0000002556	Unknown	MC_0000021387	Unknown	Unknown	0	11

VC_0000002664	Unknown	MC_0000022033	Unknown	Unknown	0	11
VC_0000002812	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.02	1
VC_0000002872	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000002886	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.03	1
VC_0000002889	Unknown	MC_0000012571	Unknown	Unknown	0.01	3
VC_0000002890	Unknown	MC_0000022142	Unknown	Unknown	2.15e-13	2
VC_0000002907	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	2
VC_0000002928	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.03	1
VC_0000002959	Unknown	MC_0000000211	*Proteobacteria*	*Methylomonas*	4.63e-4	18
VC_0000002974	Unknown	MC_0000021247	Unknown	Unknown	5.68e-3	1
VC_0000003006	Unknown	MC_0000019726	*Actinobacteria*	*Bifidobacterium*	0	21
VC_0000003014	Unknown	MC_0000022142	Unknown	Unknown	2.93e-3	2
VC_0000003018	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.04	2
VC_0000003025	Myoviridae	MC_0000019323	*Proteobacteria*	*Magnetospira*	0	44
VC_0000003212	Siphoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000003242	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.02	2
VC_0000003267	Siphoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	5.67e-3	4
VC_0000003354	Unknown	MC_0000019329	NA	*Lessievirus*	0	59
VC_0000003600	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.04	1
VC_0000003871	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.04	1
VC_0000003942	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000003953	Unknown	MC_0000032473	*Proteobacteria*	*Gluconobacter*	0	86
VC_0000003968	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000004002	Siphoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	7.63e-3	2
VC_0000004080	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000004153	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.03	2
VC_0000004207	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	8.27e-3	1

VC_0000004219	Myoviridae	MC_0000019795	*Proteobacteria*	NA	1.16e-06	1
VC_0000004314	Unknown	MC_0000032613	Unknown	Unknown	2.03e-3	1
VC_0000004333	Unknown	MC_0000022141	Unknown	Unknown	0	12
VC_0000004470	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000004518	Podoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000004529	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000004617	Unknown	MC_0000020379	Unknown	Unknown	0	19
VC_0000004619	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.02	1
VC_0000004675	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000004720	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	7.15 e-3	1
VC_0000004751	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000004899	Unknown	MC_0000012804	Unknown	Unknown	0	14
VC_0000004930	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	6.34e-3	3
VC_0000005112	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000005191	Unknown	MC_0000022142	Unknown	Unknown	6.52e-08	4
VC_0000005244	Unknown	MC_0000022142	Unknown	Unknown	0	3
VC_0000005248	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000005319	Siphoviridae	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.01	1
VC_0000005430	Unknown	MC_0000001589	*Bacteroidetes*	*Prevotella*	0	10
VC_0000005485	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.03	2
VC_0000005520	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.02	1
VC_0000005538	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000005597	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.01	1
VC_0000005617	Siphoviridae	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.01	1
VC_0000005650	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000005654	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	3
VC_0000005709	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.02	1

VC_0000005720	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	3.81e-3	1
VC_0000005897	Unknown	MC_0000019392	*Candidatus* Saccharibacteria	NA	0	37
VC_0000006135	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000006152	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.01	1
VC_0000006294	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.04	1
VC_0000006351	Unknown	MC_0000032464	Unknown	Unknown	0	46
VC_0000006420	Unknown	MC_0000020170	Unknown	Unknown	0.01	15
VC_0000006515	Myoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000006516	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.03	1
VC_0000006595	Unknown	MC_0000021265	*Candidatus* Saccharibacteria	NA	0	16
VC_0000006598	Unknown	MC_0000021604	*Proteobacteria*	*Bdellovibrio*	7.94e-3	6
VC_0000006606	Unknown	MC_0000019471	Unknown	Unknown	0	10
VC_0000006623	Podoviridae	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.02	1
VC_0000006770	Siphoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.01	1
VC_0000006796	Unknown	MC_0000012195	Unknown	Unknown	0	15
VC_0000006906	Siphoviridae	MC_0000019279	*Actinobacteria*	*Streptomyces*	0	103
VC_0000006962	Unknown	MC_0000000557	Candidatus *Saccharibacteria*	NA	0	7
VC_0000006967	Unknown	MC_0000000003	*Actinobacteria*	*Streptomyces*	0	54
VC_0000007112	Unknown	MC_0000022142	Unknown	Unknown	2.86e-07	1
VC_0000007121	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	0.03	1
VC_0000007271	Unknown	MC_0000022142	Unknown	Unknown	0	5
VC_0000007429	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.01	2
VC_0000007640	Unknown	MC_0000022142	Unknown	Unknown	2.59e-3	5
VC_0000007671	Unknown	MC_0000019267	*Actinobacteria*	*Mycobacterium*	7.80e-3	1
VC_0000007776	Unknown	MC_0000019268	*Actinobacteria*	*Actinoplanes*	0.02	1
VC_0000007875	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.01	1
VC_0000007948	Siphoviridae	MC_0000019267	*Actinobacteria*	*Mycobacterium*	3.72e-3	1

VC_0000008019	Unknown	MC_0000019263	*Proteobacteria*	*Ectothiorhodospira*	0.04	2
VC_0000008273	Unknown	MC_0000019346	*Spirochaetes*	*Spirochaeta*	0	40
VC_0000008351	Unknown	MC_0000020170	Unknown	Unknown	4.32e-3	5
VC_0000008529	Unknown	MC_0000002265	Unknown	Unknown	0	14
VC_0000008713	Siphoviridae	MC_0000019268	*Actinobacteria*	*Actinoplanes*	8.19e-3	1
VC_0000008717	Myoviridae	MC_0000019323	*Proteobacteria*	*Magnetospira*	1.66e-4	18
VC_0000008769	Unknown	MC_0000034457	Unknown	Unknown	0	6
VC_0000008859	Unknown	MC_0000000580	Unknown	Unknown	0	22

2.4.6.3. Gene homology of host-linked viruses to other prokaryotes

Of the 94 host-linked VCs (Table 2.8), only 40% of the VCs (38 VCs) had at least one homologous gene originated from the same genus as their predicted host (Table 2.9). The percentage of these homologous genes varied from 1% to 32%, except for five VC that had > 60% of genes homologous to associated taxa (Table 2.9). These homologous genes of the VCs (VC_0000001445, VC_0000001813, VC_0000001823, VC_0000006595, VC_0000006962) were matched to bacterial ribosomal proteins, indicating potential bacterial contamination. After removal of these VCs, the homologous genes included mostly hypothetical proteins (63%), viral origin genes (24%), including major capsid and portal protein, and bacterial genes required for viral replication (10%)(e.g. transcriptional regulator, DNA helicase) (data not shown).

Table 2.9. Gene homology of the viral contig (VC) with the same genus as their predicted host.

Virus ID	Predicted host by WIsH (genus)	Total number of viral genes	% of viral genes annotated	Number of homologous genes with host genus
VC_0000000009	*Mycobacterium*	37	41	3
VC_0000000149	*Mycobacterium*	33	39	2
VC_0000000155	*Mycobacterium*	41	34	0
VC_0000000215	*Nocardioides*	31	39	0
VC_0000000432	*Actinoplanes*	24	13	0
VC_0000000492	*Mycobacterium*	16	44	0
VC_0000000609	*Candidatus* Saccharibacteria	25	64	8
VC_0000000696	*Mycobacterium*	59	29	0
VC_0000001065	*Streptosporangium*	13	100	0
VC_0000001254	*Streptomyces*	39	59	6
VC_0000001313	*Mycobacterium*	60	37	0
VC_0000001327	*Ectothiorhodospira*	61	57	1
VC_0000001445	*Candidatus* Saccharibacteria	124	94	82
VC_0000001813	*Candidatus* Saccharibacteria	18	78	11
VC_0000001823	*Candidatus* Saccharibacteria	23	87	16
VC_0000001985	*Yersinia*	15	47	0
VC_0000002006	*Actinoplanes*	15	20	0
VC_0000002014	*Ectothiorhodospira*	33	24	0
VC_0000002048	*Bifidobacterium*	9	44	0
VC_0000002081	*Mycobacterium*	41	46	0
VC_0000002138	*Mycobacterium*	18	28	0
VC_0000002186	*Nocardioides*	33	18	0
VC_0000002247	*Actinoplanes*	40	33	0
VC_0000002373	*Mycobacterium*	54	17	0
VC_0000002376	*Mycobacterium*	82	24	0
VC_0000002431	*Streptomyces*	13	77	3
VC_0000002523	*Mycobacterium*	6	17	0
VC_0000002529	*Actinoplanes*	24	58	0
VC_0000002812	*Mycobacterium*	20	50	0
VC_0000002872	*Mycobacterium*	38	55	3
VC_0000002886	*Ectothiorhodospira*	21	62	0
VC_0000002907	*Mycobacterium*	17	47	0

VC_0000002928	*Ectothiorhodospira*	15	60	0
VC_0000002959	*Methylomonas*	25	88	0
VC_0000003006	*Bifidobacterium*	82	80	0
VC_0000003018	*Actinoplanes*	30	17	0
VC_0000003025	*Magnetospira*	46	15	0
VC_0000003212	*Mycobacterium*	58	48	2
VC_0000003242	*Mycobacterium*	29	28	0
VC_0000003267	*Mycobacterium*	87	41	3
VC_0000003600	*Actinoplanes*	12	25	0
VC_0000003871	*Actinoplanes*	28	54	0
VC_0000003942	*Mycobacterium*	26	46	0
VC_0000003953	*Gluconobacter*	86	31	0
VC_0000003968	*Mycobacterium*	27	22	1
VC_0000004002	*Mycobacterium*	111	36	3
VC_0000004080	*Mycobacterium*	22	41	1
VC_0000004153	*Actinoplanes*	36	11	0
VC_0000004207	*Mycobacterium*	25	20	1
VC_0000004219	*Proteobacteria*	24	33	4
VC_0000004470	*Mycobacterium*	45	51	0
VC_0000004518	*Mycobacterium*	23	39	1
VC_0000004529	*Mycobacterium*	3	33	0
VC_0000004619	*Actinoplanes*	35	26	1
VC_0000004675	*Mycobacterium*	59	27	0
VC_0000004720	*Mycobacterium*	30	40	1
VC_0000004751	*Mycobacterium*	7	57	0
VC_0000004930	*Mycobacterium*	23	57	1
VC_0000005112	*Mycobacterium*	34	44	2
VC_0000005248	*Mycobacterium*	20	35	0
VC_0000005319	*Actinoplanes*	31	26	0
VC_0000005430	*Prevotella*	48	81	0
VC_0000005485	*Actinoplanes*	32	13	0
VC_0000005520	*Mycobacterium*	54	48	1
VC_0000005538	*Mycobacterium*	30	20	1
VC_0000005597	*Actinoplanes*	37	30	1
VC_0000005617	*Ectothiorhodospira*	63	79	0
VC_0000005650	*Mycobacterium*	22	23	1
VC_0000005654	*Mycobacterium*	27	41	2
VC_0000005709	*Mycobacterium*	36	28	0
VC_0000005720	*Mycobacterium*	27	33	1
VC_0000005897	*Candidatus* Saccharibacteria	77	73	22
VC_0000006135	Mycobacterium	78	27	1
VC_0000006152	Ectothiorhodospira	27	37	0
VC_0000006294	Mycobacterium	9	44	0
VC_0000006515	Mycobacterium	79	32	4
VC_0000006516	Actinoplanes	23	61	0
VC_0000006595	*Candidatus* Saccharibacteria	39	85	29
VC_0000006598	*Bdellovibrio*	12	58	0
VC_0000006623	*Ectothiorhodospira*	79	41	0
VC_0000006770	*Mycobacterium*	122	33	4
VC_0000006906	*Streptomyces*	99	30	7
VC_0000006962	*Candidatus* Saccharibacteria	8	100	8
VC_0000006967	*Streptomyces*	51	29	12

VC_0000007121	*Mycobacterium*	23	87	0
VC_0000007429	*Actinoplanes*	18	17	0
VC_0000007671	*Mycobacterium*	29	14	0
VC_0000007776	*Actinoplanes*	75	25	0
VC_0000007875	*Ectothiorhodospira*	52	37	0
VC_0000007948	*Mycobacterium*	30	37	1
VC_0000008019	*Ectothiorhodospira*	45	27	0
VC_0000008273	*Spirochaeta*	82	78	1
VC_0000008713	*Actinoplanes*	14	79	0
VC_0000008717	*Magnetospira*	32	22	0

2.5. Discussion

Metagenomics and viromics were compared for the recovery of viral diversity from two soils of contrasting pH. In theory, metagenomics can allow for the identification and genomic characterization of all microorganisms present in a sample, including viruses (Wooley et al. 2010). Of the 125 – 182 million QC reads produced per metagenome, 2 – 3% of the contigs were viral. Previous soil metagenomic studies have reported less than 2% of reads as viral (Goordial et al. 2017; Emerson et al. 2018; Trubl et al. 2018), demonstrating the inefficiency of capturing viral signals from soil metagenomes. On average, the metagenomes produced viral contigs with a mean length of 531 bp in comparison to 21 kb for viromes, resulting in poor viral genome assembly. Short contig sizes also reduce the ability to identify viral hallmark genes and viral-host prediction. Viral contigs < 1 kb often lack major capsid proteins and terminases, which vary between 300 – 600 in amino acid size (Beitbart et al. 2007; Hurwitz et al. 2015; Brum et al. 2015; Roux et al. 2016). One of the most widely used tools for virus prediction, VIRSorter, requires long viral contigs, and virus-host linkage tools that utilize *k*-mer frequency distributions, such as VirHostMatcher, HostPhinder, and Host Taxon Predictor, are more robust when built from longer contigs (> 1 kb) (Khot et al. 2020). Another potential disadvantage of metagenomics is the underestimation of rare viruses or viruses infecting rare host cells due to insufficient sequencing deep, but, advantageously, can bypass the biases caused from size-based selection of viral particles (Roux et al. 2019). In this study, only 8% of the predicted mVCs were recovered in the viromes, with a large proportion of these mVCs (61%) identified as large viruses belonging to the *Mimiviridae* and *Phycodnaviridae* (Maruyama and Ueki 2016). As most free viruses are attached to soil particles, soil metagenomes are likely to select for the actively reproducing and temperate viruses (Trubl et al. 2018), but perhaps also those that are larger in size. In this study, only three replicate metagenomes and viromes were produced per soil pH. To increase the recovery of soil viruses and those with sufficient contig length, a much larger sequencing depth of both metagenomes and viromes would be required (Emerson et al. 2018).

Contrasting pH soils from a continuous soil pH gradient in which the microbial community structure was known to change with soil pH were utilized to investigate the relationship between

microbial and viral community structure. As shown in previous studies that have used this model soil system, microbial community structure was different between the pH 4.5 and 7.5 soil (Nicol et al. 2008; Bartram et al. 2014). The pH 4.5 soil had a relatively greater abundance of *Actinobacteria* (48%), while the pH 7.5 soil was dominated by both *Actinobacteria* (35%) and *Proteobacteria* (31%). Soil pH has previously been shown to correlate with the relative abundance of *Actinobacteria* and *Proteobacteria* (Zeng et al. 2016). The greater abundance of *Actinobacteria* in the pH 4.5 soils may be explained by their capacity to form spores for survival under adverse environmental conditions (Tamreihao et al. 2018). The pH 4.5 soils also had a greater abundance of *Thaumarchaeota*. Previous studies in these soils have detected in the acidic soil greater abundances of the *Thaumarchaeota amoA* gene, a marker gene for ammonia oxidizing archaea (AOA), and demonstrating that AOA versus ammonia oxidizing bacteria (AOB) have a stronger ability to adapt to low pH soils (Leininger et al. 2006; Nicol et al. 2008).

Viral diversity, richness and community structure were different between the contrasting pH soils. Viral diversity and richness were relatively greater in the pH 7.5 soil. Soil pH may directly affect the viral community through virus-mineral and virus-organic matter interaction, which mainly occurs between the virus protein capsids and the soil particles (Dowd et al. 1998). For example, absorption of viruses to soil decreases with increasing soil pH, making virus dissociation and extraction processes easier (Lukasik et al. 2000; Chu et al. 2003; Zhao et al. 2008; Chen et al. 2014). This phenomenon could have contributed to the greater viral richness found in the pH 7.5 soils. It was hypothesized that viral community structure would change less across soil pH than that of the prokaryotic community structure. However, the results did not support this hypothesis as structural variability was greater for viruses than prokaryotes. The proportion of pH specific VCs was 8- and 46-fold greater than the MCs for the pH 4.5 and 7.5 soil, respectively, and 28% of the host-virus linkages were pH specific. This suggests that host-ranges of viruses within these soils may be more narrow than broad.

A large number of CRISPR arrays were found (43,527 arrays). Interestingly, 98% of the spacer sequences were specific to a soil pH, indicating distinct dynamic viral populations in the pH soils. The spacer sequences (42 spacers) that were broadly present across the pH soils may represent a viral population that is continuously present in the soil, and therefore the host population might be under selection pressure to maintain these CRISPR spacer sequences (Weinberger et al. 2012; Nasko et al. 2019). There was a lower number of CRISPR arrays in the pH 7.5 soil, but this may be explained by the prevalence of prophages in these soils (pH 4.5, 112 prophages; pH 7.5, 315 prophages)(Wang et al. 2020). The different number of CRISPR arrays and spacers between the pH soils may have arisen following different selection pressures in the neutral versus acidic environments (Hargreaves et al. 2014). Furthermore, relatively large CRISPR arrays (> 1kb) were found in the pH 4.5 than pH 7.5 soil (< 500 bp). It has been suggested

that maintaining a large CRISPR array size is a high physiological cost for a prokaryotic cell but allows for better protection against abundant, diverse and evolving viruses (Snyder et al. 2010; Martynov et al. 2017). High viral diversity, possibly due to greater mutation rates engaging in frequent recombination and reassortment, creating novel genotypes, makes older spacer sequences useless (Martynov et al. 2017). This may be why shorter CRISPR arrays were found in the pH 7.5 soils, which had greater viral diversity.

Although microbial and viral community structure was influenced by soil pH, functional diversity was similar between the two pH soils. This may indicate a level of functional redundancy independently of soil pH, and the broadly present abundant phyla would display the greatest functional diversity (Wei et al. 2016). This metabolic plasticity between two discrete pH soils may explain a central role for the broadly present abundant taxa as keystone species that perform the majority of ecological transformations and facilitate the development of community richness (Chan et al. 2013). As expected, functional profiles of the metagenomes and viromes were distinct, but with strong consistency with conventional viral functions, such as mobilome:prophages, transposons functions, transcription, replication, recombination and repair, cell wall, membrane, envelope biogenesis, which are critical for the reproduction and survival of viruses (Jin et al. 2019).

A large number of viral AMGs were identified in the soil viromes. The adaptive significance of most soil viral AMGs remains unclear (Crummett et al. 2016). There might exist two types of AMGs, common and less common AMGs (Tettelin et al. 2008; Sullivan et al. 2010; Polz et al. 2013; Cordero and Polz 2014). Common AMGs among various lineages of hosts may encode metabolic functions that are essential under a range of conditions, whereas less common AMGs may adapt for only a particular set of conditions (Crummett et al. 2016). Overall, 2,071 protein families of potential AMGs were identified, and 80% were unique to a pH soil (pH 4.5, 34%; pH 7.5, 46%). Of particular interest, viral genes belonging to 19 glycoside hydrolase (GH) families, GH superfamily (0.13%) and peptidases (0.11%) were found. AMGs encoding GH and peptidases are widespread in soil, and have been previously suggested to be an important component in carbon cycling in soils (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019). The GH superfamily was found in both pH soils, and consistent with previous soil virus studies, GH 5 (cellulases, endomannases, and related enzymes), GH 19, GH 25 (lysozymes) and GH 26 (endomannases) were found to be broadly distributed. The pH 7.5 soil viruses had a greater diversity of GH families (9 families) than the pH 4.5 soil viruses (6 families), suggesting core AMGs encoded by specific soil pH viral populations. Similarly, the same pH trend was found for peptidases. These broadly presented GH families and peptidases seem to be a ubiquitous feature of soil systems and play roles in soil decomposition processes that are beyond typical viral lysis functions (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019). Several AMGs encoding for membrane transporter proteins were discovered. Virus-encoded transporter proteins have previously been reported to be common in

viral genomes (Greiner et al. 2018). However, their functional role in virus infection and replication are still unknown (Greiner et al. 2018). AMGs encoding for ABC transporters were found to be slightly more abundant in the pH 4.5 than 7.5 virus populations, and other membrane transporters, such as potassium and ammonium were also found. Results suggest that viral core AMGs of the pH 4.5 and 7.5 soils may be adapted for acidic and neutral environments, respectively (Crummett et al. 2016).

Oligonucleotide frequency analysis is an alternative approach for linking viruses with hosts when fragmented viral and host genomes lack CRISPR arrays. The WIsH tool uncovered various host-virus linkages in the soil (Galiez et al. 2017). Only 1% of the predicted host-virus linkages had homologous genes, but as viral and host genomes were partial this is not surprising. Most of the gene homologs were unknown or annotated as uncharacterized hypothetical proteins, but annotated genes, such as those encoding DNA replication (DNA polymerase, DNA-binding domain) involved in the production of phage progeny were found. Interestingly, a VC linked with mycobacterium had a homologous WhiB-like protein. Experimentally, a WhiB-like protein identified in a mycobacteriophage was shown to regulate host septation and cause superinfection exclusion (i.e. exclusion of secondary viral infections) (Rybniker et al. 2010). Overall, the gene homology analysis demonstrated gene sharing and the potential impact of viruses on host functioning.

2.6. Conclusion

Deep sequencing of metagenomes and viromes provided insight into the microbial and viral taxonomic and functional diversity and virus-host interactions in contrasting pH soils. Soil pH influenced both microbial and viral community structure, but the viral community changed more relative to the microbial community, suggesting viruses do not have host ranges greater than the constraint of host community structure imposed by soil pH. Distinct soil pH linked virus-hosts were revealed, but also predation of common soil phyla was exhibited. A large number of CRISPR arrays were uniquely present in the pH soils, demonstrating that viruses play an important role in predation of soil prokaryotic communities and therefore in controlling host abundances and compositions. Widespread common AMGs and specific core soil pH AMGs were found, supporting that viruses are significant contributors to major biogeochemical cycles, such as carbon, but also may be involved in adaptive mechanisms for acidic and neutral environments.

CHAPTER III

Diversity and abundance of viral populations across a soil pH gradient that infect an individual host

3.1. Abstract

Soil viruses have potential to influence microbial community structure and subsequent ecosystem functioning by directly affecting the abundance of host cells by lysis and through their ability to transfer genes between hosts. However, in contrast to marine environments, an understanding of the extent to which virus-bacterial host interactions regulate soil bacterial populations is lacking. While viruses will have the ability to infect a range of hosts in highly diverse bacterial soil communities, coevolutionary processes may still tightly control the susceptibility of hosts through virus-bacterial interactions and local adaptation within distinct ecological niches. This work tested the hypothesis that host bacteria are more susceptible to infection from co-localized virus populations in soil. Virus-bacterial host interactions were investigated across a continuous soil pH gradient that has been maintained for over 50 years, and which have different prokaryotic communities at pH 4.5 and 7.5. A bacterial strain (*Bacillus* sp. S4) was isolated from pH 7.5 soil and virus enrichments obtained from pH 4.5, 5.5, 6.5 and 7.5 soils were applied to the host bacteria with infectivity quantified using a plaque assay approach. The results demonstrated that infectivity (number of plaque-forming units (PFUs)) was greatest when viruses and host bacterium were not co-localized, with an increasing number of PFUs recovered with decreasing pH. Transmission electron microscopy of plaque-forming phages and hybrid sequencing data demonstrated that changes in the number of PFUs across the gradient also corresponded with changes in virus morphology, diversity and genetic composition. Six putative prophage sequences were identified in the host genome, but only very low numbers of sequences were associated with these candidate lysogenic viruses, consistent with background contamination of the host genome in the filtered virome. Evidence of coevolution between the host and viruses was demonstrated through identification of restriction modification and CRISPR-*Cas* systems, and spacer mutation or virus-encoded methyltransferase. These findings provide evidence for local adaptation in natural populations, and that virus-bacterial host interactions play an integral part in the susceptibility of a host to infection, and consequently may play an important role in regulating soil bacterial populations.

3.2. Introduction

Viruses are recognized as important drivers in influencing microbial communities by controlling host abundance or metabolic potential by the transfer of genes (Fuhrman, 1999; Weinbauer, 2004; Suttle 2007). They are ubiquitous and diverse in soils (Trubl et al. 2018) with evidence of soil viruses contributing to biogeochemical cycling by host metabolic manipulation after viral infection, via the expression of virus-encoded auxiliary metabolic genes (AMGs) (Emerson et al. 2019). While there is an increasing understanding of the scale of viral diversity and their importance in soils, there is relatively little known about host-virus interactions in soils compared

to marine environments, due to the complexity of both soil structure and microbial diversity (Williamson et al. 2017).

Viruses are able to infect a range of hosts in highly diverse soil bacterial communities (Segobola et al. 2018; Trubl et al. 2018; Graham et al. 2019), and virus-host interactions constitute a major determinant of host evolution and ecology (Mojica et al. 2009; Koskella 2014). Antagonistic virus-host coevolution is well documented in marine environments (Wommack and Colwell 2000; Martiny et al. 2014). This process is considered a main force driving the genetic diversity within host and parasite populations (Best et al. 2009; Paterson et al. 2010; Koskella 2014). The host and their viruses generally undergo antagonistic coevolution where a host cell evolves rapidly against viral infection, and subsequently the virus evolves counterdefenses in response (Weinbauer and Rassoulzadegan 2004; Martiny et al. 2014). Thus, reciprocal selective pressure may lead to rapid reciprocal adaptation of both virus and host (Martiny et al. 2014). For example, an increase in mutation rate was observed in a bacterial population evolving in the presence of phages compared to bacterial populations without phages (Pal et al. 2007; Paterson et al. 2010). This antagonistic virus-host coevolution is generally termed as the red queen hypothesis (Valen 1973; Stern and Sorek 2011). However, when bacterial mutations evolve more rapidly, and viruses are incapable of infecting the mutator bacteria, red queen dynamics may be avoided (Morgan et al. 2010). A viral nonstable state may be altered by the arrival of broad host-range viruses able to infect a range of different prokaryote hosts (Ross et al. 2016).

The arms race between virus and host occurs during virus replication, virus adsorption to cell receptors, and the entry of viral DNA into the host cell (Golais et al. 2013). During virus replication, the host cell acquires antiviral defense mechanisms, such as the restriction-modification system, the clustered regularly interspaced short palindromic repeats (CRISPR) loci with associated *Cas* genes, and abortive infections (Deveau et al. 2010; Labrie et al. 2010; Stern and Sorek 2011; tenOever 2016). Among these antiviral defense mechanisms, the CRISPR-*Cas* system and restriction-modification are well characterized. In restriction-modification, phage DNA is recognized as foreign, as it is not methylated like prokaryotic DNA, and is thus cleaved by restriction endonucleases (Wilson and Murray 1991; Tock and Dryden 2005; Vasu and Nagaraja 2013). Viruses subsequently evolve counterdefense mechanisms to evade these antiviral defense mechanisms (Golais et al. 2013; tenOever 2016). For example, some viruses can modify the restriction sites incorporating unusual bases like 5-hydroxymethyluracil instead of thymine in their genomes (Krüger and Bickle 1983). Some viruses can code for their own methyltransferase or stimulate the production of host methyltransferase, code for proteins binding to restriction sites or mimic DNA proteins that can neutralize endonuclease action (Stern and Sorek 2011; Golais et al. 2013). Viruses can also mutate the targeted spacer sequence or phosphorylate the *Cas* proteins to evade the CRISPR-*Cas* system (Horvath and Barrangou 2010; Golais et al. 2013).

Antiviral defense mechanisms and viral counterdefense mechanisms may result in lysogeny (Golais et al. 2013). Unlike lytic viruses, temperate viruses that enter the lysogenic cycle maintain a long-term association with their host cell (Feiner et al. 2015). Temperate viruses are integrated into the host chromosome, becoming prophages that are replicated along with the rest of the host genome when the cell divides. The maintenance of host-prophage relationships depends on the cost of prophage carriage, the risk of cell death and physiologically maintaining a large prophage genome (Harrison and Brockhurst 2017), and the benefits of enhanced host fitness, through horizontal gene transfer, protection by lytic infection, and lysis of competing strains through prophage induction (i.e. lysogenic conversion)(Feiner et al. 2015; Harrison and Brockhurst 2017).

Fitness of a lytic virus depends on the density of local susceptible hosts, and there should be a strong selection on virus populations to adapt to local hosts (Koskella and Brockhurst 2014). Thus, lytic viruses should be better at infecting hosts from the same ecological niche relative to those from a different niche (Gandon et al. 2008). Although local adaptation of viruses to distinct ecological niches has been demonstrated (Gorter et al. 2016), viruses often have minimal effects on bacterial population densities due to the rapid evolution of a host (Lenski and Levin 1985; Kawecki and Ebert 2004; Koskella 2014). However, viruses can have significant effects on bacterial population dynamics, diversity and bacterial fitness, and virus-host local adaptation can result in the maintenance of diversity (Bohannan et al. 2002; Kawecki and Ebert 2004; Rodriguez-Brito et al. 2010; Kuno et al. 2012).

To gain a better understanding of the extent to which virus-bacterial host interactions vary over an ecological gradient, the susceptibility of one host bacterium to infection from co-localized viruses compared to viruses from increasingly different ecological niches was determined across a soil pH gradient using a plaque assay approach. Virus-bacterial host interactions were investigated across a continuous soil pH gradient that has been maintained for over 50 years, and which have very different prokaryotic communities at the extremes of the gradient at pH 4.5 and 7.5. It was hypothesized that a host bacterium would be more susceptible to infection from co-localized virus populations in soil due to co-evolutionary processes and local adaptation to the host. Secondly, it was hypothesized that the diversity of co-localized viruses infecting an individual host would be lower compared to those from a different environment due to co-evolution and selection. This was investigated through transmission electron microscopy and hybrid metagenomic sequencing.

3.3. Materials and Methods

3.3.1. Soil sampling and physicochemical analyses

Soil was collected from the Craibstone Research Station, SRUC, Aberdeen, Scotland (57°11, 2°12) (see Chapter I, section 1.3, for site details). Triplicate surface soil samples (top 10 cm) from four discrete pH sub-plots, pH 4.5, 5.5, 6.5 and 7.5, were collected at 1 m intervals on 11th January 2019. Soil samples were sieved (2 mm) and stored at 4°C before use. Information on soil pH and moisture content is located in Supplementary Table 3.1, and methods were previously described in Chapter II, section 2.3.1.

3.3.2. Isolation of plaque-forming bacteria

In total, 120 bacterial strains were isolated on 1/10 tryptone soy agar plates (TSA) from pH 4.5 and 7.5 soil (60 strains from each). In 15 ml sterile centrifuge tubes, 1 g of sieved soil was mixed with 9 ml of 1X phosphate-buffered saline (PBS). The soil suspension was serially diluted (10^{-1}, 10^{-2} and 10^{-3}), and 100 µl of each dilution spread over the surface of a TSA plate in triplicate. Plates were incubated at 25°C for 3 days. Bacterial strains were individually isolated onto new TSA plates, and then cultured on 1/10 tryptone soy broth (TSB) (Sigma Aldrich, St. Louis, MO, USA) liquid medium, and preserved as glycerol stocks stored at –20°C.

A preliminary plaque assay experiment was performed in order to verify the capacity of each bacterial strain to form a fully confluent monolayer of growth across the plate and their susceptibility to infection of enriched virus populations from mixed soil (1:1, pH 4.5 and 7.5). In summary, only six and nine isolated bacterial strains from pH 4.5 and 7.5 soil, respectively, were able to form plaques after infection from enriched virus populations. The taxonomy of these 15 bacterial strains was determined by Sanger sequencing of the 16S rRNA genes (EurofinGenomics, Ebersberg, Germany). PCR products were derived from direct amplification from the bacterial colonies with the primer pair PA/PH (Edward et al., 1989). Each 50 µl PCR reaction contained 10X buffer, 0.4 µM of each primer, 1 µl of 100 mM dNTP, 0.2 µl of Taq DNA polymerase (Thermo Fisher, Carlsbad, CA, USA), and PCR grade water. The thermal cycling program consisted of an initial denaturation step for 10 min at 95°C, followed by 30 cycles of amplification (30 s at 95°C, 30 s at 55°C, and 30 s at 72°C) in a thermocycler (Biometra, Göttingen, Germany). The PCR products of 16S rRNA genes of all strains were blasted against the NCBI nr database, and all strains possessed > 98% identity to known *Bacillus* strains, except one strain matching to a *Rhodococcus* sp. Of the 15 bacterial strains identified, a *Bacillus* sp. strain isolated from pH 7.5, designated S4, was selected for further study. This strain was chosen as it demonstrated infectivity (i.e. plaque formation) by enriched virus populations at each pH.

3.3.3. Enrichment of virus populations

Bacillus sp. S4 was grown at 29°C in 40 ml of 1/10 TSB medium overnight and used to enrich lytic virus populations of each pH soil (pH 4.5, 5.5, 6.5 and 7.5). The lytic virus populations were enriched from 0.2 g dry-weight equivalent of soil from each pH soil, in triplicate, by adding 100 µl of overnight grown *Bacillus* sp. S4 culture to 1 ml of liquid 1/10 TSB medium in 15 ml centrifuge tubes (n = 12). Cells were then incubated at 29°C overnight without shaking. To make initial virus population stocks for each soil pH, the supernatant containing the enriched virus populations for each soil pH were filtered through a 0.2 µm syringe filter. The enriched virus populations were diluted 10^{0}, 10^{-1}, 10^{-2} and 10^{-3} stock (n = 48) with phage buffer (100 mM NaCl, 8 mM $MgSO_4$, 50 mM Tri-HCl, pH 7.5 and 0.02 µm filtered).

3.3.4. Plaque assay

The plaque assay approach was first optimized to ensure that there was no direct impact of soil pH on the growth of *Bacillus* sp. S4. The growth of *Bacillus* sp. S4 was monitored after the addition of 4.5 or 7.5 soil after addition during exponential or stationary stages of growth (Supplementary Figure 3.1). First, the optical density (OD) was adjusted to 0.05 in 500 ml of 1/10 TSB medium and 40 ml of the adjusted culture was transferred into sterile 50 ml centrifuge tubes. During the exponential phase after 4 hours (OD = 0.5 – 0.7), or after overnight growth (OD = 0.5 – 0.7), 25 ml of the culture was transferred into 50 ml tubes containing 5 g of either pH 4.5 or 7.5 soil, in triplicate. A negative control (TSB medium without *Bacillus* sp. S4) was also established in triplicate. OD was subsequently measured at one-hour intervals for 24 hours using a NanoPhotometer (Implen, Inc. CA, USA). Subsequently, the drop plate method was performed for enumerating cell numbers (Naghili et al. 2013). After 24 hours, the supernatant from each tube was diluted, 10^{-5}, 10^{-6}, 10^{-7} and 10^{-8} (n = 60), and 100 µl of each dilution was carefully added onto 1/10 TSA plates. The plates were dried for 30 min, incubated at 29°C overnight and then the number of colony-forming units (CFU) counted. There was no significant difference in CFU between the two growth stages and between soil pH (data not shown).

The virus population stocks enriched from pH 4.5, 5.5, 6.5 and 7.5 soils were applied to *Bacillus* sp. S4 culture and infectivity was quantified using a plaque assay approach. For this, 500 µl of late exponential phase *Bacillus* sp. S4 culture was added to 15 ml centrifuge tubes containing 4.5 ml of 1/10 TSB soft agar (3 g l^{-1}) with either 10 µl of the initial virus population stock (10^{0}), or diluted 10^{-1}, 10^{-2} and 10^{-3} stock, or 1/10 TSB medium as negative control. Samples were incubated for 5 min, allowing virus populations to attach to host cells. The host-virus solutions were then poured on the top of 1/10 TSA plates, completely dried, sealed with parafilm and incubated at 25°C for two days. A total of 60 plates were produced; three replicates for each virus treatment (soil and dilution) and a negative control. The 10^{-3} dilution plates were used to determine plaque

forming units (PFU), as they produced countable PFUs for all treatments. The enriched lytic virus populations for each pH soil were collectively washed from plates that had the greatest number of PFUs. Specifically, 2 ml of phage buffer was added onto the surface of plates, agitated at 50 rpm at 25°C for 2 hours and then the buffer was removed before filtering (0.2 µm) to remove *Bacillus* sp. S4 cells.

3.3.5. Visualization of virus populations by transmission electron microscopy

The virus populations from the extremes of the pH gradient, pH 4.5 and 7.5, were prepared for transmission electron microscopy (TEM) in triplicate. The six samples were absorbed for 2 min at room temperature onto TEM nickel grids with a mesh size of 200 and coated in formvar-C. The grids were stained with 2% phosphotungstic acid for 2 min and viewed with a transmission electron microscope (JEOL 1400 JEM, Tokyo, Japan) equipped with a Gatan camera (Orius 600, Pleasanton, CA, USA) and Digital Micrograph Software (Gatan, Pleasanton, CA, USA).

3.3.6. Hybrid sequencing of host bacterium and virus populations

DNA extraction was carried out in duplicate on the overnight grown *Bacillus* sp. S4 strain. The strain was grown at 29°C in liquid 1/10 TSB medium in a 15 ml centrifuge tube overnight. After centrifugation at 6,000 x *g* for 10 min and removal of supernatant from the pellet, 500 µl of TE buffer was added and the cells transferred into 2 ml microcentrifuge tubes. Bacterial cells were lysed by incubating with 100 mg ml^{-1} lysozyme at 37°C for 30 min. The lysate was then incubated at 56°C for 1 h after the addition of 10% sodium dodecyl sulfate (SDS) and 10 mg ml^{-1} of proteinase K (Promega, Madison, WI, USA). One hundred µl of 5% CTAB in 140 mM phosphate buffer was added and incubated at 65°C for 10 min. An equal volume of 24:1 (v/v) chloroform/isoamyl alcohol was added to the lysate and vortexed. Samples were centrifuged for 2 min at 4°C at 16,000 x *g*, and the supernatant added to an equal volume of 25:24:1 (v/v/v) phenol/chloroform/isoamyl alcohol and vortexed. The samples were then centrifuged at 16,000 x *g* for 2 min at 4°C. The supernatant was transferred to a 1.5 ml microcentrifuge tube and two volumes of PEG8000 in 1.4 mM NaCl solution added. The samples were incubated at 4°C for 2 h, then centrifuged for 10 min at 4°C 16,000 x *g*. Samples were then centrifuged at 16,000 x *g* for 2 min at 4°C. The PEG solution was removed, and pellets washed by adding 1 ml of 70% ice-cold ethanol. After centrifugation at 16,000 x *g* for 2 min at 4°C the ethanol was removed, and the pellet was dried at room temperature. The pellet was resuspended in 50 µl of nuclease-free water in 1.5 ml microcentrifuge tubes.

Viral DNA was extracted from the 12 filtered samples from the plaque assay plate washes (i.e. lytic virus populations). Viral particles were treated with DNase I to remove free DNA (RQ1 DNase, Promega, Madison, WI, USA) and subsequently lysed by adding 0.5% SDS and 100 ug ml^{-1}

of proteinase K. The lysate was incubated at 37°C for 1 h, followed by CTAB extraction, chloroform separation and PEG precipitation as previously described.

The quality of extracted DNA was assessed by agarose gel and quantified using Qubit dsDNA HS Assay kit (Thermo Fisher, Carlsbad, CA, USA). The DNA of each sample was normalized to 0.25 ng μl^{-1}, and triplicate DNA samples from *Bacillus* sp. S4 and the 12 virus samples were sequenced separately using the Illumina MiSeq sequencer platform (Illumina Inc). In addition, one DNA sample of *Bacillus* sp. S4 and an equal mixed pooled of DNA from the 12 viral population were sequenced separately on a MinION sequencer (Oxford Nanopore Technologies, Oxford, UK). The Nextera XT Library Prep Kit and 1D Genomic DNA by Ligation Kit (SQK-LSK109) were used for MiSeq and Nanopore library prep, respectively, following the manufacturers' protocol with minor modifications. For MiSeq sequencing, metagenomic libraries were prepared from 1 ng of DNA and sequenced with 2 x 250 cycles. For Nanopore sequencing, the end-prep was performed with 50 µl of DNA sample, 0.5 µl of DNA CS, 7 µl of Ultra II End-prep reaction buffer, and 3 µl of Ultra II End-prep enzyme mix in a 0.2 ml thin-walled PCR-tube. The tube was incubated at 20°C for 5 min and 65°C for 5 min using a thermal cycler. The product was purified with 1X AMPure XP beads (Beckman Coulter, Villepinte, France) and eluted in 31 µl of nuclease-free water. An adapter was ligated to DNA by mixing 30 µl of DNA, 50 µl of NEB Blunt/TA Ligase Master Mix (M0367, New England Biolabs, Ipswich, MA, USA), 5 µl of adapter mix (AMX) and 15 µl of nuclease-free water. The reaction was incubated for 15 min at room temperature. Purification and loading of the libraries were completed according to the manufacturers' protocols and sequenced using the MinKNOW™ (FLO-MNI106) workflow.

3.3.7. Bioinformatic analyses

3.3.7.1. Genomic analysis of host bacterium

Quality filtering of raw MiSeq sequences was performed using the Trimmomatic tool with default parameters (Bolger et al. 2014). Hybrid assembly of the *Bacillus* sp. S4 strain was performed using the Unicycler tool (Wick et al. 2017). The quality-controlled MiSeq reads were used to produce accurate long-read contigs from the Nanopore sequencing and provided the information to scaffold them together. The hybrid assembled contigs and quality-controlled MiSeq reads were used for binning (i.e. genome assembly) through the Metawrap tool in order to estimate the completeness and contamination levels (Uritskiy et al. 2018). Taxonomic classification of the strain was based on marker genes and carried out using GTDB-Tk v0.3.2 (Chaumeil et al. 2020). To investigate the presence of *Bacillus* across the pH gradient, 16S rRNA gene amplicon data derived from the soil pH gradient in 2007 (Bartram et al. 2014) was downloaded from the National Center for Biotechnology Information (NCBI) (SRR988432; https://www.ncbi.nlm.nih.gov/sra/SRX352269), and the sequence reads from each soil pH were

mapped to the 16S rRNA gene of the genome of the *Bacillus* sp. S4 strain using the Seqkit locate function (Shen et al. 2016).

To identify antiviral defense systems within the host genome, functional analysis was conducted using the protein sequences produced from the annotate bin module of the Metawrap tool. Additionally, protein sequences from the host genome were annotated using InterProScan 5 (E value < 10^{-5}), and KEGG database to infer metabolic pathways (Jones et al. 2014). Also, CRISPR arrays analysis was conducted, through screening for CRISPR arrays using the CRT tool (Bland et al. 2007). Spacer sequences from the CRISPR arrays were searched on the viral reads and contigs using Seqkit locate (Shen et al. 2016).

3.3.7.2. Metagenomic analysis of virus populations

Host contamination was checked by remapping the viromic data to the assembled host genome using Salmon mapper v0.9.1 (Patro et al. 2017; Uritskiy et al. 2018). Quality filtering of raw MiSeq sequences and hybrid assembly was performed as previously detailed above (section 3.3.7.1). The hybrid assembled contigs of the virus populations that were greater than 10 kb, henceforth referred to as metagenomic assembled viral contigs (mVCs), were used in downstream analyses. VirSorter was utilized to predict prophages within the host genome (Roux et al. 2015). The mVCs and predicted prophages were compared to reference viral genomes stored in Virus-Host DB (Mihara et al. 2016). Taxonomy of mVCs and prophages was also classified through Kaiju annotation with the NCBI Refseq viral protein database (Menzel et al. 2016). Gene prediction of the mVCs and prophages was completed using Prodigal (Hyatt et al. 2010). Additionally, homology searches were conducted using InterProScan 5 (E value < 10^{-5}) (Jones et al. 2014). To compare genomic regions between the mVCs or the prophages, all-against-all similarity scores (S_G) were computed by tBLASTx (Nishimura et al. 2017), and homology search on the mVCs was performed against the NCBI database (nr/aa) by GHOSTX software through ViPtree (Nishimura et al. 2017). Viral hallmark genes, including major capsid protein, portal, terminase large subunit, spike, tail, virion formation or coat, and viral-like genes were extracted from the GHOSTX output file (Nishimura et al. 2017). The mVCs and prophages were aligned and the similar genomic regions were highlighted through ViPtree (E-value < 10^{-3})(Nishimura et al. 2017).

To determine the distribution and abundance of each mVC and prophage across the soil pH gradient, normalized relative abundance was calculated. The Quant_bins module in Salmon v0.9.1 was used to index the mVCs and prophages and align the reads back to the mVCs and prophage (Patro et al. 2017; Uritskiy et al. 2018). The normalized relative abundance of mVCs and prophages in each sample was calculated based on contig length and coverage. Viral genome abundance was expressed as normalized genome copies per million reads (CPM). This was calculated as follows: the read per kilobase (RPK) was divided by genome length (in kb), RPK were

summed and divided by 1 million to get a scaling factor, and RPK values were divided by the scaling factor to get the CPM. A heatmap was made using the heatmaply package in R to visualize the variation in viral genome abundance across the soil pH gradient (R Core Team, 2019).

3.3.7.3. Analysis of horizontal gene transfer

To investigate gene sharing between the *Bacillus* sp. S4 strain and the lytic virus populations, gene prediction of the host and mVCs was completed using Prodigal (Hyatt et al. 2010). Protein alignment of viral and host origin proteins was conducted with BLASTp, with identity > 30%, E value < 10^{-5} and query cover > 70%. Homologous genes between host and lytic virus populations were annotated using InterProScan 5 (E value < 10^{-5})(Jones et al. 2014).

3.4. Results

3.4.1. Infectivity of virus populations across the soil pH gradient

Plaque-forming units (PFUs) were assessed for the 10^{-3} dilution plates. The size of the plaques on the pH 7.5 virus populations were visibly larger than those obtained with the soil pH 4.5 virus populations (Figure 3.1). The number of PFUs varied between the viral populations of the pH soils (Figure 3.2). Specifically, the infectivity of the soil pH 4.5 virus population (4.8×10^7 PFU ml^{-1}) was on average 600-fold greater than the pH 7.5 virus population (7×10^5 PFU ml^{-1}). Soil pH 5.5 and 6.5 virus populations yielded an average of 1.5×10^7 and 2.8×10^6 PFU ml^{-1}, respectively (Figure 3.2).

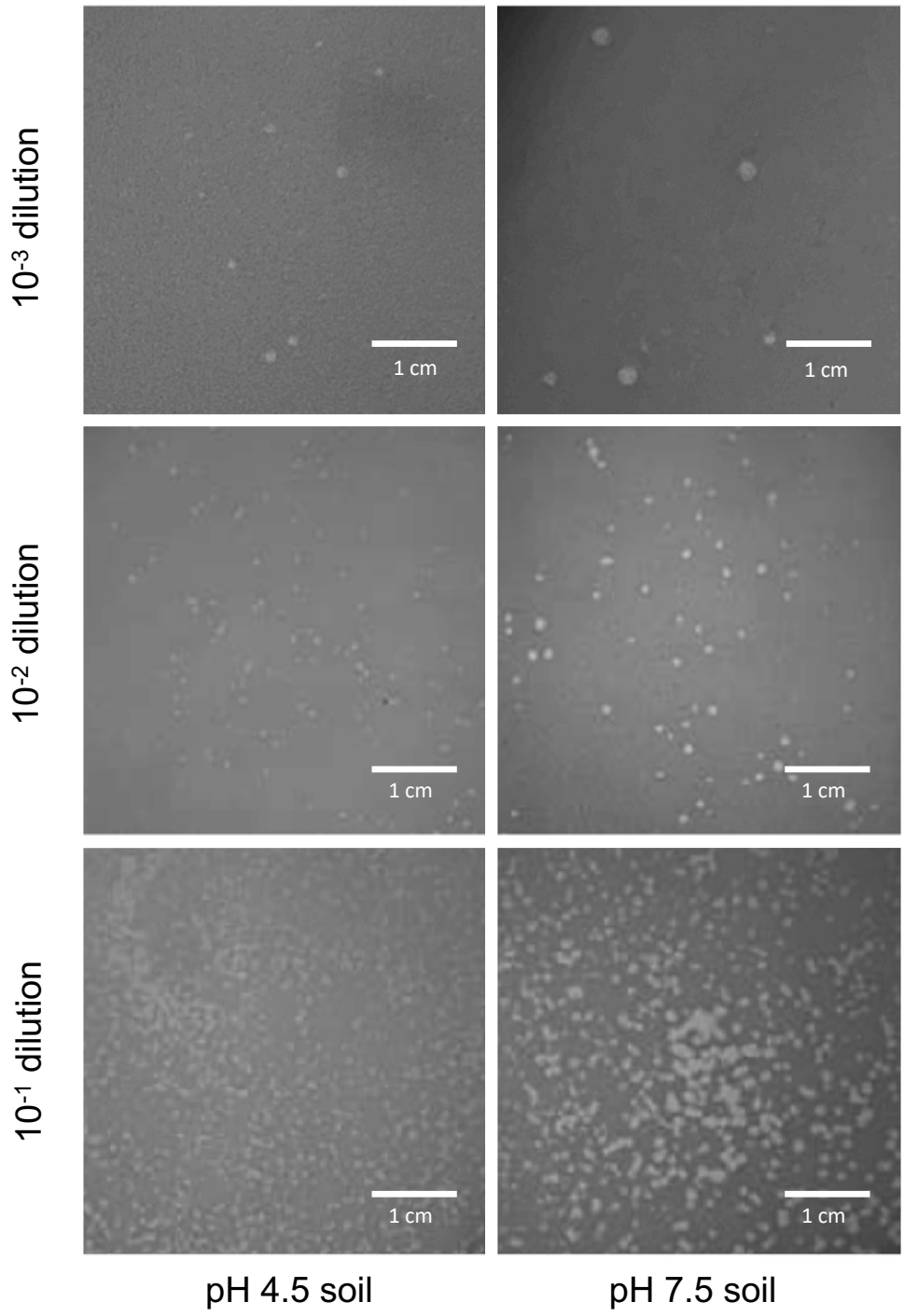


Figure 3.1. Images of plaque forming units (PFUs) derived from 10^{-1}, 10^{-2} and 10^{-3} dilutions from pH 4.5 and pH 7.5 soil.

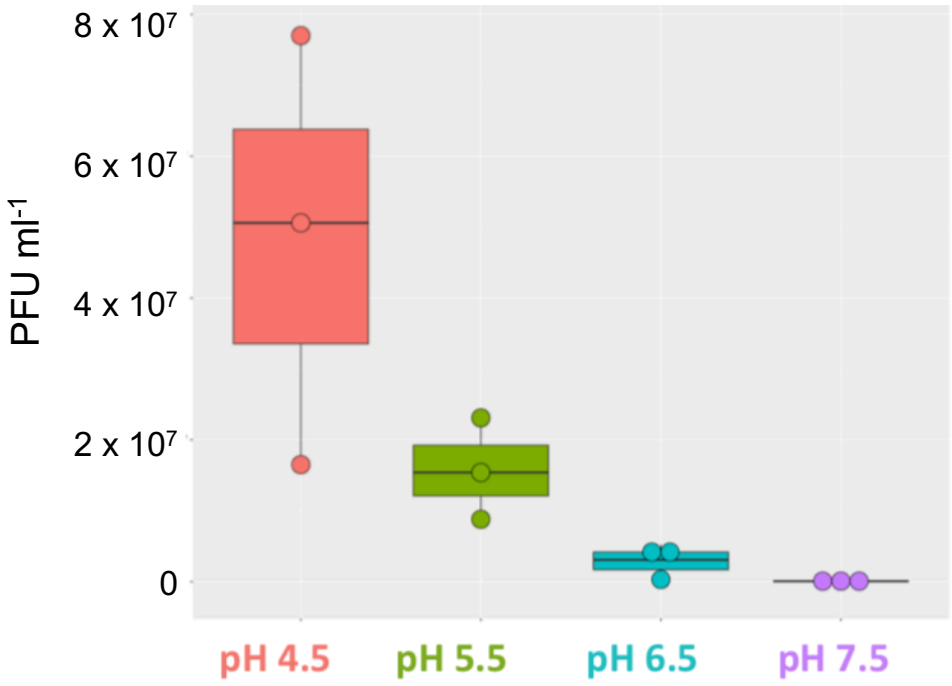


Figure 3.2. Number of plaque-forming units (PFU) derived from the virus populations of pH 4.5, 5.5, 6.5 and 7.5 soil that infected *Bacillus* sp. S4 strain (n = 3).

3.4.2. Morphology of virus populations

The morphology of the virus populations infecting *Bacillus* sp. S4 was different between the pH 4.5 and 7.5 soil (Figure 3.3). Four viral morphotypes were found in the virus populations of the pH 4.5 soil where *Myoviridae* morphotypes dominated (Figure 3.3a.i, ii and iv), with a smaller portion of *Podoviridae* or similar spherical morphotype (Figure 3.3a.iii). A single morphotype was found in the virus populations of the pH 7.5 soil and was identified as a *Siphoviridae* (Figure 3.3b).

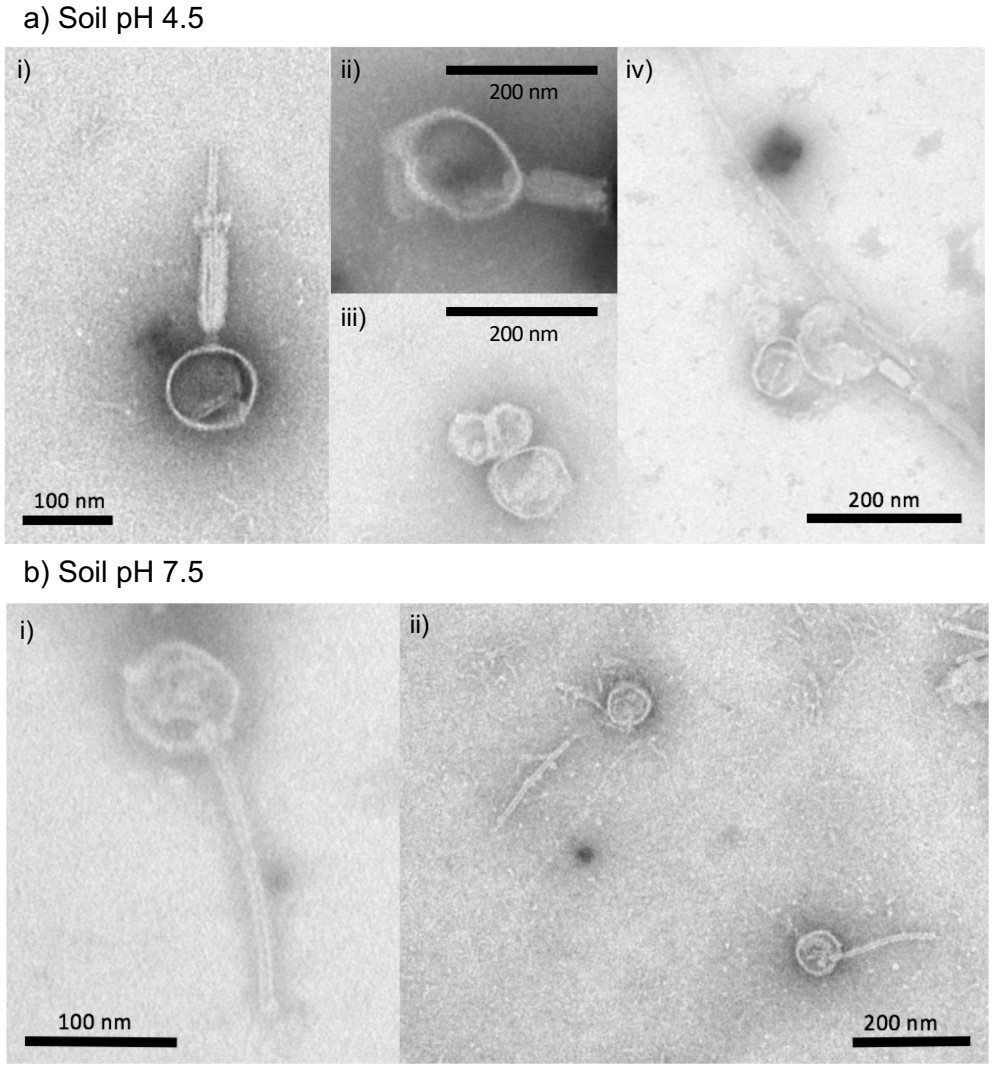


Figure 3.3. Transmission electron microscopy images of viral particles that infected *Bacillus* sp. S4 strain from two soils of contrasting pH. a) pH 4.5 soil; i) *Myoviridae*; ii) *Myoviridae*; iii) *Podoviridae* or spherical morphotype virus; iv) *Myoviridae*; and b) pH 7.5 soil; i) & 2) *Siphoviridae*.

3.4.3. *Bacillus* sp. S4 genome

Assembly of the MiSeq reads from three replicate samples of the *Bacillus* sp. S4 strain produced 735 contigs with an average length of 7,034 bp (Table 3.1). Assembly of the Nanopore sequences of the *Bacillus* sp. S4 strain yielded 12 contigs with an average length of 457,006 bp (Table 3.1). Hybrid assembly of the Nanopore and Miseq data resulted in 18 contigs with an average length of 308,477 bp (Table 3.1). Binning of the hybrid assembled contigs produced a high-quality draft genome, completeness of 98.3% with 0.5% contamination. The estimated genome size is 4,324,007 bp with 35% GC content. In total, 5,727 genes were predicted, and 40.7% of the genes (2,330 genes) were annotated using KEGG (Supplementary Figure 3.2). *Bacillus* sp. S4 had 95% ANI to a *Bacillus cereus* genome. Read mapping of the 16S rRNA gene amplicon data for each soil

pH to the 16S rRNA gene within the genome of the *Bacillus* sp. S4 strain showed that between 0.39 – 1.88% of the reads were mapped across the soil pH gradient (Table 3.2).

Within *Bacillus* sp. S4, two antiviral defense systems were identified: restriction-modification and CRISPR-*Cas* system. A total of 11 genes coding for enzymes involved in restriction-modification systems were found, including glycerate kinase and restriction endonuclease of type II, III and IV (Table 3.3). One complete CRISPR array was screened, and comprised seven direct repeat with a length of 20 bp, and interspaced by six spacers with an average length of 44 bp (Figure 3.4). The array contained CRISPR-associated protein type I, *Cas5* and *Cas3*. Only a single spacer sequence was matched to viral reads, derived from one replicate pH 4.5 and 5.5 sample (Figure 3).

Table 3.1. Hybrid sequence summary for the *Bacillus* sp. S4 strain (host).

Sample ID	Number of sequences (or contigs)	Sum of length (bp)	Minimum length (bp)	Avgerage length (bp)	Maximum length (bp)
MiSeq_Host_1-R1	260,284	41,734,891	35	160.3	251
MiSeq_Host-1-R2	260,284	41,787,066	35	160.5	251
MiSeq_Host-2-R1	270,233	44,347,792	35	164.1	251
MiSeq_Host-2-R2	270,233	44,419,326	35	164.4	251
MiSeq_Host-3-R1	200,016	30,844,555	35	154.2	251
MiSeq_Host-3-R2	200,016	30,889,425	35	154.4	251
Miseq host assembly	(735)	5,170,639	1000	7,034.9	139,968
Nanopore host assembly	(12)	5,484,075	4,505	457,006	1,838,407
Hybrid host assembly	(18)	5,552,599	1,124	308,477	4,324,007

Table 3.2. Read mapping of 16S rRNA sequences from soil pH gradient sample to host 16S rRNA gene.

	pH 4.5	pH 5.5	pH 6.5	pH 7.5
No. of read_1	651,955	1,020,309	1,075,121	835,157
No. of read_2	612,790	971,567	835,412	790,453
No. of mapped read_1	144 (0.02%)	327 (0.03%)	155 (0.01%)	348 (0.04%)
No. of mapped read_2	126 (0.02%)	256 (0.02%)	148 (0.01%)	320 (0.04%)

No., number

Table 3.3. Host genes coding for enzymes involved in restriction-modification systems.

Host gene ID	Signature description	E value
Host_00220	Restriction endonuclease type II-like	2.73e-6
Host_00943	Restriction endonuclease type II-like	2.28e-67
Host_00951	Type III restriction enzyme, res subunit (Helicase/UvrB, N-terminal)	1.90e-5
Host_01357	Restriction endonuclease type II-like	3.32e-39
Host_01555	Type III restriction enzyme, res subunit (Helicase/UvrB, N-terminal)	3.40e-18
Host_02345	Glycerate kinase, restriction-enzyme-like fold	0
Host_02679	HB1, ASXL, restriction endonuclease HTH domain	3.40e-10
Host_02830	Type III restriction enzyme, res subunit (Helicase/UvrB, N-terminal)	2.60e-10
Host_02859	Type III restriction enzyme, res subunit (Helicase/UvrB, N-terminal)	6.70e-10
Host_04126	Type III restriction enzyme, res subunit (Helicase/UvrB, N-terminal)	9.50e-17
Host_05564	Restriction endonuclease type IV, Mrr	2.50e-30

```
CRISPR Array    Range: 13694 - 14101
POSITION        DIRECT REPEAT (DR)      SPACER (S)
--------        ------------------      ------------------------------------------
13694           TTTATATCCCACTACGTTTA    AATAAAACAAATGTACATGTTTCAGAACTACGTGACATGGCCGAG
13759           TTTATATCCCACTACGTTTA    AATAATACAGCAAATGAACCCATCATTGCCCCTATTACAAACATG
13824           TTTATATCCCACTACGTTTA    AATAATACAAATACAAGCTGTTATAATAAATGTACATTGCCATG
13888           TTTATATCCCACTACGTTTA    AATAAAACGCTCAGCGTAAGCATCTACCTCAGCTTGTGTGTTCG
13952           TTTATATCCCACATGGTTCA    AATAAAACAGCAGAATTTTTTGTATCGTCTATATCTATACGGTCA
14017           TTTATATCCCATTACGTTTA    GATAAAACACCGGTCTTTCCTACTTCTATCGCATAAAGGACTTGG
14082           TTTATATCCCACTACGTTTA
--------        ------------------      ------------------------------------------
Repeats: 7      Average Length: 20              Average Length: 44
```

Figure 3.4. Predicted CRISPR array in the genome of the *Bacillus* sp. S4 strain. The CRISPR array is separated by direct repeat (DR) and spacer (S) sequences, and the position and length of the CRISPR loci are shown. The spacer sequence highlighted in blue was matched to a plaque-forming viral sequence.

3.4.4. Virus populations

MiSeq sequencing of the infecting virus populations from the pH soils yielded a total of 20 million reads, ranging between 1,549 - 2 million reads per sample (Table 3.4). Assembly of the Nanopore virome (i.e. pooled viral sample) yielded 2,183 contigs with an average length of 4,552 bp (Table 3.4). The hybrid co-assembly of the Nanopore virome with the Miseq reads resulted in 323 contigs with an average length of 1,460 bp (Table 3.4). Contigs greater than 10 kb were used for downstream analyses and resulted in nine mVCs (Table 3.4). Five of the contigs were merged into a single contig as they were highly similar (% ANI) and clearly represented five genomic

fragments of one phage when aligned to the genome of *Bacillus* virus SPO1, to which it they all possessed > 95% ANI.

Six prophages were also predicted within the host genome (Table 3.5). Their size ranged from 17 to 156 kb with an average length of 62 kb (Table 3.5). The %GC content of the viral genomes (> 40%) was greater than those of the prophages (< 40%) (Table 3.5). The most closely related reference viral genomes to the 11 viruses/prophages were *Bacillus* phages (Figure 3.5). Five of the six prophages clustered together with low similarity and one linked to *Bacillus* viruses of the same size range (35-40 kb) (Figure 3.5). Comparison of the five mVCs showed that there was little (or no) shared sequence identity, except for mVC_3 and _4, with shared genes possessing > 30% identity (Figure 3.6). Alignment of the prophages showed that Prophage_4 and _5 had some similarity between, although was low (Figure 3.7). Gene homology search revealed that all viral genomes contained at least one viral origin hallmark gene and other viral-like genes (Figure 3.7), whereas only Prophage_1, _2 and _3 contained an integrase (Figure 3.8; Supplementary Table 3.2). The mVCs contained tail proteins, whereas only two of the prophages had tail proteins (Supplementary Table 3.2). Two viruses (mVC 1 and 5), had a greater number of uncharacterized proteins (> 60%), compared to mVC_2, _3 and _4 (< 30%) (Supplementary Table 3.2)**.** Proteins involved in viral counterdefense mechanisms, endonuclease (endonuclease type II-like) and methyltransferase (methyltransferase type 12 and methyltransferase domain 25), were found in mVC _1 and _5 and Prophage_4 and _5, respectively (Table 3.6).

The distribution and abundance of the viruses varied across the soil pH gradient (Figure 3.8). Two viruses, mVC_1 and mVC_5, were found across the entire pH gradient, although mVC_1 had relatively lower abundances in pH 4.5 soil (Figure 3.8). Three viruses, mVC_2, _3 and _4 were not found in pH 4.5 soil and only in some of the pH 7.5 soil replicates (Figure 3.8). Sequences from all six prophages were found in viromes although sequence abundance was very low (< 0.001%) indicating that their recovery was consistent with host genome contamination of the filtered virome (Figure 3.8). There was therefore no evidence that these putative prophages were induced, and all plaque-forming viruses were derived from the added soil samples.

Table 3.4. Hybrid sequence summary for the infecting virus populations from the pH 4.5, 5.5, 6.5 and 7.5 soil.

Sample ID	Number of sequences (or contigs)	Sum of length (bp)	Minimum length (bp)	Average length (bp)	Maximum length (bp)
MiSeq_pH4.5_1-R1	1,834,947	314,558,419	35	171.4	251
MiSeq_pH4.5-1-R2	1,834,947	314,784,130	35	171.5	251
MiSeq_pH4.5-2-R1	1,753,904	287,813,719	35	164.1	251
MiSeq_pH4.5-2-R2	1,753,904	288,353,094	35	164.4	251
MiSeq_pH4.5-3-R1	678,512	113,038,125	35	166.6	251
MiSeq_pH4.5-3-R2	678,512	113,172,573	35	166.8	251
MiSeq_pH5.5-1-R1	2,241,809	369,951,252	35	165	251
MiSeq_pH5.5-1-R2	2,241,809	370,284,516	35	165.2	251
MiSeq_pH5.5-2-R1	1,656,251	298,277,236	35	180.1	251
MiSeq_pH5.5-2-R2	1,656,251	298,653,076	35	180.3	251
MiSeq_pH5.5-3-R1	397,238	65,628,207	35	165.2	251
MiSeq_pH5.5-3-R2	397,238	65,698,167	35	165.4	251
MiSeq_pH6.5-1-R1	83,191	14,666,660	35	176.3	251
MiSeq_pH6.5-1-R2	83,191	14,699,505	35	176.7	251
MiSeq_pH6.5-2-R1	1,650,124	290,198,605	35	175.9	251
MiSeq_pH6.5-2-R2	1,650,124	290,494,403	35	176	251
MiSeq_pH6.5-3-R1	55,663	10,289,810	35	184.9	251
MiSeq_pH6.5-3-R2	55,663	10,296,474	35	185	251
MiSeq_pH7.5-1-R1	11,439	2,130,715	35	186.3	251
MiSeq_pH7.5-1-R2	11,439	2,136,986	35	186.8	251
MiSeq_pH7.5-2-R1	1,549	245,267	35	158.3	251
MiSeq_pH7.5-2-R2	1,549	247,986	35	160.1	251
MiSeq_pH7.5-3-R1	4,600	786,175	35	170.9	251
MiSeq_pH7.5-3-R2	4,600	792,714	35	172.3	251
Nanopore virome assembly	(2,183)	9,938,408	168	4,552	57,013
Hybrid virome assembly	(323)	471,765	100	1,460	51,210
Virome assembly > 10 kb	(9)	301,384	10,575	33,487	51,210

QC, quality control

Table 3.5. Lytic viruses that infected the *Bacillus* sp. S4 strain and identified prophages.

Virus ID	Genome size (bp)	GC content (%)	Number of genes
mVC_1	51,210	41.46	80
mVC_2	49,461	42.47	71
mVC_3	41,694	40.30	55
mVC_4	41,569	40.73	56
mVC_5	117,450	40.29	134
Prophage_1	35,176	35.61	46
Prophage_2	50,691	34.35	67
Prophage_3	17,735	38.99	22
Prophage_4	70,338	34.20	90
Prophage_5	156,077	33.37	177
Prophage_6	60,654	32.55	99

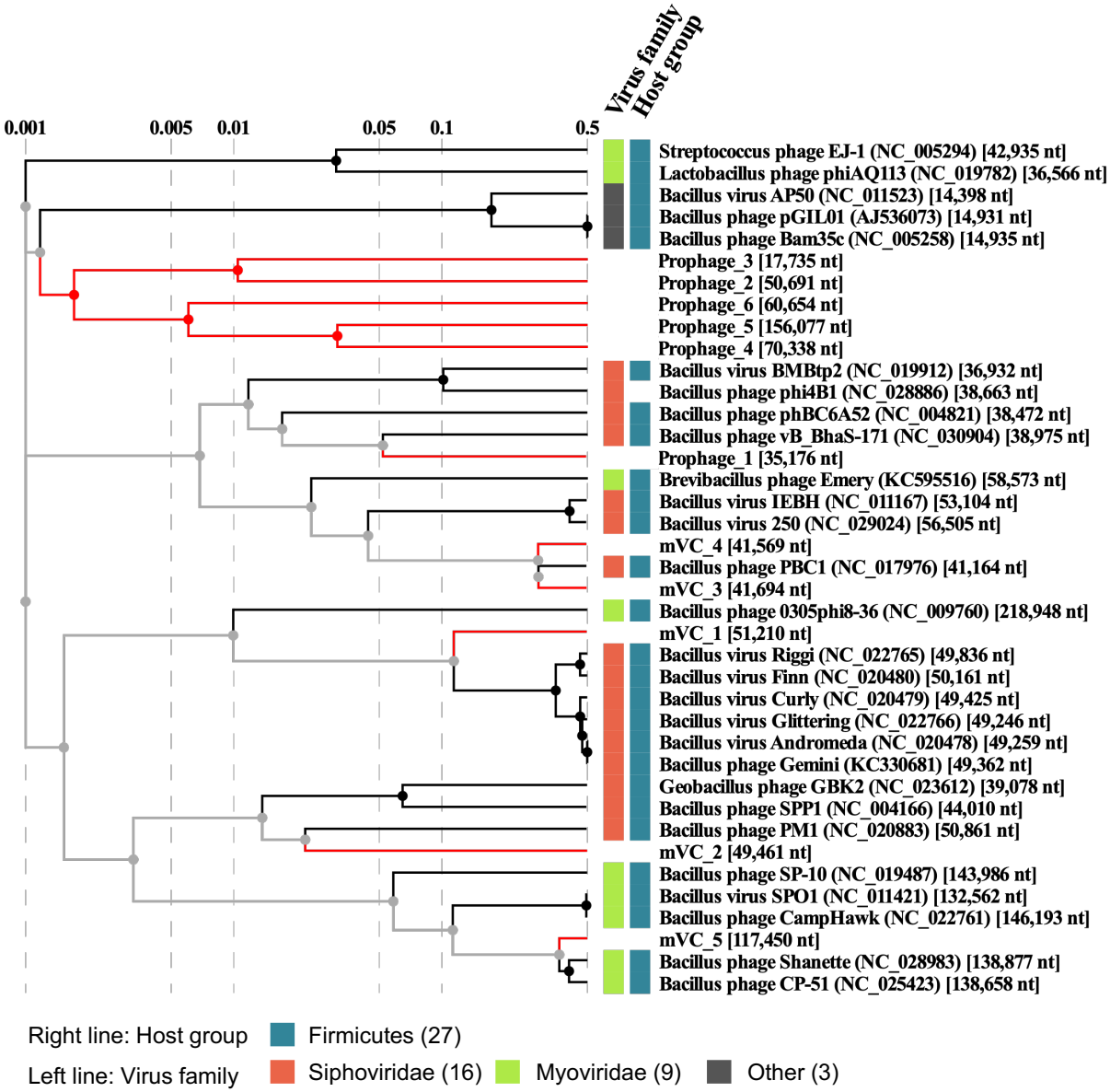


Figure 3.5. Proteomic tree showing the five lytic viruses (mVCs) and six prophages that infected the *Bacillus* sp. S4 and their relationship to reference viral genomes.

Table 3.6. Genes involved in viral counterdefense mechanisms found within viruses infecting *Bacillus* sp. S4.

Virus ID	Signature description	E value
mVC_1_4	Restriction endonuclease type II-like	9.65e-5
mVC_1_41	Restriction endonuclease type II-like	4.08e-5
mVC_5_3	Restriction endonuclease type II-like	8.71e-5
Prophage_4_73	Methyltransferase type 12	5.20e-20
Prophage_5_30	Methyltransferase domain 25	4.10e-21

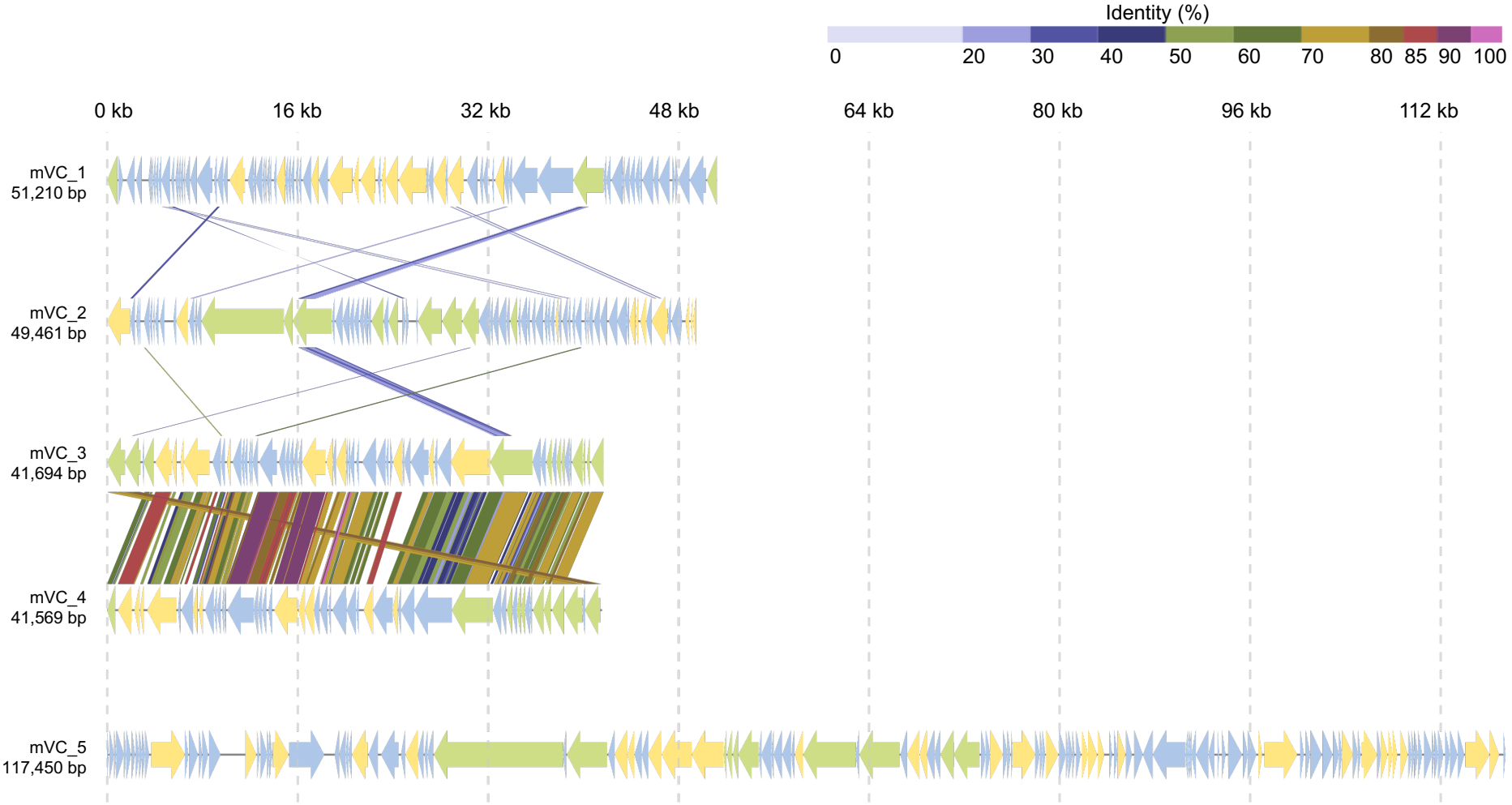


Figure 3.6. Genome map of the lytic virus populations (mVC) of *Bacillus* sp. S4. All alignments are represented by colored lines between the two genomes representing % identity. Blue arrows indicate uncharacterized genes, yellow arrows indicate *Bacillus*-originated genes and green arrows indicate viral genes.

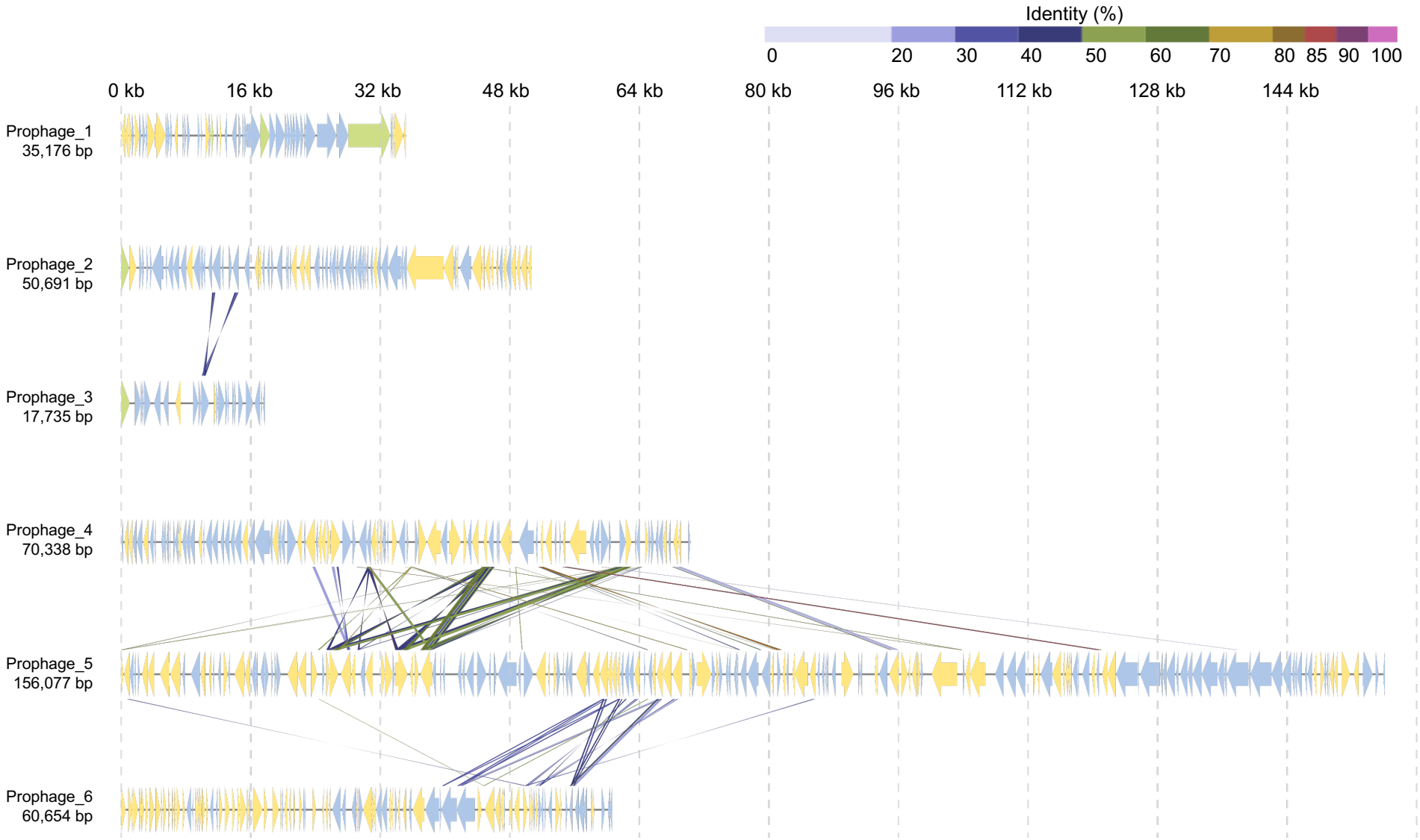


Figure 3.7. Genome map of six prophages identified in the genome of *Bacillus* sp. S4. Lines linking genes from two genomes describe the % identity. Blue arrows indicate uncharacterized genes, yellow arrows indicate *Bacillus*-originated genes and green arrows indicate viral genes.

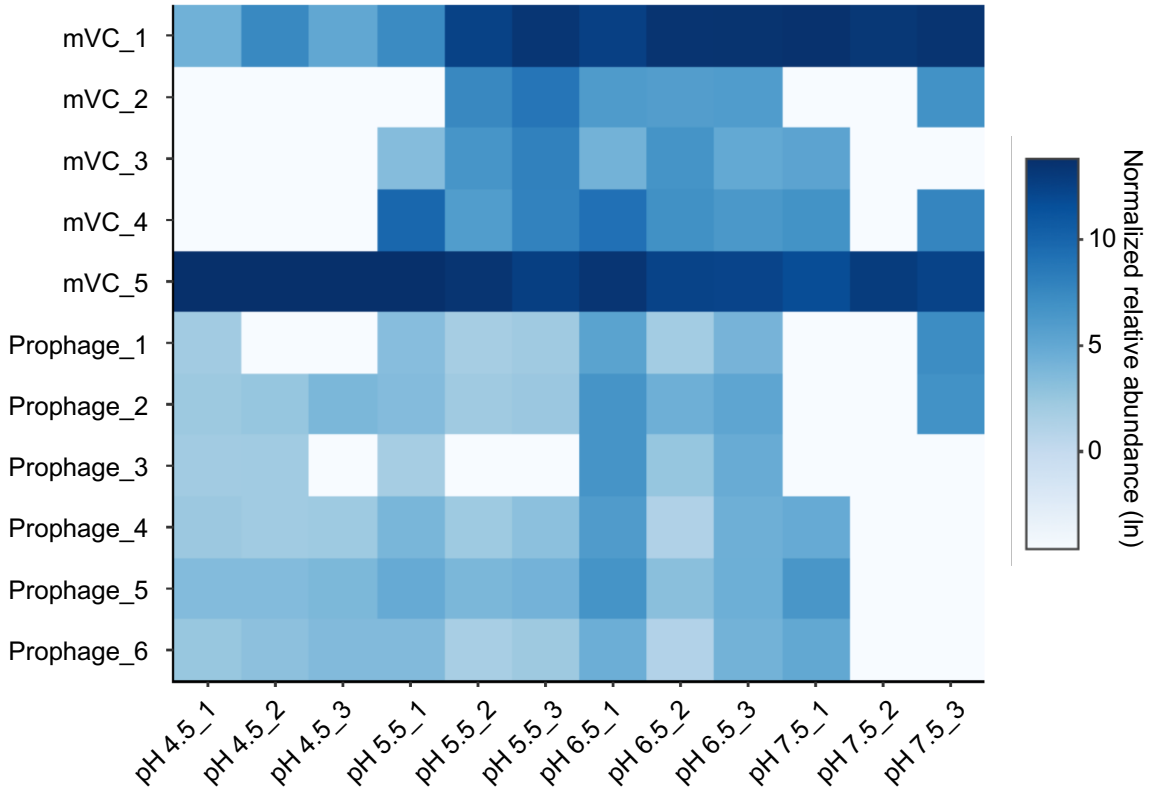


Figure 3.8. Normalized relative abundance of lytic viruses (mVCs) sequences and prophages from plaque assay washes derived from soils sampled across the pH gradient.

3.4.5. Horizontal gene transfer

In total, 5,727 genes were predicted from the *Bacillus* sp. S4 genome and 396 genes from the five lytic virus contigs (mVCs_1 to 5). Comparison of protein sequences between host and mVCs showed that mVC_5 shared the greatest number of genes (9 genes) with the host (Table 3.7). The other mVCs shared a lower number of genes with the host (< 5 genes) (Table 3.8). In particular, shared genes were annotated as DNA translocase, DNA helicase and DNA-binding protein, which are commonly used for viral genome replication. Also, genes that may represent a benefit for viral replication were identified, including thymidylate synthase (cellular dTMP synthesis), deoxyadenosine/deoxycytidine kinase (phosphorylation of several deoxyribonucleosides) and glutaredoxin (virion morphogenesis) (Table 3.7). In addition, bifunctional autolysin, exonuclease, protein deacetylase, stress proteins (SCP2 and 16U) and NAD-dependent protein deacetylase were also identified (Table 3.7).

Table 3.7. Protein assignment of viral genes that are homologous to the *Bacillus* sp. S4 strain.

Viral protein ID	Host protein ID	Id. (%)	E value	bit score	QC (%)	Function
mVC_1_3	Host_00426	37	9.07e-31	110	80	hypothetical protein
mVC_1_22	Host_05318	33	4.74e-26	96.3	88	Replication-relaxation
mVC_1_23	Host_05317	34	5.57e-65	210	85	DNA translocase FtsK
mVC_1_44	Host_03384	31	7.38e-64	218	96	
mVC_1_44	Host_04176	33	1.11e-07	51.2	90	
mVC_2_11	Host_00213	46	3.96e-111	323	98	hypothetical protein
mVC_2_59	Host_03415	32	5.50e-15	67.4	70	
mVC_2_60	Host_00308	36	2.61e-67	208	99	Thymidylate synthase
mVC_2_62	Host_02545	54	1.38e-54	168	99	Single-stranded DNA-binding protein A
mVC_2_65	Host_02550	41	2.65e-100	304	96	Replicative DNA helicase
mVC_2_65	Host_04206	37	9.13e-92	282	97	Replicative DNA helicase
mVC_2_65	Host_01544	33	1.42e-75	240	94	Replicative DNA helicase
mVC_3_40	Host_00358	56	1.09e-38	127	78	
mVC_4_11	Host_04005	33	4.76e-07	41.2	73	Glutaredoxin 3
mVC_4_37	Host_00358	60	5.38e-42	135	78	
mVC_5_33	Host_04242	39	5.57e-58	188	91	Bifunctional autolysin
mVC_5_84	Host_02078	32	1.59e-28	103	93	General stress protein 16U
mVC_5_84	Host_02080	31	1.26e-26	99.0	95	General stress protein 16U
mVC_5_84	Host_02079	31	2.71e-23	90.1	92	General stress protein 16U
mVC_5_85	Host_02077	62	1.80e-96	280	99	hypothetical protein
mVC_5_85	Host_01195	41	3.39e-40	137	72	hypothetical protein
mVC_5_103	Host_00983	37	1.04e-13	58.5	86	DNA-binding protein HU
mVC_5_103	Host_04330	38	2.74e-12	55.1	86	DNA-binding protein HU
mVC_5_103	Host_05552	37	8.74e-11	51.2	86	DNA-binding protein HU
mVC_5_111	Host_00914	32	4.51e-29	110	77	5'-3' exonuclease

mVC_5_111	Host_03401	30	3.24e-25	103	76	5'-3' exonuclease
mVC_5_120	Host_02510	32	2.62e-29	106	87	Deoxyadenosine/deoxycytidine kinase
mVC_5_132	Host_02180	35	1.92e-109	345	94	ATP-dependent DNA helicase PcrA
mVC_5_132	Host_01271	31	1.05e-83	275	93	Putative ATP-dependent DNA helicase YjcD
mVC_5_133	Host_04936	30	3.86e-31	112	94	NAD-dependent protein deacetylase

Id., Identity; QC, Query coverage

3.5. Discussion

This research tested the hypotheses that 1) a host bacterium would be more susceptible to infection from co-localized virus populations in soil due to selection for viruses infecting an endemic strain and 2) the diversity of co-localized viruses infecting an individual host would be lower compared to those from different environments due to co-evolution and selection. A host bacterium, *Bacillus* sp. S4, was isolated from pH 7.5 soil, and infectivity by virus populations across sampled across a soil pH gradient investigated a plaque assay approach. The susceptibility of the host bacterium to infection was found to change across the soil pH gradient, but the number of plaque-forming units (PFU) increased when the host bacterium was infected by the virus populations from an increasingly different pH soil to that from where the host organism was isolated. The diversity of the lytic virus populations was greatest in the pH 5.5 and 6.5 soil, which generally corresponded with the amount of host infection found across the soil pH gradient. Also, there were more morphologically diverse viruses present from the pH 4.5 than 7.5 soil. Overall, the results suggest that infectivity was lowest when viruses and the host bacterium were co-localized, and there was also a lower diversity of viruses capable of infecting the host. The results, therefore, provided support for only one of the two hypotheses.

In a resistance assay experiment that used populations of *Pseudomonas fluorescens* and its associated bacteriophage that co-evolved across a temperature gradient, it was demonstrated that when the bacteria and virus were derived from the same temperature they were more resistant and infectious, respectively (Gorter et al. 2016). However, using a cocktail composed of several different viruses versus a single virus only may have an impact on overall infectivity. It has been shown that diverse virus cocktails were more infective than expected compared to the lytic activities of single virus cocktails, suggesting that synergistic interactions may occur (Schmerer et al. 2014; Weber-Dąbrowska et al. 2016). TEM analysis demonstrated that the virus populations enriched from a different pH soil were more diverse than those enriched from the same soil. Consistent with this, infectivity was greater when the host bacterium was infected by virus populations from non-pH 7.5 soils, perhaps due to synergistic interactions of diverse virus populations (Schmerer et al. 2014; Weber-Dąbrowska et al. 2016).

Although the number of PFUs was greater when the host bacterium was infected by the virus populations from the different pH soil, virus populations exhibited discrete plaques with different sizes. For example, the size of the plaques was greater when the host bacterium was infected by virus populations that derived from the same soil niche. Plaque size can be affected by several factors, including the amount of host and or virus in the top agar layer, lysis time, incubation temperature and medium nutrients, and depending on the characteristics of the viruses, such as the efficiency of adsorption, burst size or virus morphology (Gallet et al. 2011; Kauffman and Polz 2018). Considering that the aforementioned factors were controlled in the plaque assay experiment, the variation in plaque size is more likely attributed to the characteristics of virus populations.

The host genome was analyzed for the presence of prophages to confirm that lytic viruses in plaque assays came from the soil and were not the result of lysogenic phages switching to a cycle. The analysis identified six potential prophages. However, in comparison to the lytic viruses recovered from plaque assays, these candidate viruses had a distinct lack of annotated viral genes, were genetically very distinct from the lytic viruses and may not have represented actual prophage elements in the genome.

To identify potential mechanisms between the host and viruses, host antiviral defense and viral counter-defense systems were investigated. Restriction-modification (RM) and CRISPR-Cas antiviral defense systems were both found in the host genome. The RM systems generally consist of two enzymes, a restriction endonuclease (REase) which cleaves foreign DNA at specific sites and a methyltransferase (MTase), which ensures discrimination between self and foreign DNA by the methylation of specific DNA sequences within the host's genome (Kaltz and Shykoff 1998; Vasu and Nagaraja 2013). There are four different types of RM systems, based on subunit composition, biochemical properties (sequence recognition, cleavage position and substrate specificity) and cofactor requirements (Roberts et al. 2003). Type I systems consist of a hetero-oligomeric protein complex including REase, MTase enzymes and specificity subunit that allows determining the target recognition domains (Wilson and Murray 1991). Compared to the type I system, most type II systems consist of separate REase and MTase enzymes, which independently recognize a target sequence and catalyze reactions (Roberts et al. 2003). While MTase enzymes are well conserved and their target recognition domain can be easily recognized, REase enzymes share much less similarity (Malone et al. 1995; Orlowski and Bujnicki 2008). Type II systems contain *res* and *mod* genes. The *mod* gene encodes for a protein exhibiting modification activity, and the complex of two gene products has restriction enzyme activity (Dryden et al. 2001). Type IV systems consist of a class of enzymes that cleave DNA only when the recognition site is methylated (Roberts et al. 2003). The genome of the *Bacillus* sp. S4 strain comprised three types of the RM systems: type II, III and IV systems. It has been demonstrated that the effectiveness of

RM systems has 10 to 10^8-fold protection against phage infection (Tock and Dryden 2005). The two dominant plaque-forming and broadly distributed lytic viruses (mVC_1 and _5) and two prophages (prophage_4 and _5) encoded genes for Type II-like REase and MTase, respectively. The presence of REase in the viral genomes can confer resistance to infection by more or less unrelated viruses to compete for their hosts (Lossouarn et al. 2019). MTase potentially interferes with the RM host antiviral defense system, therefore improving the infection efficiency of the viruses (Labrie et al. 2010; Koonin and Krupovic 2020; Bezuidt et al. 2020). The presence of REase and MTase in viral genomes seems to be an important mutual benefit in host-virus interactions, which may promote coexistence (Kaltz and Shykoff 1998). Generally, MTase-encoding genes are found in 20% of the currently annotated bacteriophage genomes, suggesting an important role of the gene in virus-host interaction (Kaltz and Shykoff 1998; Murphy et al. 2013). Previous studies have shown that the RM and CRISPR-*Cas* systems frequently co-occur in the host, and their combination results in increased levels of immunity and more rapid spacer acquisition (Hynes et al. 2014; Oliveira et al. 2014; van Houte et al. 2016). Interestingly, the most recently integrated spacer sequence in the CRISPR array came from a lytic virus identified in this study but was not enriched from the pH 7.5 soil. This could be because the co-localized virus populations counter-acted host defense strategies by losing complementarity to the spacers to evade the host CRISPR response (Koonin and Krupovic 2020).

Benefits to the virus and host bacterium through horizontal gene transfer were investigated. The lytic viruses found infecting *Bacillus* sp. S4 had homologous proteins involved in nucleotide metabolism, such as thymidylate synthase (i.e. cellular dTMP synthesis), deoxyadenosine/deoxycytidine kinase (i.e. phosphorylation of several deoxyribonucleosides) and glutaredoxin (i.e. cofactor for viral ribonucleotide reductase)(Rajagopal et al. 1995; Müller et al. 1998; Coulibaly et al. 2015). Other homologous proteins involved in viral replication, such as DNA translocase, DNA helicase and DNA-binding protein, which may represent a benefit for viral replication, were also identified. In addition, bifunctional autolysin, exonuclease, protein deacetylase and stress proteins (SCP2 and 16U) were present. In general, autolysins are peptidoglycan hydrolases that are involved in bacterial cell-wall remodeling in the course of bacterial cell division (Vollmer et al. 2008; Vermassen et al. 2019). The presence of autolysin in viral genomes has been shown in pneumococci-associated viruses, and is used for viral absorption, penetration and progeny release (Ackermann 1998; Frias et al. 2009). In our study, a NAD-dependent protein deacetylase and cold shock stress protein were found in both viral and host genomes and both deacetylase and cold shock stress proteins that modulate the activity of several enzymes have been identified in *Bacillus subtilis*, resulting in cellular growth with low-acetate and low-temperature conditions, respectively (Graumann et al. 1997; Gardner and Escalante-

Semerena 2009). These findings demonstrate viruses hijacking components of the host cell to support viral replication, infection and fitness through horizontal gene transfer.

This work was performed using only one host bacterium and therefore cannot be extrapolated to the thousands of different populations within a soil. In addition, while the soil samples of decreasing pH were of increasingly different physicochemical properties compared to the pH 7.5 samples, this also correlated with increasing spatial distance from the pH 7.5 soil, and the role of spatial distance in contributing to the results obtained cannot be ruled out.

Preliminary results showed that of the many isolates tested, those belonging to the genus *Bacillus* were preferentially isolated, but also formed plaques using the approach used in this study. This phenomenon requires further exploration. Based on the characteristics of the host bacterium, such as abundance, growth rate and localization, virus enrichment may require optimization. For example, longer enrichment times or concentration of the viral particles after enrichment of viruses from incubation with the host may increase plaque formation. Furthermore, the enriched virus populations from the soil samples that were used in plaque assays were added without normalization (or knowledge of) the concentration of viruses in the soil. Overall virus numbers could have changed with pH and therefore contributing to the quantified infectivity of the viral populations.

3.6. Conclusion

The plaque assay approach allowed enrichment for the viruses capable of infecting a specific host bacterium in order to address specific hypotheses regarding virus-bacterial host interactions in soil. We hypothesized that host bacteria would be more susceptible to infection from co-localized virus populations in soil due to local adaption and that coevolutionary processes may tightly control the susceptibility of hosts through virus-bacterial interactions. Overall, while greater infectivity was not observed when viruses and host bacterium were co-localized, the findings provided evidence for local adaptation in natural populations. These results demonstrate that in natural soil populations, virus-bacterial host interactions have a central role in defining the susceptibility of a host to infection, and viruses capable of infecting a host vary over an ecological gradient.

CHAPTER IV

Linking virus-host interactions in a methane-fueled trophic network using stable-isotope probing

4.1. Abstract

Methanotrophs are distributed ubiquitously in the environment and use methane as a sole carbon and energy source. They play a critical role in carbon cycling as major sink of atmospheric and soil-derived CH_4. The abundance and distribution of methanotrophs have been extensively characterized but little is known about top-down drivers of these key carbon-cycling organisms. Viruses have a major impact on their hosts by directly affecting their abundance through the lytic cycle and adaptation by transferring genes. In order to investigate both active methanotrophs and their viruses, DNA stable-isotope probing (SIP) and metagenomics were used to follow the recent transfer of carbon from host to virus. Soils from a pH gradient were sampled (pH 4.5 and 7.5) and incubated in aerobic microcosms for 30 days with a 10% ^{12}C- or ^{13}C-CH_4 headspace. Genomic DNA was extracted, subjected to isopycnic ultracentrifugation and high buoyant density DNA sequenced using the Illumina NovaSeq platform (1.5-2.0 Gbp per replicate). From both soils, 23 metagenome-assembled genomes were identified, and included primary methane oxidizers (*Methylobacter*, *Methylocystis*, *Methylosinus* and *Methylocapsa* sp.) and non-methane oxidizing methylotrophic secondary utilizers (*Gemmatimonadales, Herminiimonas*, *Hyphomicrobium*, *Rudaea* sp.) and predatory bacteria (*Bdellovibrio* and *Myxococcus*) indicating that oxidized methane fueled a trophic network. Specific host-virus interactions were identified via the transfer of virus DNA into CRISPR arrays or the presence of homologous genes shared between viral and methanotroph genomes. Three CRISPR arrays were found in the genomes of *Methylocystaceae* and *Methylococcaceae* representatives that dominated at pH 4.5 and 7.5, respectively, with spacers including those derived from six distinct ^{13}C-enriched viruses, demonstrating current methanotroph-virus interactions. In total, 720 viral contigs were predicted and were associated with both primary and secondary utilizers of CH_4-derived carbon. AMGs associated with methane metabolism were also identified with homologs to proteins from methanotrophs demonstrating that these viruses could potentially augment the functioning of methanotroph populations in soil. These results demonstrate that targeting specific functional groups via SIP facilitates analysis of individual active host-virus populations, and methanotroph viruses likely have a role in regulating methane fluxes in soil.

4.2. Introduction

After carbon dioxide (CO_2), methane (CH_4) is the second most important greenhouse gas that contributes to radiative forcing in the Earth's atmosphere (Wang et al. 2017). Despite the greater concentration of CO_2 and a longer residence time, CH_4 is 25 times more effective at trapping infrared radiation (IPCC 2013; Shindell et al. 2009). Human activities since the beginning of the industrial revolution are the main causes of increased CH_4 emissions (Ruddiman and Thomson 2001) with atmospheric concentrations 2.5x greater than preindustrial times (IPCC, 2013). While

carbon isotope data indicate that the major source of increased CH_4 emissions is fossil fuels, since 2007, biologically-derived emissions including those from agriculture and wetlands sources are responsible for recent increases in concentrations (Nisbet et al. 2016; Schwietzke et al. 2016).

Microorganisms are important regulators of the global CH_4 budget as they both produce CH_4 and also remove it from the atmosphere (Nazaries et al. 2011, 2013; Aronson et al. 2013). Methanotrophic bacteria and archaea consume at least half of all methane produced each year (Reeburgh 2003; Conrad 2009). These prokaryotes utilize CH_4 as their sole source of carbon and energy (Knief 2015). In soil, aerobic methanotrophs are estimated to be responsible for removing, 20 to 45 Tg per year (Dutaur and Verchot 2007). Aerobic methanotrophs currently consist of 26 genera in the *Gammaproteobacteria* and *Alphaproteobacteria* (Dedysh and Knief 2018) and two candidate genera in the phylum *Verrucomicrobia* (Op den Camp et al. 2009; van Teeseling et al. 2014). Based on pathways of carbon assimilation, methanotrophs can be characterized as type I that utilize the ribulose monophosphate pathway, and type II that utilize the serine pathway (Dedysh and Knief 2018). Type I belong to the family *Methylococcaceae* (*Gammaproteobacteria*) and are facultative methanotrophs, utilizing either methane or multi-carbon compounds, and type II belong to the family *Methylocystaceae* (*Alphaproteobacteria*) (Theisen et al. 2005; Knief 2015; Dedysh and Knief 2018). Methanotrophs can exhibit different affinities for methane, with type I typically having a low affinity and type II having a high affinity (Nazaries et al. 2011). Understanding the biology and ecology and regulating factors of methanotrophs is critical to our view of methane emissions and carbon cycling.

Aerobic methanotrophs oxidize methane to methanol using the enzyme methane monooxygenase enzyme (MMO). Methanotrophs can possess two different forms of this enzyme, either a soluble methane monooxygenase (sMMO) located in the cytoplasm, or a membrane-bound particulate methane monooxygenase (pMMO). The latter is found in most methanotrophs, unlike sMMO, which is only found in a more restricted number of species such as *Methylococcus capsulatus*, *Methylosinus sporium* and *Methylocystis heyeri* (Dedysh and Knief 2018). Although both enzymes perform the same chemical reaction, their structures, cofactor requirements and mechanisms are different (Sirajuddin and Rosenzweig 2015; Ross and Rosenzweig 2017). For example, sMMO utilizes NADH and H^+ as an electron donor, and does not utilize any cofactors, while the pMMO is a copper-containing enzyme that employs a higher-potential electron donor such as cytochrome C (Hanson and Hanson 1996; Sirajuddin and Rosenzweig 2015; Ross and Rosenzweig 2017). The sMMO requires three components for activity including a hydroxylase, a regulatory protein (B component) and a reductase. The hydroxylase component contains nonheme iron, and is composed of alpha, beta and gamma subunits encoded by the *mmoX*, *mmoY* and *mmoZ* genes, respectively. The B component and reductase are encoded by the *mmoB* and *mmoC* genes. The reductase contains flavin adenine dinucleotide and a Fe_2S_2 cluster (Hanson and

Hanson 1996; Sirajuddin and Rosenzweig 2015; Ross and Rosenzweig 2017). However, the pMMO is composed of the *pmoB*, *pmoA* and *pmoC* subunits, encoded by the *pmoABC* operon. The *pmoB* subunit contains both periplasmic and transmembrane domains, while the pmoA and *pmoC* subunits are composed primarily of transmembrane helices (Hanson and Hanson 1996; Sirajuddin and Rosenzweig 2015; Ross and Rosenzweig 2017). The sequences of these genes encoding the subunits are highly conserved among methanotrophs, and are used as molecular markers for detecting methanotrophs (Hanson and Hanson 1996). After oxidation of methane, methanol is subsequently oxidized to formaldehyde by a periplasmic methanol dehydrogenase (MDH) and a NAD-linked methanol dehydrogenase in gram-negative and gram-positive methylotrophs, respectively (Anthony 1986; Dijkhuizen et al. 1992; Hanson and Hanson 1996). Lastly, formaldehyde is oxidized via formate to CO_2. There are multiple enzymes involved in these steps including NAD(P)-linked aldehyde dehydrogenase and NAD-dependent formate dehydrogenase (Anthony 1986; Dijkhuizen et al. 1992; Hanson and Hanson 1996). This pathway generates reducing power for biosynthesis and the initial oxidation of methane. During the oxidation of formaldehyde, carbon is assimilated into the biomass of methanotrophs.

Methanotrophy in soil can be affected by both abiotic and biotic factors (Saggar et al. 2008; Nazaries et al. 2013). For example, soil pH plays an important role in methanotroph activity through direct effects on methanotroph physiology and also by changing the concentration of toxic elements and nutrients in soil (e.g. ammonium, aluminum and iron) (Steudler et al. 1989; Adamsen and King 1993; Dedysh et al. 1998; Benstead and King 2001). Synergistic interactions between methanotrophs and other organisms can also affect methanotrophy (Ho et al. 2016). Top-down drivers of these key carbon-cycling methanotrophs could also significantly impact methanotroph activity. In particular, viruses are recognized as an important driver of microorganisms by directly affecting host abundance through the lytic cycle, and the transferring of genes through the lysogenic cycle (Weinbauer et al. 2003; Suttle 2005; Wommack et al. 2009). Additionally, viruses can also impact carbon cycling by host metabolic manipulation during viral infection, via the expression of virus-encoded auxiliary metabolic genes (AMGs)(Breitbart et al. 2007). Recent studies have discovered a large number of AMGs encoding genes involved in carbon degradation in soils, indicating the potential impact of soil viruses on ecosystem carbon processing (Emerson et al. 2018; Trubl et al. 2018, 2020; Graham et al. 2019). Bacteriophages of methanotrophic bacteria have been isolated in several environments (Tiutikov et al. 1976; Tyutikov et al. 1980, 1983), and metagenomic assembled viruses associated with methanotroph hosts have been recently reported (Emerson et al. 2018). However, the role of viruses in the regulation of methanotrophic communities and methane metabolism and carbon cycling are unknown (Kalyuzhnaya et al. 2019).

Metagenomics involves the sequencing of mixed-community DNA extracted from any environment, allowing potential access to all the diversity present in a given environment (Handelsman 2004; Sabree et al. 2009). This enables determining the composition and structure of environmental microbial and viral communities, but also genetic characterization. While gene annotation of metagenomic data is a useful way of inferring community function, the presence of a gene does not mean that it is functionally expressed at the time of sampling (Sharpton 2014). Currently, there are no studies which have determined the interaction between active methanotrophs and their associated viruses. DNA stable-isotope probing (DNA-SIP) is a powerful method that targets active microorganisms within an environmental sample via the cellular assimilation of an added substrate enriched in a heavy isotope (Radajewski et al. 2000). As viruses are obligatory parasites and reproduce inside their host cells' using host replication machinery, consequently the associated soil viruses of the active microorganisms will be isotopically labeled (Lee et al. 2012). 'Heavy' DNA with a higher buoyant density can then be separated from unenriched DNA allowing metagenomic sequencing of isotopically-enriched DNA (Radajewski et al. 2000, 2003; Paul et al. 2017).

To target active methanotroph communities and their associated viruses, metagenomes derived from soils incubated with ^{13}C-CH_4 were analyzed, and clustered regularly interspaced short palindromic repeats (CRISPR) array and oligonucleotide frequency (ONF) analyses were used to link viruses to their hosts (Sanguino et al. 2015; Galiez et al. 2017). Acidic and near-neutral-pH soil (pH 4.5 and 7.5) was used as soil pH has previously been found to influence the community structure of methanotroph populations. This work tested the hypotheses that 1) focusing on a functionally and restricted group of microorganisms using stable isotope probing facilitates the detailed analysis of host-virus interactions, and 2) soil pH selects for distinct viral communities associated with CH_4 cycling due to the pH preference of the hosts. The aims of the study were to 1) determine the influence of soil pH on methanotrophic and methylotrophic communities and their associated viruses, 2) identify specific virus-host interactions via the presence of virus DNA in CRISPR arrays and the presence of shared homologous genes, and 3) determine the potential contribution of the methanotroph-associated viruses in methane metabolism and carbon cycling via the presence and transfer of AMGs.

4.3. Materials and Methods

4.3.1. Soil sampling and physicochemical analyses

Soil was collected from the Craibstone Research Station, SRUC, Aberdeen, Scotland (57°11, 2°12) in February 2018, with surface soil samples (top 10 cm) sampled in triplicate from pH 4.5 and 7.5 sub-plots of a continuous pH gradient. Soil samples were sieved (2-mm) and stored at 4°C prior to establishing microcosms or physicochemical analyses. Soil pH and moisture content were

determined as described previously (Chapter II, section 2.3.1.) with the measured soil pH being 5.08 (±0.01 SE) and 7.46 ((±0.03 SE), from the pH 4.5 and 7.5 sub-plots, respectively.

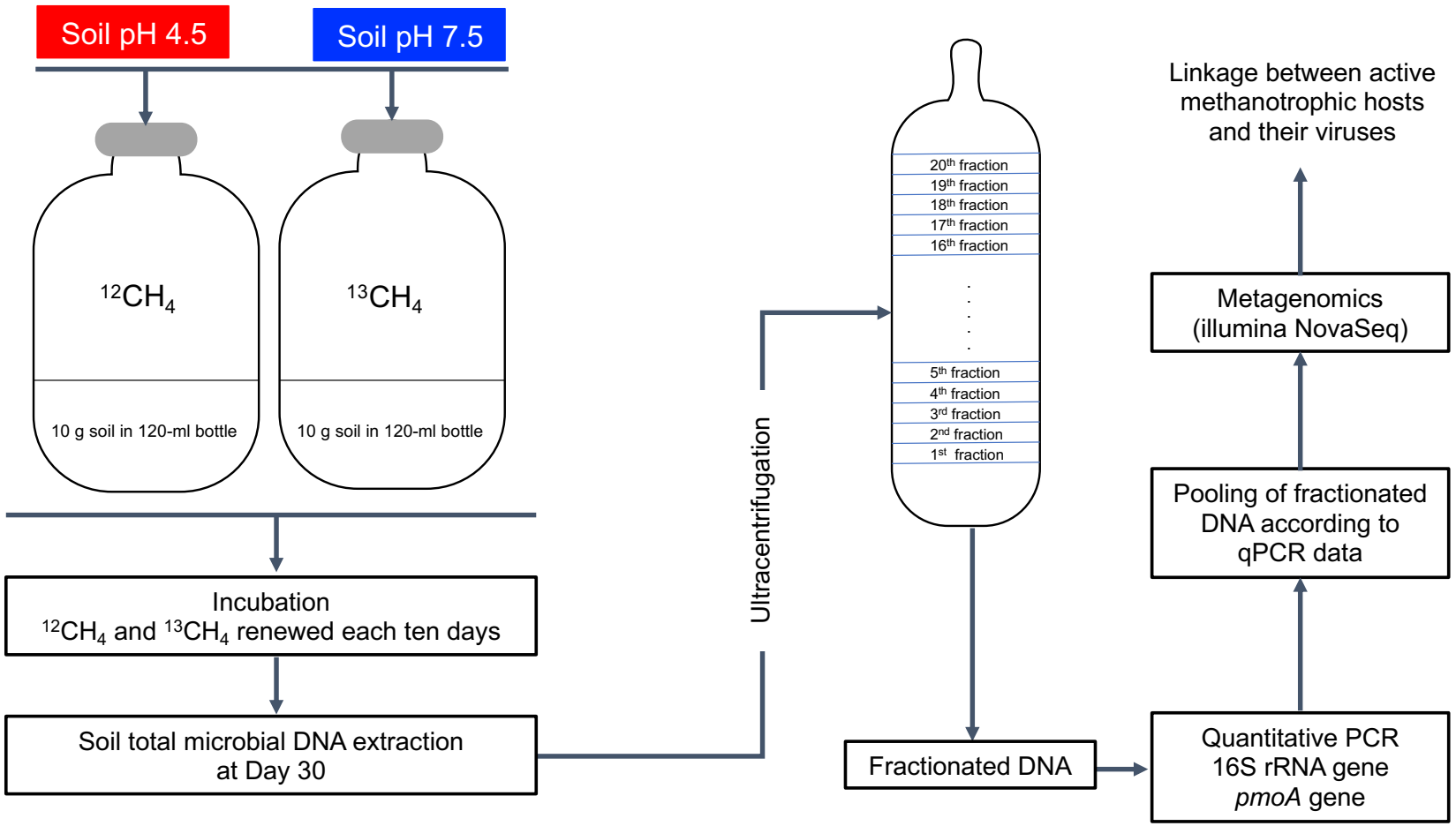


Figure 4.1. Schematic overview of the experimental workflow. Microcosms were established for destructive sampling with three time points in triplicates. At day 30, total DNA was extracted and ultracentrifuged to separated ^{12}C- and ^{13}C-DNA. Quantitative PCR was performed to verify the enriched community. According to qPCR data, heavy fractions of ^{13}C-labeled DNA were pooled and sequenced to investigate the linkage between active methanotrophs and their associated virus.

4.3.2. Soil microcosm incubations

Twelve soil microcosms were established in 120 ml serum bottles, containing 13.7 g soil (wet weight) at 27% water content (equivalent to 10 g soil dry weight) of either pH 4.5 or 7.5 soil. Bottles were tightly capped with a grey butyl stopper and sealed before the headspace was amended by injection of 99% ^{13}C-labeled CH_4 or ^{12}C-CH_4 to a final concentration of 10% (v/v) (three replicates of each isotope for each soil pH). Microcosms were incubated at 25°C in the dark for 30 days with microcosms opened and CH_4 replenished every 10 days following aeration to maintain aerobic conditions. Microcosms were destructively sampled, and soil was frozen at -20°C before analysis.

4.3.3. DNA extraction and density gradient centrifugation

An overview of the experimental workflow is shown in Figure 4.1. Soil DNA was extracted as previously described (see Section 2.3.3).

Isopycnic density gradient centrifugation of DNA was performed. For each sample, 6 µg of DNA was added to 8ml CsCl in 1 x Tris-EDTA buffer with a buoyant density of 1.696 g ml^{-1} (refractive index of 1.4004). The DNA-containing solution was added to an 8 ml polyallomer tubes which was sealed before centrifuging in a MLN-80 rotor (Beckman Coulter, Brea, CA, USA) at 151,000 x *g* (50,000 rpm) at 25°C for 72 h. The CsCl gradients were fractionated by displacing with sterile water into 20 equal fractions of 350 µl using an inhouse system consisting of a peristaltic pump (Minipuls3 peristaltic pump (Gilson, Sydney, NSW 2102, Australia) and fraction collector (Model 2110, Bio-Rad, Roanne, France). The buoyant density of each fraction was determined indirectly my measuring refractive index of a 15 µl aliquot of each fraction using an AR200 digital refractometer (Reichert Technologies, Depew, NY, USA). DNA was recovered from each fraction using PEG NaCl solution and ethanol washing as previously described. DNA pellets were dissolved in 30 µl sterile molecular grade water.

4.3.4. Quantitative PCR and metagenomic sequencing

Quantitative PCR (qPCR) was performed to determine the distribution of bacterial and methanotroph genomes through the CsCl gradients. Bacterial 16S rRNA and methanotrophic *pmoA* genes were quantified by qPCR across the entire buoyant density gradients with primer pairs P1(341f)/P2(534r) (Muyzer et al. 1993) & A189F/A682R (Holmes et al. 1995), respectively. Each 25 µl PCR contained 12.5 µl 2X QuantiFast® SYBR Green Mix (Qiagen, Germantown, MD, USA), 1 µM of each primer, 100 ng of T4 gene protein 32 (Thermo Fisher, Carlsbad, CA, USA), 2 µl of standard or 1/10 diluted DNA. No template negative controls were performed with each run. The thermal cycling program consisted of an initial denaturation step of 15 min at 95°C, followed by 30 cycles of 15 s at 94°C, 30 s at 60°C, 30 s at 72°C for the 16S rRNA gene assay or 60 s at 94°C, 60

s at 56°C, 60 s at 72°C for the *pmoA* gene assay using a Corbett Rotor-Gene 6000 real-time PCR cycler. After assessing the distribution of genomic DNA, fractions with a buoyant density between 1.73 to 1.75 g ml^{-1} were pooled. Genomic DNA samples were sequenced using the Illumina NovaSeq platform at the Joint Genome Institute (JGI, Berkeley, CA, USA).

4.3.5. Bioinformatic analyses

An overview of the bioinformatics workflow is shown in Figure 4.2. Briefly, the quality-controlled reads of each metagenome were assembled into contigs and all contigs were concatenated as co-assembled contigs. Dereplicated co-assembled contigs (> 5 kb) were used for virus prediction and binned into metagenomic assembled genomes (MAGs) for bacterial hosts followed by CRISPR array and oligonucleotide frequency (ONF) analyses to link hosts and viruses. Auxiliary metabolic genes were searched for and gene homology between host-linked viruses and their associated hosts was analyzed through protein alignment.

4.3.5.1. Sequence quality filtering, contig assembly and co-assembly

The JGI bioinformatics pipeline was utilized for sequence quality filtering and contig assembly as previously described (Chapter II, section 2.3.4.1.). The tool BBDuk v38.51 (JGI's BBTools Team) was used to remove known sequencing artifacts, contaminants and low-quality sequences (quality score = 20, minimum read length = 51 bp). To remove any contaminates reads were subjected to mapping with BBMap v38.34.1 (JGI's BBTools Team). The quality-controlled reads of each metagenome were *de-novo* assembled into contigs using MetaSPAdes v3.13.0.2 (Nurk et al. 2017). The input read set was mapped to the final assembly and coverage information was generated with BBmap (JGI's BBTools Team).

Prior to binning and virus prediction, the assembled contigs from each sample were concatenated as one co-assembled sample, and sequence redundancy was reduced using the psi-cd-hit-DNA tool with 95% identity (Fu et al. 2012). Contigs larger than 5 kb were selected and contig names were simplified using the anvi-script-reformat-fasta from anvio 5 (Eren et al. 2015). The co-assembled contigs were annotated using Kaiju (Menzel et al. 2016) with the NCBI RefSeq bacterial, archaeal and viral database. The relative abundance of the most abundant taxa (< 1%) was plotted and the Student's t-test was used to assess significant differences between soil pH for each taxon within R (R core team, 2019).

To generate %GC coverage plots, the co-assembled contigs were indexed and the reads from the six metagenomes were re-mapped with the bowtie2 mapper v2.3.0 (Langmead and Salzberg 2012). GC_cov_annotate.pl function from MetaWRAP-Bloblogy module was used to generate a blobplot file with the GC content, coverage and taxonomy of each contig (Uritskiy et al. 2018). Blobplots of the co-assembled contigs across the six metagenomes were made using the

makeblobplot.R function from MetaWRAP-Bloblogy module (Uritskiy et al. 2018) within R (R core team, 2019).

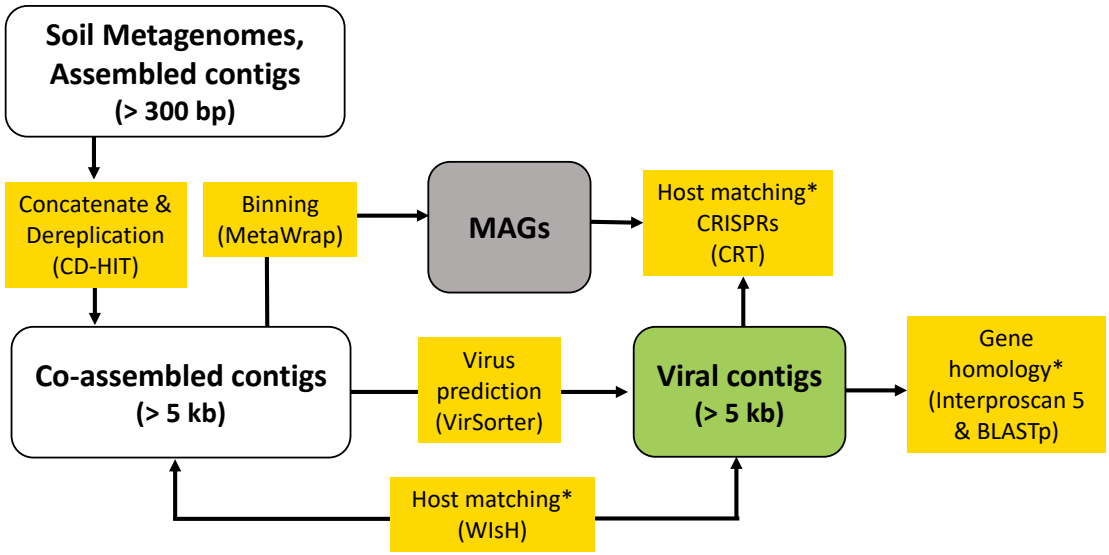


Figure 4.2. Schematic overview of the bioinformatics workflow. Square boxes in white, gray and green represent the assembled contigs and metagenome assembled genomes (MAGs) and predicted viral contigs, respectively. Square boxes in yellow represent the bioinformatics performed and the key tools are in parentheses. *Gene homology analyses between host MAGs or contigs and host-associated metagenomic viral contigs (mVCs) were realized (AMGs, Auxiliary metabolic genes).

4.3.5.2. Metagenomic assembled genomes

To analyze metagenomes at the single genome level, metagenome-assembled genomes (MAGs) were generated. Co-assembled contigs were binned using Metawrap-Binning (Uritskiy et al. 2018). Bins were de-replicated, and the completion and the contamination level of the bins were calculated by CheckM v.1.0.7 (Parks et al. 2015).

Taxonomic classification of the MAGs based on marker genes was carried out using GTDB-Tk v0.3.2 (Chaumeil et al. 2020). Functional analysis of MAGs was carried out by annotating the protein sequences using InterProScan 5 (E value < 10^{-5}) (Jones et al. 2014). For each MAG, the presence of a methane monooxygenase (MMO) was verified by screening the results.

To determine the distribution and abundance of each MAG across the samples, the relative abundance of each bin was calculated. The Quant_bins module in Salmon v0.9.1 was used to index the contigs and align reads from each metagenome back to the contigs in each bin (Patro et al. 2017; Uritskiy et al. 2018). The relative abundance of each bin in each sample was calculated based on the length of contig size and on the coverage of contigs. Bin abundance was expressed as normalized genome copies per million reads (CPM). This was calculated as follows: the read per kilobase (RPK) was divided by genome-length (in Kb), RPK was summed and divided by 1 million to get a scaling factor, and RPK values were divided by the scaling factor to get the CPM. A barplot was made using the ggplot R package (R core team, 2019). Significance was tested using either the Student's t-test or Wilcoxon-Mann-Whitney's test, according to Bartlett's test for the homogeneity of variances, with the ggpubr R package (R core team, 2019).

4.3.5.3. Virus prediction

To predict the contigs (> 5 kb) that were viral, the VirSorter tool was used (Roux et al. 2015). The contigs from VirSorter categories 1, 2 and 3 were retained (see Chapter I, section 1.3.4.1. for category explanation). Taxonomy of the predicted metagenomic assembled viral contigs (mVCs) was assigned using Kaiju (Menzel et al. 2016), with the RefSeq viral proteins database from NCBI. Taxonomic composition of the annotated mVCs was visualized using the Krona tool (Ondov et al. 2011).

To determine the distribution of the ^{13}C-enriched mVCs across the samples, the relative abundance of the mVCs across the six metagenomes were calculated. The Quant_bins module in Salmon v0.9.1 (Patro et al. 2017) was used to index the contigs and align reads from each virome back to the contigs. The relative abundance of each mVC in each metagenome was calculated based on the length of contig size and on the coverage of contigs. The abundance was expressed as normalized genome copies per million reads (CPM). This was calculated as previously described above (section 4.2.5.2.). A heatmap was made using the heatmaply R package to

visualize the variation in the relative abundance of the mVCs across the metagenomes (R Core Team, 2019).

4.3.5.4. Linking viruses to hosts using CRISPR array and ONF analysis

The CRISPR Recognition Tool (CRT) was used to identify CRISPR arrays from the MAGs and from the assembled contigs from each metagenome (Bland et al. 2007). The direct repeats (DRs) and spacer sequences (spacers) were extracted from the CRISPR arrays using Linux commands and located using the Seqkit locate function with exact match and positive and negative strand search (Shen et al. 2016). Specifically, the spacers were located by searching spacer sequences on the mVCs. The DRs from the CRISPR arrays containing the mVCs were located by searching DRs on the contigs from the MAGs or individual contigs (> 5kb) to identify the hosts. The individual contigs containing the CRISPR arrays with the mVCs were annotated using Kaiju with the NCBI RefSeq bacterial, archaeal and viral database (Menzel et al. 2016).

For ONF analysis, host contigs of mVCs were predicted using the WIsH tool (Galiez et al. 2017). Predictions of host-virus linkages with a *p*-value > 0.05 were selected (Galiez et al., 2017). The taxonomy of the potential host contigs were annotated with Kaiju with the NCBI RefSeq bacterial, archaeal and viral database (Menzel et al. 2016).

The methanotroph MAG associated mVCs were compared with reference viral genomes stored in the Virus-Host DB (Mihara et al. 2016) by calculating all-against-all similarity scores (S_G) computed by tBLASTx. The resulting phylogenetic trees were visualized using ViPtree (Nishimura et al. 2017). Relative abundance of the mVCs that were associated with the methanotroph MAGs was calculated by read remapping and normalizing by coverage and genome length as previously described in section 4.2.5.2. Normalized abundances were plotted using the boxplot function in R (R Core Team, 2019). All of the mVCs that were associated with methanotrophs, methylotrophs and potential cross-feeding hosts (e.g. predators and autotrophs) were used to manually construct a carbon flow diagram.

4.3.5.5. Analysis of gene homology

Gene prediction of the mVCs was completed using Prodigal (Hyatt et al., 2010). To identify auxiliary metabolic genes (AMGs), homology searches were conducted using InterProScan 5 (E value < 10^{-5}), using Diamond Blastp with the NCBI-nr database (E value < 10^{-5})(Jones et al. 2014; Buchfink et al. 2015). Genes involved in methane oxidation and carbon metabolism were manually identified.

Gene homology between host-linked mVCs and their associated hosts (Virus-host linkage through CRISPR analysis and ONF analysis) was investigated. Gene prediction of the host-linked mVCs and their associated hosts (MAGs or host contigs) was performed using Prodigal (Hyatt et

al. 2010). Protein alignment between viral and host origin proteins was conducted with BLASTp (identity > 30%, E value < 10^{-5} and query cover > 70%) (Madden et al. 1996). Shared protein sequences were annotated using InterProScan 5 (E value < 10^{-5})(Jones et al. 2014). Additionally, gene homology was assessed between host-linked mVC and prokaryotes using Diamond BLASTp with the NCBI-nr database (E value < 10^{-5}) (Buchfink et al. 2015). From the BLASTp output file of each host-linked mVC the number of viral genes homologous to the genera of their associated host was counted using Linux commands. A network analysis showing homologous AMGs of the methanotroph-associated mVCs with their hosts was produced using the Igraph R package (R Core Team, 2019). The bit score from BLASTp was used as the edge values.

The methanotroph MAG associated mVCs were compared to each other by calculating all-against-all similarity score (S_G) computed by results of tBLASTx, and were visualized with ViPtree (Nishimura et al. 2017). Homology search on methanotroph-associated mVCs was performed against NCBI database (nr/aa) by GHOSTX software through ViPtree (Nishimura et al. 2017).

4.4. Results

4.4.1. Distribution of prokaryotic genomes in DNA-SIP fractions

The distribution of total bacteria and *pmoA*-gene containing methanotroph communities were determined using 12 CsCl gradients. In gradients from microcosms incubated with either isotope, the maximum amount of genomic DNA was found at approximately 1.71g ml^{-1} (Figure 4.3) However, a small peak in the distribution of the 16S rRNA genes was observed in the ^{13}C-CH_4 microcosms only for both pH 4.5 and 7.5 soils (buoyant densities of 1.73 – 1.75 g ml^{-1}) demonstrating incorporation of ^{13}C into the genomic DNA of microorganisms. This incorporation was more apparent by examining the distribution of the bacterial *pmoA* gene in genomic DNA from the ^{13}C-CH_4 microcosms compared to the ^{12}C-CH_4 microcosms for both pH soils (Figure 4.3). For the pH 4.5 soil, most genomic DNA from ^{13}C-CH_4 incubations overlapped with that from the ^{12}C-CH_4 incubations, indicating that the majority of methanotroph populations were not active. However, in the pH 7.5 soil there was a clear shift in the entire profile, indicating the majority of pmoA-containing populations were incorporating methane. The qPCR data therefore demonstrated active growth on methane. These results indicate that ^{13}C-CH_4 was assimilated by methanotrophs in both pH soils. To ensure that only ^{13}C-enriched DNA was sequenced, fractions with a buoyant density between 1.73 to 1.75 g ml^{-1} were pooled for each individual gradient for sequencing.

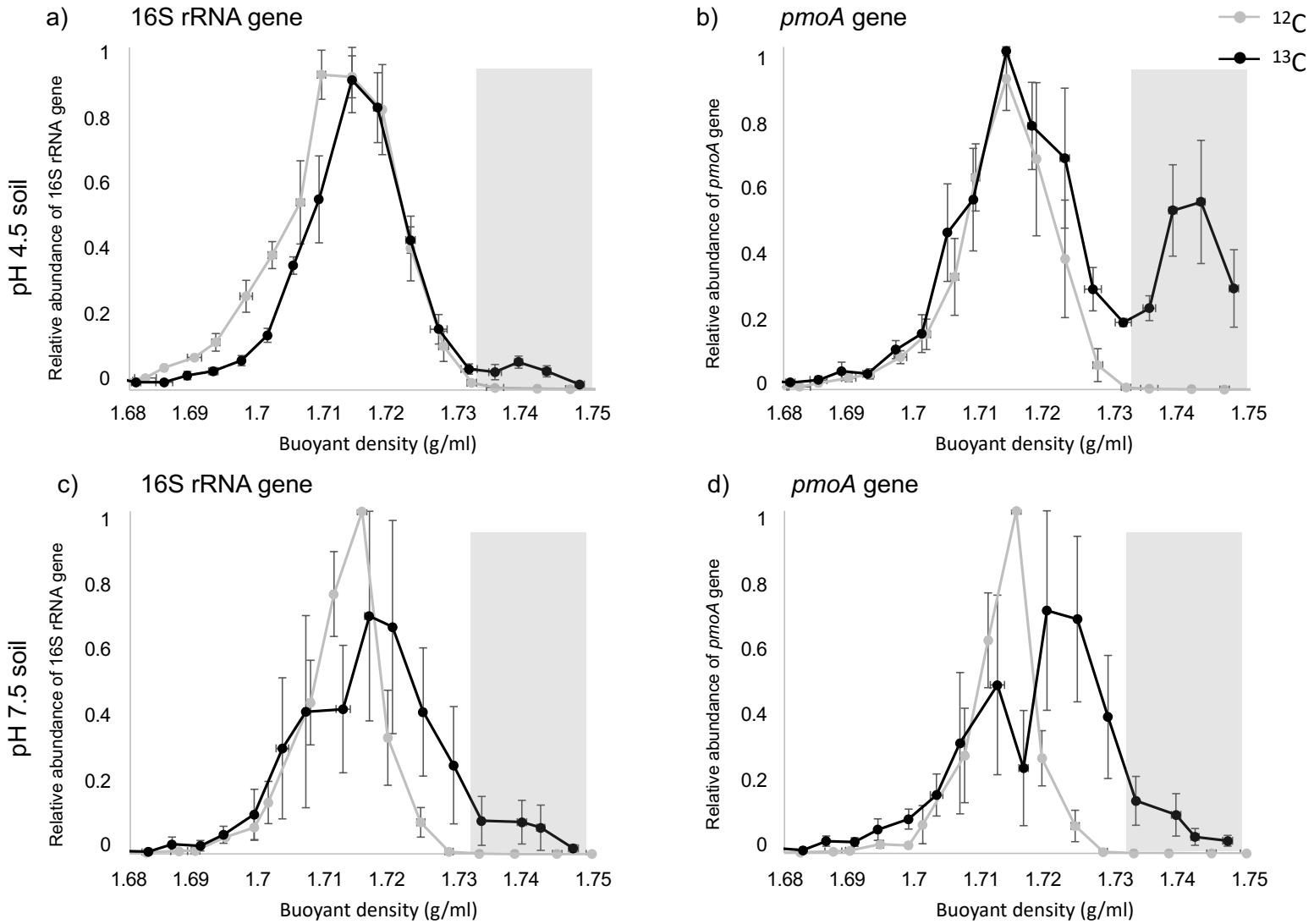


Figure 4.3. Relative abundance of bacterial 16S rRNA (a and c) and *pmoA* (b and d) genes in genomic DNA distributed in CsCl buoyant density gradients from pH 4.5 and 7.5 soils incubated with either ^{12}C-CH_4 or ^{13}C-CH_4 for 30 days. Data are normalized by ratio to the maximum abundance determined in one fraction for each gradient. Error bars represent the standard error of the mean from three replicates, each derived from an individual microcosm. The ^{13}C-labeled fractions highlighted in grey were pooled and sequenced.

4.4.2. Summary of metagenome sequencing

DNA from high buoyant densities of ^{12}C-enriched DNA were not of sufficient quantities for sequencing and therefore only DNA from ^{13}C-CH_4 incubated microcosms was sequenced. Therefore, six metagenomes ranging between 35 and 68 GB of sequence data were produced. A total of 0.9 billion sequence reads were retained after quality filtering, ranging between 100 – 190 million reads per metagenome (Table 4.1). Sequence assembly yielded in 10 million contigs, ranging between 1 – 2 million contigs per metagenome, with an average length of 556 bp (Table 4.1). The selection of contigs greater than 5 kb and the reduction of sequence redundancy resulted in 33,674 high-quality co-assembled contigs with an average length of 10,568 bp.

Table 4.1. Sequence summary for the ^{13}C-pH 4.5 and 7.5 soil metagenomes.

Sample ID	Pre-QC number of reads	Post-QC number of reads	Contigs count (> 200 bp)	Contigs count (> 5 kb)	Average contig length (bp)	Max. length (bp)
13C-CH4-pH 4.5-1	147,408,016	144,322,482	1,461,968	9,158	610.7	569,683
13C-CH4-pH 4.5-2	126,405,152	123,423,210	1,426,268	6,533	570.4	231,328
13C-CH4-pH 4.5-3	127,378,484	123,982,338	1,686,571	6,494	558.6	900,526
13C-CH4-pH 7.5-1	121,730,350	113,761,100	1,610,076	4,811	522.4	73,230
13C-CH4-pH 7.5-2	201,540,932	196,730,412	2,924,546	10,766	544.1	317,621
13C-CH4-pH 7.5-3	104,960,868	100,786,542	1,552,610	5,483	535.7	142,079

QC, quality control

4.4.3. Metagenomic community structure

Taxonomic classification of the co-assembled contigs demonstrated that on average 36% of the annotated contigs in each soil metagenome were derived from methanotrophic bacteria (Figure 4.4). Microbial community structure was significantly distinct between the two pH soils (Figure 4.4 and 5.5). The pH 4.5-derived community was mostly dominated by *Methylocystaceae* (26%), whereas *Methylococcaceae* was most abundant (23%) in the pH 7.5 soil (Figure 4.4). The relative abundance of contigs from facultative denitrifying methylotrophic *Hyphomicrobium* genus were slightly greater in the pH 4.5 soil, but overall were in low abundance (6%) (Figure 4.4). The taxon annotated %GC coverage plots across the pH 4.5 and 7.5 soil (12,335 contigs, at least 1X coverage) showed that the enrichment of methanotroph communities was reproducible between the different microcosms (Figure 4.5).

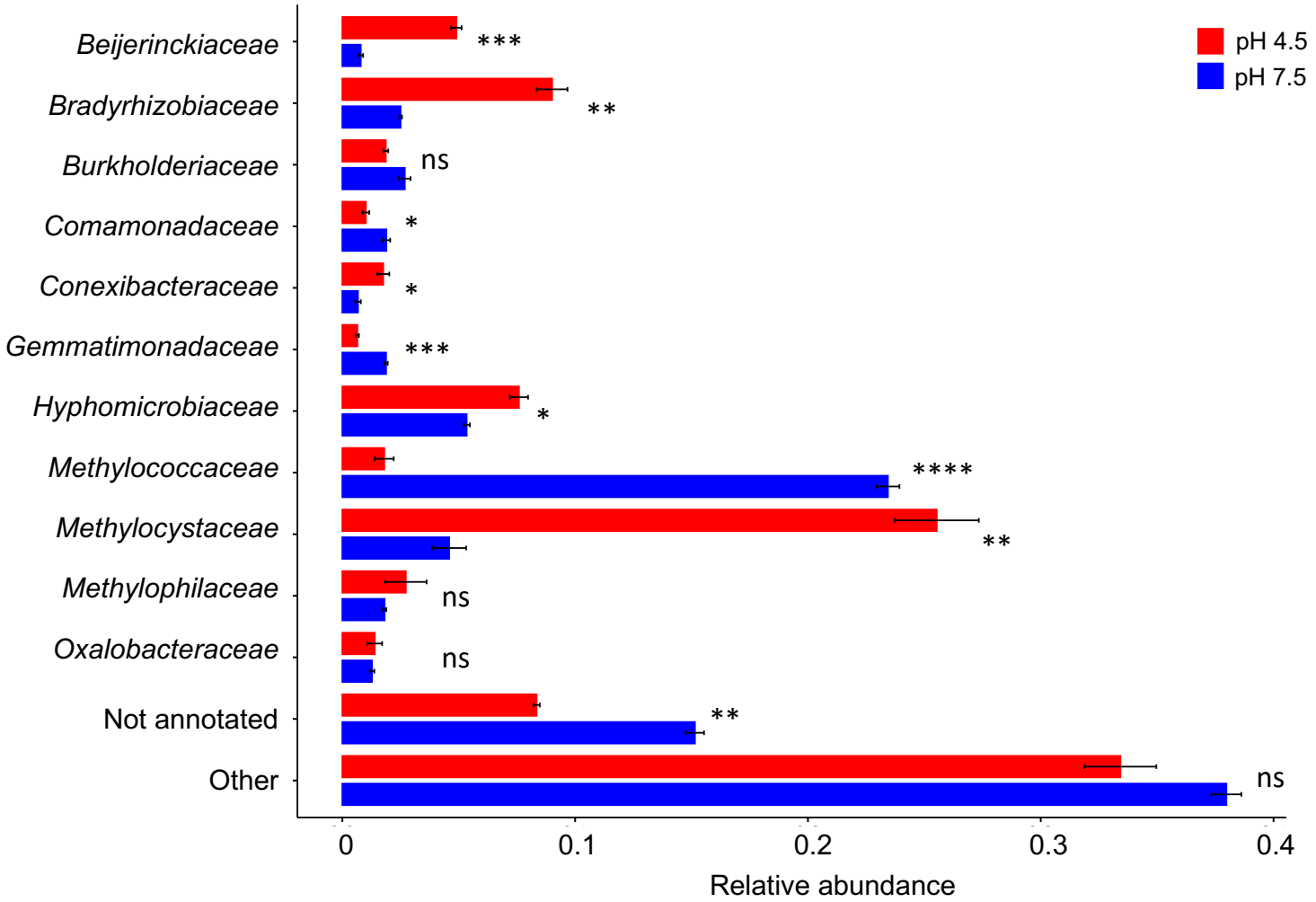


Figure 4.4. Relative abundance of the most abundant taxa (each containing > 1% of contigs) at the family level from the co-assembled contigs of the pH 4.5 (red) and pH 7.5 soil (blue) metagenomes. Families with less than 1% of contigs were grouped as "other" (310 families). Significant differences between soil pH was tested using the Student's t-test and marked as: $p > 0.05$ (ns), $p <= 0.05$ (*), $p <= 0.01$ (**), $p <= 0.001$ (***), and $p <= 0.0001$ (****). Error bars represent the standard error of the mean of three replicates.

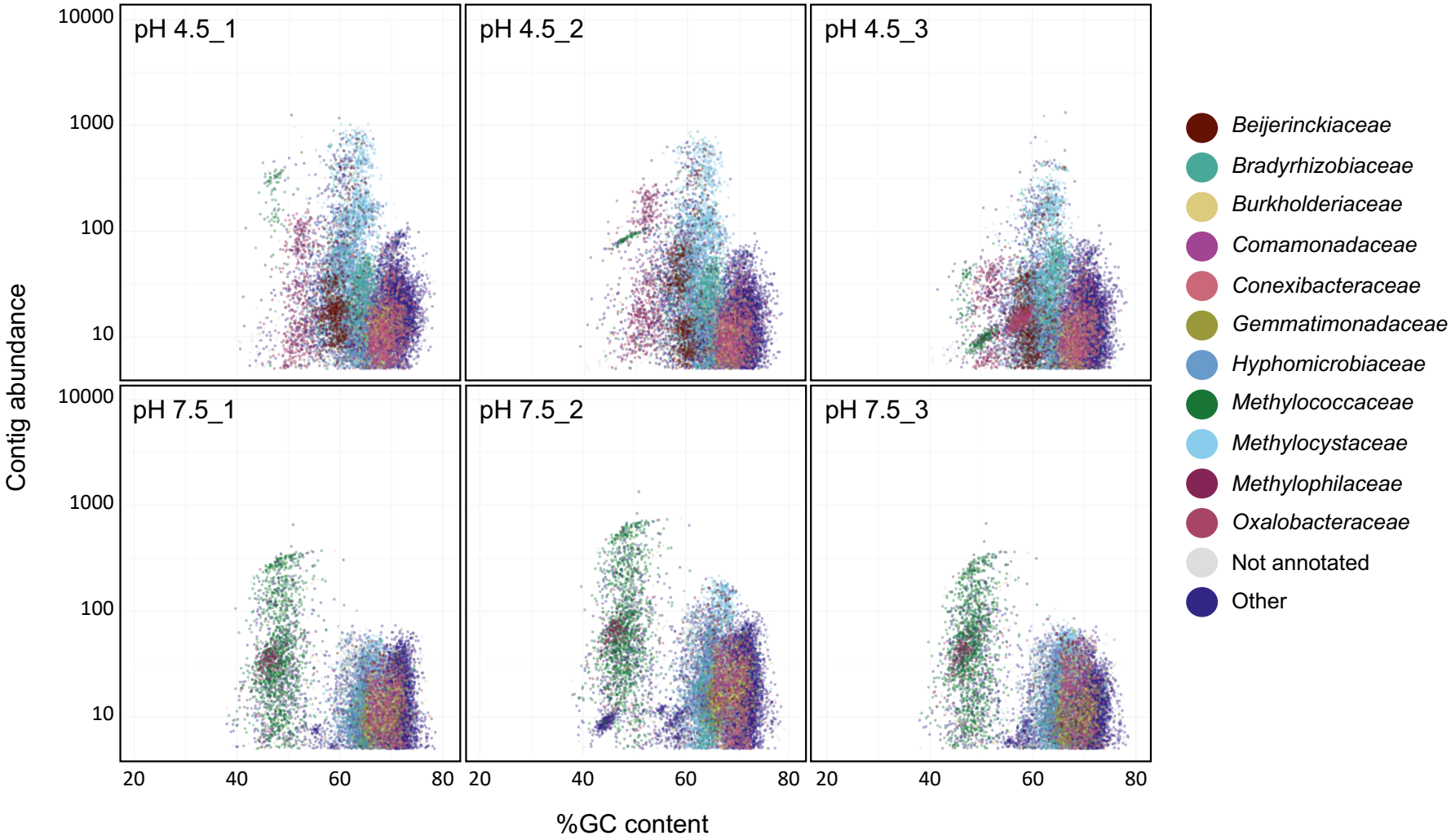


Figure 4.5. Taxon annotated %GC coverage plots of the co-assembled contigs from the pH 4.5 and 7.5 soil metagenomes. Families with less than 1% of contigs were grouped as "other" (310 families). Contig abundance was calculated from standardized read coverage in each sample.

4.4.4. Metagenomic assembled genomes

Binning of the co-assembled contigs produced 23 MAGs which had a completeness threshold of 50% with less than 10% contamination (Table 4.2). The average percent of completeness of MAGs was 72%, ranging between 99.6 - 50.3%. The average estimated level of contamination was 3%, ranging between 0 - 9.48%. Binning resulted in three high-quality draft genomes (completeness > 90%, contamination < 5%), seven medium-quality genomes (completeness > 70%, contamination < 10%), and eight partial genomes (completeness > 50%, contamination < 4%) as defined by Bower et al. (2017). The %GC content of the MAGs varied between 44% - 71% (Table 4.2).

Taxonomic assignment placed the MAGs into the following bacterial clades: one in the *Bdellovibrionales* (bin.7), three in *Burkholderiales* (bin.8, 13 and 17), two in *Gemmatimonadales* (bin.3 and 11), three in *Methylococcales* (bin.1, 18 and 19), eleven in *Rhizobiales* (bin.2, 4, 5, 6, 9, 10, 14, 15, 20, 21 and 22), one in *Myxococcales* (bin.16), one in *Verrucomicrobiota* (bin.23), and one in *Xanthomonadales* (bin.12). Of the 23 MAGs, 11 possessed genes encoding for the MMO enzyme. The distribution of the MAGs was broadly represented in both pH soils, however there were some pH specific MAGs (Figure 4.6). *Methylocystis* bin.4, 9, 20, and 21, and *Methylocapsa* bin.22 were significantly more abundant in the pH 4.5 soil, whereas *Methylobacter* bin.19, *Methyloceanibacter* bin.10 and *Hyphomicrobium* bin.15 were significantly more abundant in the pH 7.5 soil (Figure 4.6). Overall, MAGs had at least 1X coverage in all samples with 11% of all contigs binned in the 23 MAGs (Figure 4.7).

Table 4.2. Summary statistics, taxonomic classification and presence of the MMO enzyme in 23 MAGs.

MAG ID	Compl. (%)	Cont. (%)	GC (%)	N50 (Kb)	Size (Mb)	Phylum	Order	Family	Genus	MMO gene presence
bin.1	63.15	0	47.6	210	3.01	*Proteobacteria*	*Methylococcales*	*Methylococacceae*	*Methylobacter*	Yes
bin.2	72.45	1.28	66.9	80.2	2.97	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*	Yes
bin.3	53.11	5.49	67.4	26.6	2.54	*Gemmatimonadete*	*Gemmatimonadales*			No
bin.4	56.34	9.48	62.9	14.9	3.92	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*	Yes
bin.5	99.68	0.20	65.6	288	4.39	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*	Yes
bin.6	61.92	1.25	66.9	13.1	2.67	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*	Yes
bin.7	70.12	0	44.6	13.1	2.60	*Proteobacteria*	*Bdellovibrionales*	*Bdellovibrionaceae*		No
bin.8	99.52	8.73	49.7	339	4.40	*Proteobacteria*	*Burkholderiales*	*Burkholderiaceae*	*Herminiimonas*	No
bin.9	57.01	0.62	63.3	42.1	1.89	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*	Yes
bin.10	62.76	8.27	64.4	8.11	1.53	*Proteobacteria*	*Rhizobiales*	*Rhizobiaceae*	*Methyloceanibacter*	No
bin.11	50.34	3.44	65.8	8.14	1.28	*Gemmatimonadete*	*Gemmatimonadales*			No
bin.12	62.06	6.89	63.1	43.4	2.62	*Proteobacteria*	*Xanthomonadales*	*Rhodanobacteraceae*	*Rudaea*	No
bin.13	88.03	1.75	50.1	102	1.88	*Proteobacteria*	*Burkholderiales*	*Methylophilaceae*		No
bin.14	98.25	0.71	62.7	65.3	3.56	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*	Yes
bin.15	70.40	6.22	64	11.6	2.75	*Proteobacteria*	*Rhizobiales*	*Hyphomicrobiaceae*	*Hyphomicrobium*	No
bin.16	57.00	5.26	71.7	10.3	6.16	*Proteobacteria*	*Myxococcales*	*Myxococcaceae*		No
bin.17	69.67	1.18	57.6	9.64	4.02	*Proteobacteria*	*Burkholderiales*	*Burkholderiaceae*	*Herminiimonas*	No
bin.18	81.70	0.83	47.4	182	3.79	*Proteobacteria*	*Methylococcales*	*Methylococacceae*	*Methylobacter*	Yes
bin.19	99.36	1.73	48.2	111	4.70	*Proteobacteria*	*Methylococcales*	*Methylococacceae*	*Methylobacter*	Yes
bin.20	56.53	0.47	63.6	40.0	1.76	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*	No
bin.21	88.39	0.31	63.3	550	3.14	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*	Yes
bin.22	68.39	2.35	58.5	30.0	2.38	*Proteobacteria*	*Rhizobiales*	*Methylocystaceae*	*Methylocapsa*	Yes
bin.23	51.46	4.72	64.3	8.36	2.80	*Verrucomicrobiota*				No

Compl., completeness; Cont., contamination;

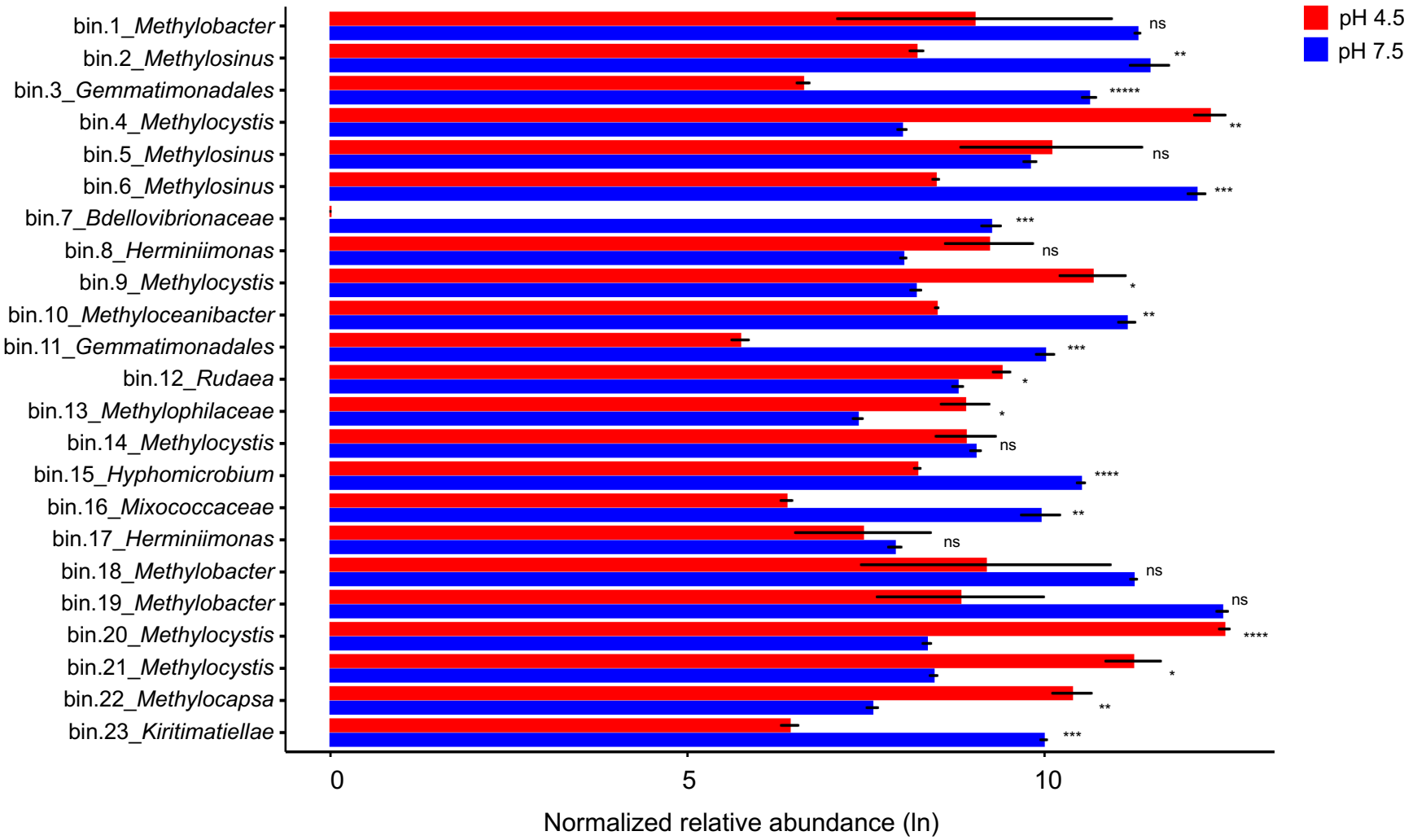


Figure 4.6. Normalized relative abundance (ln transformed) of the metagenomic assembled genomes (MAGs) at the family level for pH 4.5 (red) and pH 7.5 soil (blue). Significance differences were tested between soil pH using the Student's t-test and marked as: $p > 0.05$ (ns), $p <= 0.05$ (*), $p <= 0.01$ (**), $p <= 0.001$ (***), $p <= 0.0001$ (****), and $p <= 0.00001$ (*****). Error bars represent the standard error of the mean from three replicates.

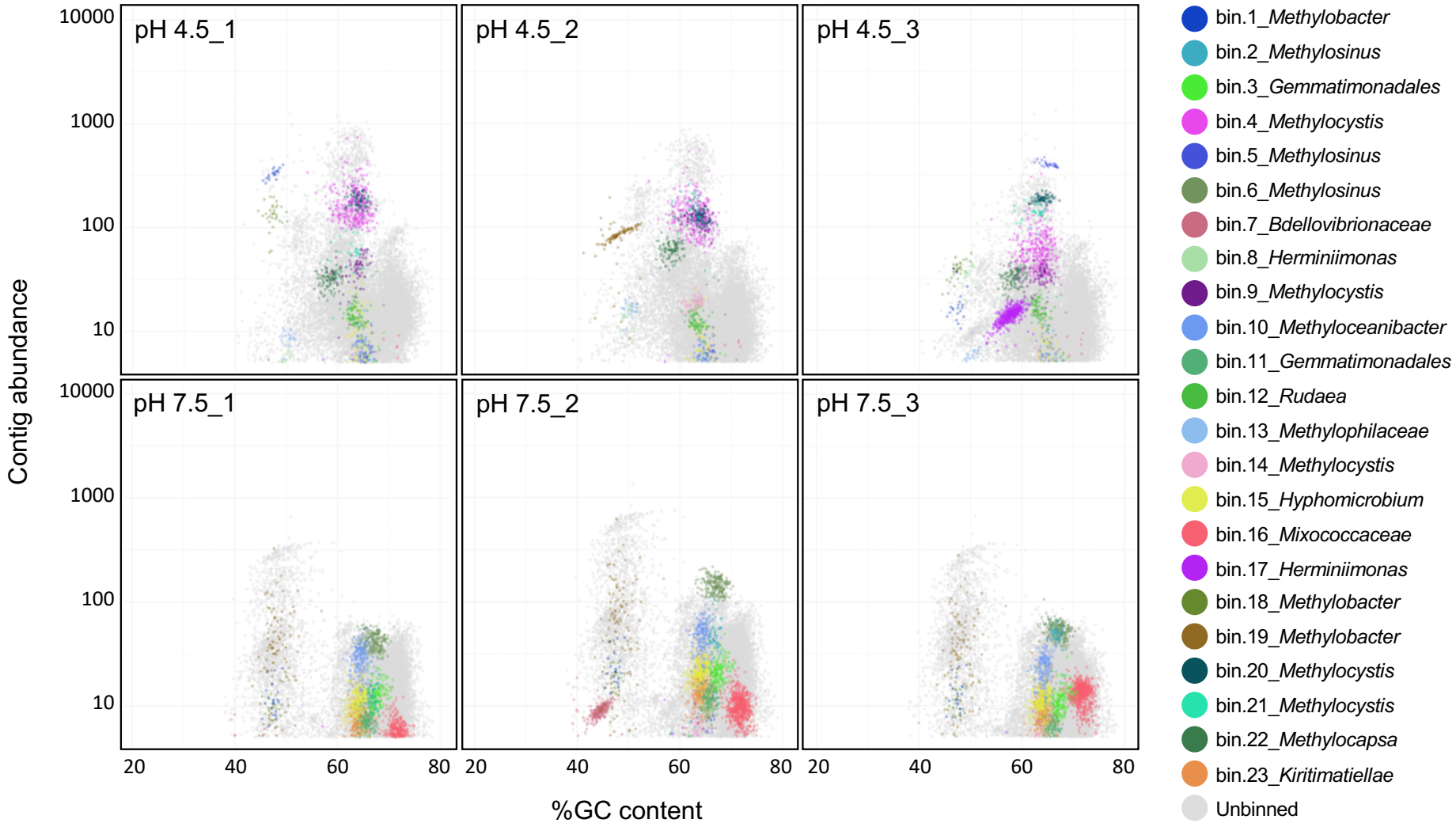


Figure 4.7. Contig-%GC coverage plots of the metagenomic assembled genomes (MAGs) from the pH 4.5 and 7.5 soil metagenomes. Contig abundance was calculated from standardized read coverage in each sample.

4.4.5. Virus prediction

In total, 650 contigs (min length = 5,003 bp; average length = 13,261 bp; max length = 96,180 bp) were predicted as metagenomic viral contigs (mVCs). Of these, 17, 205 and 428 mVCs were designated as category 1, 2 and 3, respectively, and 12 mVCs were in a circular form. In addition, 70 prophage sequences (29 and 41 prophages in category 5 and 6, respectively) were predicted. Viral community structure was distinct between soil pH (Figure 4.8). Only a small proportion of the viral contigs (14.7%) were taxonomically assigned to known viruses, with most of them belonging to *Caudovirales* (Figure 4.9).

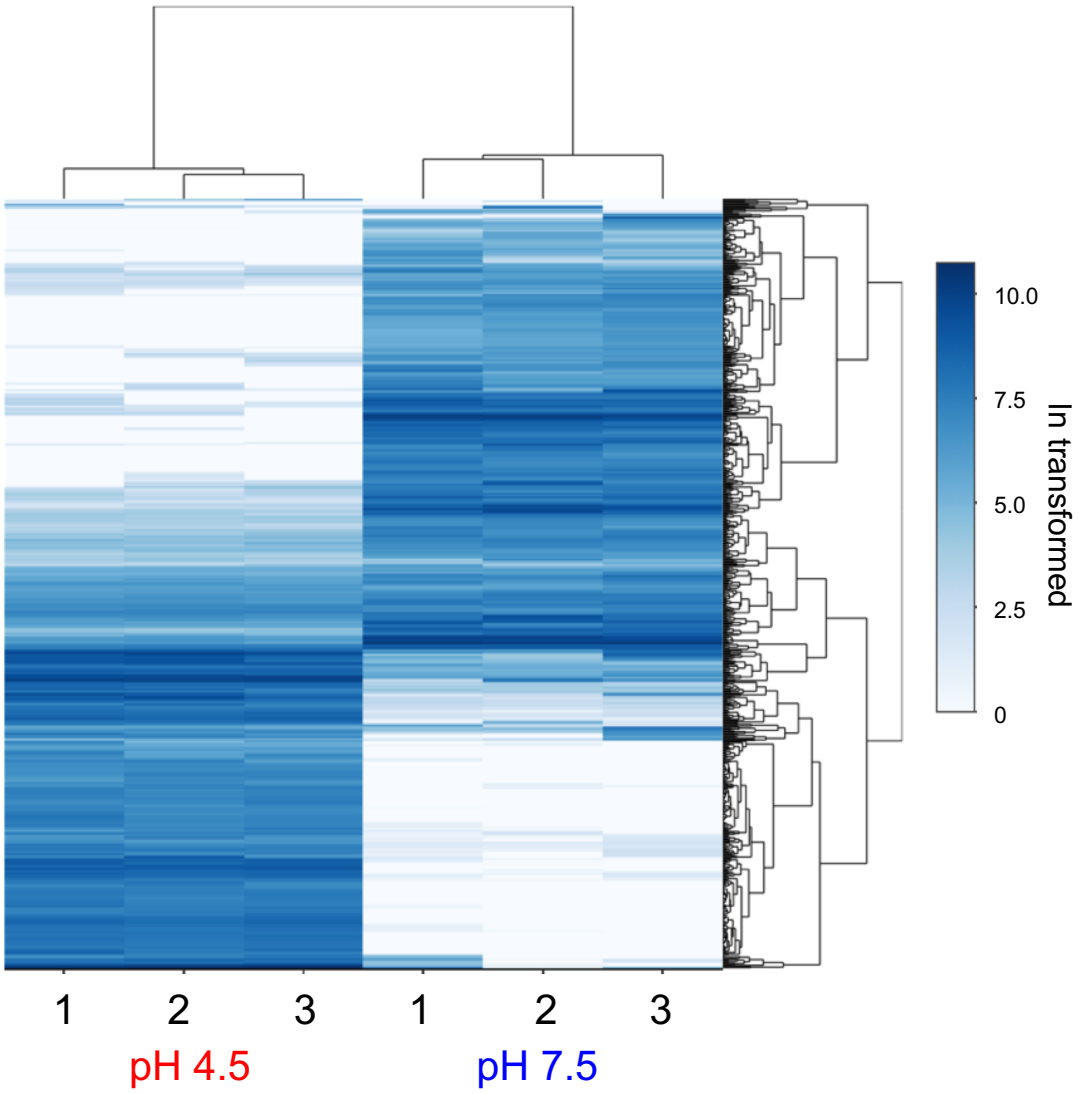


Figure 4.8. Relative abundance of ^{13}C-enriched metagenomic viral contigs (mVCs) and prophages in the pH 4.5 and 7.5 soil.

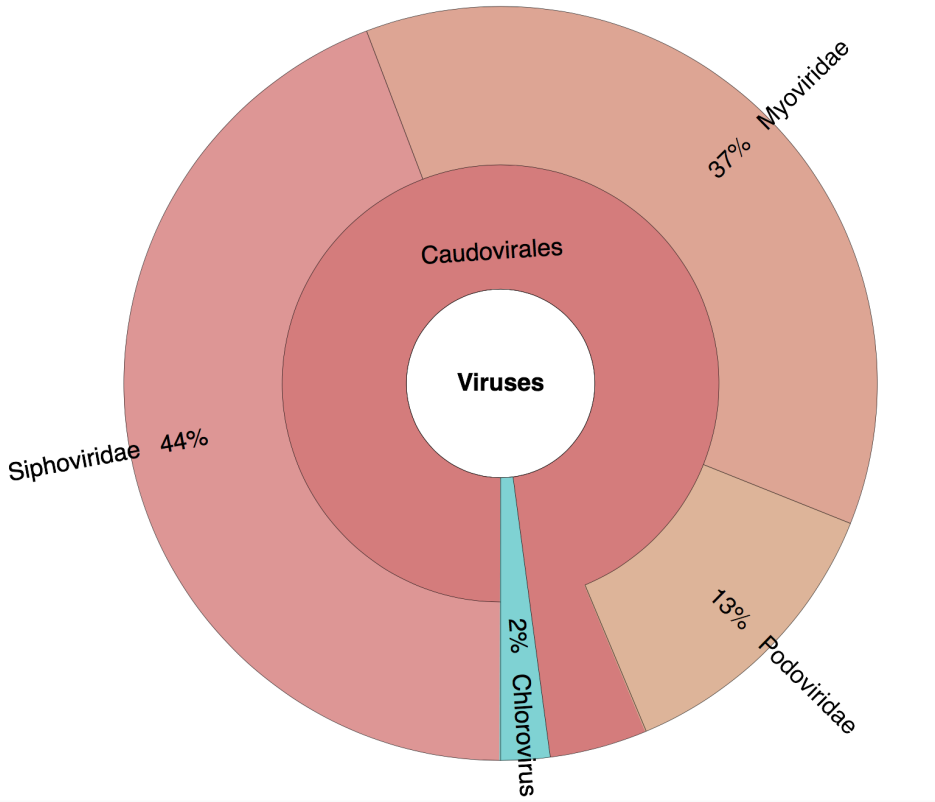


Figure 4.9. Taxonomic annotation of ^{13}C-enriched metagenomic viral contigs (mVCs) and prophages of the pH 4.5 and 7.5 soil. 85.3% of mVCs could not be assigned to a taxonomic group.

4.4.6. Host-virus linkage

4.4.6.1. CRISPR array analysis

Three of the 23 MAGs (bin.2_*Methylosinus*, bin.14_*Methylocystis* and bin.21_*Methylocystis*) contained CRISPR arrays, with the DR sequences of bin.14 and bin.21 being identical and indicating that they represent closely related populations. From the pH 4.5 and 7.5 soil metagenomes, 252 and 613 CRISPR arrays with 1,713 and 4,102 spacer sequences, respectively. were also identified in unbinned contigs. Of these, 49 CRISPR arrays from both soils contained spacers derived from mVCs (Supplementary Table 4.1) with 258 and 40 spacer sequences from pH 4.5 and 7.5 soil samples, respectively, matched to 23 mVCs (Supplementary Table 4.1). Also, DR sequences of 13 CRISPR arrays from pH 4.5 soil were the same as those found in both bin.14_*Methylocystis* and bin.21_*Methylocystis*. Analysis of contigs with CRISPR arrays with spacer sequences matched to mVCs included *Methylocystis*, *Methylobacterium*, *Methylocystis*, *Beijerinckia*, *Burkholderia, Chelatococcus, Desulfovibrio, Pannonibacter, Pseudomonas and Xanthobacter* (Supplementary Table 4.2).

In the CRISPR arrays present in the three methanotroph MAGs, spacers matched to six different mVCs (Figure 4.10). The size of the MAG-associated mVCs ranged between 15 - 60 kb

and had between 33 - 94 genes (Table 4.3). Bin.2_*Methylosinus* contained a CRISPR array (size = 24942 bp) comprising 18 DRs (37 bp) interspaced by 17 spacers (34 – 42 bp), and had six predicted *cas* genes (Type I) (Figure 4.10a). In bin.14_*Methanocystis*, the CRISPR array (size = 7805 bp) did not have identified *cas* genes due to the CRISPR array starting 96 bp from the end of the contig, it contained 115 DRs (32 bp) interspaced by 114 spacers (33 – 38 bp). In this CRISPR, spacers were derived from three mVCs (mVC_07548-cat_2, mVC _08211-cat_3 and mVC _12213-cat_2) (Figure 4.10b) with spacers from mVC_12213-cat_2 being the most abundant. Bin.21_*Methanocystis* contained a CRISPR array with 74 spacers and 75 direct repeats, the latter being identical to those in bin.14_*Methanocystis* indicating that the two populations are closely related. Spacers in the CRISPR array of bin.21_*Methanocystis* were also dominated by sequences found in mVC_12213-cat_2. However, spacers were also derived from two mVCs that were not present in bin.21_*Methanocystis*, indicating that the two populations have specific and shared viral populations infecting them.

Table 4.3. Metagenomic viral contigs associated with MAGs

Virus ID	Genome size (bp)	Number of genes	Associated host bin
mVC_20187-cat_2	41,704	49	bin.2_*Methylosinus*
mVC_12213-cat_2	60,841	94	bin.14_*Methylocystis*, bin.21_*Methylocystis*
mVC_07548-cat_2	45,200	73	bin.14_*Methylocystis*, bin.21_*Methylocystis*
mVC_08211-cat_3	15,210	33	bin.14_*Methylocystis*
mVC_12249-cat_2	41,363	75	bin.21_*Methylocystis*
mVC_00108-cat_2	40,899	65	bin.21_*Methylocystis*

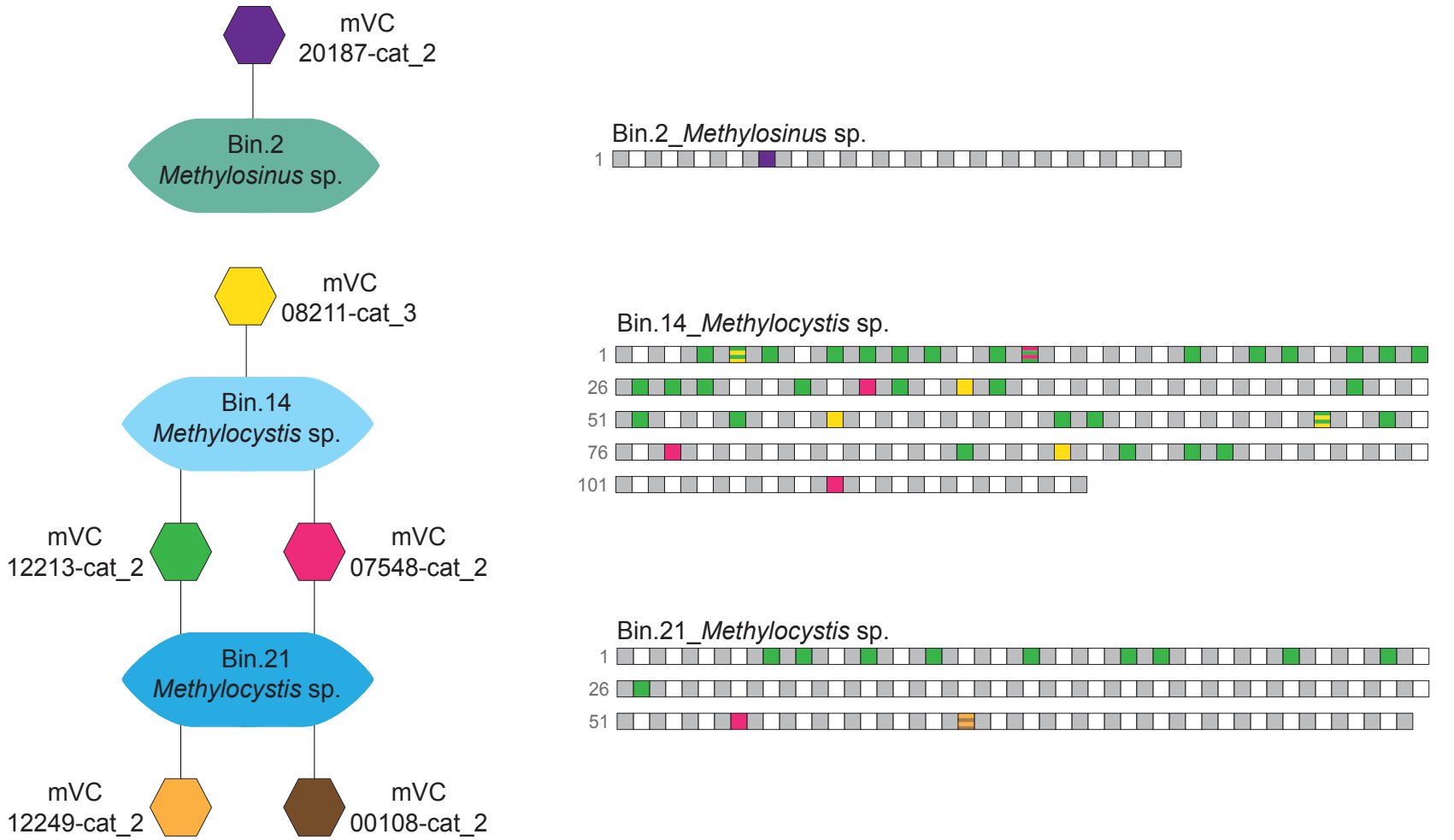


Figure 4.10. Schematic representation of mVCs derived from viruses infecting three methanotroph populations and the presence of derived spacers in CRISPR arrays. Grey squares represent direct repeats and colored squares correspond to spacers identical in sequence to one of six mVCs, with white boxes unassigned. Boxes with a striped colour represent a sequence found in two different mVCs.

Five of the six methanotroph MAG-associated mVCs formed a group distinct from other characterized viruses (Figure 4.11) with their closest association being to a virus infecting *Roseobacter*. Virus mVC_07548-cat_2 was distinct from the other five and was most closely related to a *Rhodovulum* phage, with all six viruses associated with those infecting other Alphaproteobacteria.

The abundance of the mVCs in soil was largely related to their host abundances in the respective pH soil (Figure 4.12). mVC_20187-cat_2 and its host, bin.2_*Methylosinus*, were more abundant in the pH 7.5 soil than the pH 4.5 soil. Although bin.14_*Methanocystis* had the same relative abundance in both soils, all four associated mVCs were in the pH 7.5 soil. Similarly, all viruses associated with the other *Methylocystis* MAG (bin.21) were also more abundant in the pH 4.5 soil, although this is also corresponded with the abundance of the host.

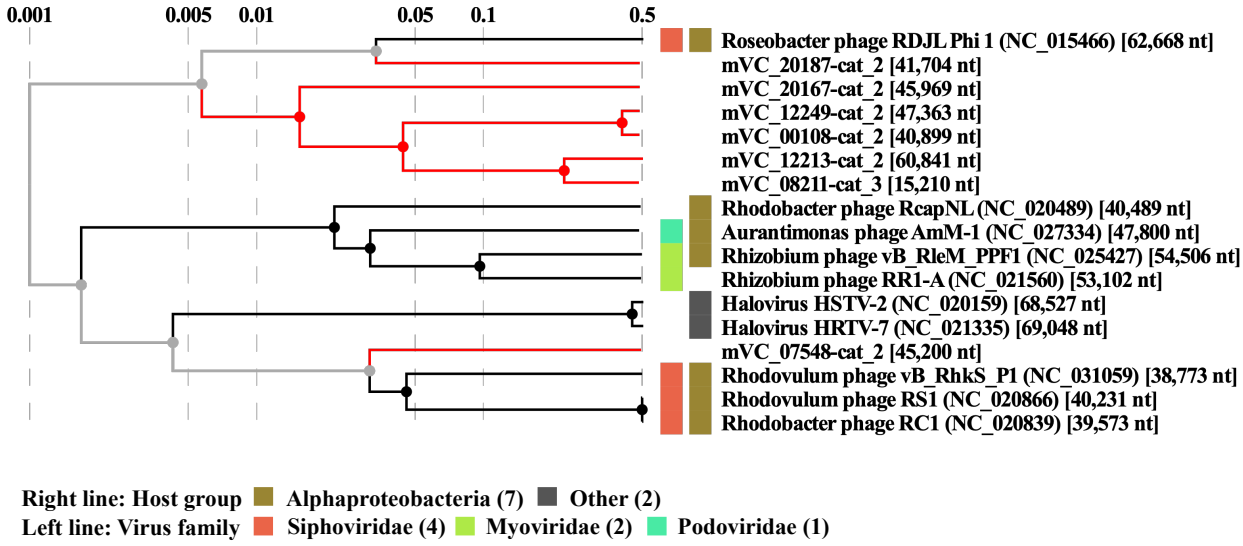


Figure 4.11. Proteomic tree showing the relationship of six methanotroph-associated metagenomic viral contigs (mVCs) (shown in red) with the most closely related characterized reference viruses and hosts.

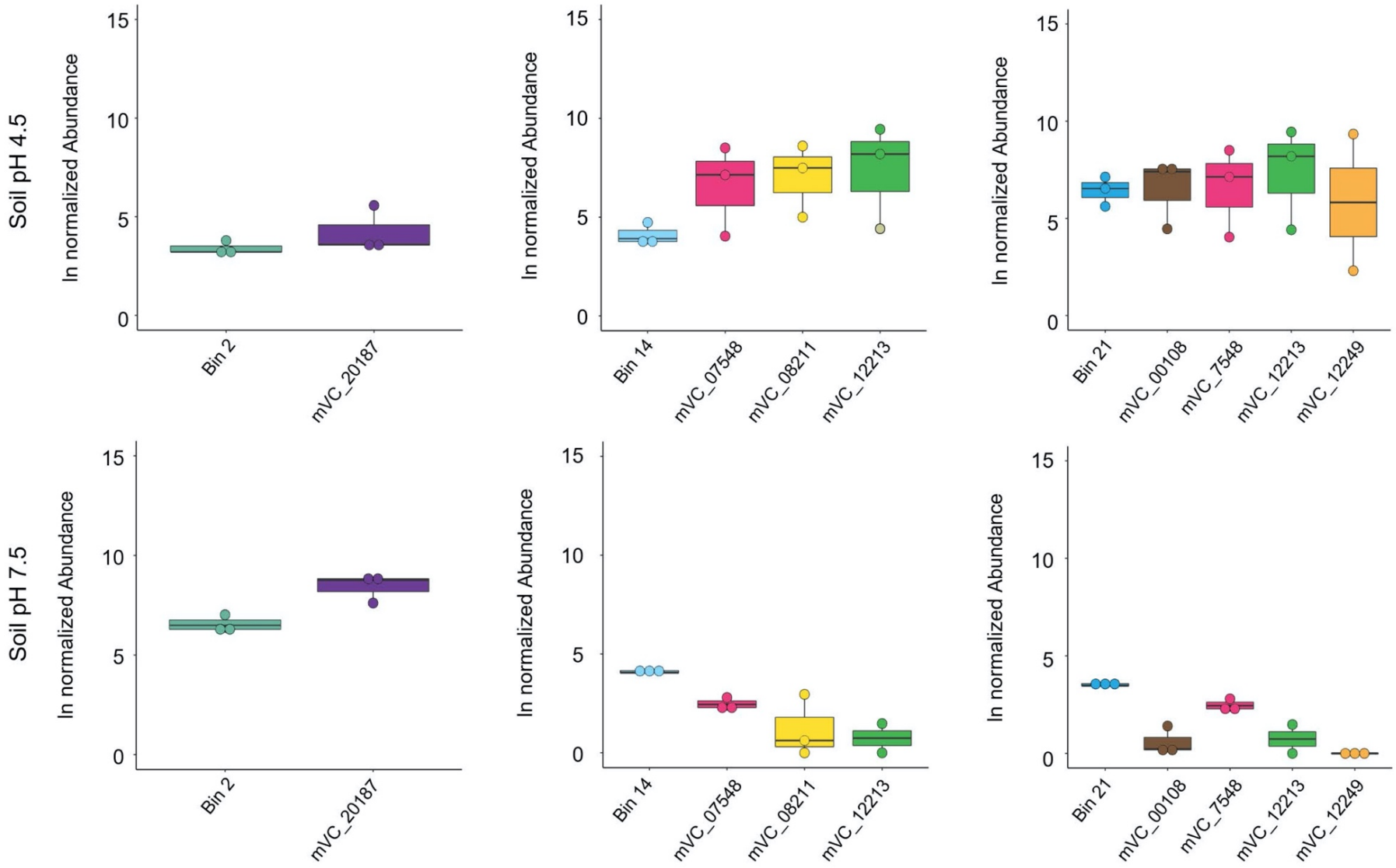


Figure 4.12. Normalized relative abundance of the three methanotroph hosts (bin) and their associated viruses (mVCs) in pH 4.5 and 7.5 soil.

4.4.6.2. ONF analysis

ONF analysis resulted in a total of 245 host-virus linkages (p-value < 0.05) (Table 4.4). Most of the viral contigs (117 out of 245) were predicted to be associated with methanotrophs, including *Methylocystaceae*, *Methylococcaceae*, *Beijerinckiaceae* and *Methylobacteriaceae* with 62 mVCs linked to methanotroph MAGs and 47 mVC linked to unbinned contigs. Nine mVCs were predicted to be associated with non-methane oxidizing methylotrophs. Four mVCs were linked to nitrifiers (*Nitrosomonas* and *Nitrospira*), and eight mVCs linked to bacterial predators, *Bdellovibrionaceae*, *Desulfovibrio*, *Lysobacter* and *Myxococcus*. Finally, 74 mVCs were linked to 51 other genera.

None of the methanotroph-associated mVCs that were linked via CRISPR arrays were found to be predicted by WIsH, with the exception of one mVC (mVC_20187-cat_2). This mVC was linked to bin.5_*Methylosinus* via WIsH, whereas it was linked to bin.2_*Methylosinus* via CRISPR array analysis.

Table 4.4. ONF host-virus linkages using the WIsH tool.

Host ID	Number of associated mVCs
Methanotrophs	
bin.1_*Methylobacter*	2
bin.4 _*Methylocystis*	5
bin.5_ *Methylosinus*	47
bin.18_*Methylobacter*	4
bin.19_*Methylobacter*	3
bin.22_*Methylocapsa*	1
Beijerinckia	2
Methylobacterium	5
Methylocella	8
Methylocystis	9
Methylomicrobium	12
Methylomonas	11
Methylotrophs	
Candidatus *Methylopumilus*	1
Hyphomicrobium	5
Methyloceanibacter	1
Methylophilus	1
Methylovorus	1
Predators	
bin.7_ *Bdellovibrionaceae*	5
Desulfovibrio	1
Lysobacter	1
Myxococcus	1
Nitrifiers	

Nitrosomonas	3
Nitrospira	1
Other	
bin.17_*Herminiimonas*	1
bin.23_*Kiritimatiellae*	1
Other (contains 51 genera)	74
Unknown contig	39
Total	245

4.4.7. Gene homology

4.4.7.1. Auxiliary metabolic genes

In total, 12,451 genes were predicted from VirSorter category 1, 2 and 3 mVCs, with 49% (6,072) annotated representing 753 unique functions. Genes encoding viral proteins accounted for 6.9% (424 genes) and included major capsid proteins, tail proteins, integrases, portal proteins and terminases. Bacterial proteins used for viral replication, such as nucleic acid synthesis and DNA repair, accounted for 4.8% (292 genes). AMGs associated with carbon metabolism included glycoside hydrolase/transferases (GH family 19, 24, 25, 46) (< 1%, 11 genes) and peptidases (< 1%, 79 genes), and for nitrogen metabolism, *nifH* and cytochrome cd1-nitrite reductase-like gene were identified. Proteins encoding for cofactors involved in methane oxidation, such as FAD- and NADH-binding domain, and [2Fe-2S] ferredoxin domain, copper binding site and cupredoxin like domain, were identified (< 1%, 44 genes), and AMGs linked to methane oxidation were *pmoC* (1 gene) and cytochrome C family gene (4 genes) (Figure 4.13). However, mVC_09622-cat_3 that contained the *pmoC* gene does not appear to have a high density of coding sequences and is atypical of virus genomes, thereby reducing support that this is a genuine viral contig.

In total, 1,797 genes were predicted from VirSorter category 5 and 6 mVCs (prophages), and 34% (616 genes) were annotated with 203 genes of unique function. Viral proteins and bacterial proteins used for viral replication accounted for 16% (101 genes) and 3.8% (24 genes) of the annotated genes, respectively. Genes involved in methane or carbon metabolism were not found.

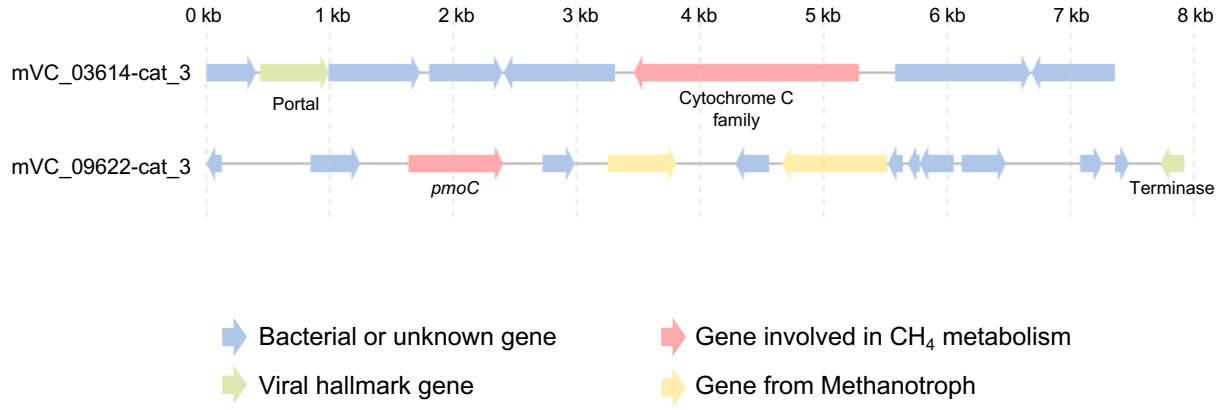


Figure 4.13. Two examples of metagenomic viral contigs (mVCs) that contain auxiliary metabolic genes (AMGs) involved in the methane oxidation.

4.4.7.2. Gene homology between viruses and their associated hosts

Gene homology between the CRISPR-containing MAGs and their associated viruses was assessed. A total of 9,142 genes, ranging between 2,679 – 3,491 genes per MAG, were predicted from the three bins (bin.2, 21 and 14), and were aligned to predicted protein sequences of the methanotroph-associated mVCs (Supplementary Table 4.3). All of the mVCs had at least one gene homologous to their host MAG (Table 4.5). The mVCs also had homologous genes to closely related MAGs, except mVC_12213-cat_2 had homologs specific to bin.14 and 21 (Figure 4.14, Supplementary Table 4.3). Contig mVC_20187-cat_2, which was linked to bin.2_*Methylosinus*, had the greatest number of homologs to its host (11 genes), with 5 genes also homologous to other MAGs (bin.12 and 14) (Supplementary Table 4.3). Although homologs were shared between the three methanotrophic MAGs and their six associated mVCs, the identity (30.6 – 99.4%) and bit score (44.7 - 831) were variable (Supplementary Table 4.3, Figure 4.14). The homologs with high identity to MAGs (> 80%) were annotated as encoding viral functions, including integrase (> 90%) and lysozyme (> 80%). Genes from mVC_12249-cat_2 and mVC_20187-cat_2 included those potentially encoding for glycoside hydrolase (GH) (Table 4.6, Figure 4.15). The identity of the GH from the mVC_12249-cat_2 was relatively high (> 80%). Viral proteins homologous to peptidase were also identified but with low identity (Table 4.6, Figure 4.14).

Of the 245 host-virus linkages identified via ONF analysis, 64% (157 mVCs) of the host-linked mVCs had at least one homologous gene to their associated host contig (Table 4.7, Supplementary Table 4.5). Specifically, 29 mVCs (1 gene), 23 mVCs (2 genes), 26 mVCs (3 genes), 18 mVCs (4 genes), 14 mVCs (5 genes), 11 mVCs (6 genes), 7 mVCs (7 genes), 6 mVCs (8 genes), 3 mVCs (9 genes) and 20 mVCs (> 10 genes) exhibited homology to their host. As expected, the prophages had a greater number of shared genes with their associated host contig (mVC_12123-cat_5, 95 genes; mVC_12120-cat_6, 34 genes; mVC_12131-cat_5, 27 genes).

Table 4.5. Gene homology of the six methanotroph-associated mVCs, determined through CRISPR array analysis, to host metagenomic assembled genomes (MAGs).

Virus ID	Host MAG ID	Number of genes	Number of homologs with host
mVC_00108-cat_2	bin.21_*Methylocystis*	65	3
mVC_07548-cat_2	bin.14_ *Methylocystis*, bin.21 *Methylocystis*	73	4
mVC_08211-cat_3	bin.14_*Methylocystis*	33	4
mVC_12213-cat_2	bin.14 _*Methylocystis,* bin.21_*Methylocystis*	94	1
mVC_12249-cat_2	bin.21_*Methylocystis*	75	4
mVC_20187-cat_2	bin.2_*Methylosinus*	49	11

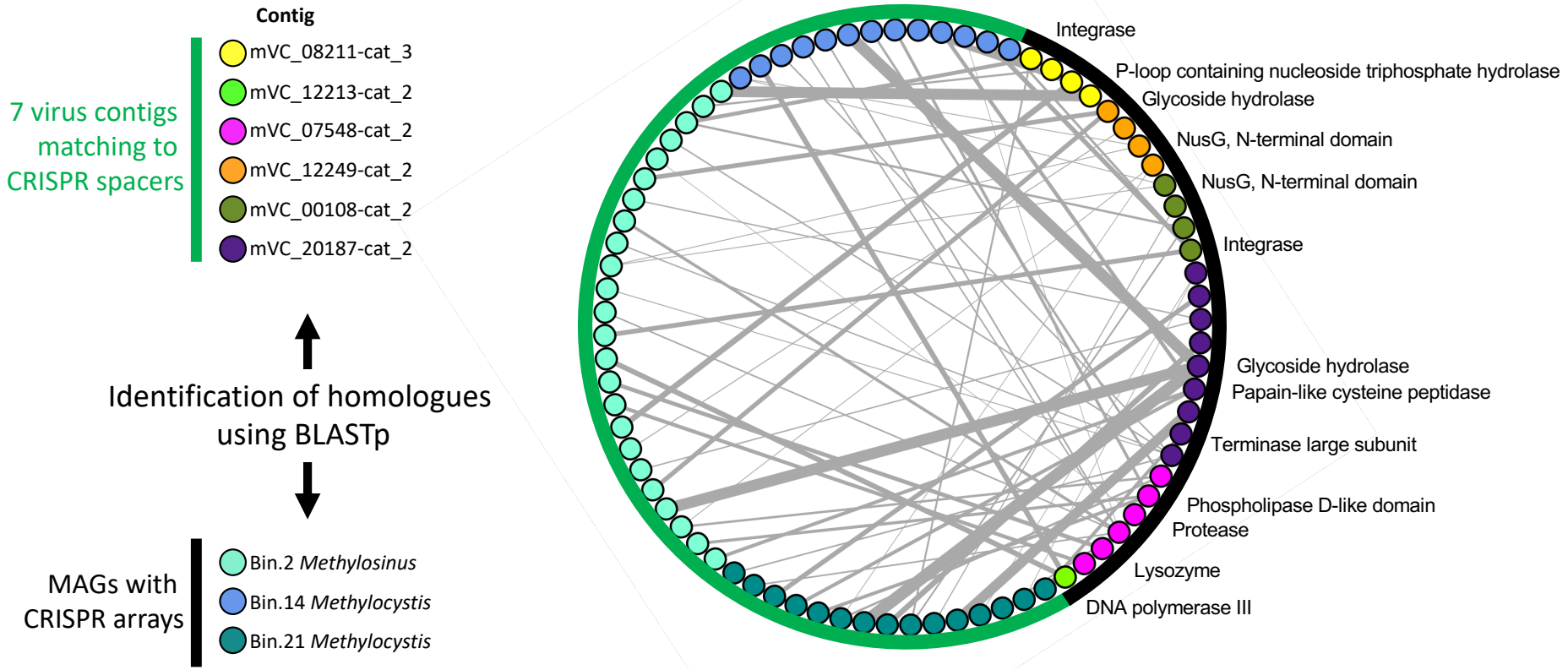


Figure 4.14. Network analysis showing homologous auxiliary metabolic genes (AMGs) linkage of homologous genes between six methanotroph-associated metagenomic viral contigs (mVCs) and three hosts (bin.2, 14 and 21). Black and green lines highlight viral and host genes, respectively. Nodes color indicates the origin of each methanotroph-associated mVC or each methanotroph metagenomic assembled genome (MAG). Edges are proportional to shared identity and calculated by the bit score divided by 100.

Table 4.6. Protein assignment of viral genes that have homologs in methanotroph metagenomic assembled genomes (MAGs). Only annotated genes are shown.

Query ID (Viral protein ID)	VP AA length[1]	Subject ID (Host protein ID)	HP AA length[2]	Identity (%)	Protein function	E value
mVC_000108-cat_2_1	379	bin.21_02902	404	48.29	Integrase, catalytic domain	1.10e-17
mVC_000108-cat_2_1	379	bin.14_01987	405	45.57		
mVC_000108-cat_2_1	379	bin.14_01627	405	35.82		
mVC_000108-cat_2_43	213	bin.21_02323	176	30.81	NusG, N-terminal domain superfamily	3.40e-15
mVC_000108-cat_2_43	213	bin.14_00936	176	31.55		
mVC_007548-cat_2_29	298	bin.21_00796	301	42.24	Bacteriophage Mu, GpT	4.40e-112
mVC_007548-cat_2_31	347	bin.21_00795	316	51.52	Protease, Mu phage/prophage I type	4.00e-18
mVC_007548-cat_2_49	170	bin.21_01541	169	63.69	Phospholipase D-like domain	8.30e-19
mVC_007548-cat_2_49	170	bin.2_02674	243	43.29		
mVC_007548-cat_2_49	170	bin.14_03309	183	43.75		
mVC_007548-cat_2_49	170	bin.2_02140	180	47.91		
mVC_007548-cat_2_49	170	bin.21_02786	180	45.13		
mVC_007548-cat_2_49	170	bin.14_03464	241	42.44		
mVC_007548-cat_2_49	170	bin.14_00002	393	30.65		
mVC_008211-cat_3_1	370	bin.21_01138	370	99.45	Integrase, catalytic domain	2.90e-14
mVC_008211-cat_3_17	379	bin.21_00915	365	54.69	P-loop containing nucleoside triphosphate hydrolase	3.57e-35
mVC_008211-cat_3_17	379	bin.14_02736	382	55.08		
mVC_008211-cat_3_3	146	bin.21_01140	139	93.47	Protein of unknown function DUF4326)	2.50e-6
mVC_008211-cat_3_3	146	bin.2_02473	150	50.00		
mVC_012213-cat_2_50	373	Bin.14_00929	372	50.93	DNA polymerase III, beta sliding clamp, C-terminal	2.20e-24
mVC_012213-cat_2_50	373	bin.21_00585	372	50.40		
mVC_012249-cat_2_12	150	bin.21_00988	145	34.00	Papain-like cysteine peptidase superfamily	1.81e-10
mVC_012249-cat_2_12	150	bin.2_02315	144	34.18		
mVC_012249-cat_2_37	213	bin.21_02323	176	30.81	NusG, N-terminal domain superfamily	3.40e-15

mVC_012249-cat_2_37	213	bin.14_00936	176	31.55		
mVC_012249-cat_2_43	182	bin.21_00974	211	32.86	HD-domain/PDEase-like superfamily	2.12e-35
mVC_012249-cat_2_9	216	bin.21_01239	276	85.22	Glycoside hydrolase, family 24	4.20e-17
mVC_012249-cat_2_9	216	bin.21_00984	272	83.05		
mVC_020187-cat_2_15	95	bin.2_01433	86	72.00	Uncharacterized protein conserved in bacteria (DUF2312)	1.90e-29
mVC_020187-cat_2_15	95	bin.14_00738	80	71.23		
mVC_020187-cat_2_15	95	bin.21_01207	84	67.12		
mVC_020187-cat_2_15	95	bin.21_02304	85	68.49		
mVC_020187-cat_2_15	95	bin.21_02731	89	41.02		
mVC_020187-cat_2_25	594	bin.2_00624	532	31.22	Terminase large subunit, Lambdalikevirus-type	1.20e-98
mVC_020187-cat_2_3	380	bin.2_02481	380	97.14	Integrase, catalytic domain	3.10e-26
mVC_020187-cat_2_39	212	bin.14_00273	212	61.50	Protein of unknown function DUF2460	2.60e-79
mVC_020187-cat_2_39	212	bin.2_02313	213	62.25		
mVC_020187-cat_2_39	212	bin.21_00990	212	57.67		
mVC_020187-cat_2_40	297	bin.2_02314	292	42.08	Bacteriophage phiJL001, Gp84, C-terminal	1.50e-29
mVC_020187-cat_2_40	297	bin.21_00989	294	40.54		
mVC_020187-cat_2_40	297	bin.14_00274	296	39.73		
mVC_020187-cat_2_41	148	bin.21_00988	145	54.54	Papain-like cysteine peptidase superfamily	1.37e-17
mVC_020187-cat_2_41	148	bin.14_00275	142	54.54		
mVC_020187-cat_2_41	148	bin.2_02315	144	53.84		
mVC_020187-cat_2_42	1368	bin.2_02316	1285	38.75	Glycoside hydrolase superfamily	4.55e-9
mVC_020187-cat_2_42	1368	bin.14_00276	1292	38.08		
mVC_020187-cat_2_42	1368	bin.21_00987	1296	37.13		
mVC_020187-cat_2_43	526	bin.2_02317	526	70.39	Protein of unknown function DUF2793	1.70e-34
mVC_020187-cat_2_44	183	bin.2_02301	188	85.95	Lysozyme-like domain superfamily	5.91e-20

Table 4.7. Shared genes between mVCs and host contigs, bins and to representatives of the same genus as the predicted host determined by ONF analysis. Predictions of taxa for the host contigs was based on analysis of homologous genes.

Virus ID	Host contig ID	Host MAG ID	Prokaryote taxa	Number of the homologs with host contig	Number of the homologs with bin	Number of homologs with genus of host	Matrix	*p*-value
mVC_12294-cat_6	c_21910	bin.7	*Bdellovibrio*	0	0	1	-1.38	1.82e-3
mVC_20207-cat_5	c_21149	bin.7	*Bdellovibrio*	0	0	16	-1.37	0
mVC_20210-cat_2	c_21787	bin.7	*Bdellovibrio*	0	0	12	-1.38	1.41e-3
mVC_06341-cat_2	c_00661	bin.4	*Beijerinckia*	5	5	1	-1.28	0
mVC_01595-cat_3	c_10118		*Hyphomicrobium*	11		1	-1.29	0
mVC_08608-cat_3	c_14956		*Hyphomicrobium*	11		1	-1.30	0
mVC_13586-cat_3	c_11733		*Hyphomicrobium*	1		1	-1.37	0
mVC_00769-cat_2	c_15791		*Methylobacterium*	6		2	-1.33	0
mVC_02890-cat_3	c_12869		*Methylobacterium*	7		1	-1.31	0
mVC_08785-cat_2	c_09673		*Methylobacterium*	5		14	-1.36	0
mVC_00204-cat_6	c_07946		*Methyloceanibacter*	8		3	-1.32	0
mVC_02377-cat_3	c_09879		*Methylocella*	0		1	-1.38	4.31e-9
mVC_08048-cat_3	c_00202		*Methylocella*	3		1	-1.37	2.38e-8
mVC_13278-cat_3	c_00202		*Methylocella*	2		5	-1.36	1.14e-12
mVC_00059-cat_3	c_00052		*Methylocystis*	15		2	-1.35	1.22e-8
mVC_02569-cat_3	c_07752		*Methylocystis*	4		2	-1.31	0
mVC_09622-cat_3	c_12264		*Methylocystis*	3		2	-1.35	1.13e-12
mVC_11036-cat_3	c_16224		*Methylocystis*	6		5	-1.30	0
mVC_11479-cat_3	c_14680		*Methylocystis*	1		1	-1.36	0
mVC_27875-cat_2	c_12160	bin.18	*Methylomicrobium*	0	0	1	-1.38	2.17e-5
mVC_28899-cat_2	c_20560		*Methylomicrobium*	0		5	-1.38	1.33e-7
mVC_00010-cat_5	c_18004		*Methylomonas*	4		4	-1.37	0
mVC_00453-cat_3	c_30908		*Methylomonas*	0		1	-1.38	0.01

mVC_09301-cat_3	c_08129		*Methylomonas*	2		2	-1.37	0
mVC_12177-cat_6	c_22473		*Methylomonas*	2		1	-1.38	6.09e-3
mVC_19312-cat_3	c_21845		*Methylomonas*	0		3	-1.38	7.45e-9
mVC_24690-cat_3	c_21670		*Methylomonas*	2		3	-1.37	0
mVC_25087-cat_2	c_07420	bin.19	*Methylomonas*	4	0	9	-1.38	5.20e-5
mVC_25450-cat_3	c_20299		*Methylomonas*	1		1	-1.38	1.69e-6
mVC_26577-cat_3	c_07420	bin.19	*Methylomonas*	5	0	4	-1.37	5.13e-7
mVC_30098-cat_2	c_07420	bin.19	*Methylomonas*	6	0	1	-1.37	6.40e-7
mVC_04118-cat_3	c_12119	bin.5	*Methylosinus*	0	0	2	-1.30	0.02
mVC_12131-cat_5	c_12119	bin.5	*Methylosinus*	27	0	6	-1.28	8.00e-3
mVC_20187-cat_2	c_12119	bin.5	*Methylosinus*	0	0	1	-1.28	7.72e-3

4.4.7.3. Gene homology of host-linked viruses to other prokaryotes

For the CRISPR array based linkages, all of the mVCs had at least one homologous gene to the genus of their of associated host (*Methylosinus* or *Methylocystis*), but they also had genes affiliated to several other methanotroph genera, including *Methylobacterium*, *Methylobrevis*, *Methylocella*, *Methyloferula* (Table 4.8; Supplementary Table 4.4). Specifically, mVC_20187 associated with bin.2_*Methylosinus* had 20 genes homologous to those of other methanotrophs, including *Methylocystis* (2 genes) and other *Methylocystaceae* (18 genes). mVC_08211, associated with bin.14_*Methylocystis*, had genes affiliated to methanotrophs, including *Methylobacterium* and *Methylosinus*, and to methylotrophic *Hyphomicrobiaceae*. mVC_12213 associated with bin.14_ and 21_*Methylocystis* had genes homologous to those of various methanotrophs, including *Methylobacterium*, *Methylosinus* and *Methylobrevis*. The mVC_07548 associated with bin.14_ and 21_ *Methylocystis* had genes homologous to those of various methanotroph genera, including *Methylobacterium* and *Methylosinus*. Lastly, the mVC_00108 and mVC_12249 associated with bin.21_*Methylocystis* had genes homologous to those of various methanotrophs, including *Methylobacterium*, *Methylocella*, *Methyloferula* and *Methylosinus*.

For the ONF based linkages, only 43 of the 206 mVCs had genes homologous to those of the same genera as their predicted host (Table 4.7; Supplementary Table 4.5). Of these, 29 mVCs were associated to methanotrophs, including the genera *Methylosinus* (3 mVCs), *Methylocystis* (5 mVCs), *Methylobacterium* (3 mVCs), *Methylocella* (3 mVCs) *Methylomicrobium* (2 mVCs), and *Methylomonas* (10 mVCs). Methylotroph-associated mVCs (4 mVCs) were linked to the genera *Beijerinckia* (1 mVC), *Hyphomicrobium* (3 mVCs) and *Methyloceanibacter* (1 mVC) were also identified (Table 4.7). In addition, although three mVCs predicted to infect the *Bdellovibriodaceae*, they did not possess homology to those in any linked contig or MAG, they possessed genes (1, 12 and 16 genes) matched to other *Bdellovibrio* (Table 4.7).

Table 4.8. Gene homology of the six methanotroph-associated metagenomic viral contigs (mVCs), determined through CRISPR array analysis, to database prokaryotes.

		Taxonomy of the database prokaryote		
Virus-host linked by CRISPR array	Number of homologs	Order	Family	Genus
mVC_08211-cat_3 and bin.2_*Methylosinus*				
	1	*Rhizobiales*	*Hyphomicrobiaceae*	*Devosia*
	2	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
	2	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*
	2	*Rhizobiales*	2 mixed families	2 mixed genera
	3	Other[1]	3 mixed families	3 mixed genera
mVC_20187-cat_2 and bin.2_*Methylosinus*				
	1	*Rhizobiales*	*Aurantimonadaceae*	*Aureimonas*

2	*Rhizobiales*	*Methylocystaceae*	*Methylocystis*
18	*Rhizobiales*	*Methylocystaceae*	NA
2	Other[1]	2 mixed families	2 mixed families
mVC_07548-cat_2 and bin.14 _ and bin.21_*Methylocystis*			
1	*Nitrosomonadales*	*Nitrosomonadaceae*	*Nitrosomonas*
1	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
4	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*
10	*Rhizobiales*	4 mixed families	7 mixed genera
12	Other[1]	8 mixed families	11 mixed genera
mVC_12213-cat_2 and bin.14_ and 21_*Methylocystis*			
24	*Rhizobiales*	4 mixed families	9 mixed genera
4	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
3	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*
1	*Rhizobiales*	NA	*Methylobrevis*
3	Other[1]	8 mixed families	14 mixed genera
mVC_00108-cat_2 and bin.21_*Methylocystis*			
1	*Nitrosomonadales*	*Nitrosomonadaceae*	*Nitrosovibrio*
1	*Nitrospirales*	*Nitrospiraceae*	*Nitrospira*
1	*Rhizobiales*	*Beijerinckiaceae*	*Methylocella*
1	*Rhizobiales*	*Beijerinckiaceae*	*Methyloferula*
11	*Rhizobiales*	2 mixed families	5 mixed genera
1	*Rhizobiales*	*Hyphomicrobiaceae*	*Devosia*
2	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
5	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*
2	*Rhizobiales*	*Methylocystaceae*	NA
16	Other[1]	8 mixed families	13 mixed genera
mVC_12249-cat_2 and bin.21_*Methylocystis*			
1	*Nitrospirales*	*Nitrospiraceae*	*Nitrospira*
2	*Rhizobiales*	*Beijerinckiaceae*	*Methylocella*
3	*Rhizobiales*	*Beijerinckiaceae*	*Methyloferula*
8	*Rhizobiales*	*Bradyrhizobiaceae*	5 mixed genera
1	*Rhizobiales*	*Hyphomicrobiaceae*	*Devosia*
4	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
7	*Rhizobiales*	*Methylocystaceae*	*Methylosinus*
2	*Rhizobiales*	*Methylocystaceae*	NA
1	*Rhizobiales*	*Phyllobacteriaceae*	*Aminobacter*

[1]Other contains 12 different orders (*Alteromonadales, Bacillales, Burkholderiales, Enterobacterales, Oceanospirillales, Planctomycetales, Pseudomonadales, Rhodobacterales, Rhodocyclales, Rhodospirillales, Sphingomonadales, Xanthomonadales*).

4.4.7.4. Gene homology between methanotroph-associated viruses

A relatively high degree of identity was found between mVC_12249 and mVC_00108 – the two single host viruses that infected bin.21_*Methylocystis* (Figure 4.15). Most of the protein sequences of these two mVCs had unknown functions (mVC_12249, 30 genes; mVC_00108, 29 genes), and six genes from mVC_12249 and four genes from mVC_00108 were homologous to those of the host. Two of these genes encode for integrases, and the other two for glycoside hydrolase and NusG, N-terminal domain superfamily (see section 4.3.7.2, Table 4.6).

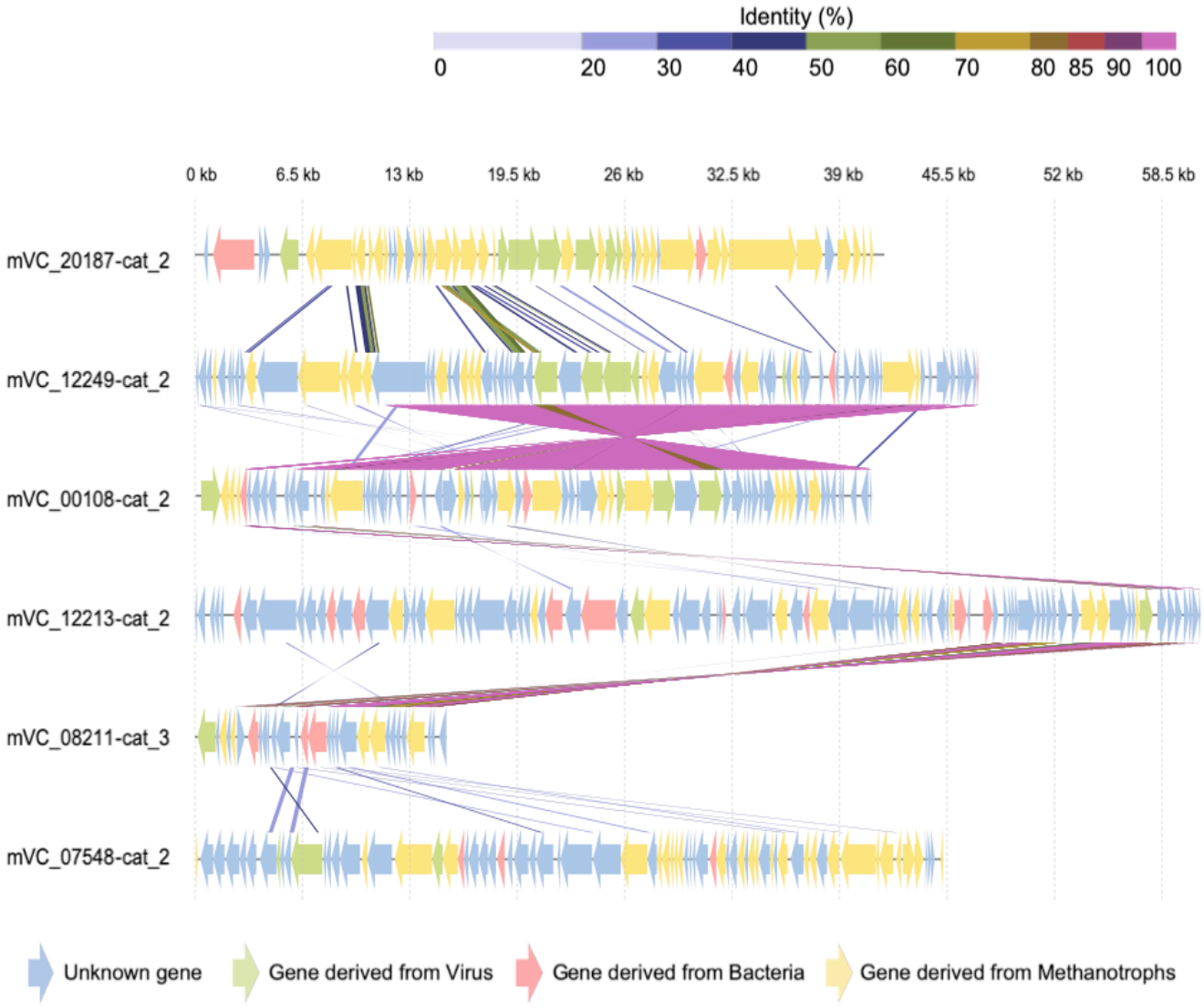


Figure 4.15. Genome map of the six metagenomic viral contigs (mVCs) that were linked to methanotroph hosts via CRISPR array analysis and highlighting the source of that gene (viral hallmark, methanotrophs, non-methanotrophic bacteria ort unknown). The color of the lines between the genes describes the percentage identity.

4.4.8. Viruses of methanotrophs and secondary users of methane carbon

A conceptual diagram of a microbial food web derived by characterizing ^{13}C-CH_4 enriched taxa (MAGs and contigs) is shown in Figure 4.17. Active methanotrophic populations oxidizing ^{13}C-methane to ^{13}C-methanol included *Methylocystaceae*, *Methylococcaceae* and *Beijerinckiaceae*. Subsequently, methylotrophic populations consuming either ^{13}C-methanol or ^{13}C-formate included *Methyloceanibacter*, *Methylophilaceae, Hyphomicrobium* (denitrifying methylotroph) and *Herminiimonas*. Finally, nitrifying populations of ammonia oxidizing archaea (AOA) and bacteria (AOB), and nitrite oxidizing bacteria (NOB), may have incorporated ^{13}C-CO_2 respired from other processes. In addition, bacterial predator populations, including *Myxococcaceae*, *Bdellovibrionaceae* and *Lysobacter*, were identified as incorporating ^{13}C by degrading active or dead ^{13}C-labeled methanotrophs or methylotrophs. In total, host-virus linkages through CRISPR array and ONF analyses, respectively, were discovered with 3 and 18 different taxonomic groups, respectively. The number of viruses linked to each taxon through CRISPR array and ONF analysis is shown in Figure 4.18. Only the mVCs that had at least one homologous gene to a host contig or bin, or to a database prokaryote of the same genera as the host (77 mVCs) were included. The abundance of viruses in each soil pH was coherent with the host relative abundance in each soil pH.

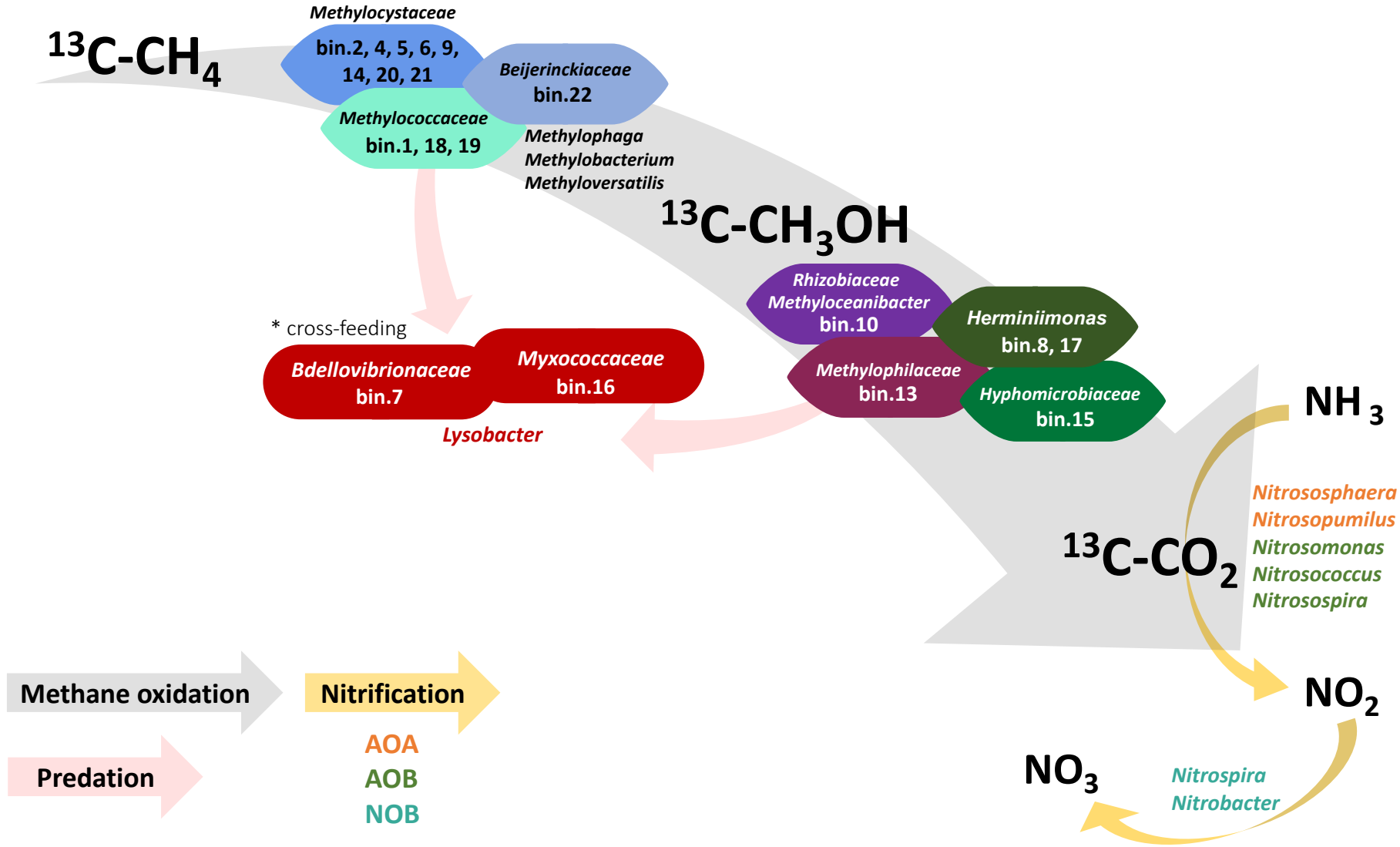


Figure 4.16. Conceptual diagram of the prokaryotic taxa involved in a microbial food web fueled by the consumption of CH_4-derived C.

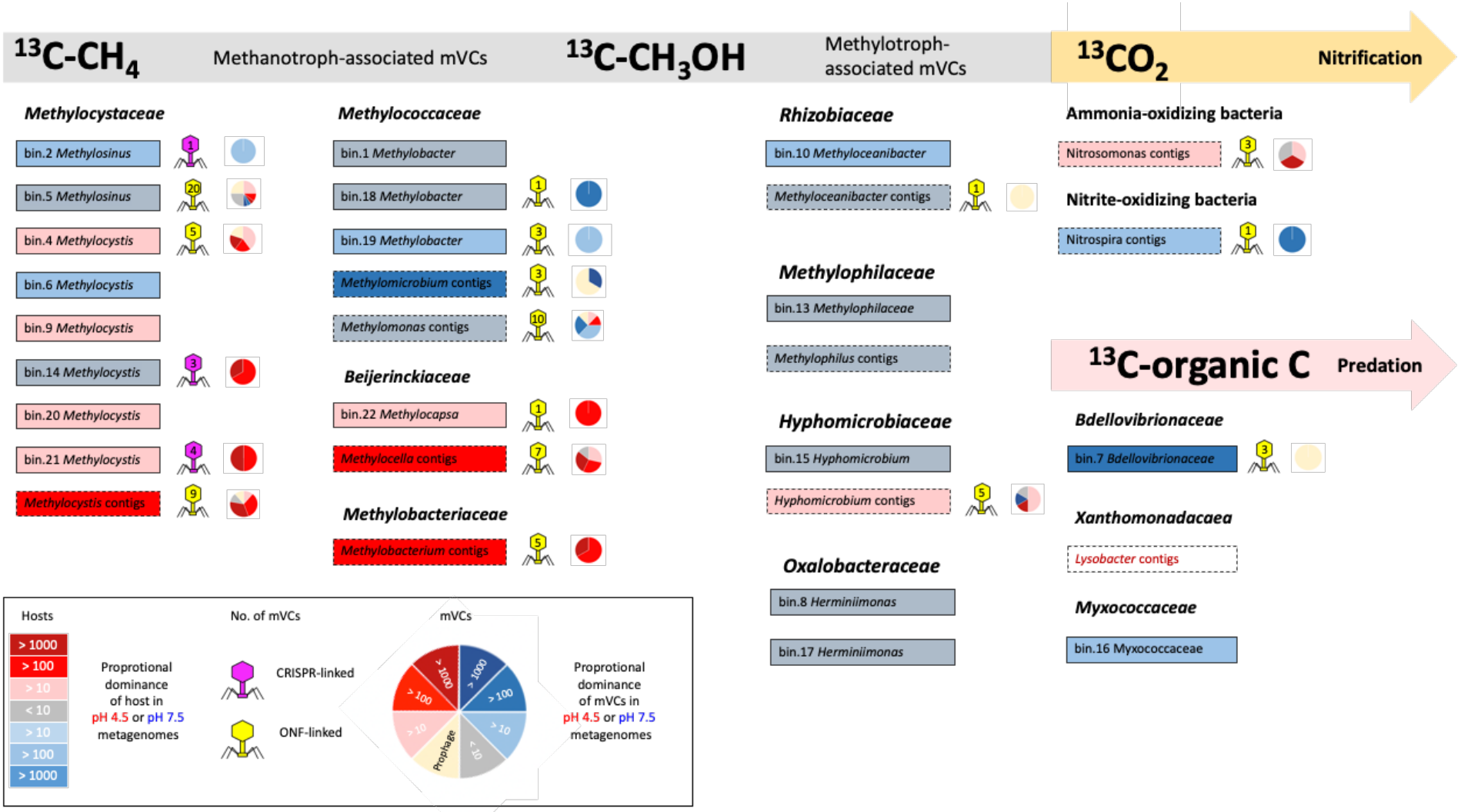


Figure 4.17. Summary of virus – host linkage based on CRISPR array or ONF analysis. The proportional distribution of each host (contig or bin) in one soil pH compared to other is color coded. The number of linked mVCs (by CRISPR array or ONF analysis) is given beside each host with the proportional distribution with respect to pH also given. The linkage of predicted prophages is also detailed,

4.5. Discussion

To target active methanotroph communities and their associated viruses, DNA-SIP was combined with metagenomics to analyze DNA enriched in CH_4-derived ^{13}C. Sequenced soil metagenomes were enriched in both methanotrophic (36%) and methanotroph-associated viral (720 mVCs) communities, including the recovery of 12 methanotroph and 3 methylotroph MAGs of medium to high quality. Furthermore, several ^{13}C-enriched viruses were linked to methanotroph hosts via the previous transfer of virus DNA into CRISPR arrays. Analysis of gene homology revealed AMGs involved with methane, carbon and nitrogen metabolism, amongst others. Overall, the results demonstrate that a combination of deep sequencing together with the targeting of a specific functional group dramatically reduces the diversity of profiled populations and facilitates detailed *in vivo* studies of virus-host interactions in soil.

As hypothesized, soil pH influenced both the growing methanotroph and methanotroph-associated viral community structure. Most of the contigs from pH 4.5 soil were annotated as *Methylocystaceae* (26%) and *Beijerinckiaceae* (4%), whereas in the pH 7.5 soil, *Methylococcaceae* dominated (23%). The *Methylocystaceae* and *Beijerinckiaceae*, within the Alphaproteobacteria, have previously been shown to be common inhabitants of acidic environments (Dedysh 2011; Shiau et al. 2018). However, *Methylococcaceae*, within the Gammaproteobacteria, have also been isolated from various acidic environments (Kolb et al. 2003; Dedysh 2011). Their relative low abundance does not necessarily indicate low abundance in the soil, but only that they were present in the pH 4.5 soil metagenomes with low abundance and did not grow under the enrichment conditions used. In addition, isolates of *Methylococcaceae*, such as *Methylobacter* and *Methylomicrobium* species, were unable to grow below pH 5 in culture (Bowman et al. 1990; Fuse et al. 1998; Sorokin et al. 2000; Wartiainen et al. 2006) and these results indicate that there were *Methylocystaceae* and *Beijerinckiaceae* were the major contributors to methane oxidation in the acidic soil microcosms. Soil viral communities have also been previously shown to be affected by soil pH (Narr et al. 2017; Adriaenssens et al. 2017). As soil pH can affect the availability of nutrients and toxic elements, acidophilic methanotrophs may potentially possess defense mechanisms against toxicity (Nguyen et al. 2018) as well as adaptations to allow growth at low pH. Comparative genomic analysis of the methanotrophic MAGs identified diverse membrane transporter systems, including potassium transporters, sodium pumps and ATP synthase that may contribute to key pH homeostasis mechanisms (Hou et al., 2008; Nguyen et al., 2018).

Through CRISPR array analysis, six viruses were linked to three methanotroph hosts, and demonstrated that a range of ^{13}C-enriched viruses was actively interacting with methylotroph hosts. The chronology of virus infections, resulting in a sequential timeline of foreign invaders, can be inferred from the CRISPR arrays (Strich and Chertow 2019). Although none of the most recently inserted spacers were matched to the six viruses, there was evidence of different virus

infection frequencies between the viruses (Martynov et al. 2017). It should also be noted that the mVCs represent incomplete viral genomes and therefore derivation of the most recent spacers being derived from these six viruses cannot be ruled out. Two of the viruses were both found to infect two *Methylocystis* hosts, whereas two other viruses were specific to only one *Methylocystis* host, suggesting there may be broad- and narrow-host range methanotroph viruses. Interestingly, the methanotroph viruses that were linked to a single host species share a greater number of homologous genes to their hosts, compared to the viruses that had multiple hosts. Also, compared to the single host viruses, the multi-host viruses had a variety of homologous genes to those found in other genera. These methanotroph viruses were not closely related to any known viruses, which is not surprising as studies on soil viruses are generally lacking (Roux et al. 2019; Trubl et al. 2020). Comparison of host-virus abundances suggested that the abundance of these viruses was coupled to their host abundance (Knowles et al. 2016). Lytic viruses infect and rapidly kill their infected host cells, thereby shaping host population dynamics. While the lysogenic life cycle may be an effective common strategy to enable viral populations to persist when the abundance of host cells are low (Kimura et al. 2008), these host and virus populations were active in an environment where there was an abundance of CH_4 substrate.

Viruses can influence biogeochemical cycling by modulating host cell metabolism during viral infection (Breitbart et al. 2007). Investigation for the presence of AMGs revealed a large number of viral encoded genes involved in carbon degradation, such as glycoside hydrolase and transferases (GH family 19, 24, 25 and 46) and peptidases, which have also been previously observed in soil viruses derived from metagenomes (Emerson et al. 2018; Graham et al. 2019). Also, two viruses had AMGs associated to nitrogen cycling, including a *nifH* gene and nitrite reductase, and four viruses had AMGs encoding for proteins involved in methane oxidation, specifically *pmoC* and cytochrome C family genes. Although there was low support that the contig with the *pmoC* gene was a genuine viral contig, viral-associated *pmoC* genes from large viral genomes from freshwater lakes was recently identified (Chen et al. 2020). Transcriptomics data demonstrated the activity of these viral-associated *pmoC* genes, suggesting the roles of methanotroph-viruses in modulating CH_4 efflux through increasing methane oxidation by viral-associated *pmoC* genes during infection (Chen et al. 2020). Several AMGs encoded a cofactor necessary for methane oxidation, such as FAD- and NADH-binding domain superfamily, ferredoxin-type iron-sulfur binding domain, ferritin-like diiron domain and cupredoxin-like domains (Sirajuddin and Rosenzweig 2015). These AMGs may benefit the viruses by temporarily enhancing host functionality through diverse functions, including electron transfer and substrate binding (Lindell et al. 2004). A few of the homologs between the viruses and their methanotroph hosts had high identity (> 80%), including integrases, lysozymes and GH protein involved in C degradation. However, most of the homologs exhibited low percentage identity suggesting more

ancestral interactions between the host and viruses (Ku and Martin 2016), i.e. sequence identity between donor and recipients in horizontal gene transfer is initially 100%, but gradually decreases over time due to undergoing genetic changes, such as mutations, gene duplications and genomic rearrangements (Ochman et al. 2000; Raz and Tannenbaum 2010; Ku and Martin 2016). Thus, recent gene transfers tend to have a higher degree of identity to homologs from the donor lineage (Shoemaker et al. 2001; Smillie et al. 2011; Ku and Martin 2016). However, recently transferred AMGs may be under strong selection to maintain the original function of the gene product (Aswad and Katzourakis 2018).

The ONF analysis recovered various host-virus associations, and conservatively only the mVCs that had at least one homologous gene to its host or host genera were considered in the food web network. This included 64 methanotroph-, 6 methylotroph-, 4 nitrifier- and 3 bacterial predator-associated viruses. The network likely includes low-affinity CH_4 utilizers as the soil microcosms were incubated with high CH_4 concentrations. In the network, viruses were mostly associated to *Methylocystaceae* (43 mVCs), indicating a predominant host-virus linkage with low-affinity CH_4 utilizers. Nitrifiers, including AOA (*Nitrososphaera* and *Nitrosopumilus*), AOB (*Nitrosomonas*, *Nitrosococcus* and *Nitrosospira*) and NOB (*Nitrospira* and *Nitrobacter*), were found, suggesting that certain nitrifiers assimilated a significant proportion of ^{13}C-CO_2. Predatory bacteria, such as *Myxococcaceae, Bdellovibrionaceae* and *Lysobacter* were found (Morgan et al. 2010; Johnke et al. 2014; Seccareccia et al. 2015). These bacteria likely incorporated ^{13}C by degrading active or dead ^{13}C-labeled methanotrophs through cross-feeding (Murase and Frenzel 2008). Interestingly, bin.8 and bin.17 MAGs are representatives of the *Herminiimonas*, a genus which has not been previously implicated in metabolism of methane-derived carbon. Both MAGs possess a putative formate dehydrogenase, with bin.17 MAG also containing a predicted oxidoreductase of the glucose-methanol-choline family. This would indicate that these organisms were therefore secondary utilizers of methane-derived carbon. Characterized isolates are heterotrophs, typically associated with the transformation of metals in contaminated environments (Koh et al., 2017).

4.6. Conclusion

This study provided a detailed analysis of active soil methanotroph-virus dynamics. Soil pH influenced both methanotroph and viral communities. Carbon flow between active methanotrophic hosts and their lytic viruses was demonstrated, and specific host-virus interactions were revealed via the transfer of virus DNA into CRISPR arrays or host DNA into viral genomes. Active viral communities exhibited potentially important ecological functions, particularly related to carbon and nitrogen cycling, and methane metabolism. These findings indicate that soil viruses are important top-down regulators of microbial communities involved

in methane cycling through host lysis and augmenting host metabolic processes involved in soil ecosystem functions.

CHAPTER V

Linking viruses to autotrophic nitrifier hosts in acidic and neutral pH soils using DNA stable-isotope probing with $^{13}CO_2$

5.1. Abstract

Nitrification is a central step in the nitrogen cycle, where ammonia is oxidized to nitrite followed by subsequent oxidation to nitrate. This process is performed by nitrifying microorganisms, including ammonia-oxidizing archaea (AOA) and bacteria (AOB) and nitrite-oxidizing bacteria (NOB). While understanding the factors that control nitrifier populations is critical to our view of nitrogen cycling in soil, there is a lack of knowledge on top-down drivers, such as viruses. The aim of this study was to determine whether nitrifier-associated viruses could be detected and therefore potentially influencing nitrifier population dynamics in soil. Soils at different ends of an established pH gradient (pH 4.5 and 7.5) were incubated in microcosms applied with either ^{12}C- or ^{13}C-CO_2 to target autotrophic nitrifying microbes and their associated lytic viruses. Extracted genomic DNA was subjected to isopycnic ultracentrifugation, with high buoyant density DNA recovered and sequenced using the Illumina NovaSeq platform (1.3-2.4 Gbp per replicate). Viruses that infect AOA (11 mVCs) and NOB (3 mVCs) were identified from the ^{13}C-enriched metagenomes, demonstrating recent carbon flow between active autotrophic hosts and their associated lytic viruses. Virus communities represented in the metagenomes were different between the two soils as previously observed for nitrifier host communities, indicating that there are different nitrifier-virus dynamics between distinct soil pH ranges. Viral-encoded Auxiliary metabolic genes for iron-sulfur clusters suggest a potential role of AOA-associated viruses to augment electron transfer, which may help in the adaptation of hosts to acidic conditions.

5.2. Introduction

The nitrogen (N) cycle is critical for supplying nitrogen to all living organisms with the activities of microorganisms central to all transformative steps in terrestrial ecosystems. One central stage of the N cycle is nitrification, where ammonia, the most reduced form of N, is oxidized to nitrate, the most oxidized form, via nitrite. In soil, aerobic autotrophic nitrification is performed by two physiologically distinct groups of nitrifying prokaryotes, ammonia oxidizers, which oxidize ammonia to nitrite, and nitrite oxidizers, which oxidize nitrite to nitrate. Additionally, the complete oxidation of ammonia to nitrate (comammox) in a single cell has been recently discovered to be performed by *Nitrospira* bacteria (Daims et al. 2015). In addition to being an essential process, nitrification also has important environmental and economic consequences through the transformation of N-based fertilizers, resulting in N loss and pollution through nitrate leaching and nitrous oxide (N_2O) emissions (Canfield et al. 2010). It is therefore imperative to understand the ecology and regulating factors of nitrifiers in soil.

Ammonia oxidizers (AO) consist of chemolithoautotrophic ammonia oxidizing archaea (AOA) of the phylum *Thaumarchaeota* and ammonia oxidizing bacteria (AOB) belonging to Beta- and Gammaproteobacteria (Purkhold et al. 2000; Tourna et al. 2011). Both AOA and AOB harbor

the enzyme ammonia mono-oxygenase (AMO) for initiating nitrification through the oxidation of ammonia to hydroxylamine (Hooper et al. 1997; Kits et al. 2017). Until recently it was widely thought that hydroxylamine is subsequently oxidized to NO_2^- by hydroxylamine dehydrogenase (HAO) in AOB, but it has now been demonstrated that hydroxylamine is oxidized to nitric oxide (NO) before subsequent oxidation to NO_2^{--}. In AOA the process of hydroxylamine oxidation is less well characterized with genomes lacking a known gene encoding HAO. However, as in AOB, NO is a central intermediate (Prosser et al. 2000). The difference in cellular biochemistry and physiology between AOA and AOB contributes to the abundance and community structure of ammonia-oxidizers in soil (He et al. 2012). Abiotic factors, such as oxygen availability, substrate concentration, and soil pH have been found to affect AO communities (Park et al. 2006; Nicol et al. 2008; de Gannes et al. 2014). The affinities of oxygen and ammonia differ between AOA and AOB (Kits et al. 2017) and AOA dominate growth in soils with low ammonium concentrations (Di et al. 2009). The abundance of AOA is relatively greater than that of AOB in acidic soils, and can contribute more than AOB to ammonia oxidation (Leininger et al. 2006; Nicol et al. 2008).

Nitrite-oxidizing bacteria (NOB), including *Nitrobacter*, *Nitrococcus*, *Nitrospira* and *Nitrospina* species, possess the enzyme nitrite oxidoreductase (NXR) (Hagopian and Riley 1998). Compared to AO, NOB have been less studied, as ammonia oxidation is usually considered the rate-limiting step in nitrification. The NXR active site in *Nitrobacter* is cytoplasmic-orientated whereas for *Nitrospira* the site faces the periplasm, and thus may contribute to the adaptation and niche differentiation among NOB species (Lebedeva et al. 2005; Sorokin et al. 2012). It was generally thought that *Nitrobacter* play the main role in nitrite oxidation in soil, however they are also the most studied NOB, and the importance of *Nitrospira*, which also perform comammox, have been studied to a lesser extent (Freitag et al. 2005; Daims et al. 2015; Koch et al. 2019). Recent studies suggest that comammox are widespread and active in soil (Li et al. 2019; Xu et al. 2020; Wang et al. 2020).

Relative to our understanding of abiotic controls on nitrifier populations, little is known about the impact of top-down drivers. Viral infection is well known to have important implications for the effects on microbial structure and function in ecosystems (Weinbauer and Rassoulzadegan 2004; Suttle 2005; Breitbart et al. 2007; Bertilsson et al. 2013). Potentially, nitrifier-associated lytic viruses could influence nitrifier population structure and rates of nitrification in soil via the control of host numbers. There is evidence of proviruses in AOA (Abby et al. 2017; Krupovic et al. 2011) and AOB (Chain et al. 2003). Studies in marine environments have reported several thaumarchaeal viruses and the presence of AMGs encoding for the ammonia monooxygenase subunit C (*amoC*), suggesting the potential role of thaumarchaeal viruses in influencing nitrification (López-Pérez et al. 2019; Ahlgren et al. 2019). Spindle-shaped viruses have also been recently isolated from seawater samples using a host AOA strain (Kim et al. 2019). These did not

display sequence similarity to other known archaeal or bacterial viruses although some genes (e.g. encoding DNA polymerase family B) were shared with other archaeal viruses. The production and release of these viruses was not accompanied by the lysis of host cell but used a mechanism similar to that of enveloped viruses infecting some hyperthermophilic archaea. Equivalent knowledge on nitrifier-associated viruses in soil is currently lacking.

A primary reason for a general lack in the understanding of prokaryote - virus interactions in soil is due to the difficulty in adequately characterizing viral communities in soils that are extremely prokaryotic rich (Trubl et al. 2020). However, reducing prokaryotic diversity within soil DNA samples by selectively targeting a key functional group may increase the ability to identify viruses and their associated hosts. DNA stable-isotope probing (DNA-SIP) is a powerful method that targets the active microorganisms within an environmental sample via the cellular assimilation of an added substrate enriched in a heavy isotope (Radajewski et al. 2003). As viruses are obligatory parasites and reproduce inside their host cells using host replication machinery, the associated soil viruses of the active microorganisms will also be isotopically labeled (Lee et al. 2012). Extracted soil DNA is separated by buoyant density by CsCl density gradient centrifugation (Radajewski et al. 2003), and the heavy DNA sequenced from soils incubated with, for example, isotopically enriched carbon dioxide ($^{13}CO_2$) will potentially include autotrophic microbes and their associated viruses.

As soil pH is a critical abiotic factor that influences nitrifier community structure, a long-term agricultural soil pH gradient was utilized to investigate nitrifiers and their associated viral communities using two soils that contain distinct nitrifier populations. To select for autotrophic prokaryotes and their associated viral communities, soils were incubated in microcosms with ^{12}C- or ^{13}C-CO_2, and with additions of urea to encourage the growth and activity of AO and NOB (Zhao et al. 2020) with high buoyant density DNA recovered after isopycnic ultracentrifugation and metagenomic sequencing. The aims of the study were to 1) grow isotopically enriched nitrifier communities from soils at the extreme ends of a pH gradient (pH 4.5 and 7.5), 2) identify isotopically-enriched viruses containing CO_2-derived C, and 3) determine active nitrifier–virus interactions.

5.3. Materials and Methods

5.3.1. Soil sampling and physicochemical analyses

Soil was collected from the Craibstone Research Station, SRUC, Aberdeen, Scotland (57°11, 2°12) as previously described (see Chapter IV, section 4.3.1.).

5.3.2. Soil microcosm incubations

Soil microcosms were established as previously described (see Chapter IV, section 4.3.2.) except that a 5% (v/v) ^{12}C- or ^{13}C-CO_2 headspace was established. In total, 30 microcosms were

established for destructive sampling at three time points (day 0, 15, and 30) in triplicate. Microcosms were amended with 100 µg Urea-N g^{-1} soil (dry weight equivalent) at day 0 and 15, and the control received water. CO_2 was replenished every three days following aeration to maintain aerobic conditions in the soil. Microcosms were incubated with a water content of 30 – 32% (w/w) and at 28°C in the dark. Microcosms were destructively sampled on day 0, 15 and 30, and soil stored at –20°C.

5.3.3. Nitrification assay

Ammonium (NH_4^+), nitrite (NO_2^-) and nitrate (NO_3^-) concentrations of each soil sample were determined immediately after destructively sampling microcosms. For NH_4^+ and NO_3^- measurements, 1 M potassium chloride (KCl) solution was used for extraction, with water extraction was used for NO_2^-. In a 15 ml centrifuge tube, 1.5 g of soil (wet weight) and 10 ml 1 M KCl or deionized water was added and shaken using a SB rotator for 30 min, and then centrifuged at 6000 x *g* for 10 min to pellet the soil. A 2 ml subsample of the supernatant was then stored at –20°C before colorimetric determination of N concentrations. For measuring NH_4^+, 50 µl of a 1:10 diluted KCl extract was mixed with 50 µl of color reagent in a 96-well plate. The color regent was freshly-made by mixing 1:1:1 (v/v/v) 0.3 M NaOH solution, sodium salicylate solution (0.17 g ml^{-1} sodium salicylate and 1.28 mg ml^{-1} sodium nitroprusside) and deionized water. The 96-well plate was shaken at 95 rpm for 30 min at room temperature. Subsequently, the absorbance was measured at 660 nm using a Multiscan GO Microplate Spectrophotometer (Thermo Scientific, Waltham, MA, USA). A standard curved was produced using a dilution series of ammonium chloride prepared in deionized water over the range of 0 to 500 µM. For determination of NO_2^- and NO_3^- concentrations, in a 96-well plate 100 µl of KCl extract or water extract was mixed with 20 µl of diazotizing solution (5 mg sulfanilamide in 2.4 M HCl), and 20 µl of coupling reagent (3 mg N-(1-naphthyl)-ethylenediamine in 0.12 M HCl). A standard curve was produced using a dilution series of sodium nitrite and sodium nitrate prepared in KCl solution over the range of 0 to 100 µM. The absorbance of NO_2^- concentration was measured at 540 nm. Prior to NO_3^- measurement, 20 µl of vanadium chloride solution (70 mg ml^{-1} vanadium chloride in 1 M HCl) was added and incubated overnight at 37°C prior to absorbance measurement.

5.3.4. DNA extraction and density gradient centrifugation

Total DNA extraction, CsCl ultracentrifugation and DNA recovery were as described in Chapter IV, section 4.3.3.

5.3.5. Quantitative PCR and metagenomic sequencing

qPCR was performed as described previously (Chapter IV, section 4.3.4.) except primers amoA1F/amoA2R (Rotthauwe et al. 1997) and crenamoA23F/crenamoA616R (Tourna et al. 2008), were used for AOB and AOA amoA genes, respectively. The thermocycling conditions were as used for bacterial 16S rRNA genes.

5.3.6. Bioinformatic analyses

An overview of the bioinformatics workflow is shown in Figure 5.1. For each metagenome, quality-controlled reads were assembled into contigs and all contigs were concatenated as co-assembled contigs. Dereplicated co-assembled contigs (> 5 kb) were used for virus prediction and binned into metagenomic assembled genomes (MAGs). To focus on the ^{13}C-enriched nitrifier populations, microbial populations were selected by separating the assembled contigs based on %GC content. Populations with %GC content less than 50 and between 50 - 63 were selected from each metagenome, respectively. To link viruses to hosts and explore host-virus interactions, CRISPR array and oligonucleotide frequency (ONF) analysis and gene homology analyses were performed.

5.3.6.1. Sequence quality filtering, contig assembly and co-assembly

Quality filtering and assembly steps were performed with the JGI pipeline as previously described in Chapter II, section 2.3.4.1. For quality trimming, BBDuk v38.51 was used to remove known Illumina artifacts, and bad quality sequences (quality score = 20, minimum read length = 51 bp). To remove any contaminating DNA sequences, reads were subjected to mapping with BBMap v38.34.1. The quality-controlled reads of each metagenome were *de novo* assembled into contigs using MetaSPAdes v3.13.0.2 (Nurk et al. 2016). The input read set was mapped to the final assembly and coverage information was generated with BBMap v38.34.1.

Prior to binning and virus prediction with VirSorter, assembled contigs from each metagenome were concatenated as one co-assembled sample, and sequence redundancy was reduced using psi-cd-hit-DNA tool with 95% identity(Fu et al. 2012). The contigs larger than 5 kb were selected and the contig name was simplified using anvi-script-reformat-fasta from anvio 5 (Eren et al. 2015).

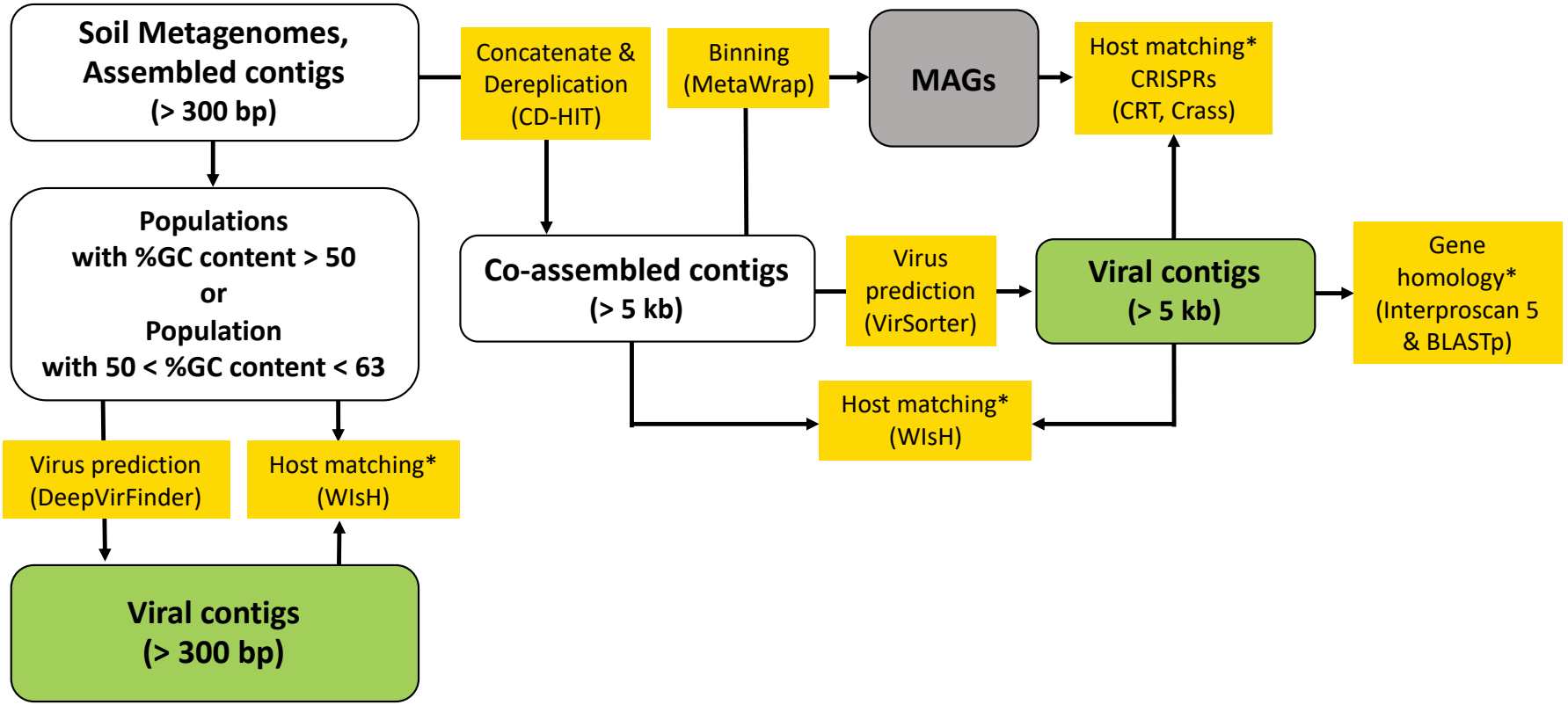


Figure 5.1. Schematic overview of the bioinformatics workflow. Square boxes in white, gray and green represent the assembled contigs and metagenome assembled genomes (MAGs) and predicted viral contigs, respectively. Square boxes in yellow represent the bioinformatics performed and key tools used are in parentheses. *Gene homology analyses between host MAGs or contigs and host-associated metagenomic viral contigs (mVCs) were realized (AMGs, Auxiliary metabolic genes).

5.3.6.2. Metagenomic assembled genomes

Metagenomic assembled genomes (MAGs) were generated using the MetaWRAP tool (Uritskiy et al. 2018). The co-assembled contigs were binned with the Metawrap-Binning module, follow by de-replication and calculation of bin completion and contamination level by CheckM v.1.0.7 (Parks et al. 2015). Taxonomic classification of the MAGs was carried out using GTDB-Tk v0.3.2 (Chaumeil et al. 2020). Functional analysis of MAGs was processed by annotating the protein sequences using InterProScan 5 (E value < 10^{-5})(Jones et al. 2014). To determine the distribution and abundance of each MAG across the samples, the relative abundance of each bin was calculated. Briefly, the Salmon v0.9.1 in Quant_bins module was used to index the contigs from the MAGs and align reads from each sample back to the contigs (Patro et al. 2017; Uritskiy et al. 2018). As previously described in Chapter IV, section 4.3.5.2., relative abundance of each contig in each MAG was calculated and the abundance was expressed as normalized genome copies per million reads (CPM). A heatmap was made using the heatmaply R package to visualize the variation in relative abundance of the MAGs across the metagenomes (R core team, 2019).

5.3.6.3. Analysis of %GC coverage and selection of ^{13}C-enriched populations

Taxon annotated GC – coverage plots of the co-assembled contigs for each of the 12 metagenomes were generated. The co-assembled contigs were annotated with Kaiju using the NCBI RefSeq bacterial, archaeal and viral database (Menzel et al. 2016). The co-assembled contigs were indexed and the reads from the metagenomes were re-mapped with bowtie2 mapper v2.3.0 (Langmead and Salzberg 2012). The GC_cov_annotate.pl function from MetaWRAP-Bloblogy module was used to generate a blobplot file with the GC content, coverage and taxonomy of each contig (Uritskiy et al. 2018). Blobplots of the co-assembled contigs across the 12 metagenomes were made using the makeblobplot.R function within R (R core team, 2019). Nitrifying communities were selected from the Blobplots and normalized, and visualized using R (R core team, 2019).

To focus on the ^{13}C-enriched prokaryotic and viral populations, contigs were separated based on two ranges of %GC content. Populations that had a %GC content less than 50 and those with a %GC content between 50 and 63 were selected from the assembled contigs. The two ranges in %GC content were chosen based on the presence of unique contigs that were found in the ^{13}C metagenomes and not in the ^{12}C metagenomes (see Figure 5.6). To compare community structure between the ^{12}C- and ^{13}C communities with that of the %GC < 50 and 50 < %GC < 63 communities, non-metric multidimensional scaling (NMDS) was carried out on Bray-Curtis dissimilarity matrices, based on the relative abundances, using the vegan R package (R core team, 2015). The taxonomy of the contigs with %GC < 50 and 50 < %GC < 63 were annotated using Kaiju with the

NCBI RefSeq bacterial, archaeal and viral database, and the normalized raw abundances were calculated (Menzel et al. 2016).

5.3.6.4. Virus prediction

From the assembled contigs that were greater than 300 bp, and those of the specific %GC content populations, DeepVirFinder (Ren et al. 2018) was used to predict metagenomic viral contigs (mVCs). Also, the co-assembled contigs that were greater than 5 kb were subjected to viral prediction using VirSorter (Roux et al. 2015). The taxonomy of the VirSorter predicted mVCs was assigned using Kaiju with the RefSeq viral proteins database from NCBI (Madden et al. 1996; Menzel et al. 2016), and the taxonomic composition of mVCs was visualized with the Krona tool (Ondov et al. 2011).

The distribution of the ^{12}C and ^{13}C derived mVCs (predicted with VirSorter) were compared. The Quant_bins module in Salmon v0.9.1 (Patro et al. 2017) was used to index the contigs and align reads from each metagenome back to the contigs. The relative abundance of each mVC in each metagenome was calculated based on the length of contig size and on the coverage of contigs. The abundance was expressed as normalized genome copies per million reads (CPM). This was calculated as previously described in Chapter IV, section 4.3.5.2. A heatmap was made using the heatmaply R package to visualize the variation in the relative abundance of mVCs across the metagenomes (R core team, 2019).

5.3.6.5. Linking viruses to hosts using CRISPR array and ONF analysis

The CRISPR Recognition Tool (CRT) was used to identify CRISPR arrays from each MAG and from the assembled contigs from each metagenome (Bland et al. 2007). CRISPR arrays were assembled using the Crass tool with unassembled reads longer than 60 bp (Skennerton et al. 2013). The direct repeats (DR) and spacer sequences were extracted using linux commands and located using the Seqkit locate function with exact match and positive and negative strand search (Shen et al. 2016). The spacer sequences were located to predict associations with ^{13}C-CO_2-enriched mVCs. CRISPR sequences were annotated using Kaiju with the NCBI-nr database (Menzel et al. 2016).

For oligonucleotide frequency analysis (ONF), the WIsH tool (Galiez et al. 2017) was used to predict host contig matches to the mVCs that were greater than 5 kb and to mVCs of the two specific %GC content ranges (i.e. %GC < 50 and 50 < %GC < 63). Predicted host-virus linkages with a p-value > 0.05 were considered potential matches (Galiez et al. 2017). The taxonomy of the potential host contigs were annotated with Kaiju using the NCBI RefSeq bacterial, archaeal and viral database (Menzel et al. 2016). Also, the nitrifier host-linked mVCs were compared to reference viral genomes stored in the Virus-Host DB (Mihara et al. 2016) by calculating all-

against-all similarity scores (S_G) computed by tBLASTx (Nishimura et al. 2017). Resulting phylogenetic trees were visualized using ViPtree (Nishimura et al. 2017).

5.3.6.6. Analysis of gene homology

Gene prediction of the mVCs was completed using Prodigal (Hyatt et al., 2010). To identify auxiliary metabolic genes (AMGs), homology searches were conducted using InterProScan 5 (E value < 10^{-5}), and using Diamond Blastp with the NCBI-nr database (E value < 10^{-5})(Jones et al. 2014; Buchfink et al. 2015).

Gene homology between host-linked mVCs and their associated hosts was investigated. Gene prediction of the mVCs and their associated host MAGs or contigs was performed using Prodigal (Hyatt et al. 2010). Protein alignment between viral and host origin proteins was conducted with BLASTp (identity > 30%, E value < 10^{-5} and query cover > 70%) (Madden et al. 1996). Shared protein sequences were annotated using InterProScan 5 (E value < 10^{-5})(Jones et al. 2014). Additionally, gene homology was assessed between host-linked mVC and prokaryotes using Diamond BLASTp with the NCBI-nr database (E value < 10^{-5}) (Buchfink et al. 2015). From the BLASTp output file of each host-linked mVC, the number of viral genes homologous to the genera of their associated host was counted using Linux commands.

The nitrifier-associated mVCs were compared between each other by calculating all-against-all similarity score (S_G) computed by results of tBLASTx, and were visualized with ViPtree (Nishimura et al. 2017).

5.4. Results

5.4.1. Nitrification in soil microcosms

Urea is rapidly hydrolyzed to NH_4^+ in soil and after 15 days incubation all added urea had been stoichiometrically oxidized through to NO_3^- in both soils (Figure 5.2). After 30 days, all urea added at day 15 was also oxidized in pH 7.5 microcosms, but only approximately half was oxidized in the pH 4.5 microcosms indicating a decrease in activity rate, potentially due to a decrease in pH inhibiting NH_4^+ oxidation. Nitrite (NO_2^-) concentration was always below the level of detection in all of the soil samples. Although there was a significant difference in NH_4^+ concentrations between the two soils at day 30, this was not reflected by statistically significant differences in measured NO_3^- concentrations due to larger variation.

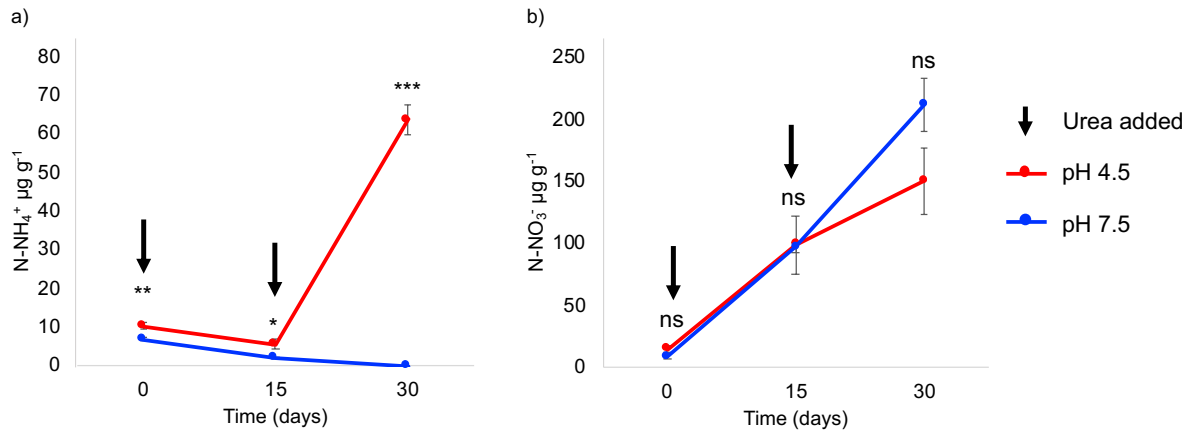


Figure 5.2. Concentration of a) NH_4^+ and b) NO_3^- in the pH 4.5 and 7.5 soil microcosms during 30 days of incubation. Error bars are the standard error of the mean of three replicates. Significance was tested between soil pH using Student's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.001$ (***).

5.4.2. Distribution of prokaryotic communities in DNA-SIP fractions

After ultracentrifugation of DNA extracted after 30 days incubation there was no measurable difference in the distribution of bacterial 16S rRNA genes in the microcosms incubated with either 5% ^{12}C-CO_2 and ^{13}C-CO_2 for both the pH 4.5 and 7.5 soils, with the maximum abundance of genomic DNA being present at a buoyant density of approximately 1.715 g ml^{-1} (Figure 5.3). This indicates that enriched autotrophs represent a very small proportion of the total prokaryotic community. Differences in the distribution of archaeal and bacterial *amoA* genes was observed between ^{12}C- and ^{13}C-CO_2 incubations. For both soils, the maximum abundance of AOA *amoA* genes occurred at a buoyant density of approximately 1.69 g ml^{-1}, reflecting the relatively low %GC of most soil AOA genomes. In the pH 4.5 soil, while the majority of DNA from the ^{13}C-CO_2 had the same buoyant density as in the ^{12}C-CO_2 microcosms, there was a clear enrichment of ^{13}C-labeled DNA peaking at approximately 1.73 g ml^{-1}. In the pH 7.5 soil, the majority of AOA genomes had a higher buoyant density than that from the ^{12}C-CO_2 microcosms, with two peaks at approximately 1.71 and 1.73 g ml^{-1}, indicating potential enrichment of two groups of AOA with different %GC contents or different levels of ^{13}C-enrichment. Due to the low %GC content of fully ^{13}C-enriched AOA genomes, there was overlap in the distribution with unlabeled bacterial genomic DNA as profiled by the 16S rRNA gene analysis. For the AOB, the distribution of unlabeled genomic DNA was at a higher buoyant density compared to AOA, and with a clear increase in the buoyant density of the majority of AOB genomic DNA from the ^{13}C-CO_2 microcosms. To obtain enriched DNA from both AOA and AOB, fractions > 1.72 g ml^{-1} were pooled in each replicate for sequencing.

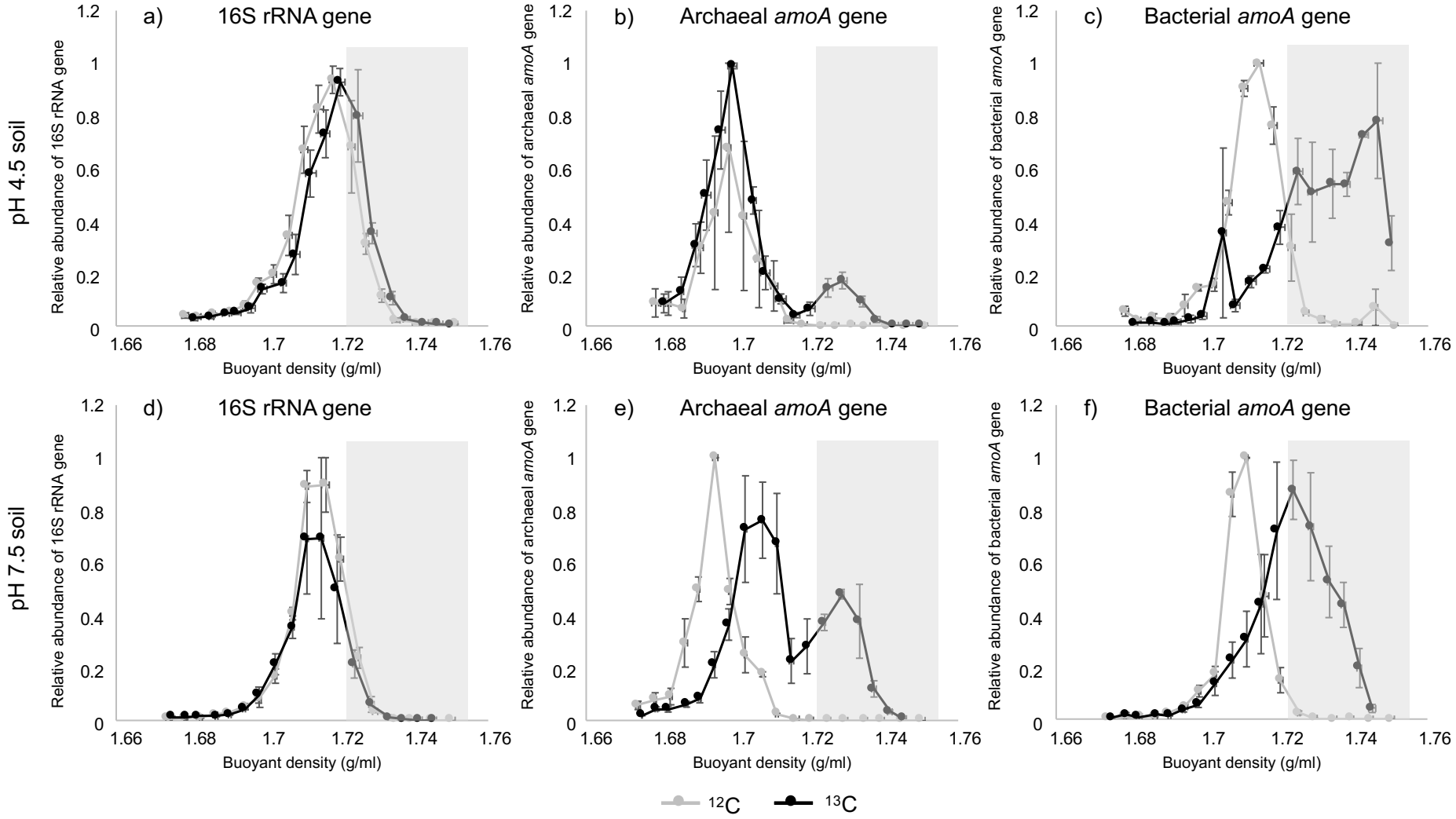


Figure 5.3. Distribution of genomic DNA in CsCl gradients from ^{12}C- and ^{13}C-CO_2 microcosm incubations of pH 4.5 and 7.5 soil. The relative abundance of bacterial 16S rRNA (a and d), archaeal *amoA* (b and e) and bacterial amoA (c and f) genes from fractionated DNA derived from pH 4.5 (a, b and c) and 7.5 (d, e and f) was determined by qPCR. Data are normalized by the ratio of gene abundance in each fraction to the maximum quantity obtained in any one fraction per replicate. Error bars represent the standard error of the mean of three replicates. The ^{12}C- and ^{13}C-enriched fractions highlighted in the shaded area were pooled and sequenced.

5.4.3. Summary of metagenome sequencing

The 12 metagenomes generated between 47 and 84 GB of sequence data, totaling 684 GB. In total, 1.9 billion quality-controlled sequence reads were retained, ranging between 125 – 245 million reads per metagenome (Table 5.1). Sequence assembly yielded a total of 35 million contigs, ranging between 2 – 4 million contigs per metagenome with an average length of 567 bp (Table 5.1). The selection of contigs greater than 5 kb and the reduction of sequence redundancy resulted in 61,447 high-quality co-assembled contigs with an average length of 10,568 bp.

Table 5.1. Sequence summary for the ^{12}C- and ^{13}C-pH 4.5 and 7.5 soil metagenomes.

Sample ID	Pre-QC number of reads	Post-QC number of reads	Contigs count (> 200 bp)	Average contig length (bp)	Maximum length (bp)
12C-pH 4.5-1	140,070,754	137,927,826	2,702,971	579	489,948
12C-pH 4.5-2	168,118,921	165,499,900	3,158,661	601.4	251,541
12C-pH 4.5-3	164,428,434	161,786,148	2,893,235	621.6	490,918
13C-pH 4.5-1	159,166,954	157,372,294	2,665,960	647	632,029
13C-pH 4.5-2	142,462,948	140,157,446	2,621,517	592.5	632,066
13C-pH 4.5-3	155,980,938	153,263,184	2,633,463	639	632,128
12C-pH 7.5-1	175,160,938	172,033,988	3,144,204	497.3	179,515
12C-pH 7.5-2	186,701,904	184,944,656	3,302,811	527.3	402,384
12C-pH 7.5-3	175,052,494	173,060,618	3,047,820	506.3	356,185
13C-pH 7.5-1	136,844,724	135,351,318	2,029,573	522.9	681,469
13C-pH 7.5-2	247,952,572	245,188,756	4,475,134	548.6	809,791
13C-pH 7.5-3	158,055,930	156,686,154	2,434,331	532	812,893

QC, quality control

5.4.4. Metagenomic assembled genomes

A total of 13 metagenomic assembled genomes (MAGs) surpassed a completeness threshold of 50% with less than 10% contamination (Table 5.2). The average percent completeness of MAGs was 61.7%, ranging between 87.4 - 51%, and the average estimated level of contamination was 3.3%, ranging between 0 - 8.9%. Nine MAGs belonged to the *Actinobacteria*, and three MAGs each belonged to the *Acidobacteria*, *Proteobacteria* and *Candidatus* Eremiobacerota (Table 5.2). Only one MAG, bin.12_*Thaumarcheota*, was annotated as an nitrifier (*Nitrososphaera*) (Table 5.2). Taxon annotated %GC coverage plots showed distinct differences between the pH soils, and only bin.12_*Thaumarcheota* was sufficiently ^{13}C-enriched in two of the three pH 7.5 soil metagenomes (contigs > 1X coverage) (Figure 5.4). The relative abundance of the MAGs were significantly different (p-value < 0.05) between soil pH (data not shown). In the pH 7.5 soil, the relative abundance of bin.12_*Thaumarcheota* was greater in the ^{13}C than the ^{12}C-samples (Figure 5.5b).

Table 5.2. Summary statistics and taxonomic classification of the metagenomic assembled genomes (MAGs).

MAG ID	Completeness (%)	Contamination (%)	GC (%)	N50	Size (bp)	Phylum	Class	Order
bin.1	75.43	1.724	65.5	184,710	1,770,496	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.2	59.31	0.862	68.4	9,904	1,164,422	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.3	56.83	1.293	70.7	13,67	1,438,546	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.4	55.74	0.443	73.3	8,775	3,522,350	*Actinobacteriota*	*Actinobacteria*	*Mycobacteriales*
bin.5	56.08	8.26	61.8	33,788	2,273,982	*Proteobacteria*	*Alphaproteobacteria*	
bin.6	53.44	1.724	65.6	11,217	1,104,670	*Ca.* Eremiobacterota	*Ca.* Eremiobacteria	*Ca.* Eremiobacterales
bin.7	51.01	7.758	68.2	7,883	1,523,824	*Actinobacteriota*	*Thermoleophilia*	
bin.8	54.31	0	68.2	14,651	1,485,041	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.9	87.4	8.974	66	20,616	2,914,588	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.10	68.96	1.724	68.2	336,085	1,827,759	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*
bin.11	51.72	4.31	69.5	8,608	4,230,577	*Acidobacteriota*	*Vicinamibacteria*	
bin.12	66.5	2.427	50.2	9,121	1,193,146	*Thaumarcheota*	*Nitrososphaeria*	*Nitrososphaerales*
bin.13	65.51	3.448	66	240,363	2,164,288	*Actinobacteriota*	*Thermoleophilia*	*Solirubrobacterales*

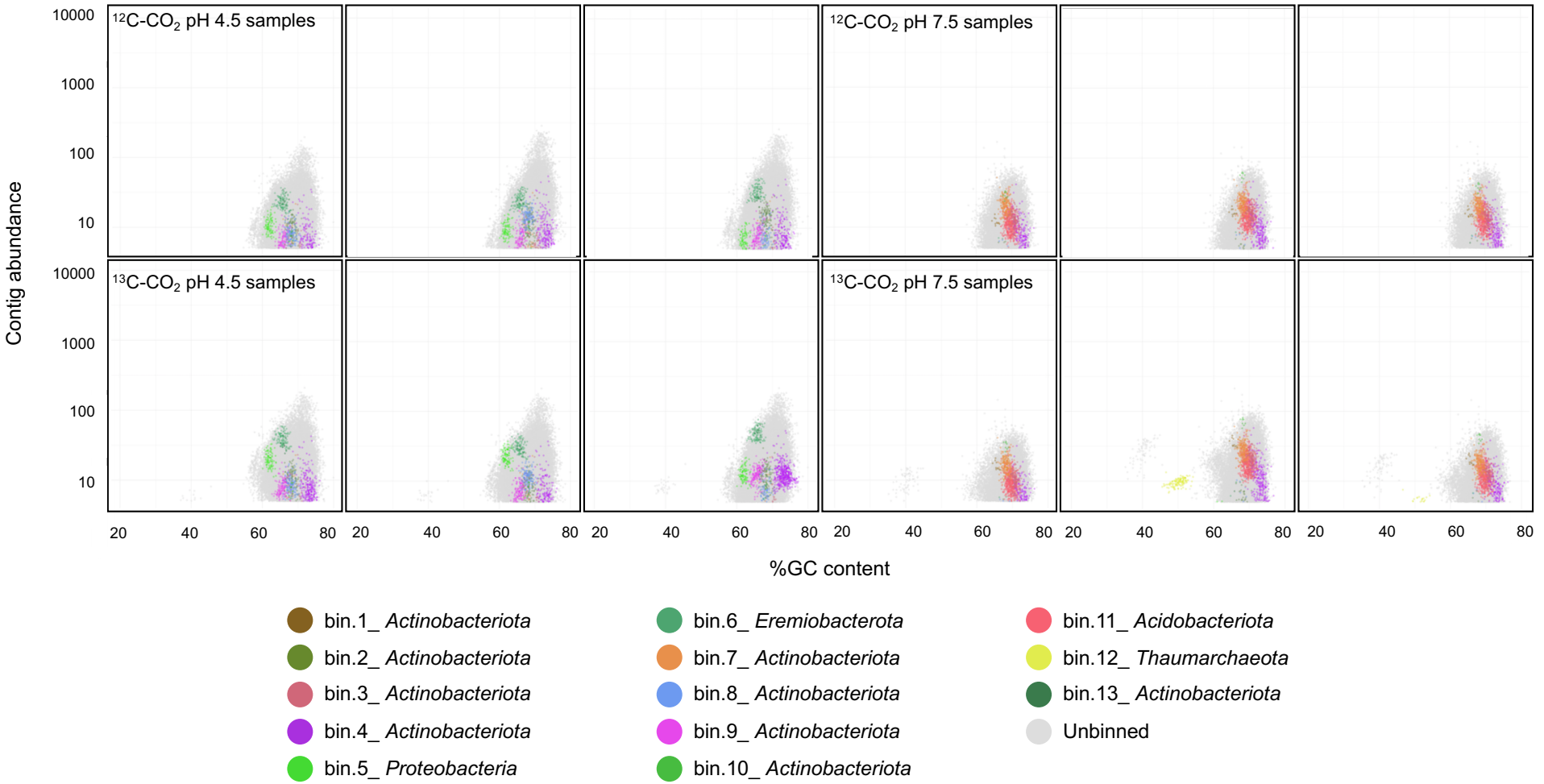


Figure 5.4. GC – coverage plots of the metagenomic assembled genomes (MAGs) for the ^{12}C- and ^{13}C-pH 4.5 and 7.5 soil. Contig abundance was calculated from standardized read coverage in each sample.

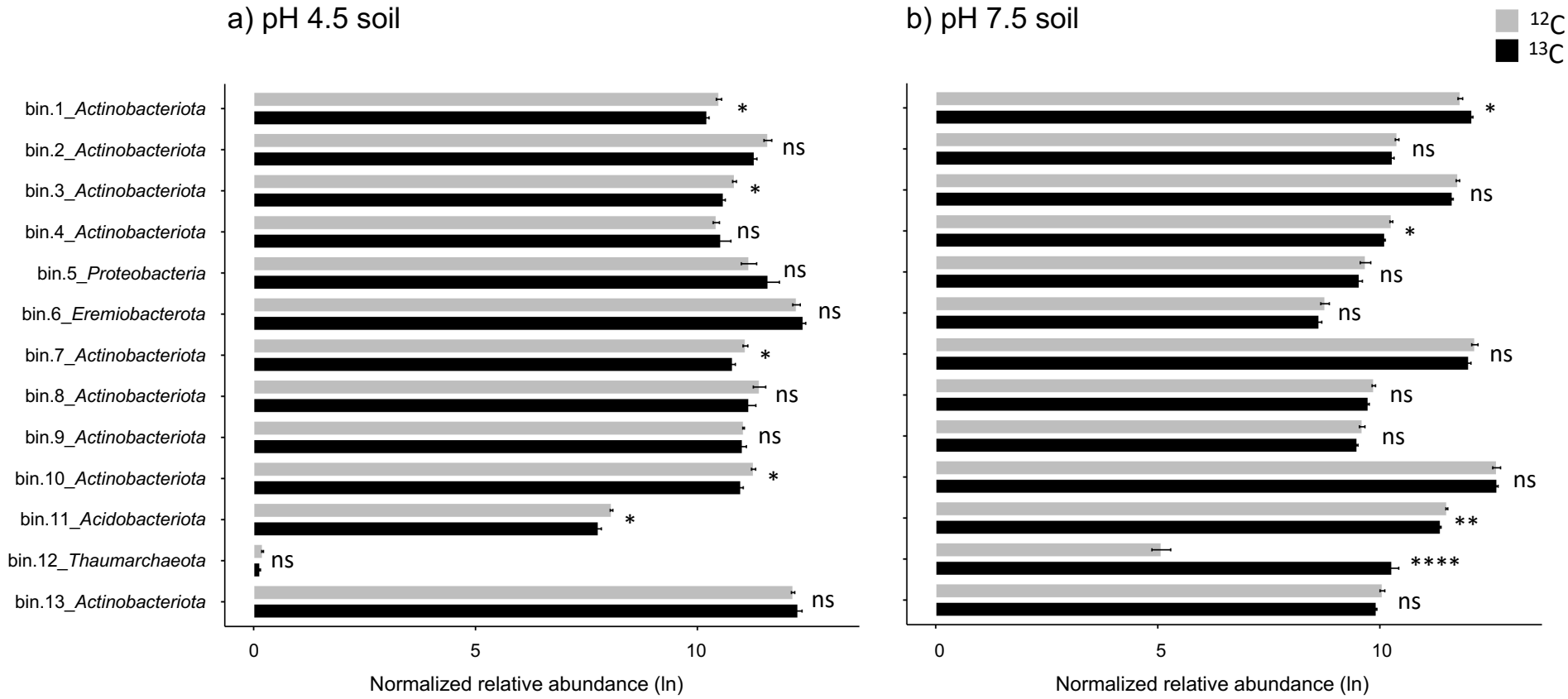


Figure 5.5. Normalized relative abundance (ln) of the metagenomic assembled genomes (MAGs) of the ^{12}C- (grey) and ^{13}C-samples (black) of the a) pH 4.5 and b) 7.5 soil. The error bars are based on three replicates (± SE). Significance was tested between ^{12}C and ^{13}C-samples using StudentStudent's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.0001$ (****).

5.4.5. ^{13}C-enriched populations

The taxon annotated %GC coverage plots showed that the metagenomes were mostly composed of heterotrophic bacteria (Figure 5.6). Only a few contigs (822 contigs, 1%) were specific to ^{13}C-samples. In both soils and in both ^{12}C- and ^{13}C incubations the distribution of genomic DNA was between 55 and 75%, with a small number of low %GC contigs in all ^{13}C incubations only. The relative abundances at the family level were generally similar between the ^{12}C- and ^{13}C-samples for both pH soils (Figure 5.7). However, there was evidence of ^{13}C-enriched nitrifier communities in both pH soils, and despite their overall low abundance (< 2% of contigs) (Figure 5.8). NOB *Nitrobacter* and *Nitrospira*, AOB *Nitrosococcus, Nitrosomonas* and *Nitrosospira*, and AOA *Nitrosopumilus*, *Nitrososphaera* and *Nitrosopelagicus* were identified (Figure 5.8). In the pH 4.5 soil, the relative abundance of *Nitrospira* was slightly greater (1.07-fold) in the ^{13}C- than ^{12}C-samples (Figure 5.8a). In the pH 7.5 soil, the relative abundance of *Nitrobacter* was greater (1.9-fold) in the ^{13}C- than ^{12}C-samples (Figure 5.8b). The relative abundances of Candidatus *Nitrosopumilus* and *Nitrosopelagicus* were greater in the ^{13}C- than ^{12}C-samples for both pH soils (311- and 215-fold, and 1.7- and 11.1-fold greater for pH 4.5 and 7.5 soil, respectively) (Figure 5.8). Conversely, the relative abundances of AOB genera were not significantly different between the ^{12}C- and ^{13}C-samples for both pH soils (Figure 5.8).

As there was relatively small differences in the relative distribution of populations between ^{12}C- and ^{13}C incubations, likely due to the abundance of contigs from unlabeled heterotrophic populations, metagenomic contigs were separated into two specific %GC groups to potentially increase the relative proportion of ^{13}C populations in the analysis (e.g. AOA). Specifically, contigs were separated into those with a %GC < 50 and 50 < %GC < 63. For the %GC < 50 population, there was a 3- and 14-fold greater number of sequences in the ^{13}C- than ^{12}C-samples of the pH 4.5 and 7.5 soil, respectively (Table 5.3), and contig length was greater in the ^{13}C- than ^{12}C-samples for both pH soils (5- and 16-fold greater in pH 4.5 and 7.5 soil, respectively). The number of sequences and length of contigs of the 50 < %GC < 63 population were similar between the ^{13}C and ^{12}C-samples (Table 5.4). The %GC < 50 population consisted of 161 contigs with an average size of 7,891 bp, ranging between 5 - 30 kb. The 50 < %GC < 63 population consisted of 4,346 contigs with an average contig size of 9,094 bp, ranging between 5 - 95 kb. Comparing NMDS plots between the ^{12}C- and ^{13}C-communities of the total metagenomes versus the two specific %GC content ranges, suggested increased resolution of enriched populations for the %GC < 50 only (Figure 5.9). In the %GC < 50 metagenomes, based on relative abundance, AOA, including *Nitrososphaera, Nitrosopumilus, Ca.* Nitrosocosmicus, *Ca.* Nitrosotenuis and *Ca.* Nitrosotalea, were enriched in both pH 4.5 and 7.5 soil (Figure 5.10). Additionally, *Nitrosospira* (AOB) was enriched in the pH 4.5 soil (Figure 5.10a), and *Ca. Nitrosoarchaeum* (AOA), and NOB, including *Nitrobacter* and *Nitrospira*, were enriched in the pH 7.5 soil (Figure 5.10b). In the 50

< %GC < 63 metagenomes, based on relative abundance, *Nitrospira* in both pH 4.5 and 7.5 soil was enriched, and *Ca.* Nitrosocosmicus in the pH 4.5 soil, and *Nitrobacter*, *Nitrosopumilus*, *Ca*. Nitrosotenuis, *Nitrosovibrio* and *Nitrosococcus* in the pH 7.5 soil (Figure 5.11).

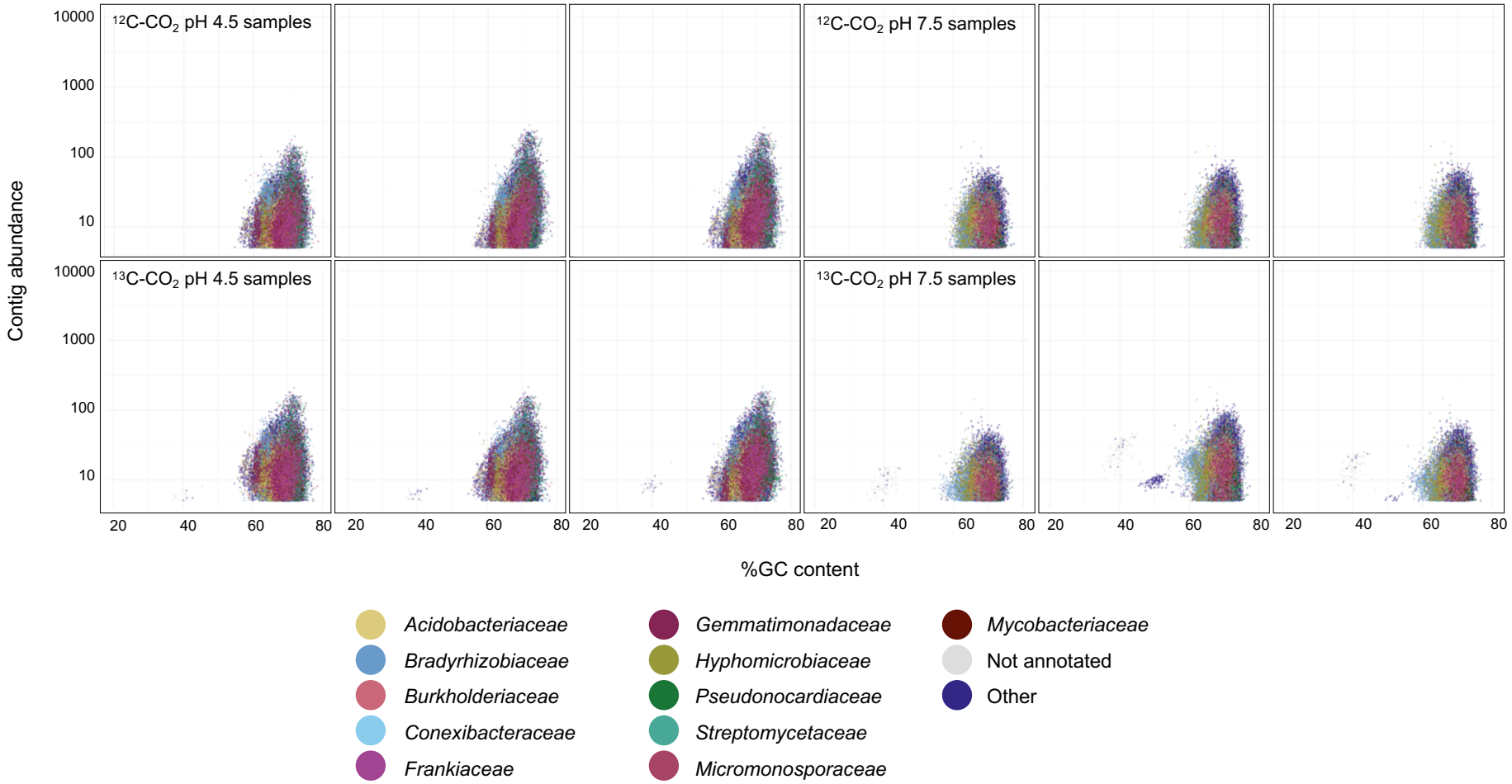


Figure 5.6. Taxon annotated GC – coverage plots of the co-assembled contigs across the ^{12}C and ^{13}C-samples of the pH 4.5 and 7.5 soil. Families with less than 1% of contigs were grouped as "Other" (contains 325 families). Contig abundance was calculated from standardized read coverage in each sample.

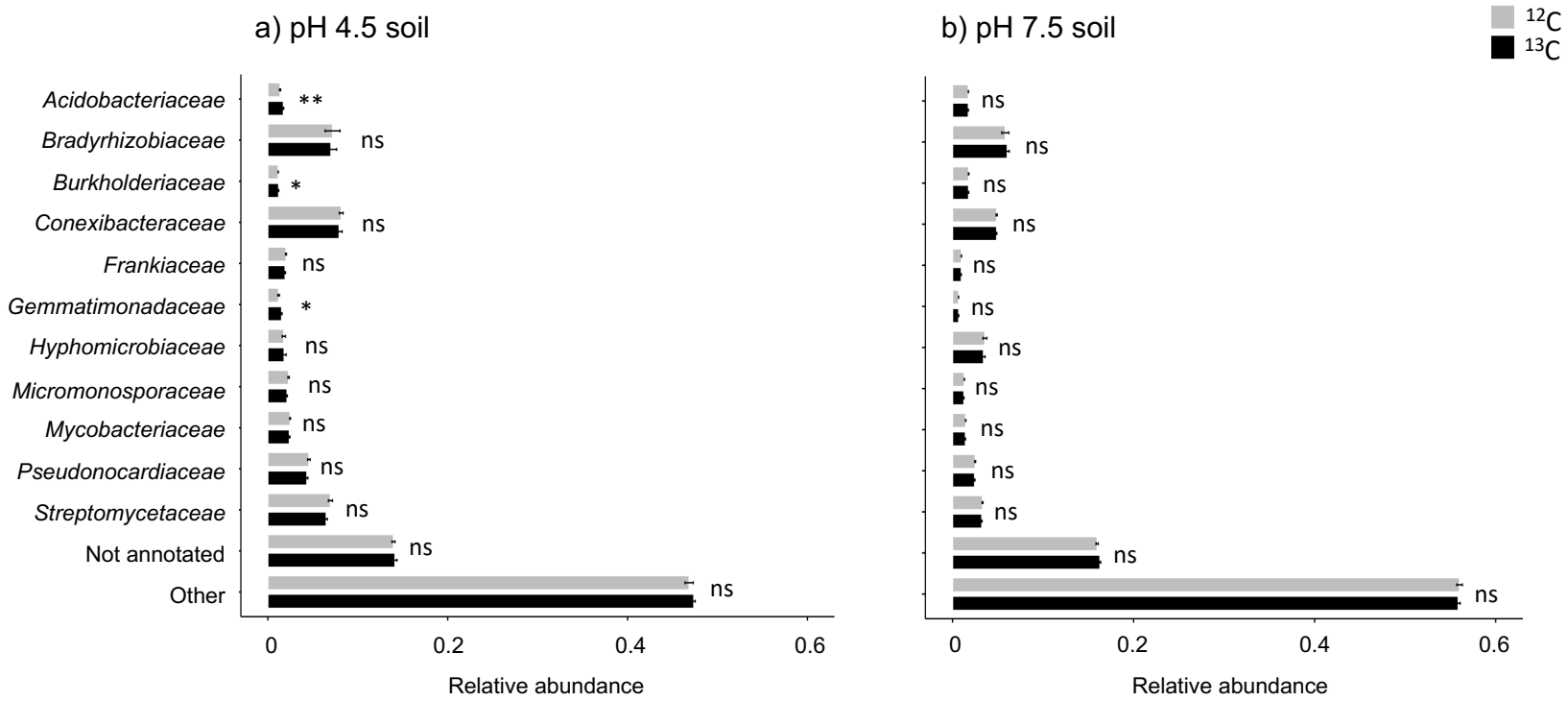


Figure 5.7. Relative abundance of each family with greater than 1% of contigs for the ^{12}C- (grey) and ^{13}C-samples (black) of the a) pH 4.5 and b) 7.5 soil. Families with less than 1% of contigs were grouped as "Other" (contains 325 families). The error bars are based on three replicates (± SE). Significance was tested between ^{12}C- and ^{13}C-samples using Student's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.001$ (***).

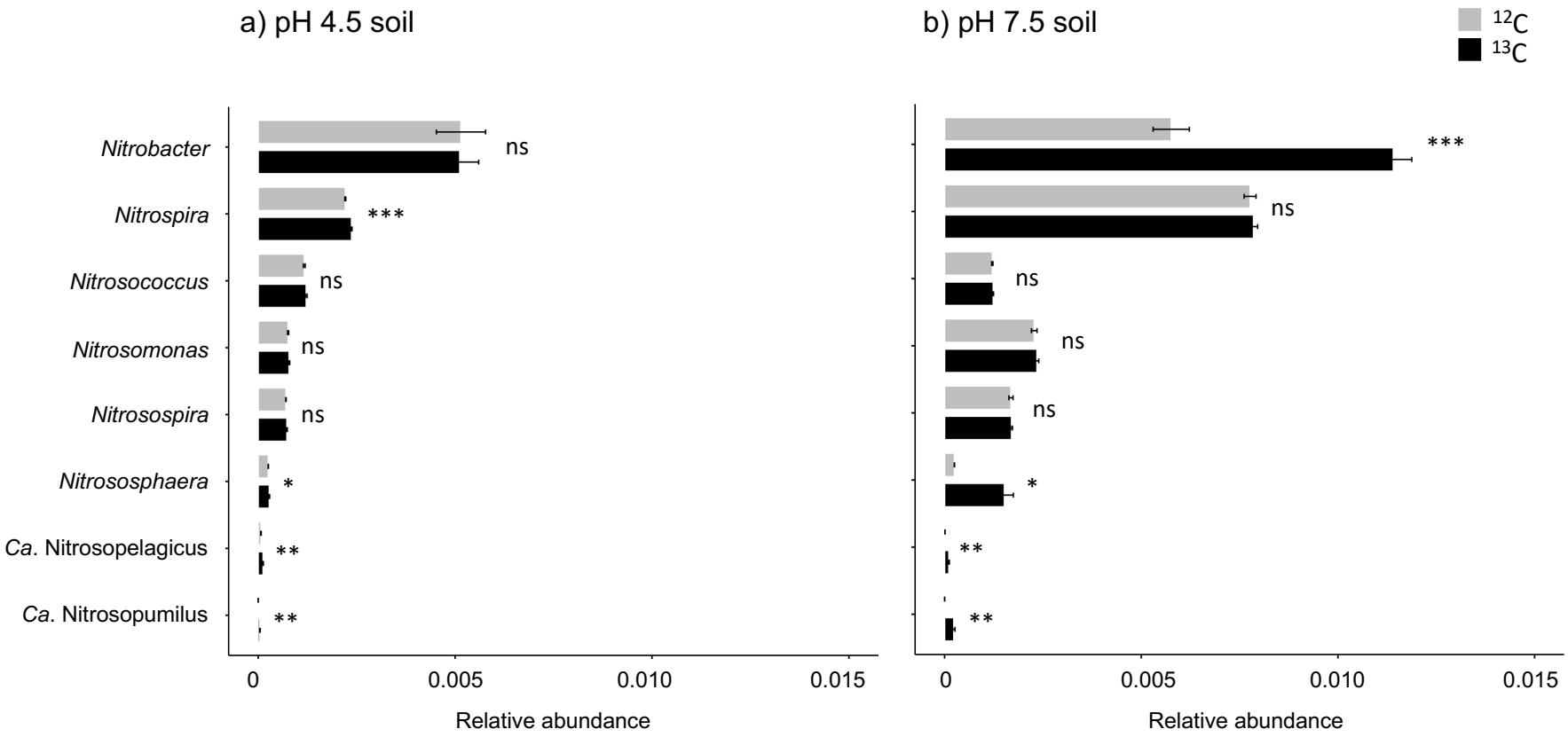


Figure 5.8. Relative abundance of the nitrifying community at the genus level of ^{12}C- (grey) and ^{13}C-samples (black) of the a) pH 4.5 and b) 7.5 soil. The error bars are based on three replicates (± SE). Significance was tested between ^{12}C and ^{13}C-samples using Student's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**); $p <= 0.001$ (***).

Table 5.3. Sequence summary information for the %GC < 50 contigs.

Sample ID	Number of sequences	Summary length	Minimum length	Avgerage length	Maximum length	Number of mVC
GC_50_12C_pH 4.5_1	1,288	417,384	217	324.1	1,378	0
GC_50_12C_pH 4.5_2	805	255,992	249	318	1,055	0
GC_50_12C_pH 4.5_3	898	288,252	227	321	863	0
GC_50_13C_pH 4.5_1	3,187	1,438,745	207	451.4	4,557	0
GC_50_13C_pH 4.5_2	3,906	1,850,145	200	473.7	5,132	0
GC_50_13C_pH 4.5_3	2,507	2,096,369	212	836.2	9,418	2
GC_50_12C_pH 7.5_1	646	197,792	245	306.2	1,378	0
GC_50_12C_pH 7.5_2	505	159,549	245	315.9	1,385	0
GC_50_12C_pH 7.5_3	626	193,266	245	308.7	1,037	0
GC_50_13C_pH 7.5_1	4,199	2,939,521	200	700.1	13,413	0
GC_50_13C_pH 7.5_2	14,144	8,223,381	200	581.4	30,939	5
GC_50_13C_pH 7.5_3	6,889	4,406,211	200	639.6	19,891	1

mVC, metagenomic viral contig

Table 5.4. Sequence summary information for the 50 < %GC < 63 contigs.

Sample ID	Number of sequences	Summary length (bp)	Minimum length (bp)	Avgerage length (bp)	Maximum length (bp)	Number of mVC
GC_50-63_12C_pH 4.5_1	344,918	169,989,431	200	492.8	55,053	0
GC_50-63_12C_pH 4.5_2	276,804	131,467,126	200	474.9	33,955	2
GC_50-63_12C_pH 4.5_3	276,503	137,124,061	200	495.9	30,878	2
GC_50-63_13C_pH 4.5_1	329,093	186,461,291	200	566.6	77,004	3
GC_50-63_13C_pH 4.5_2	296,669	153,029,878	200	515.8	95,220	4
GC_50-63_13C_pH 4.5_3	237,416	124,919,398	200	526.2	78,946	3
GC_50-63_12C_pH 7.5_1	241,570	94,409,305	200	390.8	31,244	4
GC_50-63_12C_pH 7.5_2	188,434	73,106,509	200	388	12,015	3
GC_50-63_12C_pH 7.5_3	232,557	91,479,097	200	393.4	52,275	6
GC_50-63_13C_pH 7.5_1	161,327	68,875,216	200	426.9	13,672	2
GC_50-63_13C_pH 7.5_2	342,677	152,840,879	200	446	32,996	16
GC_50-63_13C_pH 7.5_3	178,477	76,870,784	200	430.7	17,755	4

mVCs, metagenomic viral contigs

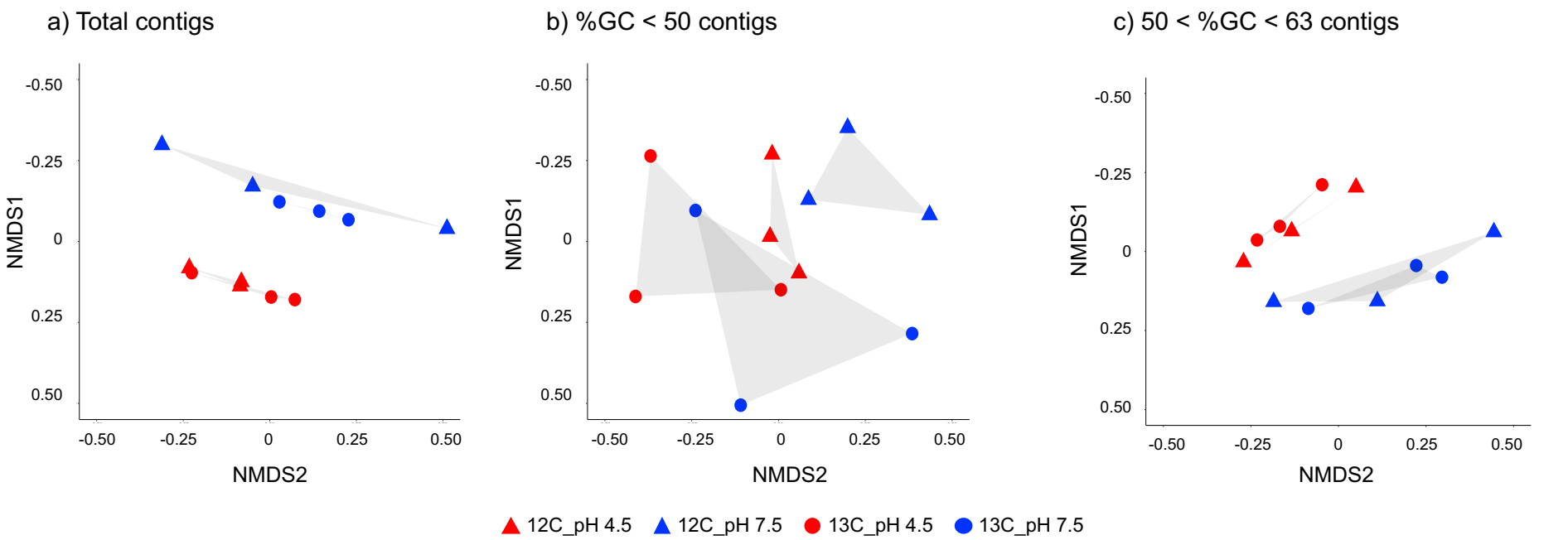


Figure 5.9. Non-metric multidimensional scaling (NMDS) plots of a) total contigs, b) %GC < 50 contigs and c) 50 < %GC < 63 contigs. NMDS plots were derived using Bray-Curtis distance of relative abundance.

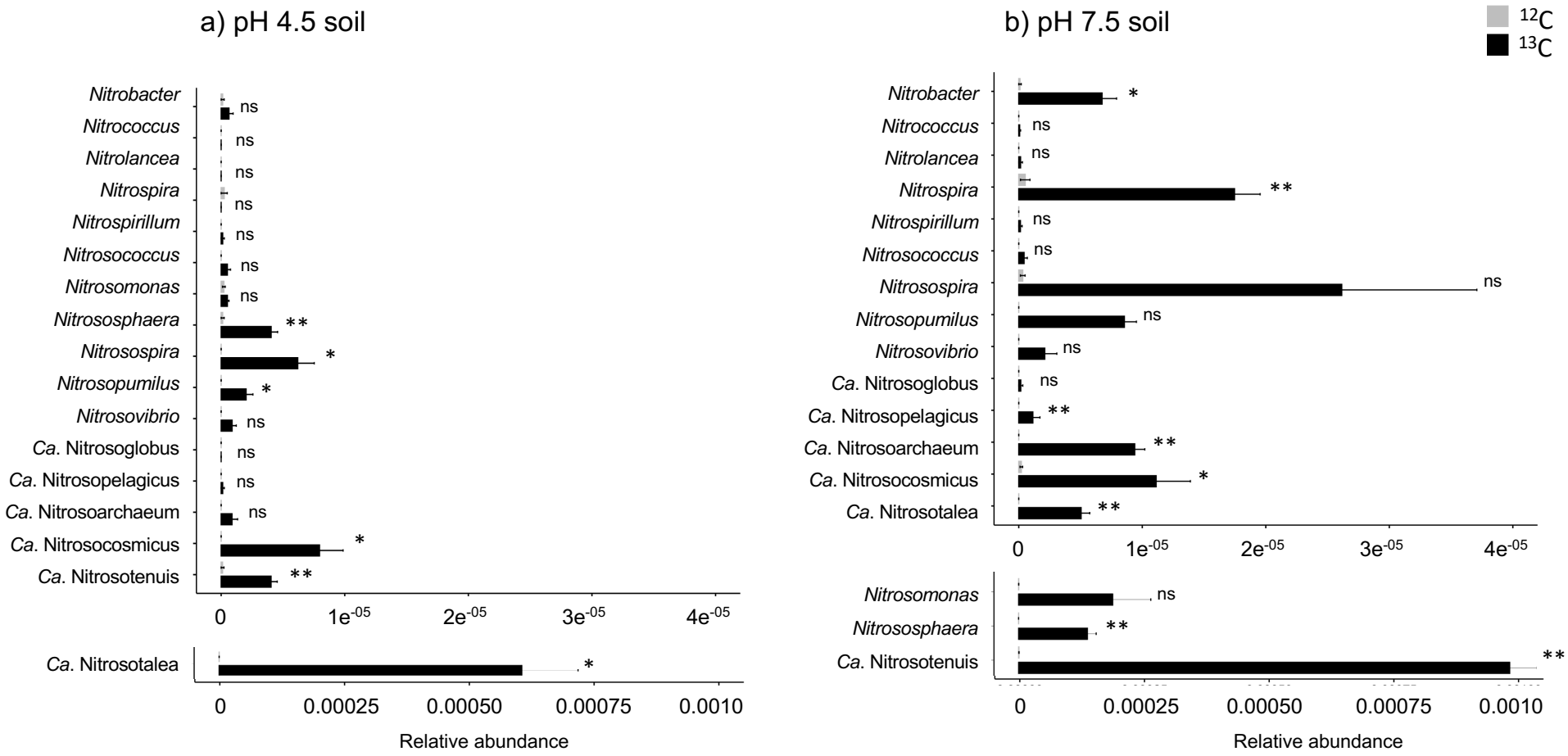


Figure 5.10. Relative abundance of nitrifiers at the genus level in the %GC < 50 contigs of ^{12}C- (grey) and ^{13}C-samples (black) of the in a) pH 4.5 and b) pH 7.5 soil. The error bars are the standard error of the mean of three replicates. Significance was tested between ^{12}C- and ^{13}C-samples using Student's t-test and marked as: $p < 0.05$ (*) and $p <= 0.01$ (**).

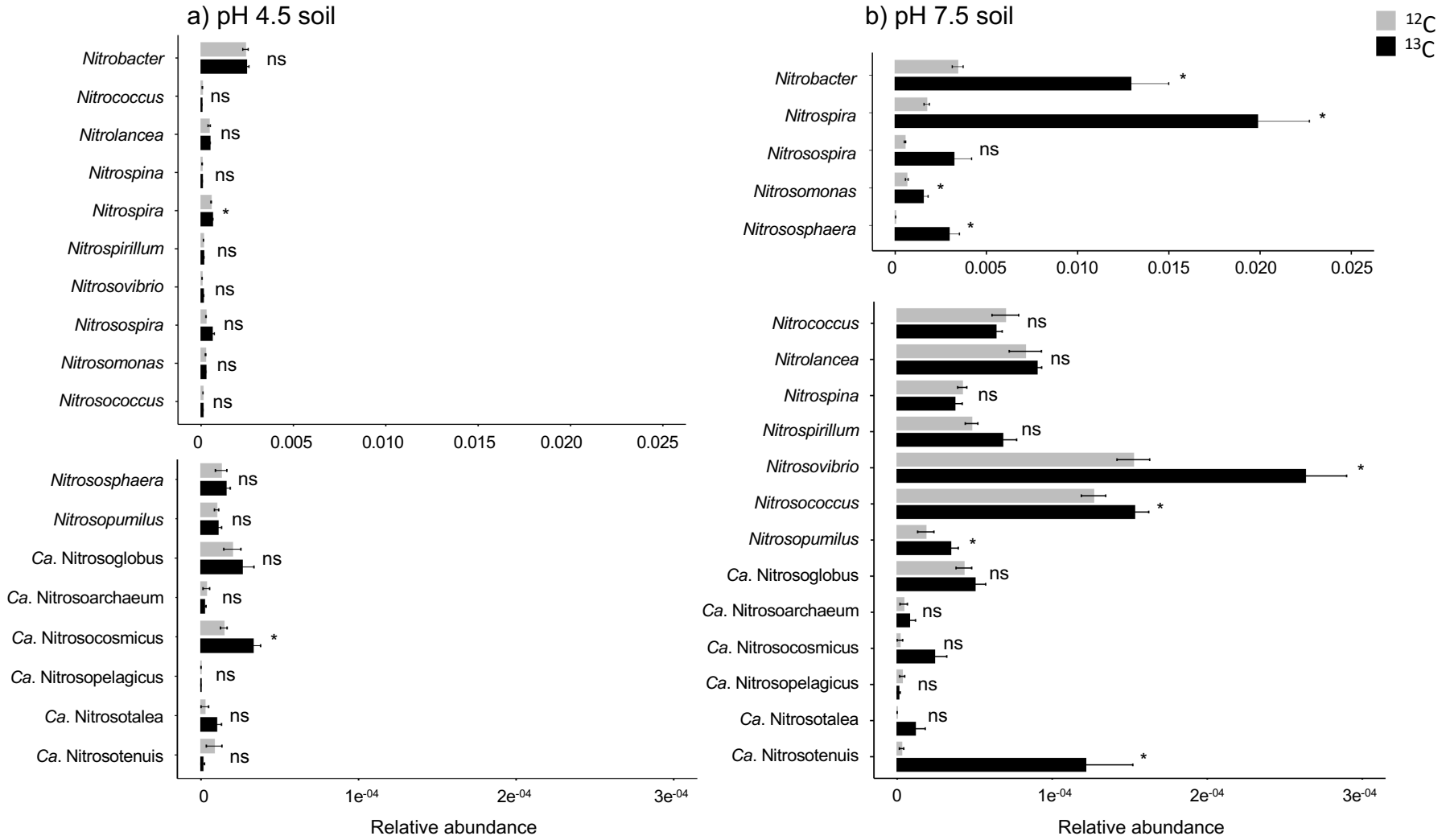


Figure 5.11. Relative abundance of nitrifiers at the genus level in the 50 < %GC < 63 contigs of the ^{12}C- (grey) and ^{13}C-samples (black) of the a) pH 4.5 and b) 7.5 soil. Significance was tested between ^{12}C- and ^{13}C-samples using Student's t-test and marked as: $p < 0.05$ (*).

5.4.6. Virus prediction

Overall, 378 metagenomic viral contigs (mVCs) were predicted using VIRSorter (min. length of 5,003 bp, average length of 9,785 bp and max. length of 46,052 bp). Of these, 6, 108 and 264 mVCs were designated as category 1, 2 and 3 viruses, respectively (see Chapter I, section 1.3.4.1. for category explanations). Seven mVCs were in circular form, and 42 prophage sequences were predicted. Viral community structure was reproducibly distinct between the pH 4.5 and 7.5 soil (Figure 5.12). Only 8% of the mVCs were taxonomically assigned to known viruses, with the majority belonging to *Caudovirales* (Figure 5.13).

From the two separated %GC categories, no viral contigs were predicted using VirSorter. However, using DeepVirFinder, eight and 49 mVCs were predicted from the %GC < 50 and 50 < %GC < 63 populations, respectively (Table 5.3 and 5.4). Specifically, from the %GC < 50 contigs, two mVCs were predicted from one replicate ^{13}C-sample in pH 4.5 soil, and five and one mVC from two replicates of the ^{13}C-samples in pH 7.5 soil (Table 5.3). From the %GC < 50 and 50 < %GC < 63 populations, 14 mVCs were predicted from the pH 4.5 soil (4 and 10 mVCs in the ^{12}C- and ^{13}C-samples, respectively) and 35 mVCs from the pH 7.5 soil (13 and 22 mVCs in the ^{12}C- and ^{3}C-samples, respectively) (Table 5.4).

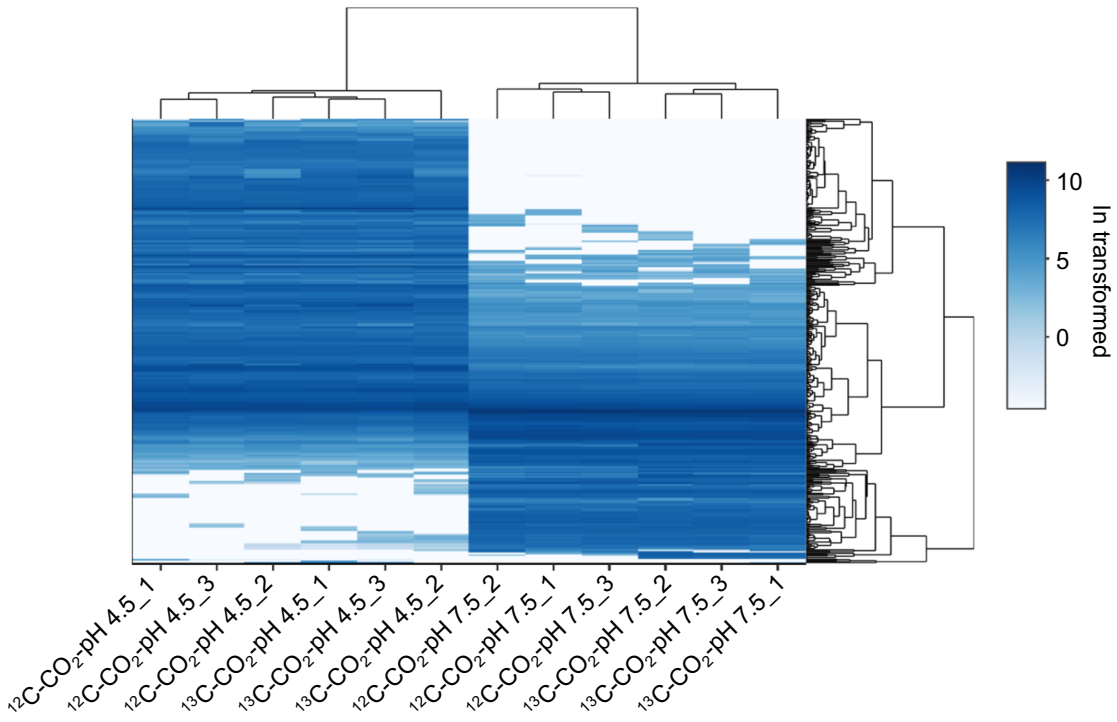


Figure 5.12. Normalized relative abundance (ln transformed) of metagenomic viral contigs (mVCs) of the ^{12}C and ^{13}C-samples of the pH 4.5 and 7.5 soil.

Figure 5.13. Taxonomic annotation of the metagenomic viral contigs (mVCs) of the ^{12}C and ^{13}C-samples of the pH 4.5 and 7.5 soil.

5.4.7. Host-virus linkage

5.4.7.1. CRISPR array analysis

The CRT and Crass tools were both used for CRISPR array analysis. With CRT, ten CRISPR arrays were predicted from the 76,311 co-assembled contigs (> 5 kb). Also, a total of 505 and 217 CRISPR arrays were predicted from 8,180,712 and 8,939,038 contigs (> 300 bp), respectively, from the ^{13}C-samples of the pH 4.5 and 7.5 soils, respectively. Of the 3,468 spacer sequences of the predicted CRISPR arrays, only 12 spacers matched to seven mVCs. These seven mVCs were specific to the ^{13}C-samples of the pH 4.5 soil. One contig was annotated as *Citrobacter freundii,* with the other six annotated as related to *Salinispora arenicola.* A CRISPR array that was predicted from a MAG belonged to *Solirubrobacterales* (bin.10), however, the associated spacer sequences did not match to any mVCs. With Crass, 1,357 and 907 CRISPR arrays were assembled from the ^{13}C-samples of the pH 4.5 and 7.5 soil, respectively. Of the 23,994 spacer sequences, 19 mVCs of the ^{13}C-samples of the pH 4.5 soil were linked to hosts. Among these, 14 mVCs were also host-linked by the CRT tool. Six of the mVCs were linked to *Salinispora arenicola*, and 2 mVCs to each of *Salinispora arenicola*, *Citrobacter freundii*, *Paraburkholderia* caribensis and *Serratia sp.*

5.4.7.2. ONF analysis

For ONF analysis, the WIsH tool was used to predict host-virus linkage between 420 viruses (378 mVCs, 42 prophages) and 75,891 co-assembled prokaryotic contigs (> 5 kb). A total of 105 host-virus linkages had a *p*-value equal to or less than 0.05 (Supplementary Table 5.1). A large proportion of the mVCs (44%) were predicted to be associated with *Actinobacteria* (31 mVCs)

and *Acidobacteria* (15 mVCs). Seven nitrifier-associated mVCs (four AOA- and three NOB-associated mVCs) were also predicted (Table 5.5). For the AOA-associated mVCs, two host contigs belonged to *Ca.* Nitrosotenuis, and one host contig linked to *Nitrososphaera* and *Ca*. Nitrosopelagicus (Table 5.5). The contig size of the AOA hosts ranged between 7 – 19 kb, and the mVCs contig size ranged between 5 – 27 kb (Table 5.5).

For the %GC < 50 population, among the 161 co-assembled prokaryote contigs, eight contigs were each predicted as a host and a total of seven host-virus linkages had a *p*-value equal to or less than 0.05 (Table 5.6). The host contigs were taxonomically annotated as Archaea, with seven of the eight contigs being AOA (Table 5.6). Specifically, three contigs each belonged to *Ca*. Nitrosotalea devanaterra and *Ca*. Nitrosotenuis chungbukensis, and one contig belong to *Ca*. Nitrososphaera evergladensis. The contig size of the AOA hosts ranged between 5 – 11 kb, and the mVCs contig size ranged between 397 – 2,578 bp (Table 5.6). For the 50 < %GC < 63 population, among the 4,346 co-assembled prokaryotic contigs, hosts for each of the 33 mVC were predicted. None of the mVCs were linked to nitrifier hosts, and the majority of the mVCs (31 out of 33 mVCs) were linked to *Alphaproteobacteria*.

The 12 AOA-associated mVCs and three NOB-associated mVCs were compared to 1,450 reference viral genomes (Figure 5.14). Eleven of the AOA-associated mVCs clustered together with no association of any characterized viruses, with one mVC (mVC_473879_LOW_GC) showing a low amount of association with a *Crenarchaeota*-associated *Sulfolobus* virus (Figure 5.14). The NOB-associated mVCs shared some similarity with viral genomes, with mVC_17296-cat_2-NOB, mVC_71296-cat_3-NOB and mVC_71803-cat_3-NOB related to a *Caulobacter* phage, a *Haloarcula hispanica* icosahedral virus and a *Natrinema* virus, respectively (Figure 5.14).

Table 5.5. Summary information of the WIsH predicted nitrifier-associated metagenomic viral contigs (mVCs) and host contigs (all contigs) .

Virus ID	Virus size (bp)	Virus GC (%)	Host ID/ Host annotation	Host size (bp)	Host GC (%)	Log likelihood	*p*-value
mVC_58257-cat_2-AOA	5,707	39.58	c_073106/ *Ca.* Nitrosotenuis cloacae	19,891	40.58	-1.38407	8.18678e-13
mVC_61037-cat_3-AOA	10,230	35.87	c_074537/ *Ca.*Nitrosopelagicus brevis	7,008	37.34	-1.28576	0
mVC_67361-cat_3-AOA	7,077	46.01	c_065610/ *Nitrososphaera viennensis*	8,632	48.95	-1.38332	1.36557e-14
mVC_62847-circular-cat_3-AOA	27,239	38.60	c_073203/ *Ca.*Nitrosotenuis cloacae	15,794	43.40	-1.38441	1.30066e-09
mVC_17296-cat_2-NOB	7,416	64.21	c_042479/ *Nitrobacter hamburgensis*	5,414	61.60	-1.2923	0
mVC_71296-cat_3-NOB	5,423	58.40	c_074499/ *Nitrobacter winogradskyi*	7,100	59.96	-1.27976	0
mVC_71803-cat_3-NOB	5,297	55.88	c_065314/ *Nitrobacter winogradskyi*	9,021	59.08	-1.38051	4.26444e-09

Table 5.6. Summary information of the WIsH predicted ammonia-oxidizing archaea (AOA)-associated metagenomic viral contigs (mVCs) and their host contigs of %GC < 50 contigs.

Virus ID	Virus size (bp)	Virus GC (%)	Best hit among provided hosts through WIsH	Host size (bp)	Host GC (%)	Log likelihood	*p*-value
mVC_51245_LOW_GC	2,578	41.8	c_068591/ *Ca.* Nitrososphaera evergladensis	6,406	47.9	-1.384	0.028
mVC_131962_LOW_GC	1,655	42.8	c_061037/ *Ca.* Nitrosotenuis chungbukensis	10,230	35.8	-1.381	0.015
mVC_92579_LOW_GC	1,956	40.3	c_075866/ *Ca.* Nitrosotenuis chungbukensis	5,326	42.4	-1.384	0.042
mVC_36288_LOW_GC	2,115	36.5	c_064273/ *Thaumarchaeota*	11,072	49.9	-1.383	0.031
mVC_179812_LOW_GC	1,429	40.9	c_056233/ *Ca.* Nitrosotalea devanaterra	6,565	38.8	-1.384	0.043
mVC_186378_LOW_GC	1,405	41	c_074802/ *Ca* .Nitrosotenuis chungbukensis	6,544	40.5	-1.384	0.140
mVC_473879_LOW_GC	783	40.8	c_053890/ *Ca.* Nitrosotalea devanaterra	8,429	39.1	-1.381	0.030
mVC_1363578_LOW_GC	397	42.8	c_058074/ *Ca.* Nitrosotalea devanaterra	5,776	37.8	-1.379	0.010

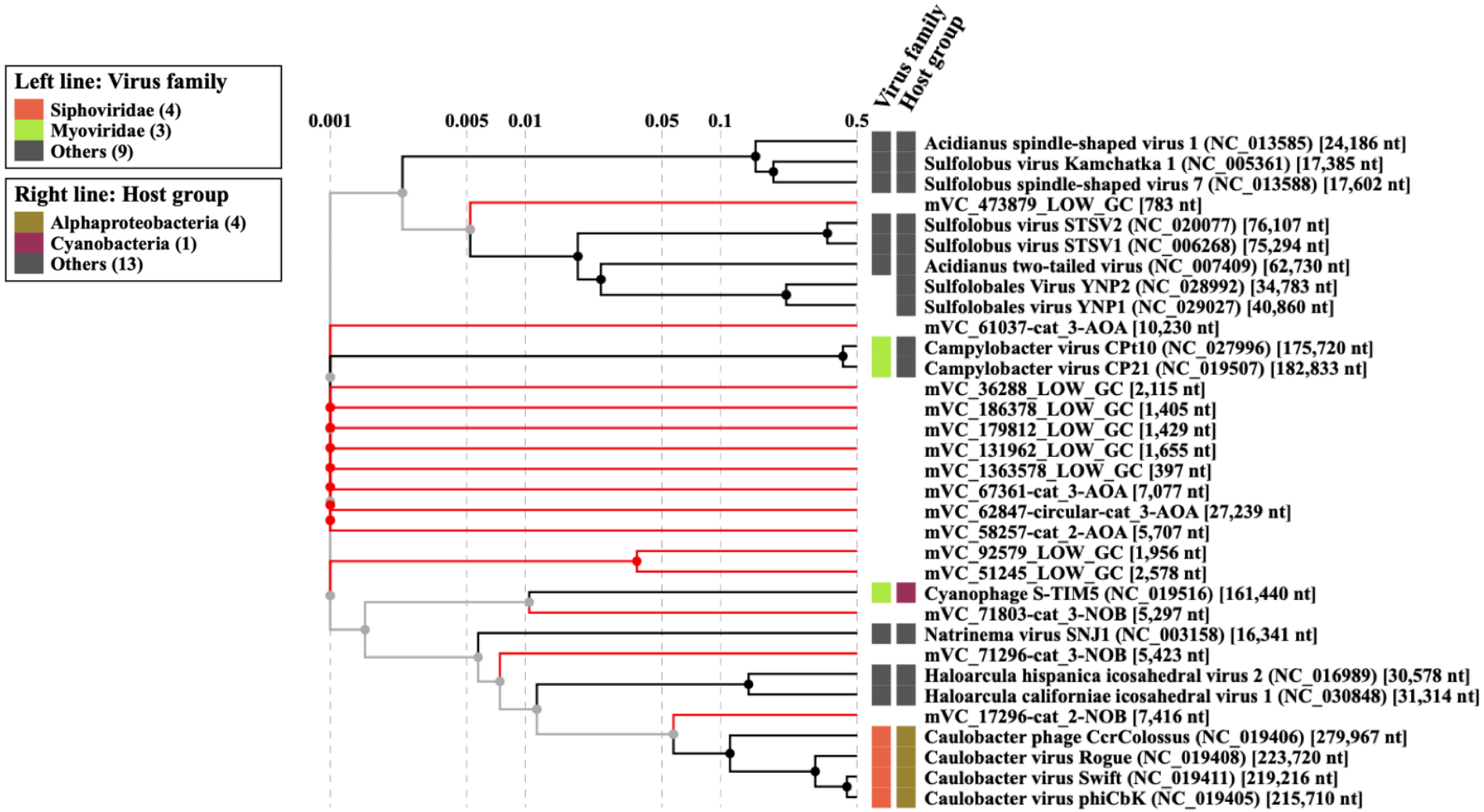


Figure 5.14. Proteomic tree containing the 15 nitrifier-associated metagenomic viral contigs (mVCs) (red) with the most closely related reference viral genomes (black).

5.4.8. Gene homology

5.4.8.1. Auxiliary metabolic genes

In total, 5,375 genes were predicted from VirSorter category 1, 2 and 3 mVCs (378 mVCs), and 49% (2,666 genes) were annotated, with 465 genes of unique function. Viral proteins accounted for 5.1% (137 genes) of the annotated genes, and included major capsid proteins, tail proteins, integrases, portal proteins and terminases Bacterial proteins used for viral replication, such as nucleic acid synthesis and DNA repair, accounted for 4.3% (116 genes) of the annotated genes. After the removal of known viral-associated protein families, the total number of unique protein families that were possible AMGs was 377 (90%, 2,413 genes). In greater detail, AMGs associated with carbon metabolism; glycoside hydrolase and transferases (GH family 25) (< 1%, 14 genes) and peptidases (1%, 29 genes) were identified. Proteins encoding for ABC transporter (< 1%, 20 genes) and ATPase (< 1%, 11 genes) were also identified.

In total, 1,377 genes were predicted from VirSorter category 5 and 6 mVCs (42 prophages), and 65% (896 genes) were annotated, with 240 genes of unique function. Viral and bacterial proteins used for viral replication accounted for 1.5% (14 genes) and 4.1% (37 genes) of the annotated genes, respectively. The total number of unique protein families that were possible AMGs was 209 (61%, 845 genes), with proteins encoding for peptidases (11 genes), ABC transporter (16 genes) and ATPase (5 genes).

5.4.8.2. Gene homology between viruses and their associated nitrifier hosts

A total of 78 genes, ranging between 8 – 20 genes per host, were predicted (c_073106, 17; c_074537, 8; c_065610, 7; c_042479, 8; c_074499, 8; c_065314, 10; c_073203, 20), and aligned to the genes of the nitrifier-associated mVCs (Table 5.7). Overall, a high level of identity was observed between viral and host genes (Table 5.7). Two out of the 12 AOA-associated mVCs and all three of the NOB-associated mVCs were found to have acquired genes from their hosts. Specifically, seven genes from mVC_61037-cat_3-AOA were 100% identical to genes of its host (c_074537). Also, one gene of mVC_17296-cat_2-NOB, four genes of mVC_17296-cat_2-NOB and one gene of mVC_71296-cat_3-NOB were 100% identical to genes of their hosts (Table 5.7). Ten of these AOA and NOB mVC genes were functionally annotated and included genes encoding for a SNase-like, dimeric alpha-beta barrel, lambda repressor-like, DNA-binding domain and an integrase-like, catalytic domain (Table 5.7).

Table 5.7. Gene homology between nitrifier host contigs and associated metagenomic viral contigs (mVCs).

Viral protein ID	VP AA length[1] (bp)	Host protein ID (bp)	HP AA length[2] (bp)	Identity (%)	E value	bit score	QC[3] (%)	Protein function	E value
mVC_61037-cat_3-AOA_2	102	c_074537_2	102	100	1.24e-75	211	100		
mVC_61037-cat_3-AOA_3	96	c_074537_3	96	100	7.67e-72	201	100	SNase-like, OB-fold superfamily	8.3e-6
mVC_61037-cat_3-AOA_4	84	c_074537_4	84	100	2.14e-60	171	100		
mVC_61037-cat_3-AOA_5	35	c_074537_5	35	100	2.09e-22	71.6	100		
mVC_61037-cat_3-AOA_6	82	c_074537_6	82	100	2.20e-58	166	100	Dimeric alpha-beta barrel	2.68e-9
mVC_61037-cat_3-AOA_7	302	c_074537_7	302	100	0	610	100	Transcription factor TFIIB, conserved site	8.14e-8
mVC_61037-cat_3-AOA_8	955	c_074537_8	960	100	0	1931	100		
mVC_67361-cat_3-AOA_15	231	c_065610_1	450	83.8	4.67e-91	267	70		
mVC_17296-cat_2-NOB_2	234	c_042479_3	234	100	1.86e-172	466	100	Outer membrane protein beta-barrel domain	8.3e-21
mVC_17296-cat_2-NOB_3	144	c_042479_4	144	99.3	1.51e-105	290	100	Putative phage cell wall peptidase	2.4e-62
mVC_17296-cat_2-NOB_4	299	c_042479_5	299	100	0	601	100	Phage conserved hypothetical protein BR0599	3.4e-31
mVC_17296-cat_2-NOB_5	80	c_042479_6	80	100	2.91e-56	160	100	Toxin-antitoxin system	1.83e-13
mVC_17296-cat_2-NOB_6	125	c_042479_7	125	100	2.85e-88	244	100	Lambda repressor-like, DNA-binding domain superfamily	1.01e-10
mVC_17296-cat_2-NOB_7	213	c_042479_8	213	100	1.05e-157	427	100	Protein of unknown function DUF2460	5.8e-81
mVC_71296-cat_3-NOB_3	259	c_074499_1	194	96.7	1.01e-132	365	71		
mVC_71296-cat_3-NOB_4	190	c_074499_2	190	99.4	2.43e-141	384	100		
mVC_71296-cat_3-NOB_5	101	c_074499_3	101	99.0	3.90e-72	202	100	Integrase-like, catalytic domain superfamily	6.3e-13
mVC_71296-cat_3-NOB_6	158	c_074499_4	158	100	2.37e-120	328	100		
mVC_71296-cat_3-NOB_7	247	c_074499_5	237	99.5	0	488	96		
mVC_71296-cat_3-NOB_8	259	c_074499_6	314	99.6	0	502	98		
mVC_71803-cat_3-NOB_6	92	c_065314_1	75	80.5	4.58e-32	99.8	73		

[1]VP AA= Viral Protein, Amino Acid ; [2]HP AA= Host Protein, Amino Acid ; [3]QC= Query coverage

5.4.8.3. Gene homology of host-linked viruses to other prokaryotes

Analysis of gene homology between the 15 predicted nitrifier-associated mVCs and prokaryotes revealed that at least one gene exhibited similarity to a previously characterized nitrifier gene (of mostly unknown function), with the exception of four AOA-associated mVCs from the %GC < 50 population (mVC_131962_LOW_GC, mVC_36288_LOW_GC, mVC_179812_LOW_GC and mVC_186378_LOW_GC) (Table 5.8, Supplementary Table 5.2). In particular, the genes of two mVCs (mVC_58257-cat_2-AOA, mVC_473879) associated with Ca. *Nitrososphaera evergladensis* and Ca. *Nitrosotalea devanaterra*, had high identity (> 96%) to genes of Ca. *Nitrosotalea devanaterra* (an uncharacterized protein and metal-sulfur cluster biosynthetic enzyme) (Supplementary Table 5.2). The genes of other AOA-associated mVCs had similarities to predicted proteins ranging between 25 - 93.7% identity (average 58.4%) (Supplementary Table 5.2). These genes were mostly homologous to proteins of *Thaumarchaeota*, but some AOA-mVCs contained genes homologous to proteins of other prokaryotes, such as Ca. *Wolfebacteria bacterium*, *Phialocephala subalpina*, *Bifidobacterium breve* (mVC_62847-circular-cat_3), *Clostridium* sp. (mVC_92579_LOW_GC), *Planctomycetia bacterium* (mVC_179812_LOW_GC), and *Spirosoma radiotolerans* (mVC_186378_LOW_GC) (Supplementary Table 5.2).

Predicted AMGs of the AOA-associated mVCs included a metal sulfur cluster biosynthetic enzyme (mVC_58257-cat_2-AOA), methylthioribose-1-phosphate isomerase (mVC_58257-cat_2-AOA), structural maintenance of chromosomes protein (mVC_62847-circular-cat_3-AOA), prepilin peptidase (mVC_61037-cat_3-AOA), putative UbiA prenyltransferase (mVC_67361-cat_3-AOA), AAA family ATPase (mVC_67361-cat_3-AOA) and 6-bladed beta-propeller (mVC_51245_LOW_GC) (Supplementary Table 5.2). Only one AOA-associated mVC (mVC_58257-cat_2-AOA) contained a viral hallmark gene (a terminase large subunit) (Figure 5.15). The mVC_61037-cat_3-AOA harbored a gene encoding N-6-adenine-methyltransferase that originated from a phage (Supplementary Table 5.2).

The genes of the predicted NOB-associated mVCs exhibited sequence similarity to proteins with mostly unknown functions of previously characterized NOB (Table 5.8). Gene identity of homologs of the NOB-associated mVCs ranged between 39 – 91% (average of 77%) (Supplementary Table 5.2). Genes were mostly homologous to proteins of *Nitrobacter*, but some NOB-associated mVCs had genes homologous to known proteins of *Bradyrhizobium* (mVC_17296-cat_2-NOB, mVC_71296-cat_3-NOB, mVC_71803-cat_3-NOB), and *Pseudorhodoplanes* and *Pseudolabrys* (mVC_17296-cat_2-NOB) (Supplementary Table 5.2). The mVC_71296-cat_3-NOB that was predicted to be associated with *Nitrobacter hamburgensis* (Table 5.8), had a gene coding for phage integrase that was homologous to *Nitrobacter winogradskyi*, with 91% identity (Supplementary Table 5.2). Predicted AMGs of the NOB-associated mVCs were of various functions, including outer membrane immunogenic protein, peptidase, type II toxin-antitoxin

system (mVC_17296-cat_2-NOB), and cold-shock protein and molecular chaperone IbpA (mVC_71803-cat_3-NOB) (Supplementary Table 5.2). Two of the NOB-associated mVCs (mVC_17296-cat_2-NOB, mVC_71296-cat_3-NOB) had viral hallmark genes (Figure 5.17). Specifically, mVC_17296-cat_2-NOB had a phage tail tube protein, gene transfer agent, major tail protein and tail completion protein, and mVC_71296-cat_3-NOB had a phage integrase (Supplementary Table 5.2).

Table 5.8. Gene homology between the nitrifier-associated metagenomic viral contigs (mVCs) and database prokaryotes with the same taxa as the host and those of ammonia oxidizers (AO).

Virus ID	Host annotation	Number of homologs with host genera	Number of homologs with AO
mVC_58257-cat_2-AOA	*Ca.* Nitrosotenuis cloacae	0	6
mVC_61037-cat_3-AOA	*Ca.*Nitrosopelagicus brevis	7	10
mVC_67361-cat_3-AOA	*Nitrososphaera viennensis*	1	7
mVC_62847-circular-cat_3-AOA	*Ca.*Nitrosotenuis cloacae	0	2
mVC_17296-cat_2-NOB	*Nitrobacter hamburgensis*	6	1
mVC_71296-cat_3-NOB	*Nitrobacter winogradskyi*	6	4
mVC_71803-cat_3-NOB	*Nitrobacter winogradskyi*	1	7
mVC_51245_LOW_GC	*Ca.* Nitrososphaera evergladensis		1
mVC_131962_LOW_GC	*Ca.* Nitrosotenuis chungbukensis		1
mVC_92579_LOW_GC	*Ca.* Nitrosotenuis chungbukensis		1
mVC_36288_LOW_GC	*Thaumarchaeota*		0
mVC_179812_LOW_GC	*Ca.* Nitrosotalea devanaterra		0
mVC_186378_LOW_GC	*Ca* .Nitrosotenuis chungbukensis		0
mVC_473879_LOW_GC	*Ca.* Nitrosotalea devanaterra		0
mVC_1363578_LOW_GC	*Ca.* Nitrosotalea devanaterra		1

5.4.8.3. Gene homology between nitrifier-associated viruses

Gene homology analysis amongst the AOA-associated mVCs showed that there was no similarity between AOA-associated mVCs, except for between mVC_51245 and mVC_92579 from the% GC < 50 contigs (Figure 5.15). Comparison of these AOA-associated mVCs to an AOA-associated viral contig derived from viromes of the pH 4.5 and 7.5 soil reported in Chapter II, showed sequence similarities of 30 – 40% identity (Figure 5.16). There was no similarity between NOB-associated mVCs (Figure 5.17).

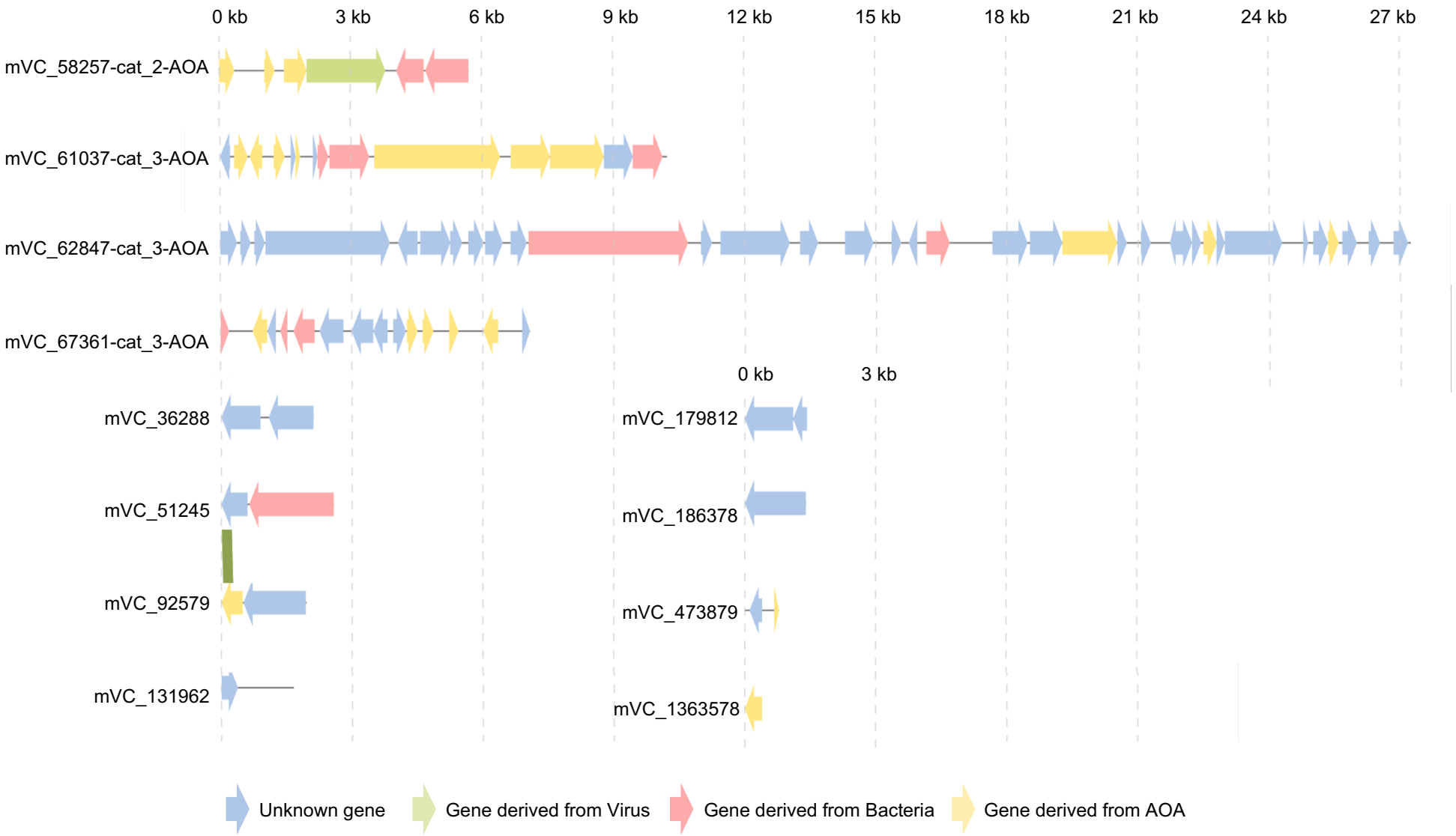


Figure 5.15. Genome map of the AOA-associated metagenomic viral contigs (mVCs) and their predicted encoded proteins. Bacterial and unknown genes (blue), genes derived from Thaumarchaeota (yellow) and auxiliary metabolic genes (red) are shown.

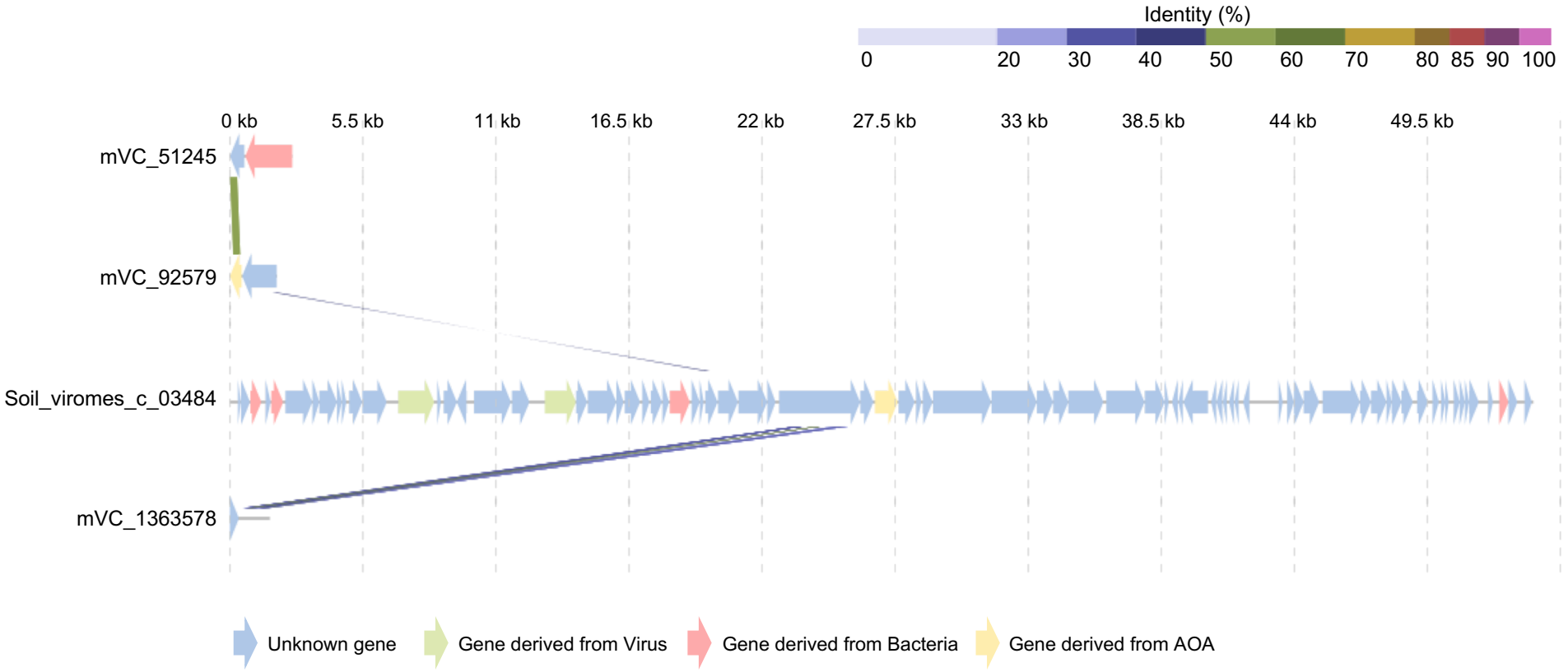


Figure 5.16. Genome map comparing an AOA-associated viral contig derived from a direct soil virome (c_3484; see Chapter III) with short AOA-associated metagenomic viral contigs (mVCs). Bacterial and unknown genes (blue), genes derived from Thaumarchaeota (yellow) and auxiliary metabolic genes (red) are shown.

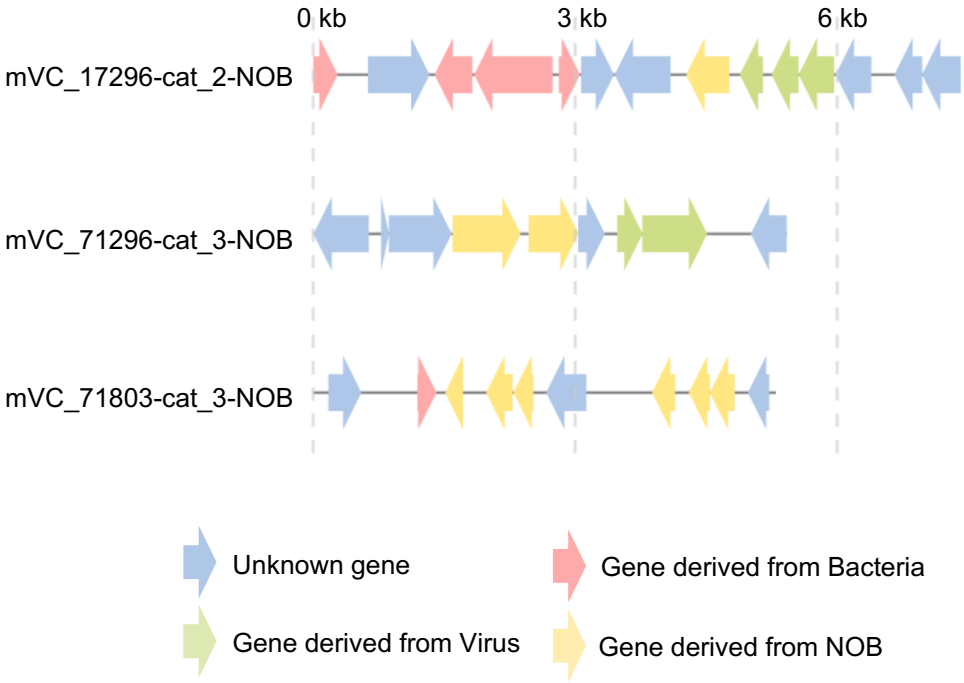


Figure 5.17. Genome map of the NOB-associated metagenomic viral contigs depicting predicted encoded proteins. Bacterial and unknown genes (blue), genes derived from NOB (yellow) and auxiliary metabolic genes (red) are shown.

5.4.9. Distribution of nitrifier-associated viruses across soil pH

The nitrifier-associated viral populations were all sufficiently enriched (at least 1X coverage) in all pH 4.5 and 7.5 soil replicates, with the exception of mVC_17296-cat_2-NOB (Figure 5.18). All of the AOA-associated mVCs were present in the pH 7.5 soil, with the exception of mVC_58257-cat_2 that was only detected in the pH 4.5 soil (Figure 5.18). The predicted host of mVC_58257-cat_2, *Ca*. Nitrosotenuis, was found to be enriched in both pH soils of the %GC < 50 population (Figure 5.10), but was only enriched in the 50 < %GC < 63 population of the pH 7.5 soil (Figure 5.11b). The predicted host of mVC_61037-cat_3, Ca. *Nitrosopelagicus*, was enriched in the pH 7.5 soil in the 50 < %GC < 63 population (Figure 5.11b), and the predicted host of mVC_67361-cat_3, *Nitrososphaera*, was enriched in both pH soils of the %GC < 50 population (Figure 5.10). The NOB-associated mVC_71296-cat_3 and mVC_71803-cat_3 were both enriched in the pH 7.5 soil (Figure 5.18b), and similarly as its predicted *Nitrobacter* host (Figure 5.10b).

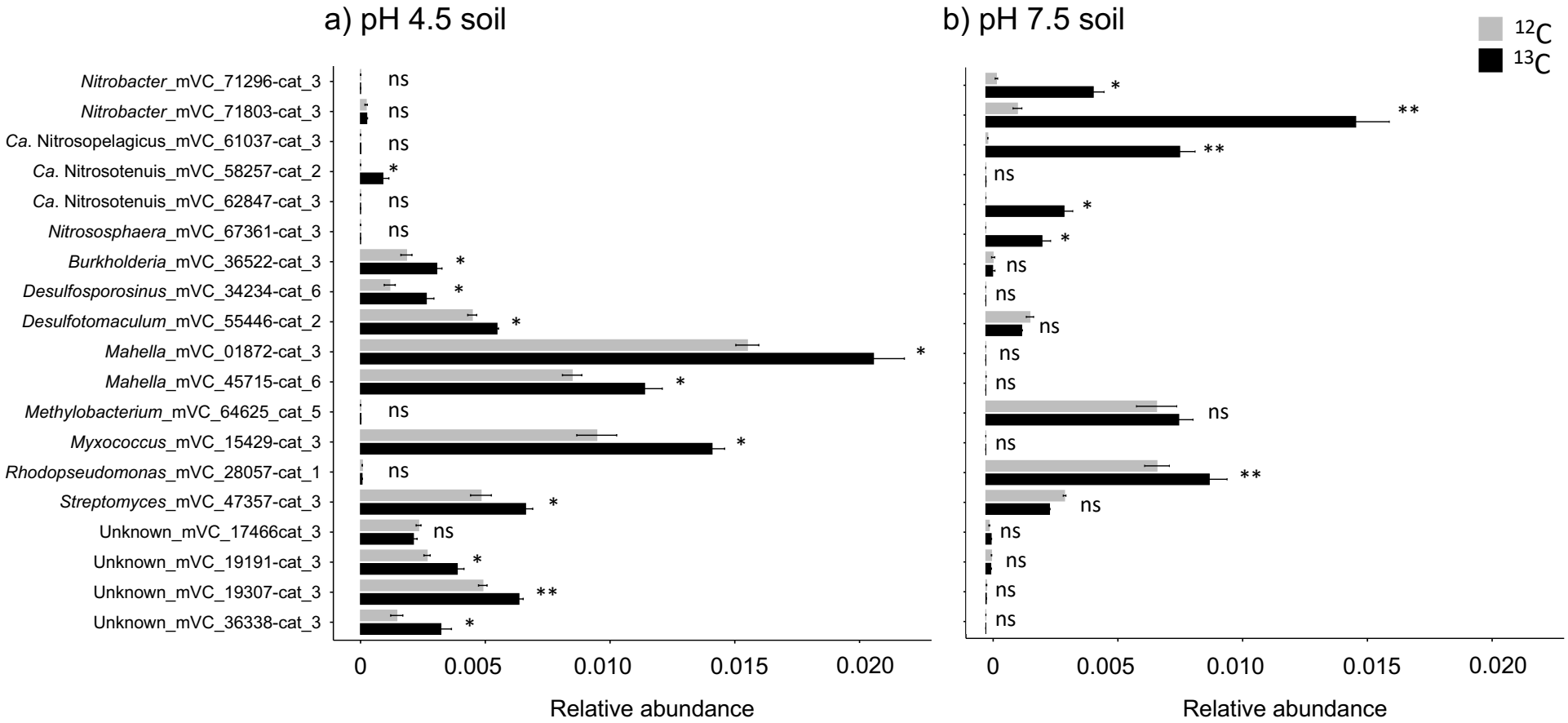


Figure 5.18. Relative abundance of representative enriched metagenomic viral contigs (mVCs) of the ^{12}C- (grey) and ^{13}C-samples (black) of the a) pH 4.5 and b) 7.5 soil. Error bars represent the standard error of the mean of three replicates. Significance was tested between ^{12}C- and ^{13}C-samples using Student's t-test and marked as: $p > 0.05$ (ns); $p <= 0.05$ (*); $p <= 0.01$ (**).

5.5. Discussion

To investigate the potential interaction between viruses and nitrifying prokaryotes, DNA-SIP was combined with deep metagenomic sequencing to identify populations of both host and virus. Specifically, using $^{13}CO_2$, and two soils of contrasting pH known to contain distinct nitrifier communities, autotrophic nitrifiers and associated viruses were identified, demonstrating potential carbon flow between hosts and their associated lytic viruses.

The quantitative analysis of archaeal and bacterial *amoA* gene distribution in CsCl gradients revealed ^{13}C-CO_2 assimilation by both ammonia-oxidizing archaea (AOA) and ammonia-oxidizing bacteria (AOB) within both pH 4.5 and 7.5 soils. Taxonomic assignment of contigs also demonstrated that representatives of both AOA, AOB and NOB had significantly higher relative abundances in samples derived from ^{13}C incubation demonstrating enrichment in metagenomes of a wide variety of nitrifier populations. However, despite this substantial enrichment, ^{13}C-enriched nitrifier communities in both pH soils never represented more than 2% of all contigs, the majority being derived from heterotrophic bacteria. This was consistent with the quantitative analysis of the distribution of total bacterial 16S rRNA genes, where there was no difference between ^{12}C- and ^{13}C-samples in both pH soils. This was also consistent with similar taxonomic profiles of contigs from ^{12}C- and ^{13}C-samples in both pH soils, also indicating the vast majority of sequenced DNA came from unlabeled heterotrophic bacteria. The lack of major differences in the relative abundance between autotrophic nitrifier vs heterotrophic DNA in high buoyant density samples was also compounded by differences in the %GC content, with an increasing %GC content also increasing the buoyant density of DNA. Soil AOA often have a low %GC content of < 40% (Lehtovirta-Morley et al. 2011; Jung et al. 2014; Zhalnina et al. 2014), with fully ^{13}C-enriched AOA DNA having a similar range of buoyant density as the numerically dominant unlabeled heterotrophs. These two factors likely resulted in the low representation of active nitrifiers' sequences (Howe et al. 2014).

To increase the resolution of ^{13}C-labeled autotrophic populations, contigs within specific ranges of %GC content (i.e. %GC < 50 and 50 < %GC < 63) were analyzed separately. Although 33 viruses were predicted from the 50 < %GC < 63 range, no nitrifier-associated virus was found. In comparison, the %GC < 50 population was less diverse and mostly enriched in thaumarchaeal sequences. All eight of the predicted viruses from the %GC < 50 contigs were predicted to be associated with AOA hosts including those related to *Candidatus* Nitrososphaera evergladensis, *Ca*. Nitrosotenuis chungbukensis, *Ca*. Nitrosotalea devanaterra and other *Thaumarchaeota.* These related hosts were all related to strains isolated from soil environments. For example, the strain *Ca*. N. evergladensis was derived from a highly enriched culture obtained from an agricultural soil (Zhalnina et al. 2014). Strain *Ca*. N. chungbukensis, was derived from a deep oligotrophic soil horizon (Jung et al. 2014), and the obligate acidophilic Ca. *N. devanaterra* was isolated from the

pH 4.5 soil of the pH gradient used in this study (Lehtovirta-Morley et al. 2011). In addition, of the 378 predicted viruses (contigs > 5 kb) from all contigs (i.e. without focusing on specific %GC ranges), seven were linked to nitrifier hosts. Predicted AOA hosts were to *Ca*. Nitrosotenuis cloacae, a strain derived from a wastewater treatment plant (Li et al. 2016), *Ca*. Nitrosopelagicus brevis, a strain derived from the open ocean (Santoro et al. 2015) and *Nitrososphaera viennensis*, a strain isolated from a garden soil (Tourna et al. 2011). The NOB-annotated predicted hosts were related to *Nitrobacter hamburgensis* and *Nitrobacter winogradskyi,* both of which were isolated from soil (Bock et al. 1983, 1990). Most of the AOA-associated viruses were closely related and did not cluster together with any other known viruses. However, one virus was most closely related to a *Sulfolobus* virus isolated from an acidic hot spring (Xiang et al. 2005). This suggests that these viruses are a novel clade of AOA-infecting viruses, which are evolutionarily distinct from other previous archaeal viruses that have been mostly isolated from hyperthermophilic or hyperhalophilic environments where *Crenarchaeota* or *Euryarchaea* dominate (Snyder et al. 2015; López-Pérez et al. 2019). Most of the AOA-associated viruses had a similar %GC content to that of their host, with most having lower %GC compared to their host. Viruses and hosts typically have similar %GC content (Rocha and Danchin 2002). However, viruses could have %GC contents different than their hosts due to horizontal gene transfer (HGT) from other organisms (Rocha and Danchin 2002). The %GC analyses may also be limited due to the virus and host genomic fragments being partial and incomplete.

Comparison of genes in predicted hosts- and viruses were investigated and gene sharing identified. Viral genes recently transferred from AOA and NOB hosts were homologous to genes encoding for proteins that may be used for viral genome replication and the lytic cycle. These proteins included SNase-like gene (*Staphylococcus aureus* nuclease, cleavage of extracellular DNA)(Kiedrowski et al. 2014), a dimeric alpha-beta barrel (Bussiere et al. 1998), a lambda repressor-like, DNA-binding domain (regulation of transcription of cl or Cro protein allowing the determination of the life cycle of lambda phages) (Burz et al. 1994; Hernandez-Doria and Sperandio 2018), and an integrase-like, catalytic domain (establishment of lysogeny) (Fogg et al. 2014). Interestingly, an iron-sulfur cluster assembly protein (Justino et al. 2006; Ayala-Castro et al. 2008) was found to be homologous between the acidophilic AOA host, *Ca*. N. devanaterra, and an associated virus, and was only found in the pH 4.5 soil. Iron-sulfur clusters are known to participate in a wide array of essential physiological pathways, including electron transfer, catalysis and regulatory processes (Rouault and Tong 2005; Rouault 2012). It has been proposed that the Fe-S clusters may be involved in the protection of aerobic and thermo-acidophilic archaea growing at extremely low pH by scavenging reactive oxygen species and by adjusting their intracellular pH to an acceptable value against a large proton gradient (Iwasaki 2010). This

finding suggests a potential function of viruses linked to electron transfer, and contribution to the adaptation of host cells in acidic environments (Hurwitz et al. 2015).

The nitrifier-associated viruses also had gene homologs present in the genomes of other previously characterized nitrifiers. For example, gene homologs between AOA-associated viruses and non-pH gradient-associated AOA had a higher similarity (88 - 97%) to the genes linked with *Ca*. N. devanaterra, which was derived from the acidic soil used in this study (Lehtovirta-Morley et al. 2011). Although most of homologous genes were annotated as uncharacterized proteins, one included a metal-sulfur cluster biosynthetic enzyme. This finding may suggest that there was a recent gene transfer between this virus and another host (Ku and Martin 2016). Confidence in the host-virus linkages were verified by the presence of at least one homologous gene originating from their associated host or to that of the same genera as the host within a database.

Associated viral genes that may help to prevent infection by other invading viruses were identified in the NOB-associated viruses. These included genes encoding for the toxin-antitoxin system (antiviral defense system) and putative phage cell wall peptidase (cleavage of the amide bonds between amino acids in peptidoglycan) (Holtje et al. 1995; Anantharaman et al. 2003). In addition, an AOA-associated virus had a gene which encoded for N-6-adenine-methyltransferase, a hallmark of viral evasion against the native host antiviral immune system (i.e. restriction modification) (Dupuis et al. 2013; Koonin and Krupovic 2020). These findings are indicative of active interactions between host and virus (Labrie et al. 2010; Bezuidt et al. 2020). Viral AMGs belonging glycoside hydrolase (GH) families and peptidases were also identified in the predicted viral contigs. These AMGs are common in soil, suggesting the potential impact of soil viruses in carbon cycling (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019). Unlike in marine environments, where the viral *amoC* gene was found to be widespread, this AMG was not found in any of our soil viral contigs (Ahlgren et al., 2019). However, this may be due to the low number of nitrifier-associated viruses that were recovered, and with only partial genomes.

Unlike in Chapter IV, the analysis of MAGs was unsuccessful. The binning of contigs into MAGs yielded only 11 MAGS, with only one of relatively low completeness (66.5%) identified as a an AOA (*Nitrososphaera*-like). All other MAGs were equally present in both the ^{12}C- and ^{13}C-samples, indicating that ^{13}C-CO_2 was not assimilated into their biomass and were all classified as heterotrophic bacteria. The low recovery of the ^{13}C-labeled MAGs was therefore due to the significant overlap between the unlabeled heterotrophic bacteria with the ^{13}C-labeled nitrifiers and the maintenance of a very high level of prokaryotic diversity that prevented the assembly of MAGs. Furthermore, no CRISPR arrays were identified in MAGs preventing this analysis as an approach for examining host-virus linkages.

5.6. Conclusion

Despite the inability to fully separate unlabeled heterotrophic DNA from ^{13}C-enriched DNA of autotrophic nitrifiers, this study demonstrated that targeting specific functional groups via stable isotope probing did facilitate the identification of putative host-virus populations with potential carbon flow between active nitrifying hosts and their lytic viruses in two contrasting pH soils. Nitrifier-associated viruses, including 11 AOA and three NOB, were discovered. Nitrifier-associated viruses appeared to exhibit different abundance patterns between the two soil pH, reflecting host pH preference. Evidence of dynamic nitrifier-virus interactions was found, including potential recent gene transfer between a virus and its associated nitrifying host.

CHAPTER VI

General discussion:
Host-virus interactions in soil

6.1. Overview

Microorganisms play a central role in soil where they are involved in a vast range of processes that facilitate biogeochemical cycling (Prosser and Nicol 2008; Chaparro et al. 2012; Aislabie and Deslippe 2013; Philippot et al. 2013). Most research focuses on understanding the 'bottom-up' factors that control microbial community structure and activity (e.g. substrate availability, abiotic factors, etc.), but comparatively little is known about the top-down controllers of diversity and abundance, particularly with respect to viruses. Viral infection has important implications for microbial community structure and function in ecosystems (Weinbauer and Rassoulzadegan 2004; Suttle 2005; Breitbart et al. 2007; Bertilsson et al. 2013). Host-associated viruses can influence the population size of their host(s) and their rates of performing enzymatic processes via the lysis of host cells (i.e. decreasing the rates of the function) or by providing (or augmenting) enzymatic processes via the provision of virus-encoded auxiliary metabolic genes (AMGs), respectively. Recent studies have discovered a large number of AMGs involved in carbon degradation, demonstrating the potential impact of soil viruses on ecosystem carbon processing (Emerson et al. 2018; Trubl et al. 2018, 2020; Graham et al. 2019).

The overall aim of this thesis was to gain insights into soil virus-host interactions at both the community and individual scale *in situ*. In Chapter II, the impact of an abiotic physicochemical gradient (soil pH) on the distribution of microbial hosts and subsequently on viral community structure was investigated using total community and virus-targeted metagenomics. In Chapter III, using a culture-based approach with a single host (*Bacillus* sp. S4), a plaque assay approach was combined with metagenomics to investigate how co-localization of soil host-virus populations effect the diversity of infecting viruses compared to viruses from a different niche (i.e. different soil pH). Finally, to focus on soil viruses infecting prokaryotes with central roles in carbon (C) and nitrogen (N) cycling, a DNA-SIP approach was combined with metagenomics to identify active viruses of hosts using CH_4-derived C (Chapter IV), and autotrophic hosts transforming inorganic N by following the incorporation of CO_2-derived C (Chapter V). Collectively the work presented in this thesis highlights the challenges encountered with the use of metagenomics for the study of soil viruses, but also provides insights on virus-host dynamics and host ranges through linking viruses and hosts via CRISPR array and oligonucleotide frequency (ONF) analysis, and analyses of gene homology to uncover AMGs and horizontal gene transfer. Finally, it demonstrates that using stable carbon isotopes allows detailed, high resolution analysis of active interactions *in situ* by following transfer of carbon from host to virus.

6.2. Challenges in soil virus metagenomics

As microbial diversity is vast within soil and the size of prokaryote genomes are larger than that of viruses, deep sequencing of soil metagenomes is required to capture both microbial and viral

diversity (Papudeshi et al. 2017; Maurier et al. 2019). Additionally, deep sequencing of soil viral filtrates (i.e. viromes) can aid in the recovery of viral sequences. With support from the Department of Energy's Joint Genome Institute (JGI), high throughput sequencing using the Illumina NovaSeq platform was utilized to determine the soil prokaryotic and viral diversity in a series of experiments using soils from a long term pH gradient (pH 4.5 and 7.5) in Aberdeen, Scotland. The NovaSeq sequencing generated 1.6 TB of sequencing data and 4.7 billion high quality reads from 26 metagenomes and 6 viromes. Considering the metagenomic results from Chapters II – V, the number of metagenomic viral contigs (mVCs) and metagenomic assembled genomes (MAGs) recovered increased with decreasing microbial community diversity. Analyzing whole datasets generated by NovaSeq sequencing required a large amount of RAM and computational time. For example, the co-assembly of the six viromes (334 GB) using 384 GB of RAM, 32 CPU, and > 4 TB of storage space for output files took 7 days to complete. Soil viral diversity studies are challenging as deep sequencing and high computational power are basic requirements. However, technological advances within these areas evolve quickly allowing for increased opportunities at reduced costs.

Another challenge is identifying viral signals within metagenomes that contain both viral and host contigs derived from a vast diversity of organisms. In this thesis, two virus prediction tools were routinely used: gene-based similarity approaches (VirSorter)(Roux et al. 2015) and a deep machine learning tool based on *k*-mer frequencies (DeepVirFinder) (Ren et al. 2017, 2018). Although the first approach can give high confidence predictions by identifying the presence of viral hallmark genes, the predictions are also constrained by the limitations imposed by alignment biases, including the lack of previously identified viral hallmark genes, the fractionation of viral genomes resulting in genome portions without hallmark genes, and low sequence similarity with poorly represented or uncharacterized reference viral genes. The *k*-mer frequency-based approach using DeepVirFinder provides alignment free methods and is notably advantageous for short viral contigs (> 300 bp) compared to VirSorter which is limited to the analysis of contigs that have at least three coding genes, and is recommended for use with contigs >10 kb. However, while the use of DeepVirFinder would be clearly beneficial for the analysis of smaller contigs, predicted viral contigs often appeared to be derived from bacterial genomes. Thus, VirSorter was preferred for analyses performed in this thesis, and when DeepVirFinder was used for short viral contigs, the genes of the predicted viral contigs were carefully verified. However, the use of both tools allowed the discovery of a large number of novel viruses from the soil metagenomes (Figure 6.1).

Overall, deep high-throughput sequencing and virus prediction tools enabled the identification of 8,801 virome viral contigs (VCs) > 10 kb, 1,070 DNA-SIP-derived metagenomic viral contigs (mVC) > 5 kb, and 45 prokaryotic metagenome-assembled genomes (MAGs) which

surpassed a completeness threshold of 50% with less than 10% contamination. The mVCs from total soil metagenomes (i.e. with no virus-enrichment prior to sequencing) could only be predicted using DeepVirFinder due to their short length, with an average length of 531 bp, and contrasted with VCs which had an average length of 21 kb. The short mVC size was due to the low proportion of viral sequences within the metagenomes due to the vast diversity and comparatively large genome size of prokaryotes in soil (Papudeshi et al. 2017; Maurier et al. 2019). Virus prediction between metagenomes and viromes also revealed size selection methodology biases. Specifically, the metagenomes included predicted VCs of large viruses (> 0.2 μm) belonging to the *Mimiviridae* and *Phycodnaviridae* (Maruyama and Ueki 2016) which were not present in the size-selected viromes. Therefore, the utilization of both metagenomes and viromes for capturing soil viral diversity is generally required.

In Chapters IV and V, DNA stable isotope probing (SIP) was utilized to target specific microbial groups, decreasing microbial diversity through the selection of high buoyant density DNA for specific ^{13}C-enriched populations. The more diverse CO_2-SIP derived metagenomes resulted in 420 mVCs and 9 MAGs compared to the CH_4-SIP derived metagenomes with 650 mVCs and 23 MAGs. The DNA-SIP approach was largely successful allowing for the investigation of individual virus-host interactions of key soil microbial functional groups. The CH_4-SIP metagenomes that were enriched in ^{13}C were all active primary or secondary utilizers of CH_4, with 12 methanotrophic and 6 methylotrophic MAGs. However, compared to the CH_4-SIP metagenomes, the CO_2-SIP metagenomes generated only one nitrifier MAG, with relatively low completeness (66.5%), due to the high microbial complexity remaining in the 'enriched' DNA as a result of the overlap in the buoyant density and distribution of genomic DNA from unlabeled heterotrophic bacteria and low %GC ^{13}C-labeled nitrifier DNA (Howe et al. 2014). An alternative approach would be to perform DNA-SIP with virome-enriched (i.e. filtered) DNA extracts. While this would also include unenriched heterotrophic viruses, it would remove the dominating prokaryotic genomes which resulted in the relatively low recovery of nitrifier and nitrifier-associated viral metagenomes. However, the approach used had the theoretical advantage of not only identifying viruses that have been recently produced in host cells, but also linking them to ^{13}C-enriched hosts. Nevertheless, the approaches used in the thesis provided direct evidence of contemporary virus-host activity in soil.

As with other soil metagenomic studies, the analyses presented expanded our knowledge of soil viruses and highlighted that the majority of sequences (> 90%) have no significant similarity to reference databases (Figure 6.1), and is referred to as "viral dark matter" (Roux et al. 2015). Of the annotated viruses recovered in this study, prokaryotic viruses presented the largest component (94% Caudovirales), consistent with the conclusion that tailed bacteriophages are the dominant virus component in soil habitats (Zablocki et al. 2014). Viral gene-based network

analysis can be used to infer prokaryotic virus relatedness by clustering viral protein sequences to overcome issues associated with low sequence similarity between viral genomes and the lack of a universal marker gene common to all viral genomes that would allow for taxonomic discrimination (Edwards and Rohwer 2005; Bolduc et al. 2017; Bin Jang et al. 2019). Clustering of the pH 4.5 and 7.5 VCs with the co-assembled VCs demonstrated that co-assembly appears to recover those viral contigs obtained from individual assemblies (i.e. individual pH 4.5 and 7.5 VCs) (Figure 6.1). Only 69 viral clusters containing the VCs were shown to be associated with known reference viruses that were commonly found in soils, such as *Arthrobacter*, *Bacillus*, *Burkholderia*, *Flavobacterium*, *Pseudomonas*, *Ralstonia*, *Rhodococcus* and *Vibrio* phage. Several singletons and clusters containing only pH gradient origin viruses were not linked to any reference viral genomes, indicating that these soil viruses are mostly novel (Figure 6.1). Although most of the CO_2-SIP derived mVCs were clustered together into small clusters, the CH_4-SIP derived mVCs were found to be grouped into relatively large clusters. This likely reflects the greater diversity of host potential in the CO_2- than CH_4-SIP communities, resulting in diverse small unique viral clusters verses large clusters of viruses that infect similar host cells. On average, 11.5% of all SIP derived mVCs (56 (13%) CO_2-SIP derived mVCs, 77 (10%) CH_4-SIP derived mVCs) were grouped with VCs, demonstrating the presence of active viruses or closely related viruses to those in the soil viromes. The recapitulative network also demonstrates pH niche preferences of viruses with shared genes within the related viruses specific to soil pH. With the increase in soil metagenomic studies and reference viral genomes, the taxonomic annotation of viral contigs will soon improve.

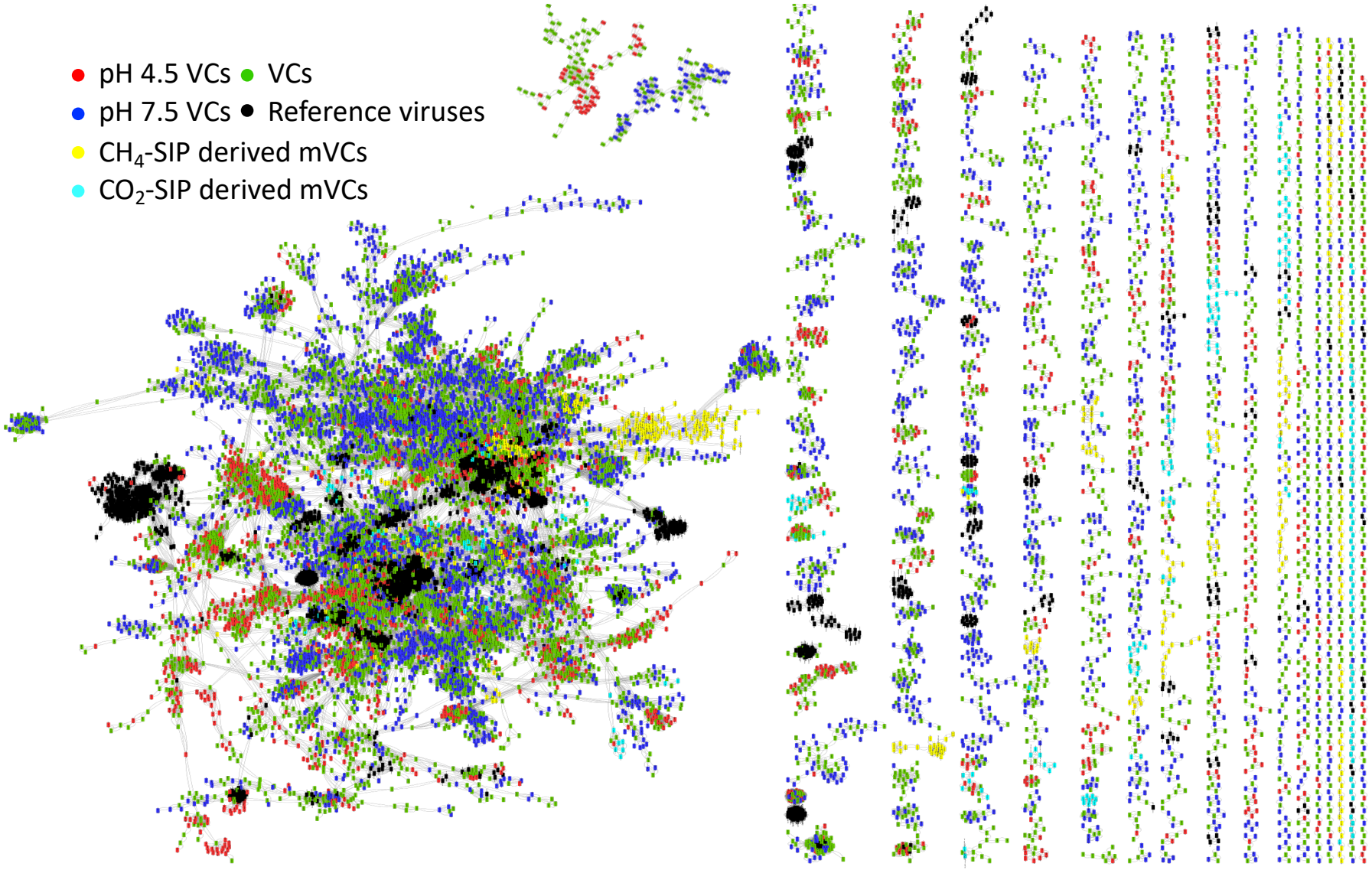


Figure 6.1. Network of shared predicted protein content among pH gradient co-assembled viromes (VCs, green), viral contigs from pH 4.5 (pH 4.5 VCs, red nodes) and pH 7.5 (pH 7.5 VCs, dark blue nodes), predicted metagenomic viral contigs from CH_4-SIP metagenomes (yellow nodes) and CO_2-SIP metagenomes (CO_2-SIP derived mVCs, light blue nodes), and RefSeq prokaryotic viral genomes (Reference viruses, black nodes). The network was produced using vConTACT (Bolduc et al. 2017).

6.3. Virus-host dynamics in soil

Virus-host interactions constitute a major determinant of host evolution and ecology (Mojica et al. 2009; Koskella 2014). During virus infection, the host cell may use antiviral defense mechanisms, such as the restriction-modification system (RM), the CRISPR-*Cas* systems, and abortive infections (Abi) (Deveau et al. 2010; Labrie et al. 2010; Stern and Sorek 2011; tenOever 2016). Among these, the CRISPR-*Cas* system can be used to infer direct evidence of virus-host interaction via the analysis of viral spacer sequences that are incorporated into host genomes. As CRISPR arrays provides a sequence 'memory' of virus infection, analysis of the spacer in CRISPR arrays from host can be used not only for host-virus linkages, but also for inferring virus characteristics (i.e. virus infection frequency and virus host-range) (Deveau et al. 2010; Fineran and Charpentier 2012; Strich and Chertow 2019). However, when analyzed as part of a metagenomic analysis, CRISPR arrays are often not found within MAGs or large host contigs due to a low completion level of MAGs and fragmentated host contigs, respectively. Also, although the CRISPR arrays can be analyzed, spacer sequences may not match to identified viral contigs due to limitations of viral genome assembly (e.g. incomplete sequencing due to high microbial diversity) or rapid evolution of viral sequences (i.e. mutation of loci of viral genomes represented in the spacer sequences or within the virus counter-defense mechanism itself). Oligonucleotide frequency (ONF) frequency analysis can be used as an alternative approach for host-virus linkage, especially for fragmented viral and host genomes that lack CRISPR arrays (Galiez et al. 2017).

In this thesis, CRISPR array and ONF analysis were used for host-virus linkage. For CRISPR array analysis, the CRT tool was used to identify CRISPR arrays directly from host contigs (Bland et al. 2007). However, as this approach can also recover false CRISPR arrays (including tandem repeats and STAR-like elements), the tool CRISPRCasFinder was also used in parallel to detect upstream *Cas* genes for confirming CRISPR arrays predicted via the CRT tool (Couvin et al. 2018).

The WIsH tool based on ONF analysis can predict hosts for short viral sequences (5 kb) and is reported to work with good accuracy (Galize et al. 2017). However, linkages may not be supported with further direct evidence, such as the presence of spacer sequences in a CRISPR array, or the presence of shared homologous genes. In the analysis of CRISPR-predicted host-virus linkages there was always at least one homologous gene shared between the virus and linked host. Therefore, if accurate, it would seem logical a linkage via gene homology would also be reflected in WIsH-predicted host-virus linkages and an additional gene homolog analysis was performed in order to validate the WIsH predicted host-virus linkages. For example, in the analysis of methane-fueled networks, of 245 WIsH predicted host-virus linkages, 157 shared at least one homologue. However, it is interesting to note that with CRISPR spacer-defined linkages (which would be considered 'high confidence'), WIsH did not predict the same linkage. Therefore, the use of ONF analyses should perhaps be considered with caution and require additional curation.

In associating identified host-virus linkages, a taxonomic annotation of whole contigs was used. Comparison with the NCBI-nr database was used to identify the possible taxonomic affiliation of genes in viral contigs, and subsequently compared to the taxonomic affiliation of the linked contig and putative host (previously annotated using the Kaiju tool). However, mis-annotation of host contigs was occasionally observed and misprediction of viral contigs was caused by a lack of related homologs in databases or a high proportion of horizontally acquired genes that are shared amongst closed related species, potentially leading to false validation criteria (Bolotin and Hershberg 2017).

Overall, the results presented in this thesis demonstrated dynamic virus-host interactions across the soil pH gradient. A relatively greater number and broader size range of CRISPR arrays were detected in the pH 4.5 soil compared to pH 7.5, suggesting contrasting virus infection frequencies across the pH gradient (Bezuidt et al. 2020). While the host-virus linkages determined by CRISPR arrays and WIsH-predicted associations were distinct between the two soils, they largely involved the same phyla. In addition, it appeared that the same host contigs were found to be infected by several different viruses, possibly leading to competition between viruses. In particular, evidence of virus-virus competition was found with the presence of genes encoding for the restriction endonuclease (REase) in most viruses (1,842 VCs, 20%) (Chapter III). REase in the viral genomes can confer resistance in hosts to infection by other phages that can also infect it (Lossouarn et al. 2019). Similarly, several REases were found in the soil viromes, suggesting that virus-virus competition might be common in soil environments. Furthermore, a number of methytransferases (MTase) were also found (446 VCs, 5%), potentially interfering with the RM host antiviral defense system, therefore improving the infection efficiency of the viruses (Labrie et al. 2010; Koonin and Krupovic 2020; Bezuidt et al. 2020). Generally, MTase-encoding genes are found in 20% of all currently annotated bacteriophage genomes, suggesting an important role in virus-host interaction (Kaltz and Shykoff 1998; Murphy et al. 2013).

Host-virus associations of C and N-cycling prokaryotes were identified using both CRISPR array and ONF analysis. Specifically, viruses interacting with methanotrophs (6 mVCs via CRISPRs; 64 mVCs via WIsH), methylotrophs (7 mVCs via WIsH) and nitrifiers (4 AOA-associated and 3 NOB-associated mVCs via WIsH) were identified in the CH_4 and CO_2-SIP metagenomes, respectively. The identification of ^{13}C-enriched CRISPR-linked methanotroph-associated viruses demonstrated active interaction within the soil system through the presence of spacer sequences in MAGs from *Methylosinus* and *Methylocystis* sp. Different host range viruses were shown with two of the viruses that were both found to infect two *Methylocystis* hosts, whereas other viruses were specific to a single host. Also, one virus appeared to infect more frequently than other viruses due to the greater number of spacers found in CRISPR arrays. However, analysis of MAGs derived from metagenomes of either native soil (Chapter II) or CO_2-SIP (Chapter V) metagenomes did not

contain any identifiable CRISPR arrays, and therefore this approach could not be used to provide linkage to viruses. While the WIsH analysis revealed a large number of methanotroph and methylotroph-associated viruses (64 and 7 mVCs, respectively) from CH_4-SIP metagenomes, only seven nitrifier-associated viruses were recovered. The selection of exclusively ^{13}C-enriched populations to reduce microbial diversity with the use of DeepVirFinder tool, helped to uncover eight more nitrifier-associated viruses, but these were derived from a very small fraction of viral genomes. The abundant methanotroph-associated mVCs in the CH_4-SIP metagenomes might reflect the abundance of methanotroph communities enriched in these metagenomes.

Abundance profiling demonstrated the habitat specificity of both hosts and viruses along a pH gradient, as shown in other previous studies (Nicol et al. 2008; Adriaenssens et al. 2017). However, unexpectedly, comparison of community structures of both viruses and prokaryotic hosts between pH 4.5 and 7.5 soils indicated that the viral community structure was comparatively more distinct, suggesting viruses do not have the same ranges as their hosts, and while constrained by the host community structure, soil pH itself may directly effect viral populations.

6.4. Virus host-ranges

In the analyses performed in this thesis, most interactions appeared to be relatively specific. In the analysis of viruses associated with a *Bacillus* strain (Chapter III), although there was a diverse range of infecting viruses (genetic and morphological), comparison with reference viruses demonstrated that they were all related to other *Bacillus* viruses and therefore likely represented *Bacillus*-specific viruses. In the analysis of viruses infecting two closely-related strains of *Methylocystis*, three viruses were specific to one strain, with two viruses infecting both. It may be that narrow-host specificity is common in soil. For example, in an analysis of phages infecting *Rhizobium* in rhizosphere soil, (Santamaría et al. 2014) it was demonstrated that they exhibited a narrow host range for infecting 48 *Rhizobium* sp. tested. However, the method of determining host range is limited by the fact that not every host will generate plaques and there is no standard for the number of strains or species to be tested (Ross et al. 2016). Generally, it is thought that narrow host-range viruses are prevalent when their host are abundant, whereas broad host-range viruses are assumed to infect low abundant hosts (Woolhouse et al. 2001; Sullivan et al. 2003; Elena et al. 2009; Dekel-Bird et al. 2015; Doron et al. 2016). This might be the same for soil viruses, however, as metagenomic datasets often represent the most abundant organisms in a sample (Rodriguez-R and Konstantinidis 2014), analysis of host-virus linkages is performed using the abundant organisms, narrow host-range viruses are likely selected. In addition, the rapid evolution of the spacer sequences within viral genomes can result in no matching of spacers to viral genomes which may reduce the ability to determine linkage (Koonin and Krupovic 2020). However,

genomic analysis of CRISPR-linked methanotroph-viruses demonstrated that the number of homologous genes between virus and host seems to be related to their host range specificity. Narrow-host viruses may share higher number of homologous genes to their specific ancestral methanotrophic host cells, allowing the narrow specificity of the virus while broad range specificity viruses seem to contain a variety of genes homologous to those found in other methanotrophs of different strains or genera level. However, as both viral and host genomes were not complete in our analyses and host-range was defined through bioinformatic evidence (e.g. the presence of the spacers from one virus in different hosts), it is unclear how well this hypothesis is supported from the evidence presented here. Future work could involve culture-based experiments to characterize broad- and narrow-host range viruses together with analysis of complete genomes of both host and virus to gain a better understand of the genomic features defining host range.

6.5. Auxiliary metabolic genes

Soil viruses contribute to the biogeochemical cycling by augmenting a host's metabolic potential after viral infection via the expression of virus-encoded AMGs (Breitbart et al. 2007). Host-derived AMGs can be acquired from their immediate previous host or from more distantly-related/ancestral hosts (Sharon et al. 2009; Sullivan et al. 2010; Kelly et al. 2013; Crummett et al. 2016). Analysis of the viromes and SIP-derived viruses identified a large number of AMGs involved in carbon metabolism, such as glycoside hydrolases and peptidases, which have been observed previously in soil viruses (Emerson et al. 2018; Trubl et al. 2018; Graham et al. 2019). In this study, methanotroph-associated viruses contained AMGs encoding for proteins involved in methane oxidation (*pmoC* and cytochrome C-related genes). Recently, several virus-encoded *pmoC* genes from freshwater lakes were identified (Chen et al. 2020) with transcriptomic data demonstrating their potential activity and a possible role for methanotroph-viruses modulating the efflux of CH_4 through changing methane oxidation rates of methanotrophs infected by *pmoC*-carrying phages (Chen et al. 2020). Similarly, a recent study reported that *amoC* genes associated with *Thaumarchaeota* are widely distributed in viruses from marine environments (Roux et al. 2016; López-Pérez et al. 2019; Ahlgren et al. 2019). Although virus-encoded *pmoC* and *amoC* genes were found to be abundant in aquatic environments, only one virus-encoded *pmoC* (and no virus-encoded *amoC*) were found in this study. Greater sequencing depth or approaches increasing the proportion of nitrifier-associated viruses (e.g. sequencing of individual fractions in DNA-SIP that were known to contain relatively high amounts of *amo* genes) may facilitate obtaining complete methanotroph- or nitrifier-associated viruses. Although viruses infecting methanotroph and nitrifiers were quite distinct, it is interesting that viruses of each functional group preferentially acquire *pmoC* or *amoC* genes (rather than *pmo*/*amoA* or *pmo*/*amoB*) to

enhance the fitness of their hosts. Potentially, methanotroph- and nitrifier-associated viruses could influence the rates of methane oxidation and nitrification in soil via the lysing of hosts or the expression of the virus-encoded AMGs (e.g. *amoC* and *pmoC* genes), respectively.

Generally, sequence identity between donor and recipients in horizontal gene transfer is initially 100%, but gradually decreases over time due to genetic change, such as mutations, gene duplications and genomic rearrangements (Ochman et al. 2000; Raz and Tannenbaum 2010; Ku and Martin 2016). Thus, recent gene transfers tend to have a higher degree of identity to homologs from the donor lineage (Shoemaker et al. 2001; Smillie et al. 2011; Ku and Martin 2016). Homologues shared between a host and virus generally had a higher identity when they encoded for proteins involved in virus replication, suggesting more recent horizontal gene transfer. Homologues encoding for other functions were generally more variable (30 – 90%).

6.6. Critique of experiments and future work

Between the two soil of contrasting pH, viral diversity and richness were relatively greater at pH 7.5 soil. However, it is important to consider that the data presented in this thesis does not reflect the complete viral diversity. Both virus recovery and sequencing methodologies used only detect DNA viruses and exclude the detection of RNA viruses. The procedures used would also likely be biased towards those viruses that are easily extracted, and virus-organic matter interactions will likely vary with a changing soil pH (Dowd et al. 1998; Lukasik et al. 2000; Chu et al. 2003; Zhao et al. 2008; Chen et al. 2014).

A plaque assay approach combined with metagenomics recovered evidence of the effects of co-evolutionary interactions between host and virus populations, such as the presence of host antiviral mechanisms (e.g. RM and CRISPR-*Cas*-system) and virus counterdefense mechanisms (e.g. mutation of the spacers, presence of the methyltransferases). Also, the diversity and infectivity of virus populations (e.g. number of PFUs) were found to be greater when virus populations were derived from a different soil niche. However, only one bacterial strain isolated from the pH 7.5 soil was used in the experiment. The plaque assay approach relies on cultivable bacterial strains, and those that are capable of forming confluent cell monolayers and are susceptible to virus infections. This leads to a limitation in the ability to analyze the greater uncultured majority. Using additional bacterial isolates of the same species and those of different genera to test whether there is a reciprocal observation for a bacterial strain isolated from the pH 4.5 soil is a future goal. The virus enrichment and bacterial isolation used a sieved composite (0.2 g and 1 g, respectively) of intact soils (50 g). Although the sieving of soil helps to reduce spatial heterogeneity by removing roots and stones and results in producing representative homogenous samples, it may reduce the multiple microhabitats harbored by microbes that are associated with soil aggregates. As soil aggregates are known as microhabitats that promote parallel microbial

evolution trajectory, this may represent a hotspot for virus-host interactions (Rilling et al. 2017; Pratama et al. 2018). It therefore would be more realistic to maintain the natural soil structure and compare viruses in individual aggregates at different distances and scales. Rhizosphere soil aggregates (e.g. root-adhering soil) would be an interesting model to examine viral populations that control soil microorganisms compared to those found in aggregates from bulk-soil (e.g. root-free soil aggregates).

DNA-SIP combined with deep sequencing resulted in the discovery of diverse ^{13}C-enriched active viruses, and allowed a relatively comprehensive analysis of a methane-fueled microbial-viral food web. While these analyses identified recent interaction at a single time-point, they did not examine rates of infection or actual changes in CRISPR arrays, nor temporal changes in population and associated virus numbers. Samples were only sequenced after 30 days of soil incubation, and not periodically, which does not allow conclusions to be drawn about the rate of CRISPR evolution. As such, a time-series across a longer time span would enable analysis into the rates of CRISPR evolution, and infection rates of viral populations associated with different trophic levels and cross-feeders could be examined. In the CH_4-SIP experiment, a high concentration of methane (10%) was used in comparison to atmospheric concentrations of methane 0.00017%. In the complex soil environment, in a well-drained soil, microsites with low concentrations of oxygen or anoxia (e.g. after rainfall) may allow conditions that result in methanogenesis and high methane concentrations in microsites. The experimental conditions used likely resulted in the growth low-affinity methanotrophs, and did allow examination of high-affinity methanotrophs that are important for removing large quantities of atmospheric methane (Kneif et al. 2005). The use of different methane concentrations would allow examination of other member of the methanotroph community. In the CO_2-SIP experiment, the obvious limitation was the lack of separation of fully $^{13}CO_2$-enriched DNA due to the overlap with unlabeled high %GC DNA from heterotrophic bacteria. Individually sequencing each DNA fraction and that of a viral extraction from the ^{13}C-enriched soil microcosms would overcome these limitations. Alternatively, nitrifier-associated viruses could be studied by using cultures of nitrifiers (AOA, AOB and NOB) using a culture-based approach.

6.7. Conclusion

This work expanded our knowledge on soil viruses. Viral community structures were tightly constrained by their prokaryotic hosts and viruses appeared to have narrow-host ranges. Following carbon flow from host to virus allowed high resolution analysis of virus communities and even allowed identification of individual interactions *in situ*. This approach may therefore be suitable for understanding the functional importance and dynamics of viruses associated with

many different stages of the C and N biogeochemical cycles in soil through the use of appropriate isotopically-enriched substates.

Synthèse en français

Résumé

Les virus du sol sont capables d'influencer la structure de la communauté microbienne et le fonctionnement de l'écosystème en affectant l'abondance des cellules hôtes par lyse et par leurs caractéristiques à transférer des gènes entre les hôtes. Bien que notre compréhension sur la diversité et la fonction virales s'améliore, la connaissance des rôles des virus et des interactions hôte-virus dans le sol reste limitée. Pour mieux comprendre les interactions virus-hôtes dans le sol, un système de sol avec un gradient de pH contrôlé sur le long terme a été utilisé dans lequel la structure de la communauté microbienne varie selon le gradient pH du sol. Les objectifs principaux de cette thèse étaient premièrement, de déterminer l'influence de la structure de la communauté microbienne et du pH du sol sur les virus en utilisant l'approche métagénomique et viromique (Chapitre II), puis en second lieu de déterminer l'infectiosité des populations virales natives (isolées à partir de niches de sol co-localisées et non-natives (isolées à partir de niches de sol différents) en utilisant une approche d'essai de plaque combinée à un séquençage hybride (Chapitre III) et dans un troisième temps, d'identifier les populations virales infectant des groupes fonctionnels microbiens spécifiques du sol, en particulier les méthanotrophes (Chapitre IV) et les nitrifiants (Chapitre V), à l'aide d'une sonde isotopique stable à l'ADN combinée à un séquençage métagénomique profond. Nos premiers résultats ont montré que la structure des communautés virales changeait en fonction du pH du sol, ce qui montre que les communautés virales sont étroitement liées aux populations hôtes, mais qu'elles peuvent aussi avoir des gammes d'hôtes étroites. L'analyse de CRISPRs a révélé des interactions dynamiques entre virus et hôtes, le nombre et la taille des CRISPRs étant distincts dans des sols au pH différent. Le profil des liens entre l'hôte et le virus entre le pH du sol suggère que les virus jouent un rôle essentiel dans la composition et la fonction de la communauté procaryote du sol. De manière surprenante, une plus grande infectivité d'une bactérie hôte par des populations de virus a été constatée lorsque les virus et la bactérie hôte n'étaient pas co-localisés dans le même sol de pH. Les processus de coévolution entre les populations de l'hôte et du virus, tels que la modification de restriction/méthyltransférase codée par le virus et la mutation du spacer dans le système CRISPR-*Cas*, fournissent la preuve d'une adaptation locale, et que les interactions virus-bactérie hôte jouent un rôle intégral dans la sensibilité d'un hôte à l'infection et, par conséquent, dans la régulation des populations bactériennes du sol. Le ciblage de groupes fonctionnels microbiens spécifiques par les isotopes stables a permis d'analyser des populations individuelles de virus hôtes. Le suivi du flux de carbone dans les populations procaryotes et virales a révélé des interactions actives entre les virus et les méthanotrophes et nitrifiants, ainsi que des préférences de niche en matière de pH du sol. Les preuves de transfert horizontal de gènes et de gènes métaboliques auxiliaires codés par les virus, tels que les familles de glycosides hydrolases, les peptidases, la sous-unité de la méthane monooxygénase particulaire (*pmoC*), la nitrogénase (*nifH*) et le cytochrome cd1-nitrite réductase, confirment que les virus contribuent de manière significative au fonctionnement de l'hôte et au cycle du carbone et de l'azote dans le sol. Dans l'ensemble, ces travaux ont démontré que les virus du sol sont d'importants régulateurs des communautés microbiennes par la lyse spécifique de l'hôte et aux interactions hôte-virus dynamiques.

Introduction générale :
Les virus, les interactions hôte-virus et l'approche métagénomique pour comprendre leur écologie

1.1. Contexte

Les virus font partie intégrante de tout environnement et peuvent infecter toutes cellules des organismes de tous les domaines du monde vivant (archées, bactéries, et eucaryotes) (Clokie et al. 2011). Les virus marins sont bien connus pour être des acteurs majeurs dans la régulation de la structure des communautés microbiennes marines et du fonctionnement des océans, car ils affectent directement l'abondance des cellules hôtes par lyse qui constitue la base de tous les réseaux trophiques dans l'océan et par leur capacité à transférer des gènes entre les hôtes favorisant les échanges génétiques au sein des communautés microbiennes marines (Wommack et Colwell 2000 ; Weinbauer et al. 2003 ; Weinbauer et Rassoulzadegan 2004 ; Suttle 2005).

Dans les océans, les virus sont responsables de la mortalité d'environ un tiers des cellules procaryotes par jour, et modifient ainsi les cycles biogéochimiques (carbone) par un relargage important de matière organique (Fuhrman 1999). En outre, les virus marins peuvent avoir un impact sur la productivité par l'expression des gènes auxiliaires de métabolisme codés par le virus (AMG) (Breitbart et al. 2007). Par rapport à l'environnement marin relativement homogène, les sols sont considérés comme des habitats plus diversifiés aux micro-organismes en raison de la grande variation de leur composition, hétérogénéité spatiale et propriétés physicochimiques (Williamson et al. 2017). Dans un gramme de sol, il y a jusqu'à 10 milliards de procaryotes (Raynaud et Nunan 2014). De même, les virus présents dans le sol peuvent être aussi abondants et divers, et donc potentiellement jouer un rôle important en influençant la structure des communautés microbiennes et les fonctions des écosystèmes (Williamson et al. 2017). Bien que notre compréhension sur la diversité et le fonctionnement des virus dans le sol progressent, les connaissances sur les interactions hôtes-virus dans le sol restent limitées (Han et al. 2017 ; Trubl et al. 2018 ; Graham et al. 2019 ; Emerson 2019). Ce chapitre présente la biologie et les interactions des virus avec les procaryotes, l'état actuel des connaissances sur les virus du sol, les approches métagénomiques utilisées pour étudier les communautés virales et le système modèle de gradient de pH du sol utilisé pour les expérimentations ainsi que les objectifs spécifiques de la recherche.

1.1.1. Biologie des virus et cycle de vie des virus

Les virus sont des parasites intracellulaires obligatoires, ce qui signifie qu'ils sont dépourvus de métabolisme intrinsèque qui dépendent entièrement de leur hôte vivant pour se multiplier (Gelderblom 1996). Tous les virus sont constitués d'une information génétique (acide nucléique) entourée d'une structure complexe, la capside (King et al. 2011). Les capsides sont constituées d'une série d'éléments structuraux donc les protéines monomères identiques (capsomères) qui s'auto-assemblent. Ensemble, la capside et le génome composent le virion, généralement appelé la particule libre infectieuse (King et al. 2011). La majorité des capsides caractérisées ont une

symétrie cubique (icosaèdre) ou hélicoïdale. Cependant, certains virus, en particulier, ceux qui infectent des bactéries ou des archées, peuvent avoir des capsides plus complexes (King et al. 2011). Les génomes des virus sont soit à ADN, soit à ARN (Gelderblom 1996 ; King et al. 2011). Les virus à ADN peuvent être soit simple brin (ADN sb), soit double brin (ADN db) et être linéaires ou circulaires. Les virus à ARN peuvent également être simple brin (ARN ss) ou double brin (ARN db) et segmentés ou non segmentés (Lodish et al. 2000). Les virus qui infectent les bactéries sont généralement appelés bactériophages (ou phages), alors que les virus qui infectent les archées sont appelés les virus d'archées. Selon l'International Committee on Taxonomy of Viruses (ICTV), les virus sont taxonomiquement classés en ordres, familles, sous-familles, genres, espèces, isolats et variantes (Kuhn et al. 2010).

Il existe deux cycles de multiplication, un cycle lytique et un cycle lysogène (figure 1.1). Tout d'abord, le virus s'attache à la surface de la cellule hôte par une liaison spécifique entre les glycoprotéines de surface du virus et des molécules réceptrices spécifiques de l'hôte (Aswad et Katzourakis 2018). Le virus insère ensuite son matériel génétique dans la cellule hôte par une endocytose ou d'autres mécanismes. Les enzymes du virus ou de l'hôte dégradent la capside virale. Dans le cycle lytique, le génome viral induit la synthèse des constituants viraux, l'assemblage de nouvelles particules virales et la lyse de la cellule infectée, libérant ainsi de nouveaux virus qui se répandront et infecteront d'autres cellules hôtes (Clokie et al. 2011). Dans le cycle lysogène, le génome du virus peut s'intégrer dans le génome de la cellule hôte (prophage), ou rester comme un plasmide (pseudo-lysogénie) (Weinbauer et al. 2003 ; Weinbauer et Rassoulzadegan 2004).
Il s'agit d'une forme latente, dans laquelle les gènes viraux sont présents dans l'hôte sans provoquer de perturbation de la cellule (Clokie et al. 2011). Les prophages sont répliqués et maintenus dans les générations suivantes jusqu'à ce qu'un stress environnemental déclenche un passage au cycle lytique (Weinbauer et al. 2003). La lysogénie est une stratégie efficace pour que les populations virales persistent lorsque l'abondance des cellules hôtes est faible (Williamson et al. 2002 ; Weinbauer et al. 2003 ; Mann 2003 ; Chibani-Chennoufi et al. 2004 ; Kimura et al. 2008) ou lorsque la survie de l'hôte dépend de périodes d'inactivité (Pantastico-Caldas et al. 1992 ; Kimura et al. 2008).

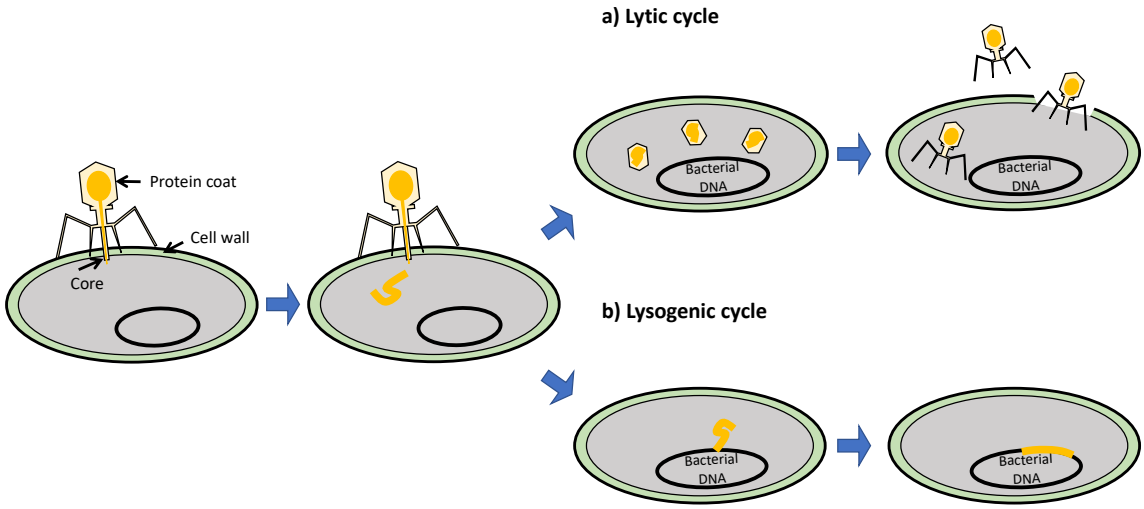


Figure 1.1. Représentation schématique a) le cycle lytique et b) le cycle lysogène. Le cycle lytique se termine par la libération des virus matures alors que le génome du virus est incorporé dans le génome hôte lors du cycle lysogène.

1.1.2. Abondance et diversité des virus infectant les procaryotes dans le sol

Il a été démontré que les virus infectant les procaryotes sont plus abondants que les procaryotes, et qu'ils constituent les entités biologiques les plus abondantes et les plus diverses de la biosphère (Fuhrman 1999 ; Williamson et al. 2013). Il a été estimé que la virosphère peut contenir jusqu'à ~10^{31} particules virales à l'aide du comptage direct de particules virales dans différents environnements (Edwards et Rohwer 2005 ; Breitbart et Rohwer 2005 ; Suttle 2005 ; Silveira et Rohwer 2016). Plusieurs études mesurant l'abondance virale dans les sols par microscopie électronique à transmission ou par microscopie à épifluorescence ont montré un grand nombre de particules virales allant de 10^7 à 10^{10} dans un gramme de sol (Williamson et al. 2003, 2017 ; Han et al. 2017). Dans une étude récente, l'analyse des particules virales dans quatre différents types de sol a révélé que l'abondance des particules virales était similaire, mais les différentes morphologies de virus ayant des abondances relatives différentes selon les types de sol ont été observées (Reavy et al. 2015). Il a été démontré que l'abondance virale et bactérienne augmente avec la productivité des écosystèmes, généralement la plus faible dans les sols secs et arides et la plus abondante dans les sols humides et riches en matière organique (Williamson et al. 2017). Cependant, malgré le fait que l'abondance virale varie entre des sols de composition et de localisation géographique différentes (Williamson et al. 2005), l'abondance virale des sols est relativement stable par rapport aux écosystèmes marins. Il a été démontré que l'abondance virale dans les environnements marins change plus de 2000 fois à travers la colonne d'eau (Srinivasiah

et al. 2008). Cependant, l'estimation de l'abondance virale dans le sol peut être biaisée par divers paramètres, notamment les méthodes d'extraction et la détection, mais aussi les caractéristiques du sol (Trubl et al. 2016 ; Williamson et al. 2017). L'abondance virale peut être largement sous-estimée en raison de la difficulté d'extraire tous les virus présents dans le sol. Le nombre total de particules virales libres dans le sol est probablement supérieur de 1 à 2 ordres de grandeur à celui des populations hôtes, et donc leur abondance relative par rapport au nombre de cellules procaryotes peut être comparable à celle des écosystèmes marins (Watt et al. 2006 ; Trubl et al. 2018).

Les virus peuvent être classés par morphologie à l'aide d'un microscope électronique à transmission qui est la technique la plus utilisée pour caractériser les virus en morphotypes (Ackerman et al. 1978 ; van Regenmortel et al. 2000). Les phages caudés constituent l'ordre des Caudovirales, représentant 95% de tous les phages et forment probablement la majorité des virus de la planète (Ackermann 1998 ; Maniloff et Ackermann 1998). L'ordre des Caudovirales a été sous-divisé en trois grandes familles : les Myoviridae, les Siphoviridae et les Podoviridae. Les Myoviridae ont une longue queue contractile, les Siphoviridae ont une longue queue non contractile et les Podoviridae ont une queue courte non contractile (Fauquet et al. 2005). L'analyse de la morphologie et des métagenomes viraux (viromes) ont démontré que les virus appartenant à l'ordre des Caudovirales dominent dans le sol (Zablocki et al. 2014 ; Ballaud et al. 2016 ; Han et al. 2017). D'après l'ICTV et la Virus-Host database, il existe actuellement 21 familles, 57 sous-familles et 797 genres de virus infectant les bactéries, et 19 familles et 24 genres de virus infectant les archées (Adriaenssens et al. 2020 ; Mihara et al. 2016). En ce qui concerne les virus cultivés, bien que les virus des archées soient sous-représentés, ils sont morphologiquement plus divers que les bactériophages (Pietilä et al. 2014 ; Snyder et al. 2015). L'analyse des viromes suggère que la diversité virale environnementale est vaste et que la majorité de la diversité virale n'a pas encore été caractérisée (Angly et al. 2006 ; Mokili et al. 2012 ; Roux et al. 2015b). Généralement, entre 60 et 99% des séquences virales obtenues n'ont aucun homologue dans les banques de données actuelles, et sont appelées "viral dark matter" (Brum et al. 2015 ; Roux et al. 2015b). Bien que la plupart des recherches se soient concentrées sur les virus à ADN db, des travaux récents ont également révélé la présence de virus à ADN sb circulaire distinct, et de virus à ARN divers et abondants dans le sol (Reavy et al. 2015 ; Starr et al. 2019).

1.1.3. Interactions virus-hôte

Les virus infectent tous les types de vie cellulaire présents dans le sol, y compris les eucaryotes et les procaryotes (Weinbauer et Rassoulzadegan 2004). Par le cycle lytique et le cycle lysogène, les virus interagissent avec leurs hôtes, ce qui engendre une régulation sur les communautés microbiennes et une modification des cycles biogéochimiques. Comme les virus provoquent la

lyse des cellules et la libération de protéines et d'acides nucléiques, ils peuvent jouer un rôle important dans le cycle du carbone, de l'azote, du soufre et du phosphore dans le sol (Williamson et al. 2017). Les virus du sol peuvent augmenter la quantité de carbone disponible, ce qui peut influencer la production et la respiration des microbes (Williamson et al. 2017). Le relargage important de matière organique par lyse des cellules, réutilisée par les micro-organismes fait partie des réseaux trophiques microbiens communément appelée "shunt viral" (Suttle 2005). Dans les systèmes marins, il a été estimé que jusqu'à 40% des bactéries marines sont lysées quotidiennement par les virus contribuant au shunt viral (Suttle, 2005). Bien que des rôles similaires aient été suggérés pour les virus du sol, des informations sur les contributions du shunt viral dans les réseaux trophiques microbiens du sol sont méconnues (Kuzyakov et Mason-Jones, 2018 ; Emerson et al. 2019).

Le transfert de matériel génétique entre les virus et les hôtes, puis entre les micro-organismes, a des conséquences importantes. Au cours du cycle lysogène, les prophages peuvent modifier le métabolisme et le phénotype de l'hôte, ce qui entraîne un changement de fitness, et l'expression des gènes des prophages peut même protéger leurs hôtes d'une infection d'autres phages (Williamson et al. 2017). Un autre rôle des virus est la transduction qui est une forme d'échange génétique entre les bactéries, médiée par un virus (Canchaya et al. 2003a). Dans les sols, il s'agit d'un mécanisme important de transfert de gènes, qui entraîne une diversification et une spéciation des hôtes (Wiedenbeck et Cohan 2011). La transduction entre les bactéries introduites et les phages a été observée dans les microcosmes du sol, bien qu'il n'ait pas encore été démontré que la transduction se produisait *in situ* entre les bactéries indigènes du sol, probablement en raison des difficultés techniques de détection des rares événements de transduction dans le sol (Elsas et al. 2003).

1.1.3.1. Interactions entre les bactéries et les bactériophages

L'étude des interactions bactériophage-hôte reste un défi car de nombreuses bactéries ne peuvent pas être cultivées et les techniques utilisées pour l'isolation et la caractérisation des bactériophages se limitent à celles associées aux organismes qui sont cultivés (de Jonge et al. 2019). D'après les tests d'infection, la gamme d'hôtes des bactériophages peut se situer sur un continuum, allant de large à extrêmement étroit (Ross et al. 2016). Par exemple, une étude antérieure a montré que les bactériophages isolés de *Vibrio parahaemolyticus* n'infectaient pas d'autres souches de cette espèce ou d'autres espèces de *Vibrio* (Weinbauer et Rassoulzadegan 2004). Cependant, des études sur les tests d'infection ont identifié les bactériophages infectant à la fois des souches multiples de la même espèce et des espèces multiples (Greene et Goldberg 1985 ; Vinod et al. 2006 ; Uchiyama et al. 2008 ; Gupta et Prasad 2011 ; Khan Mirzaei et Nilsson

2015 ; Yu et al. 2016). Par exemple, le bactériophage Mu pouvait infecter des espèces d'*Escherichia coli*, *Citrobacter freundii*, *Shigella sonnei*, *Enterobacter* et *Erwinia* (Ross et al. 2016).
Dans les sols, la compréhension des taux d'infectiosité et des interactions entre des virus et des principaux groupes microbiens fonctionnels, tels que les méthanotrophes et les nitrifiants qui régulent directement le cycle du carbone et de l'azote, restent méconnues. Cependant, quelques virus infectant les méthanotrophes ont été isolés dans plusieurs environnements (Tiutikov et al. 1976 ; Tyutikov et al. 1980, 1983) et une récente étude de métagénomique du sol a identifié des virus qui sont associés à des hôtes méthanotrophes (Emerson et al. 2018).

Sur les banques de données de génomes bactériens séquencés, on a estimé qu'environ 70% des génomes bactériens contiennent des prophages (Paul 2008). Il a été proposé que la lysogénie se produit à des moments où les nutriments sont limités et où la taille des populations d'hôtes est faible (Kimura et al. 2008). Lorsque les facteurs environnementaux deviennent favorables, le phage peut s'extraire du génome de l'hôte et entrer dans la voie lytique (Williamson et al. 2002 ; Kimura et al. 2008). Au cours de la lysogénie, une relation symbiotique entre les prophages et son hôte peut favoriser le fitness des prophages et de l'hôte en exprimant des gènes qui augmentent le fitness de la cellule hôte (Canchaya et al. 2003b). Ce processus est connu sous le nom de conversion lysogénique (van Houte et al. 2016). Par exemple, les phages tempérés peuvent influencer la colonisation des nodules racinaires, l'efficacité de la fixation de l'azote et la productivité des cultures en modifiant les phénotypes du rhizobium par conversion lysogénique (Kimura et al. 2008). Les hôtes et les phages développent aussi respectivement des mécanismes de défense antivirale et des mécanismes de contre-défense virale (Abedon 2012 ; Vasu et Nagaraja 2013). De nombreuses bactéries possèdent également des systèmes CRISPR « Clustered Regularly Interspaced Palindromic Repeats » ou « courtes répétitions palindromiques regroupées et régulièrement espacées », couplés aux protéines Cas comme défense adaptative contre les phages (Bhaya et al. 2011). Ces mécanismes sont présentés plus en détail dans la section 1.1.3.3.

1.1.3.2. Interactions entre les archées et les virus d'archées

La plupart des connaissances actuelles sur les virus d'archées sont fondées sur les extrêmophiles, les virus caractérisés infectant des Crenarchaeota hyperthermophiles ou des Euryarchaeota halophiles ou méthanogènes (Snyder et al. 2015 ; Albers 2016 ; Quemin et al. 2016).
Récemment, des génomes de virus d'archées marins tels que les virus infectant des Euryarchaeota et des Thaumarchaeota ont été assemblés (Uchiyama et al. 2008 ; Philosof et al. 2017 ; Prangishvili et al. 2017 ; López-Pérez et al. 2019 ; Ahlgren et al. 2019). Par exemple, le magrovirus infectant le groupe marin II Euryarcheota qui est omniprésent mais non cultivé, a été récemment découvert (Philosof et al. 2017). En outre, des virus infectant les archées oxydantes d'ammoniac (AOA) appartenant au phylum Thaumarchaeota, ont également été assemblés à partir d'échantillons

métagénomiques du système marin (López-Pérez et al. 2019 ; Ahlgren et al. 2019). Ces virus infectant les AOA sont liés aux membres de l'ordre des Caudovirales (Suttle 2005 ; Labonté et al. 2015 ; López-Pérez et al. 2019 ; Ahlgren et al. 2019). Dans certains génomes de virus infectant les AOA, l'AMG codant pour la sous-unité C de l'ammoniac monooxygénase (amoC) a été identifié, suggérant le rôle potentiel des virus infectant les AOA dans le processus de nitrification dans les océans (Roux et al. 2016 ; Ahlgren et al. 2019). En outre, des provirus ont été retrouvés dans le génome du Candidatus Thaumarchaeota marin *Nitrosomariuns catalina* SPOT01, et le génome de *Nitrososphera viennensis* EN76, le premier AOA du sol isolé (Krupovic et al. 2011 ; Ahlgren et al. 2019). Récemment, trois virus infectant le *Nitrosopumilus* en forme de fuseau (NSV) ont été isolés à partir d'échantillons d'eau de mer riches en particules en suspension avec une souche d'AOA hôte (Kim et al. 2019). Ces NSV ont une gamme d'hôtes étroite et présentent des taux d'absorption élevés sur leurs cellules hôtes, ce qui indique une infectiosité efficace de la souche hôte de *Nitrosopumilus* (Kim et al. 2019). Des études de métagénomique ont suggéré que les virus infectant le Thaumarchaeota ont un impact majeur sur le fonctionnement et la mortalité des AOA par lyse cellulaire, et donc sur la régulation des cycles de l'azote et du carbone (Danovaro et al. 2016 ; López-Pérez et al. 2019 ; Ahlgren et al. 2019). Néanmoins, la plupart des virus infectant les AOA marins n'ont pas encore été cultivés en raison de la difficulté à obtenir des cellules hôtes en culture pure. Actuellement, aucun virus infectant les AOA n'a été découvert dans le sol.

1.1.3.3. "Course à l'armement" entre les cellules hôtes et les virus

Les interactions hôte-virus au cours de la réplication du virus impliquent l'adsorption du virus sur les récepteurs cellulaires et l'entrée du génome viral dans la cellule hôte, ce qui entraîne une course à l'armement entre les cellules hôtes et les virus pour survivre à l'infection virale et une évolution subséquente des virus pour contrecarrer ces adaptations défensives (Golais et al. 2013). Pendant la réplication du virus, la cellule hôte peut utiliser divers mécanismes de défense antiviraux, notamment le système de restriction-modification, le système CRISPR-Cas et le système d'infection abortive (Deveau et al. 2010 ; Labrie et al. 2010 ; Stern et Sorek 2011 ; tenOever 2016). Le système de restriction-modification se trouve chez les procaryotes et fournit une défense contre l'ADN (ou l'ARN) étranger par l'action de l'endonucléase de restriction et des méthyltransférases (Tock et Dryden 2005 ; Vasu et Nagaraja 2013). Les endonucléases sont des protéines qui reconnaissent l'ADN étranger et le clivent à des séquences d'ADN palindromique spécifiques (c'est-à-dire des sites de restriction) (Vasu et Nagaraja 2013). Contrairement à l'ADN bactérien de l'hôte, l'ADN viral n'est pas méthylé, donc non protégé et clivé par les endonucléases (Wilson et Murray 1991). Par la suite, certains virus déploient des stratégies de contournement du système de restriction modification. Par exemple, l'incorporation de bases inhabituelles, comme le 5-hydroxyméthyluracile au lieu de la thymine, dans le génome viral peut modifier les

sites de restriction (Krüger et Bickle 1983). Certains phages peuvent coder pour leur propre méthyltransférase, inciter à la production de la méthyltransférase de l'hôte ou posséder des gènes codant pour des protéines qui se lient aux sites de restriction ou imiter les protéines qui peuvent neutraliser l'action des endonucléases (Stern et Sorek 2011 ; Golais et al. 2013).

La quasi-totalité des archées et environ la moitié des bactéries possèdent des CRISPR systèmes couplés aux protéines Cas (codées par les gènes *cas*) qui sont la base du système immunitaire adaptatif CRISPR-Cas des procaryotes (Figure 1.2) (Jansen et al. 2002 ; Terns et Terns 2011). Les cellules hôtes acquièrent continuellement les séquences d'espacement « spacers » des virus pour faciliter la reconnaissance et éviter une future infection virale. Par la suite, les virus peuvent muter la séquence d'espacement ciblée ou phosphoryler les protéines Cas pour échapper aux systèmes CRISPR-Cas (Horvath et Barrangou 2010 ; Golais et al. 2013). Inversement, les répétitions CRISPR et les protéines Cas évoluent pour échapper à un mécanisme d'arrêt du système CRISPR codé par les virus (Wang et al. 2020). Ainsi, le virus et son hôte sont enfermés dans une course à l'armement. La coévolution entre le virus et son hôte dans le même habitat se produit sans cesse et constitue le régulateur clé des processus écologiques et évolutifs des communautés microbiennes (Koskella et Brockhurst 2014). La course à l'armement peut avoir des conséquences évolutives à long terme sur la population hôte, et la lysogénie qui est considéré comme un état compromis peut se produire (Golais et al. 2013 ; Koskella et Brockhurst 2014).

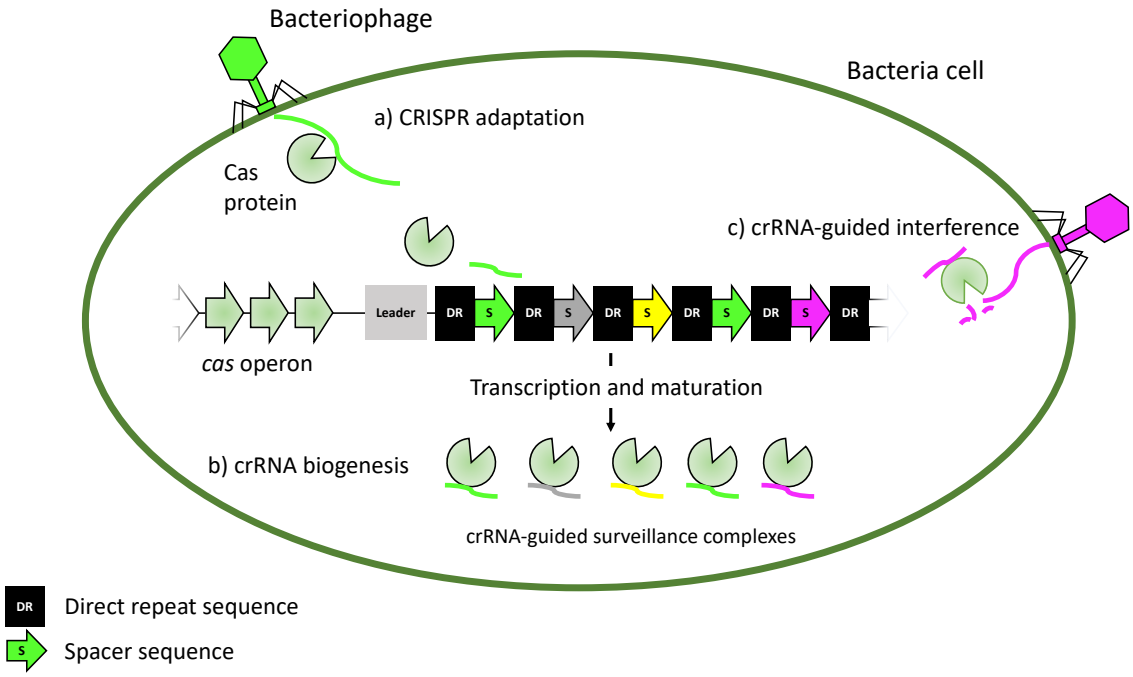


Figure 1.2. Représentation schématique du mécanisme CRISPR-Cas. a) Adaptation de CRISPR où une nouvelle séquence d'espacement (S) provenant d'un génome de phage est incorporée dans le système CRISPR-Cas ; b) Biogénèse des CRISPR ARN (crARN) où les séquences d'espacement de

CRISPR sont transcrites en ARN et la maturation en crARN ; et c) Interférence guidée par crARN où l'ADN viral est reconnu et dégradé par le complexe de crARN.

1.1.3.4. Gènes métaboliques auxiliaires

Les virus acquièrent souvent des gènes d'hôtes facilitant le transfert horizontal de gènes entre les hôtes (Hendrix et al. 2000 ; Miller et al. 2003 ; Lindell et al. 2004). Les génomes viraux comprennent des gènes codant pour des protéines impliquées dans la production des virus, y compris la production de nucléotides, la réplication de l'ADN et la transcription de l'ARN (Lindell et al. 2004). Les gènes viraux qui ne contribuent pas à la réplication virale, mais qui ont des fonctions qui modifient le métabolisme de l'hôte et qui peuvent augmenter le fitness des virus sont appelés gènes métaboliques auxiliaires (AMG) (Breitbart et al. 2007 ; Crummett et al. 2016 ; Jin et al. 2019). Particulièrement, les AMG des cyanophages marins ont été les plus étudiés (Mann 2003 ; Lindell et al. 2004 ; Sullivan et al. 2005 ; Millard et al. 2010). Les AMG peuvent être considérés comme communs ou rares. Les AMG communs à diverses lignées d'hôtes codent pour des fonctions métaboliques essentielles dans toute une série de conditions, tandis que les AMG rares peuvent n'être impliquées que pour des conditions particulières (Crummett et al. 2016).

Par exemple, une analyse récente de viromes de l'Océan Pacifique a identifié des AMG spécialisés dans des niches qui contribuent à l'adaptation des hôtes stratifiés en profondeur, comme celui qui permet la survie en haute pression en haute mer (Hurwitz et al. 2015). Un grand nombre d'AMG ont également été identifiés. Ils sont associés au métabolisme du carbone dans les analyses métagénomiques du sol, y compris les gènes codant pour les hydrolases glycosidiques, les endomannanases et les chitosanases, ce qui suggère l'impact potentiel des virus sur le cycle du carbone dans les écosystèmes du sol (Emerson et al. 2018 ; Trubl et al. 2018 ; Graham et al. 2019 ; Emerson 2019 ; Li et al. 2020). Cependant, la signification adaptative de la plupart des AMG virales du sol n'est généralement pas claire.

1.2. Approche métagénomique pour l'étude des communautés virales du sol

1.2.1. Métagénomique

L'étude sur la diversité et l'écologie des virus dans les écosystèmes a été révolutionnée par les approches moléculaires (Mokili et al. 2012). La métagénomique peut être décrite comme l'analyse de l'ADN génomique des communautés environnementales (Riesenfeld et al. 2004). La caractérisation de la diversité virale dans le sol est difficile car la plupart des hôtes procaryotes ne sont pas cultivables et il n'existe pas de gène marqueur universel commun à tous les génomes viraux, contrairement aux communautés procaryotes et eucaryotes, qui permettent une discrimination taxonomique (Edwards et Rohwer 2005). Toutefois, la métagénomique peut être utilisée pour décrire la majorité des organismes non cultivés et, en combinaison avec le développement d'outils bioinformatiques, elle peut servir à identifier et à fournir des informations génétiques sur les virus présents dans un échantillon environnemental (Breitbart et

al. 2002 ; Edwards et Rohwer 2005 ; Roux et al. 2015a ; Ren et al. 2017). Cependant, en raison de la complexité des sols qui contiennent une grande diversité microbienne, la métagénomique ne récupère généralement que des génomes microbiens complets ou partiels, les organismes les plus abondants, à condition que le séquençage soit suffisamment approfondi (Liolios et al. 2008 ; Mende et al. 2012). Par conséquent, la production d'échantillons enrichis en virus avant de procéder à la métagénomique (c'est-à-dire des viromes) peut faciliter une analyse plus approfondie de la composition de la communauté virale. Afin de produire des viromes du sol, les virus sont généralement extraits du sol dans une solution tamponnée, concentrés par filtrage pour éliminer les grandes cellules procaryotes, puis précipités (Trubl et al. 2016).

1.2.2. Séquençage à haut débit

Le séquençage à haut débit permet de séquencer parallèlement de millions de fragments d'ADN (Tucker et al. 2009). Au cours des deux dernières décennies, de nombreux systèmes de séquençage à haut débit ont été développés en utilisant une variété d'approches différentes, telles que les diverses technologies Illumina (par exemple MiSeq, HiSeq et NovaSeq) et Oxford Nanopore, couramment utilisées aujourd'hui, et qui sont utilisées dans cette thèse.

Le séquençage d'Illumina est basé sur le séquençage par synthèse, où des bases uniques sont détectées par des terminateurs de colorant réversibles lorsqu'elles sont incorporées dans des brins d'ADN (Ambardar et al. 2016). En revanche, le séquençage Oxford Nanopore détecte directement les nucléotides sans synthèse active de l'ADN, et mesure la variation du courant électrique d'un pore lors du passage d'un brin d'ADN simple (Branton et al. 2008 ; Feng et al. 2015). Les systèmes de séquençage varient dans le nombre de séquences et la longueur des fragments d'ADN qui peuvent être traités. Par exemple, les systèmes Illumina MiSeq et NovaSeq peuvent être utilisées pour le séquençage d'amplicons ou l'extraction d'ADN génomique, générant respectivement, jusqu'à 15 Go ou 6 TB de données via 50 millions ou 20 milliards de paires de lectures. En revanche, le système Oxford Nanopore est capable de produire de longues lectures à partir d'un nombre très réduit de fragments d'ADN. Le séquençage par Nanopore présente un taux d'erreur relativement élevé par rapport au séquençage à lecture courte, c'est pourquoi la combinaison de ces deux approches où la correction des erreurs des lectures longues est effectuée en utilisant des données de séquences à lecture courte est couramment réalisée (Kono et Arakawa 2019).

1.2.3. Bio-informatiques utilisées pour les analyses métagénomiques

L'analyse des données du séquençage métagénomique se fait en cinq étapes fondamentales : 1) le contrôle de la qualité des séquences, 2) l'assemblage des séquences, 3) l'assemblage des génomes

(binning), 4) l'annotation fonctionnelle et taxonomique des contigs ou des génomes (bins) et 5) la quantification des bins ou des gènes spécifiques.
Dans cette section, chaque étape est décrite, et sont présentés les outils bioinformatiques associés qui ont été utilisés tout au long de cette thèse.

1.2.3.1. Contrôle de la qualité des données de séquençage

Les lectures séquencées contiennent des séquences d'adaptateur (ajoutées aux fragments d'ADN pour identifier l'échantillon) et la qualité des séquences diminue souvent au fil de la lecture. De plus, dans le cas d'un séquençage en paire, la qualité de la deuxième lecture peut être faible (Tan et al. 2018). Ces erreurs dans les lectures de séquences pouvant compromettre l'analyse en aval, il est nécessaire de filtrer les données pour éliminer les lectures de mauvaise qualité et les séquences d'adaptateur. Le programme Phred est souvent utilisé pour calculer les scores de qualité (Ewing et al. 1998). Un score est attribué à chaque base, ce qui permet d'estimer la probabilité d'erreurs du nucléotide. En général, un score de qualité Phred Q médian supérieur à 20 (i.e. la probabilité d'identification d'une base incorrect de 1 chance pour 100) est considéré comme acceptable, et au-dessus de 30 (i.e. la probabilité d'identification d'une base incorrect de 1 chance pour 100) est considéré comme adéquat. Deux outils de ligne de commande ont été utilisés pour le contrôle de qualité, le FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) et le Trimmomatic (Bolger et al. 2014). Le premier a l'avantage de fournir des mesures de contrôle de qualité sous forme graphique mais peut entraîner un nombre de séquences différent entre deux lectures (lors du séquençage en paire) après la qualité de contrôle, contrairement au second qui maintient le même nombre des paires de lecture et peut trouver efficacement les séquences d'adaptateur.

1.2.3.2. Assemblage des séquences

L'assemblage des séquences implique la fusion des lectures courtes ou longues du même génome en séquences plus longues, appelées contigs. Comme l'assemblage *de-novo* n'est pas biaisé vers un génome de référence, il est généralement utilisé pour assembler de nouveaux génomes (Baker 2012). Trois outils de ligne de commande ont été utilisés : MEGAHIT (Li et al. 2016), metaSPAdes (Nurk et al. 2017) et UniCycler (Wick et al. 2017). MEGAHIT est optimisé pour des lectures de métagénomique complexes et de grande taille, et le paramètre "meta-large" a été utilisé. Comme MEGAHIT, metaSPAdes construit un graphe de *De Bruijn* (i.e. l'identification de chevauchements de toutes les lectures), et à partir du graphe d'assemblage reconstruit les longs contigs dans un métagénome (Bankevich et al. 2012 ; Nurk et al. 2013, 2017). MetaSPAdes est utilisé dans le pipeline bioinformatique du Joint Genome Institute (JGI). L'efficacité des assembleurs MEGAHIT et MetaSPAdes a été testée à l'aide de données de séquençage sur les viromes au Chapitre II.

L’assembleur UniCycler a été utilisé pour un assemblage hybride en utilisant à la fois les lectures courtes d’Illumina et lectures longues de Nanopore au Chapitre III. Unicycler produit d'abord un graph d'assemblage des lectures courtes d’Illumina en utilisant l’assembleur SPAdes, puis utilise les lectures longues de Nanopore pour construire des ponts qui peuvent résoudre les répétitions dans le génome, ce qui permet de produire un assemblage du génome complet.

1.2.3.3. Binning

L’assemblage des génomes (binning) implique le regroupement de contigs ayant des attributs de séquence similaires, tels que la composition *k*-mer, l'utilisation de codons et une couverture d’assemblage similaire dans le même génome (Uritskiy et al. 2018). Dans cette thèse, les modules implémentés dans l'outil MetaWRAP ont été largement utilisés car cet outil réalise toutes les principales procédures de l'analyse métagénomique (contrôle de la qualité, assemblage, binning, annotation taxonomique et fonctionnelle, visualisation et quantification). En outre, MetaWRAP utilise trois logiciels de binning : MaxBin2, metaBAT2 et CONCOCT, et effectue ensuite une approche hybride en considérant les trois ensembles de bins (génomes assemblés) pour produire un ensemble de bins amélioré consolidé. L'outil CheckM a été utilisé pour évaluer la qualité des génomes assemblés à partir de métagénomes (Parks et al. 2015). Il fournit des estimations de la complétude et de la contamination de chaque génome en utilisant des ensembles de gènes collectés qui sont omniprésents et des gènes à copie unique au sein d'une lignée phylogénétique. Les bins peuvent être considérés comme un génome assemblé à partir de métagnomes (MAG) de haute qualité, complétude > 90% et contamination < 5%, un MAG de qualité moyenne, complétude > 70% et contamination < 10%, et un génome partiel, complétude >50% et contamination < 4%. Cependant, il est important de noter que l'utilisation des MAGs peut entraîner la perte d'informations, car seule une proportion relativement faible de lectures est assemblée et intégrée avec succès dans des ensembles de données complexes sur le métagénome (Maguire et al. 2020). Les MAGs peuvent être visualisés à l'aide de l'outil Anvi'o (Eren et al. 2015).

1.2.3.4. Annotation taxonomique et fonctionnelle des contigs ou des bins

Pour l'analyse des contigs, l'outil BLAST (Basic Local Alignment Search Tools) peut être utilisé pour effectuer des alignements locaux afin de rechercher des régions similaires entre les séquences. BLASTn et BLASTp peuvent être utilisés pour comparer les séquences de nucléotides ou de protéines à des bases de données de séquences nucléotidiques ou protéiques, respectivement, et fournir la signification statistique des correspondances. Cette méthode présente l'inconvénient d'un long temps de traitement, et même avec des serveurs de calcul puissants, elle peut prendre de nombreux jours à des semaines. Par ailleurs, l’outil Diamond BLASTp peut être une alternative pour aligner rapidement les séquences protéiques sur une base

de données de séquences non redondantes (nr) (Madden et al. 1996 ; Buchfink et al. 2015). Le site NCBI (National Center for Biotechnology Information) héberge une collection de séquences de nucléotides (nt) et de protéines (nr) non redondantes qui est couramment utilisée comme une base de données de référence dans les recherches BLAST. En outre, la base existante de données de séquences protéiques annotées manuellement, Swiss-Prot qui fait partie de la collection de bases de données UniProt (Universal Protein Resorce) a été utilisée (Tableau 1.1).

Tableau 1.1. Informations sur les bases de données de référence utilisées dans le cadre de cette thèse.

Database	Host	Type	Number of sequences	Length of sequences (bp)
nt	NCBI	Nucleotide	53,777,267	237,410,501,766
nr	NCBI	Protein	115,570,790	42,364,384,627
Refseq	NCBI	Protein	49,770,189	15,556,247,796
Swiss-Prot	UniProt	Protein	560,823	201,585,439
viruses	NCBI	Nucleotide	12,156	317,115,877
viruses	NCBI	Protein	315,213	79,514,978

Pour annoter la taxonomie et la fonction des contigs, les outils Kaiju et InterProScan 5 sont couramment utilisés, respectivement (Jones et al. 2014 ; Menzel et al. 2016). Kaiju est un programme de classification taxonomique des contigs à partir d'un séquençage métagénomique (Menzel et al. 2016). Chaque contig est traduit en une séquence d'acides aminés, qui est comparé à la base de données de référence. La collection de RefSeq (Séquences de référence), hébergée par le site NCBI, fournit un ensemble de séquences protéiques, inclusives et non redondantes (Tableau 1.1). InterProScan 5 utilise les séquences de nucléotides et d'acides aminés pour comparer à une collection de bases de données de signatures de protéines, y compris CATH-Gene3D, CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, SFLD et SMART (Jones et al. 2014). Cet outil fournit des familles et des domaines de protéines, ainsi que l'ontologie et les voies métaboliques (KEGG, MetaCyc et Reactome).

L'analyse taxonomique et fonctionnelle des MAGs est différente de celle des contigs. L'identification taxonomique peut être effectuée à l'aide de l'outil GTDB-Tk (The Genome Taxonomy Database Toolkit) (Chaumeil et al. 2020). Cet outil utilise un ensemble de 120 gènes marqueurs bactériens et 122 gènes marqueurs archéens ainsi que l'outil FastANI pour estimer ANI (Average Nucleotide Identity) entre des gènes communs à un MAG et à un génome de référence. Si l'ANI entre le MAG et le génome de référence est supérieur à 95% et que la fraction d'alignement est supérieure à 0.65, le MAG est classé comme appartenant à une espèce de génome de référence. La prédiction des gènes à partir d'un contig ou MAG peut être effectuée avec l'outil Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm)(Hyatt et al. 2010) et l'analyse fonctionnelle avec l'outil Diamond BLASTp, KEGG (Kyoto Encyclopedia of Genes and Genomes)(Kanehisa et Goto 2000) et InterProScan 5 (Jones et al. 2014). L'outil d'annotation de

contig (CAT) et l'outil d'annotation de bin (BAT) sont également utilisés pour classer la taxonomie des longues séquences d'ADN et des MAG (Meijenfeldt et al. 2019).

1.2.3.5. Quantification des bins ou des gènes spécifiques

Le processus d'alignement des lectures courtes sur une séquence de référence d'un génome complet ou d'un assemblage *de-novo* est appelé « read mapping » (Schatz et al. 2010). De nombreux programmes ont été développés pour aligner les lectures avec une séquence de référence, et qui varient dans leur algorithme (Eren et al. 2015 ; Uritskiy et al. 2018). Les abondances qui résultent de l'alignement de lectures peuvent être normalisées par la couverture et la taille du génome. Ici, les outils Salmon et Bowties2 implantés dans metaWRAP et Anvi'o, ont été utilisés respectivement (Patro et al. 2017; Langdon 2015).

1.2.4. Analyses des séquences virales

Comme les séquences virales peuvent avoir une homologie plus élevée avec les gènes de procaryotes ou eucaryotes (Breitbart et al. 2002, 2003 ; Angly et al. 2006 ; Schoenfeld et al. 2008 ; Blomström et al. 2010), lors du traitement des viromes, il est important d'éliminer les séquences contaminées par des séquences microbiennes en les alignant avec les génomes de référence procaryotes ou eucaryotes. Après l'assemblage des données de viromes, les contigs viraux peuvent être identifiés par une approche basée sur les références ou sans référence. Les contigs viraux identifiés peuvent ensuite être classés dans des arbres phylogénétiques viraux basés sur la similarité à l'échelle du génome, et différentes approches peuvent être utilisées pour associer les virus à leurs hôtes, comme l'analyse des CRISPRs, la similarité génomique et l'homologie des gènes (Sanguino et al. 2015 ; Ahlgren et al. 2017 ; Galiez et al. 2017).

1.2.4.1. Prédiction des contigs viraux

Les séquences génomiques virales peuvent être prédites grâce à des approches basées sur la similarité des gènes (par exemple, l'outil VirSorter) et sur la fréquence d'utilisation de *k*-mer (par exemple, les outils VirFinder et DeepVirFinder). L'approche basée sur la similarité des gènes utilise des modèles HMM (Hidden Markov Model) basés sur l'annotation des gènes pour reconnaître les régions codantes homologues aux gènes de l'origine virale. L'approche basée sur la fréquence d'utilisation de *k*-mer est utile pour les contigs courts avec peu de gènes prédits ou lorsque les gènes n'ont pas de similarité avec ceux des virus connus. Dans le cadre de cette thèse, les deux outils VirSorter et DeepVirFinder ont été utilisées (Roux et al. 2015a; Ren et al. 2017).

VirSorter est l'outil le plus récent développé pour prédire le signal viral, y compris les prophages et les virus lytiques (Roux et al. 2015a). VirSorter identifie les séquences virales enrichies en gènes ayant une similarité avec les bases de données de viromes, et sont prédites

dans l'une des trois catégories selon la présence ou l'absence de gènes caractéristiques d'origine virale annotés telles que protéines majeures de la capside, portales, terminase, et gènes de type viral basés sur les bases de données de viromes et la délétion de gènes affiliés à la famille de protéines (Pfam). La catégorie 1, les prédictions "les plus sûres", sont des séquences qui contiennent un enrichissement en gènes de type viral ou non-Caudovirales et au moins un gène viral caractéristique détecté. La catégorie 2, les prédictions "probables", sont des séquences qui contiennent soit un enrichissement en gènes de type viral ou non-Caudovirales, soit un gène viral caractéristique, et qui présentent au moins une des mesures suivantes : déplétion dans l'affiliation Pfam, enrichissement en gènes non caractérisés, enrichissement en gènes courts ou déplétion dans l'inversion de brin. La catégorie 3, prédictions "possibles", est constituée de séquences qui n'ont pas de gène caractéristique viral ni d'enrichissement en gènes de type viral ou non-Caudovirales, mais qui ont au moins deux des mesures susmentionnées, mais dont l'une au moins exige un score de signification considérable. En outre, un contig est déterminé viral si une séquence prédite possède plus de 80% des gènes prédits sur un contig, (Roux et al. 2015a). Les prophages identifiés sont classés de la même manière que les virus, dans les catégories 4 ("le plus sûr"), 5 ("probable") et 6 ("possible") (Roux et al. 2015a).

Inversement, DeepVirFinder est une approche sans référence et sans alignement pour la détection de séquences virales, basée sur des méthodes d'apprentissage automatique en profondeur. DeepVirFinder a été formé sur la base d'un grand nombre de génomes viraux de référence, et a appris un réseau de neurones convolutifs qui identifie les séquences virales à toutes les longueurs de contig avec précision (Ren et al. 2017, 2018b). L'utilisation de cet outil a permis la découverte d'un grand nombre de nouveaux virus à partir de métagénomes, ce qui a conduit à des progrès marqués dans notre connaissance des interactions virus-hôte.

1.2.4.2. Populations virales

La classification des virus est difficile en raison de l'absence de gènes marqueurs universels (Edwards et Rohwer 2005). Cependant, les récents développements des technologies de séquençage ont permis l'augmentation de la taille des séquences de génome viral, ce qui facilite la classification des virus (Paez-Espino et al. 2016 ; Roux et al. 2016). Plusieurs approches pour la classification des virus ont été proposées, basées sur des stratégies comparatives génomiques, protéomiques et spécifiques aux gènes (Quan et al. 2016). En général, les génomes ou séquences virales (> 10 kb) avec un seuil d'ANI de 95% sont classés au rang d'espèce (Quan et al. 2016 ; Roux et al., 2019). Il a été suggéré de considérer ces populations comme des unités de population virale (vOTUs)(Hurwitz et al., 2015). La similarité des gènes codant pour les protéines dans les génomes viraux, et qui sont utilisés pour générer des arbres phylogénétiques viraux, a démontré que cette méthode peut être utilisée comme base d'un système taxonomique basé sur le génome

(Nishimura et al. 2017). Cette méthode permet d'étudier la biodiversité des virus en regroupant les virus en taxons qui permettent de prévoir plusieurs aspects de la biologie des virus (Rohwer et Edwards 2002). Cependant, cette approche est limitée lorsque la détermination de la similarité basée sur BLASTp est confrontée à des séquences de protéines éloignées (Chibani et al. 2019).

Dans cette thèse, le serveur VipTree (The viral proteomic tree server) est utilisé pour la classification des contigs viraux basée sur la similarité à l'échelle du génome (Nishimura et al. 2017). Les génomes des virus associés à l'hôte sont comparés aux génomes viraux de référence stockés dans la base de données DB (Virus-Host database) qui contient un total de 2,687 virus à ADNdb associés aux procaryotes et 1,119 virus à ADNdb associés aux eucaryotes (Mihara et al. 2016). Les scores de similarité sont calculés à partir des résultats du tBLASTx. En outre, l'outil vConTACT (the viral contig automatic clustering and taxonomy) a été utilisé avec la base de données RefSeq du NCBI (Bolduc et al. 2017 ; Bin Jang et al. 2019). Il s'agit d'une analyse de réseau basée sur le génome du contenu protéique viral partagé avec une base de données de référence. La taxonomie des virus de procaryotes est déduite en regroupant les séquences de protéines virales par l'outil BLASTp (E value < 10^{-4} et bit score > 50). Sur la base du nombre de groupes de protéines partagées entre les génomes viraux et les génomes de référence, un score de similarité pour chaque paire est calculé sur la base d'une *P*-value générée par le nombre total de comparaisons de génomes par paires. Le réseau de paires de génomes qui en résulte (score de similarité > 1) a été visualisé avec le logiciel Cytoscape (source : http://cytoscape.org/).

1.2.4.3. Lien entre le virus et l'hôte

L'approche la plus sûre pour associer les virus à leurs hôtes est l'analyse des CRISPRs.
Les CRISPRs sont composées de répétitions directes (DR), qui sont une succession de 24 à 47 pb d'origine microbienne, intercalées par des spacers (S) dérivés d'envahisseurs (Figure 1.2) (Mojica et al. 2009). Après l'insertion du spacer dans l'extrémité leader riche en AT du CRISPR, les séquences spacer sont transcrites et maturées en petits ARN interférents (ARNc) assurant l'immunité contre les envahisseurs (Jansen et al. 2002 ; Terns et Terns 2011). L'analyse des CRISPRs présentes dans les génomes des hôtes et les spacers pour identifier les fragments du génome viral permet de relier les hôtes et les virus qui leur sont associés (Andersson et Banfield 2008 ; Mojica et al. 2009). Il existe plusieurs logiciels développés qui utilisent différents algorithmes afin de rechercher les séquences de CRISPRs. Dans cette thèse, trois outils CRT (Bland et al. 2007), Crass (Skennerton et al. 2013) et CRISPRCasFinder (Couvin et al. 2018) ont été utilisés : l'outil CRT recherche une série de courtes répétitions exactes (Bland et al. 2007), l'outil Crass assemble les CRISPRs en utilisant les recherches *k*-mer (Skennerton et al. 2013) et CRISPRCasFinder détecte les gènes Cas (Couvin et al. 2018).

Comme on ne trouve pas souvent de CRISPRs dans les contigs des hôtes, d'autres approches ont été développées pour relier les virus aux hôtes, telle que l'analyse de la fréquence des oligonucléotides (ONF), les régions génomiques partagées, la corrélation de l'abondance des virus et des hôtes et les correspondances d'ARNt. Dans cette thèse, une approche ONF a été utilisée avec l'outil WIsH (Who Is the Host) (Galiez et al. 2017). L'hôte potentiel d'un virus peut être prédit en identifiant l'hôte avec lequel il présente la plus grande similitude de fréquences *k*-mer. L'outil WIsH a été développé en adoptant une approche probabiliste adaptée (Galiez et al. 2017). Un modèle de Markov homogène d'ordre k (k = 8, car la précision est maximale pour l'ordre 8) peut être formé pour chaque contig d'hôte potentiel. La probabilité d'un contig viral sous chacun des modèles de Markov entraînés est alors calculée, et l'hôte dont le modèle donne la plus grande probabilité est prédit *de novo*. En outre, des P-values sont calculées en utilisant les paramètres des distributions nulles gaussiennes de chaque modèle de Markov. Contrairement aux autres outils basés sur *k*-mer, WIsH peut être utilisé avec des contigs courts (< 3 kb). En combinaison avec l'analyse ONF, l'homologie des gènes entre les hôtes et les virus a été évaluée en utilisant les différents outils décrits dans la section 1.2.3.4.

1.2.5. Calcul haute performance

Les bioinformatiques nécessitent souvent des exigences matérielles élevées, comme des serveurs multi-cœurs, ainsi qu'un niveau élevé d'expertise et de "savoir-faire". En utilisant une interface graphique pour interagir avec le shell (c'est-à-dire une interface en ligne de commande avec le système d'exploitation) un ordinateur peut être connecté à un serveur par des programmes client SSH (secure shell), par example, Terminal pour MacOS, ou PuTTY pour Windows ou Linux. Dans les études de cette thèse, les bioinformatiques ont été réalisée en utilisant le système d'exploitation Mac (MacOS) et deux grappes de calculateur (c'est-à-dire des serveurs), NEWTON et MUSCLOR. Le serveur NEWTON possède un nombre élevé de CPU (unités centrales de traitement) et a été utilisé pour analyser les plus grands ensembles de données (par exemple, assemblage *de-novo* et > 200 Go de RAM). NEWTON est composé de 84 nœuds de 1,704 CPU et de 12 TB de RAM. Les nœuds sont séparés par plusieurs réseaux en fonction des besoins en termes de CPU et de RAM. MUSCLOR a été utilisé pour analyser les ensembles de données plus petits car il est composé de 24 CPU et 40 GB de RAM. Avec Terminal, les lignes de commande shell ont été utilisées pour exécuter diverses tâches, telles que l'installation d'outils. Les outils ont été installés dans des environnements créés à l'aide du programme Anaconda en ligne de commande basée sur Python (source : https://www.anaconda.com/). L'outil Seqkit a été utilisé (Shen et al. 2016) pour le traitement efficace des fichiers de séquence FASTA et FASTQ (conversion, recherche, filtrage, déduplication, division, etc.).

1.3. Modèle de gradient de pH du sol

L'utilisation de gradients environnementaux, tels que la température, les nutriments, la salinité, l'humidité, l'oxygène et le pH, pour répondre à diverses questions sur la distribution, les interactions et l'assemblage des communautés microbiennes est une pratique de longue date dans le domaine de l'écologie microbienne et a permis de mieux comprendre la dynamique microbienne. Il a été démontré à plusieurs reprises que le pH du sol influence la diversité et la composition microbiennes, la diversité bactérienne étant plus importante dans les sols neutres que dans les sols acides (Lauber et al. 2009 ; Rousk et al. 2010 ; Zhalnina et al. 2014 ; Ren et al. 2018a). Ainsi, un gradient de pH du sol bien établi et étudié a été utilisé dans cette thèse pour étudier l'impact de la structure des communautés procaryotes et du pH du sol sur les virus et leurs interactions avec différents hôtes. Le campus de Craibstone du SRUC (Scottish Rural College), situé à Aberdeen, en Écosse, au Royaume-Uni (57°11, 2°12), dispose d'un certain nombre de systèmes expérimentaux agricoles établis de longue date, dont une série de huit parcelles parallèles dans lesquelles chaque parcelle présente un gradient de pH allant de 4.5 à 7.5 à des intervalles de 0.5 unité de pH. Les parcelles ont fait l'objet d'une rotation des cultures sur 8 ans (blé d'hiver, pommes de terre, orge de printemps, houblon, avoine de printemps, 3 ans d'herbe sans réensemencement), et ont toutes reçues un engrais minéral NPK modéré avant les semis, à l'exception des années 2 et 3 de la rotation des graminées, et sont maintenues depuis plus de 50 ans (Figure 1.3). Le pH de chaque sous-parcelle est contrôlé par l'ajout de chaux ou de sulfate d'aluminium (Bartram et al. 2014). Comme la rotation des plantes cultivées se fait sur une base annuelle, l'effet d'une espèce de plante sur la communauté microbienne n'est pas un facteur majeur de la structure de la communauté. Le carbone et l'azote total, et la matière organique ne changent pas de manière significative à travers le gradient (Nicol et al. 2008). Étant donné que le rendement, la santé et la qualité des cultures sont soutenus par le maintien de la fonction de l'écosystème agricole du sol sous l'impulsion des microorganismes du sol, l'étude des virus qui contrôlent l'abondance des hôtes microbiens présente un intérêt agronomique. Par exemple, plusieurs études ont été menées pour identifier les phages associés à *Rhizobium* sp. qui peuvent améliorer la productivité des cultures (Kimura et al. 2008 ; Santamaría et al. 2014).

Des études antérieures menées sur le gradient de pH du sol à Craibstone ont montré que le pH du sol influençait la diversité microbienne et la composition des communautés (Nicol et al. 2008 ; Bartram et al. 2014). Par exemple, sur la base des gènes 16S rRNA et *amoA*, il a été démontré que la structure des bactéries et des archées oxydant l'ammoniac changeait en fonction du pH du sol, avec des populations distinctes de nitrifiants dans le sol acide et neutre (Stephen et al. 1998 ; Nicol et al. 2008 ; Gubry-Rangin et al. 2011). 2011). En particulier, les *Acidobactéries* ont augmenté dans les sols à pH élevé, tandis que les *Actinobactéries* étaient proportionnellement plus nombreuses dans les sols à faible pH. Les genres *Burkholderia* et *Paucibacter* (*Bêta-*

Protéobactéries) n'ont été trouvés que dans les sols à faible pH, tandis que les genres *Nitrosospira*, *Denitratisoma*, *Paucimonas*, *Herbaspirillum*, *Tepidimonas* et *Polaromonas* (*Bêta-Protéobactéries*) étaient associés à un pH élevé : Chez les *Alphaprotéobactéries*, le genre *Phenylobacterium* était associé aux sols à faible pH, tandis que *Devosia*, *Roseomonas*, *Labrys*, *Methylosinus*, *Fulvimarina*, *Filomicrobium*, *Rhodobacter*, *Hyphomicrobium*, *Bartonella* et *Mesorhizobium* n'étaient associés qu'aux sols à fort pH. Chez les *Gamma-Protéobactéries*, les genres *Dyella* et *Rhodanobacter* n'ont été trouvés que dans les sols à faible pH, tandis que le genre *Lysobacter* n'a été observé que dans les sols à pH élevé et moyen (Bartram et al. 2014). En tant que tel, ce gradient de pH du sol est un excellent environnement modèle pour l'étude des interactions virus-hôtes dans différentes niches de pH du sol.

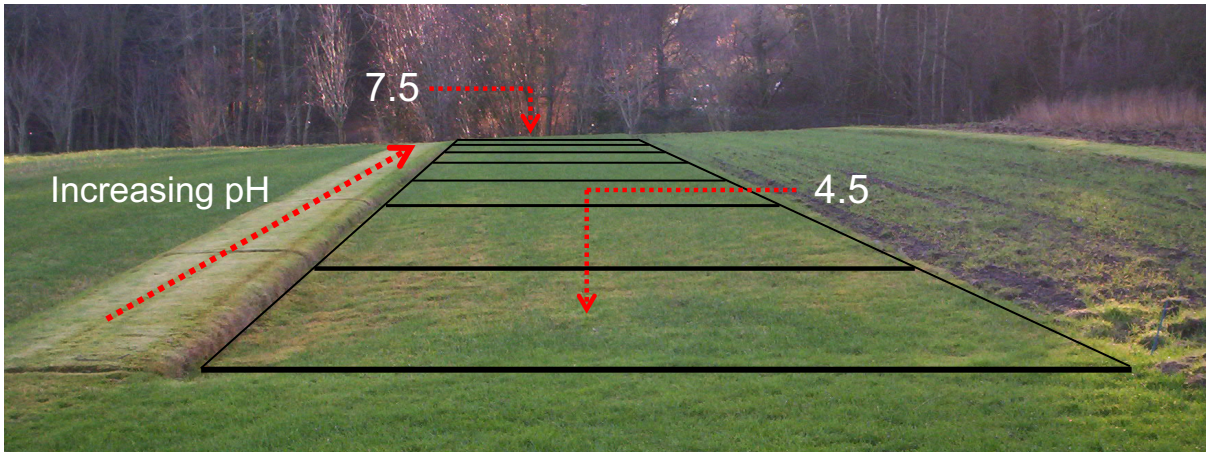


Figure 1.3. Parcelles à pH contrôlé échantillonnées au Scottish Agricultural College, Craibstone, Écosse.

1.4. Objectifs de la recherche

Dans cette thèse, les quatre études qui ont utilisé les sols du gradient de pH (Chapitre II - V) sont présentées et se terminent par une discussion générale sur les virus du sol dans le contexte des conclusions de la thèse (Chapitre VI). La diversité virale et les interactions hôte-virus à travers le pH du sol n'ont pas encore été suffisamment décrites. C'est pourquoi l'étude présentée au Chapitre II vise à déterminer l'influence de la structure de la communauté microbienne et du pH du sol sur les virus en utilisant la métagénomique et la viromique. Contrairement aux milieux marins, il n'existe pas de compréhension générale de la mesure dans laquelle les interactions virus-hôte procaryote régulent les populations procaryotes dans le sol. Alors que certains virus ont la capacité d'infecter une série d'hôtes dans des communautés procaryotes très diverses du sol, des processus coévolutifs au sein de niches écologiques peuvent contrôler étroitement la susceptibilité des hôtes. Afin de mieux comprendre dans quelle mesure les interactions virus-hôte procaryote régulent les populations procaryotes du sol, le Chapitre III présente une étude qui

évalue l'infectivité d'une bactérie hôte par rapport à des sources co-localisées et de plus en plus allochtones de populations virales isolées à travers le gradient de pH du sol en utilisant une approche d'essai de plaque basée sur la culture, suivie d'une microscopie à transmission électronique et d'un séquençage métagénomique hybride pour déterminer la diversité virale et les interactions virus-bactérie hôte.

Enfin, deux études sont présentées et visent à identifier les populations de virus infectant des groupes fonctionnels microbiens spécifiques dans un sol au pH contrasté, en particulier les méthanotrophes (Chapitre IV) et les nitrifiants (Chapitre V), à l'aide de sondes d'isotopes stables (SIP) combinées à un séquençage métagénomique profond. Actuellement, il n'existe pas d'études ayant déterminé les relations actives entre les communautés microbiennes du sol et les virus qui leur sont associés *in situ*. La SIP implique l'assimilation de substrats enrichis en isotope lourd dans la biomasse microbienne cellulaire d'échantillons environnementaux (Radajewski et al. 2000). L'analyse moléculaire de l'ADN isotopiquement marqué fournit des informations phylogénétiques et fonctionnelles sur les micro-organismes actifs responsables du métabolisme d'un substrat (Radajewski et al. 2000 ; Chen et al. 2008 ; Prosser et Nicol 2012). Comme les virus sont des parasites intracellulaires obligatoires qui utilisent le matériel génétique de leur cellule hôte, les virus associés des micro-organismes actifs seront également marqués de manière isotopique (Gelderblom 1996 ; Lee et al. 2012). En combinant le séquençage métagénomique profond avec le DNA-SIP et en déterminant les liens hôte-virus par le CRISPR array et l'analyse ONF, les interactions virus-hôte actives dans un système de sol complexe ont été étudiées et l'impact des virus sur les communautés microbiennes a été évalué en identifiant les gènes hôtes qui ont été potentiellement transférés parmi leurs virus associés (Chapitre IV et V).

L'étude de l'impact des virus sur les communautés méthanotrophes et nitrifiantes du sol a également d'importantes ramifications écologiques. Les nitrifiants et les méthanotrophes du sol jouent un rôle crucial dans la production et la consommation de gaz à effet de serre en oxydant l'ammonium et en produisant du N_2O (Prosser et al. 2020) et en consommant du méthane (Dedysh et Knief 2018), respectivement. Le dioxyde de carbone (CO_2), le méthane (CH_4) et l'oxyde nitreux (N_2O) sont des gaz à effet de serre importants qui contribuent au changement climatique (GIEC 2013). Les activités humaines, telles que la gestion des sols agricoles, ont stimulé la production de ces gaz à effet de serre par les microbes du sol (Canfield et al. 2010). Le méthane est le deuxième gaz à effet de serre le plus abondant après le CO_2, et représente, selon les estimations, 20% du réchauffement climatique (Tate 2015 ; Nisbet et al. 2016). Les bactéries aérobies oxydant le méthane sont omniprésentes dans le sol, et utilisent le CH_4 comme seule source d'énergie et de carbone (Dedysh et Knief 2018). L'oxydation de l'ammoniac par les nitrifiants est couplée à une fixation autotrophe du CO_2. L'oxydation de l'ammoniac génère elle-même du N_2O et produit le nitrate de substrat qui est ensuite utilisé pour produire du N_2O par dénitrification (Nicol et al.

2008 ; Prosser et Nicol 2012). Par conséquent, il est essentiel de comprendre les contrôles et les influences sur les communautés de méthanotrophes et de nitrifiants du sol pour contribuer aux efforts d'atténuation du changement climatique.

Discussion générale :
Interactions hôte-virus dans le sol

2.1. Vue d'ensemble

Les micro-organismes jouent un rôle central dans le sol où ils sont impliqués dans une vaste gamme de processus qui facilitent le cycle biogéochimique (Prosser et Nicol 2008 ; Chaparro et al. 2012 ; Aislabie et Deslippe 2013 ; Philippot et al. 2013). La plupart des recherches se concentrent sur la compréhension des facteurs abiotiques qui contrôlent la structure et l'activité des communautés microbiennes (par exemple, la disponibilité du substrat, les facteurs abiotiques, etc.), mais l'effet des facteurs biotiques sur de la diversité et de l'abondance microbiennes, notamment les virus est relativement peu connu. L'infection virale a des implications importantes sur la structure et la fonction des communautés microbiennes dans les écosystèmes (Weinbauer et Rassoulzadegan 2004 ; Suttle 2005 ; Breitbart et al. 2007 ; Bertilsson et al. 2013). Les virus associés à l'hôte peuvent influencer la taille de la population de leur hôte et leur taux des processus enzymatiques via la lyse des cellules hôtes (c'est-à-dire en diminuant le taux de la fonction) ou en fournissant (ou en augmentant) les processus enzymatiques via l'expression de gènes métaboliques auxiliaires (AMG), respectivement. Des études récentes ont découvert un grand nombre de AMG impliqués dans la dégradation du carbone, démontrant l'impact potentiel des virus du sol sur le traitement du carbone dans les écosystèmes (Emerson et al. 2018 ; Trubl et al. 2018, 2020 ; Graham et al. 2019).

L'objectif général de cette thèse était de mieux comprendre les interactions entre les virus du sol et leurs hôtes, à la fois à l'échelle communautaire et individuelle *in situ*. Dans le Chapitre II, l'impact d'un gradient physico-chimique abiotique (pH du sol) sur la distribution des hôtes microbiens et, par la suite, sur la structure de la communauté virale a été étudié à l'aide de la métagénomique de la communauté totale et de la métagénomique ciblée sur les virus. Dans le Chapitre III, en utilisant une approche basée sur la culture avec un seul hôte (*Bacillus* sp. S4), une approche de test par plage de lyse (plaque assay) a été combinée avec la métagénomique pour étudier comment la co-localisation des populations de virus hôtes du sol affecte la diversité des virus infectants par rapport aux virus provenant d'une niche différente (c'est-à-dire un pH du sol différent). Enfin, pour se concentrer sur les virus du sol qui infectent les procaryotes et jouent un rôle central dans le cycle du carbone (C) et de l'azote (N), une approche DNA-SIP a été combinée à la métagénomique pour identifier les virus actifs des hôtes méthanotrophes incorporant du C dérivé du $^{13}CH_4$ (Chapitre IV), et les hôtes autotrophes transformant l'azote inorganique, nitrifiants incorporant C dérivé du $^{13}CO2$ (Chapitre V). L'ensemble des travaux présentés dans cette thèse met en évidence les défis rencontrés avec l'utilisation de la métagénomique pour l'étude des virus du sol, mais fournit également des indications sur la dynamique virus-hôte et les gammes d'hôtes en reliant les virus et les hôtes par le biais de l'analyse de CRISPR array et de la fréquence des oligonucléotides (ONF), et des analyses de l'homologie des gènes pour découvrir les AMG et le transfert horizontal de gènes. Enfin, il démontre que l'utilisation d'isotopes stables

du carbone permet une analyse détaillée et à haute résolution des interactions actives *in situ* en suivant le transfert du carbone de l'hôte au virus.

2.2. Défis de la métagénomique des virus du sol

Comme la diversité microbienne est vaste dans le sol et que la taille des génomes des procaryotes est supérieure à celle des virus, un séquençage profond est nécessaire pour capturer la diversité microbienne et virale (Papudeshi et al. 2017 ; Maurier et al. 2019). En outre, le séquençage en profondeur des filtrats viraux du sol (c'est-à-dire des viromes) peut faciliter la récupération des séquences virales. Avec le financement du Joint Genome Institute (JGI), le séquençage à haut débit utilisant la plateforme Illumina NovaSeq a été effectué pour déterminer la diversité procaryote et virale du sol dans une série d'expériences utilisant des sols à partir d'un gradient de pH à long terme (pH 4.5 et 7.5) à Aberdeen, en Écosse. Le séquençage NovaSeq a généré 1.6 TB de données de séquençage et 4.7 milliards de lectures de haute qualité à partir de 26 métagénomes et 6 viromes. Compte tenu des résultats de la métagénomique des Chapitres II à V, le nombre de contigs viraux (VC) et de génomes assemblés (MAG) a augmenté avec la diminution de la diversité des communautés microbiennes. L'analyse d'ensembles de données générés par le séquençage NovaSeq a nécessité une grande quantité de RAM et un temps de calcul important. Par exemple, le co-assemblage des six viromes (334 Go) en utilisant 384 Go de RAM, 32 CPU a pris 7 jours. Les études sur la diversité virale des sols sont un défi car un séquençage profond et une puissance de calcul élevée sont des exigences de base. Toutefois, les progrès technologiques dans ces domaines évoluent rapidement, ce qui permet d'accroître les possibilités à des coûts réduits.

Un autre défi consiste à identifier les signaux viraux dans les métagénomes qui contiennent à la fois des contigs viraux et des contigs hôtes provenant d'une grande diversité d'organismes. Dans cette thèse, deux outils de prédiction de virus ont été utilisés : une approche basée sur la similarité des gènes (VirSorter) (Roux et al. 2015) et une autre approche utilisant la méthode d'apprentissage profond (deep learning) pour un apprentissage supervisé basée sur la fréquence d'utilisation de *k*-mer (DeepVirFinder) (Ren et al. 2017, 2018). Bien que la première approche puisse donner des prédictions très fiables en identifiant la présence de gènes viraux, les prédictions sont également limitées par les biais d'alignement, notamment l'absence de gènes marqueurs viraux dans une base de données, les génomes viraux incomplets résultant en des portions de génome sans gènes marqueurs viraux, et une faible similarité de séquence avec des gènes viraux de référence peu représentés ou non caractérisés. L'approche basée sur la fréquence *k*-mer utilisant DeepVirFinder permet l'étude sans alignement et est notamment plus avantageuse pour les contigs viraux courts (> 300 bp) par rapport à VirSorter qui est limité à l'analyse des contigs qui ont au moins trois gènes codants, et est recommandé pour utiliser des contigs >10 kb. Cependant, alors que l'utilisation de DeepVirFinder serait clairement bénéfique pour l'analyse des

petits contigs, les contigs viraux prédits semblent souvent être dérivés de génomes bactériens. Ainsi, VirSorter a été utilisé préférentiellement pour les analyses effectuées dans le cadre de cette thèse, et lorsque DeepVirFinder a été utilisé pour des contigs viraux courts, les gènes des contigs viraux prédits ont été manuellement vérifiés. Cependant, l'utilisation de ces deux outils a permis de découvrir un grand nombre de nouveaux virus à partir des métagénomes du sol (Figure 6.1).

Dans l'ensemble, le séquençage profond à haut débit et les outils de prédiction des virus ont permis d'identifier 8,801 contigs viraux à partir des viromes (VC, > 10 kb), 1,070 contigs viraux à partir des métagénomes (mVC, > 5 kb) dérivées de l'ADN-SIP expériences, et 45 génomes procaryotes par assemblage métagénomique (MAG) qui ont dépassé un seuil de complétion de 50% avec moins de 10% de contamination. Les mVCs provenant des métagénomes totaux du sol (c'est-à-dire sans enrichissement des particules virales avant le séquençage) n'ont pu être prédits qu'à l'aide du DeepVirFinder en raison de leur courte longueur en moyenne de 531 bp, et par contraste avec les VCs qui avaient une longueur moyenne de 21 kb. La petite taille des mVCs était due à la faible proportion de séquences virales dans les métagénomes en raison de la grande diversité et de la taille relativement large du génome des procaryotes dans le sol comparant ceux des virus (Papudeshi et al. 2017 ; Maurier et al. 2019). La prédiction des virus entre les métagénomes et les viromes a également révélé des biais dans la méthodologie comme la sélection de la taille. Plus précisément, les mVCs comprenaient de grands virus (> 0,2 μm) appartenant aux Mimiviridae et Phycodnaviridae (Maruyama et Ueki 2016) qui n'étaient pas présents dans les viromes sélectionnés en fonction de leur taille. Par conséquent, l'utilisation des métagénomes et des viromes pour capturer la diversité virale du sol est généralement nécessaire.

Dans les Chapitres IV et V, l'ADN des isotopes stables de sondage (DNA-SIP) a été utilisé pour cibler des groupes microbiens spécifiques, diminuant la diversité microbienne par la sélection d'ADN pour des populations spécifiques enrichies en ^{13}C. Les métagénomes dérivés du CO_2-SIP, plus diversifiés, ont généré 420 mVC et 9 MAG par rapport aux métagénomes dérivés du CH_4-SIP avec 650 mVC et 23 MAG. L'approche DNA-SIP a été largement couronnée de succès et a permis d'étudier les interactions virus-hôtes individuelles des principaux groupes fonctionnels microbiens du sol. Les métagénomes CH_4-SIP enrichis en ^{13}C-CH_4 contenaient tous des utilisateurs primaires ou secondaires actifs de CH_4, avec 12 MAG méthanotrophes et 6 MAG méthylotrophes. Toutefois, par rapport aux métagénomes CH_4-SIP, les métagénomes CO_2-SIP n'ont généré qu'un seul MAG nitrifiant, avec un seuil de complétion relativement faible (66.5 %), en raison de la grande complexité microbienne qui subsiste dans l'ADN "enrichi" du fait du chevauchement de la densité et de la distribution de l'ADN génomique des bactéries hétérotrophes non marquées et de l'ADN nitrifiant marqué au ^{13}C à faible pourcentage de GC (Howe et al. 2014). Une autre approche consisterait à effectuer un DNA-SIP avec des extraits d'ADN enrichis en virome (c'est-à-dire filtrés). Bien que cela inclurait également des virus hétérotrophes non enrichis, cela permettrait

d'éliminer les génomes procaryotes dominants, ce qui se traduirait par une récupération relativement faible de l'ADN nitrifiant et des métagénomes viraux associés aux nitrifiants. Cependant, l'approche utilisée avait l'avantage théorique non seulement d'identifier les virus qui ont été récemment produits dans les cellules hôtes, mais aussi de les relier à des hôtes enrichis en ^{13}C. Néanmoins, les approches utilisées dans cette thèse ont fourni des preuves directes de l'activité des virus-hôtes dans le sol.

Comme pour les autres études de métagénomique des sols, les analyses présentées ont permis d'élargir nos connaissances sur les virus du sol et ont mis en évidence que la majorité des séquences (> 90 %) n'ont pas de similarité significative avec les bases de données de référence (Figure 6.1), et sont désignées sous le nom de "matière noire virale" (Roux et al. 2015). Parmi les virus annotés récupérés dans cette étude, les virus procaryotes présentaient la composante la plus importante (94 % de Caudovirales), ce qui est conforme à la conclusion selon laquelle les bactériophages à queue sont la composante virale dominante dans les habitats du sol (Zablocki et al. 2014). L'analyse des réseaux de gènes viraux peut être utilisée pour déduire la parenté des virus procaryotes en regroupant les séquences des protéines virales afin de surmonter les problèmes liés à la faible similitude des séquences entre les génomes viraux et à l'absence d'un gène marqueur universel commun à tous les génomes viraux qui permettrait une discrimination taxonomique (Edwards et Rohwer 2005 ; Bolduc et al. 2017 ; Bin Jang et al. 2019). Le regroupement des VCs de pH 4.5 et 7.5 avec les VCs co-assemblés a démontré que le co-assemblage permet de récupérer les contigs viraux obtenus à partir d'assemblages individuels (c'est-à-dire les VCs de pH 4.5 et 7.5) (Figure 6.1). Seuls 69 groupes de virus contenant les VCs ont été associés à des virus de référence connus que l'on trouve couramment dans les sols, tels que *Arthrobacter*, *Bacillus*, *Burkholderia*, *Flavobacterium*, *Pseudomonas*, *Ralstonia*, *Rhodococcus* et *Vibrio* phage. Plusieurs singletons et groupes ne contenant que des virus à gradient de pH n'ont été liés à aucun génome viral de référence, ce qui indique que ces virus du sol sont nouveaux pour la plupart (Figure 2.1). Bien que la plupart des mVCs dérivés du CO_2-SIP aient été regroupés en petits groupes, les mVCs dérivés du CH_4-SIP se sont avérés être regroupés en groupes relativement importants. Cela reflète probablement la plus grande diversité du potentiel d'hôtes dans les communautés dérivées du CO_2-SIP que dans les communautés dérivées du CH_4-SIP, ce qui a donné lieu à de petits groupes viraux uniques par rapport à de grands groupes de virus qui infectent des cellules hôtes similaires. En moyenne, 11.5% de tous les mVCs dérivés du SIP, 56 mVCs dérivés du CO_2-SIP (13%), 77 mVCs dérivés du CH_4-SIP (10%) ont été regroupés avec des VCs, ce qui démontre la présence de virus actifs ou de virus étroitement apparentés à ceux des viromes du sol. Le réseau récapitulatif démontre également les préférences de niche de pH des virus ayant des gènes partagés au sein des virus apparentés spécifiques au pH du sol. Avec

l'augmentation des études de métagnomes du sol et des génomes viraux de référence, l'annotation taxonomique des congénères viraux s'améliorera bientôt.

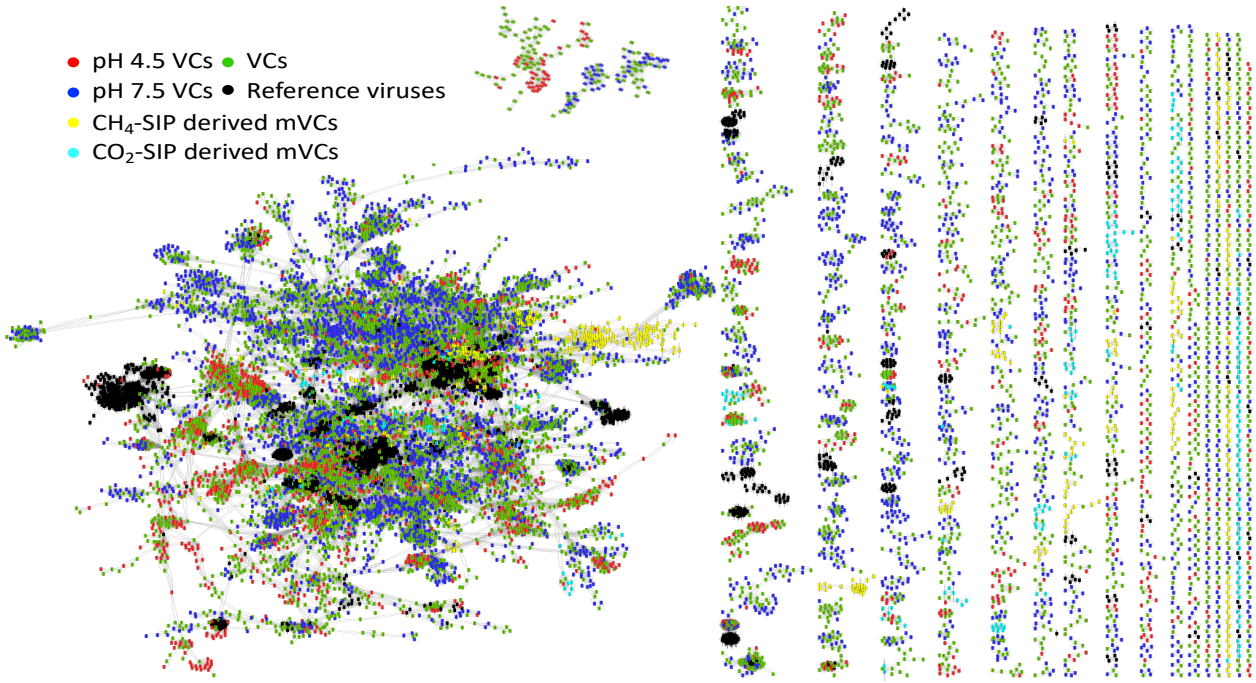


Figure 2.1. Réseau de gènes viraux partagé entre les viromes co-assemblés à gradient de pH (VC, en vert), les contigs viraux à partir de pH 4.5 (VC de pH 4.5, nœuds rouges) et de pH 7.5 (VC de pH 7.5, nœuds bleu), les contigs viraux prédits à partir des métagénomes CH_4-SIP (mVC dérivé de CH_4-SIP, nœuds jaunes) et CO_2-SIP (mVC dérivé de CO_2-SIP, nœuds bleu clair), et les génomes viraux procaryotes RefSeq (virus de référence, nœuds noirs).
Le réseau a été produit en utilisant vConTACT (Bolduc et al. 2017).

2.3. Dynamique des virus-hôtes dans le sol

Les interactions virus-hôte constituent un déterminant majeur de l'évolution et de l'écologie de l'hôte (Mojica et al. 2009 ; Koskella 2014). Lors d'une infection virale, la cellule hôte peut utiliser des mécanismes de défense antiviraux, tels que le système de restriction-modification (RM), le système CRISPR-Cas et les infections abortives (Abi) (Deveau et al. 2010 ; Labrie et al. 2010 ; Stern et Sorek 2011 ; tenOever 2016). Parmi ces derniers, le système CRISPR-Cas peut être utilisé pour déduire des preuves directes de l'interaction virus-hôte par l'analyse des séquences virales (spacers) qui sont incorporées dans les CRISPRs des génomes des hôtes. Comme les CRISPRs fournissent une "mémoire" des séquences d'infection virale, l'analyse de spacers dans les CRISPRs de l'hôte peut être utilisée non seulement pour les liens entre l'hôte et le virus, mais aussi pour déduire les caractéristiques du virus (c'est-à-dire la fréquence d'infection et la gamme d'hôtes du virus) (Deveau et al. 2010 ; Fineran et Charpentier 2012 ; Strich et Chertow 2019). Cependant, lorsqu'ils sont analysés dans le cadre d'une analyse métagénomique, les CRISPRs ne sont souvent

pas trouvés dans les MAG en raison d'un faible niveau de complétion des MAG et des contigs hôtes fragmentés. En outre, bien que les CRISPRs puissent être analysées, les séquences de spacers peuvent ne pas correspondre aux contigs viraux identifiés en raison des limites de l'assemblage du génome viral (par exemple, un séquençage incomplet dû à une grande diversité microbienne) ou de l'évolution rapide des séquences virales (c'est-à-dire la mutation des loci des génomes viraux représentés dans les séquences de spacers ou dans le mécanisme de contre-défense du virus lui-même). L'analyse de la fréquence des oligonucléotides (ONF) peut être utilisée comme une approche alternative pour la liaison hôte-virus, en particulier pour les génomes viraux et hôtes fragmentés qui ne disposent pas de CRISPR (Galiez et al. 2017).

Dans cette thèse, le CRISPR et l'analyse de l'ONF ont été utilisés pour la liaison hôte-virus. Pour l'analyse des CRISPRs, l'outil CRT a été utilisé pour identifier les CRISPRs directement à partir des contigs hôtes (Bland et al. 2007). Cependant, comme cette approche permet également de récupérer de CRISPRs erronés (y compris les répétitions en tandem et les éléments de type STAR), l'outil CRISPRCasFinder a également été utilisé en parallèle pour détecter les gènes Cas afin de confirmer les CRISPRs prédites par l'outil CRT (Couvin et al. 2018).

L'outil WIsH basé sur l'analyse d'ONF peut prédire les hôtes pour des séquences virales courtes (5 kb) et fonctionne avec une bonne précision (Galize et al. 2017). Cependant, les liens peuvent ne pas être étayés par d'autres preuves directes, comme la présence de séquences de spacers dans un CRISPR, ou la présence de gènes homologues partagés. Dans l'analyse des liaisons hôte-virus prévues par le CRISPR, il y avait toujours au moins un gène homologue partagé entre le virus et l'hôte lié. Par conséquent, s'il était exact, il semblerait logique qu'un lien par homologie de gènes soit également reflété dans les liaisons hôte-virus prévues par le WIsH et une analyse supplémentaire des homologues de gènes a été effectuée afin de valider les liaisons hôte-virus obtenues par le WIsH. Par exemple, dans l'étude des métagénomes CH_4-SIP, sur 245 liaisons hôte-virus obtenues par la WIsH, 157 liaisons partageaient au moins un homologue. Cependant, il est intéressant de noter qu'avec les liens définis par l'analyse de CRISPRs (qui seraient considérés comme "de haute confiance"), le WIsH n'a pas prédit le même lien. Par conséquent, l'utilisation des analyses d'ONF devrait peut-être être envisagée avec prudence et nécessiter un traitement supplémentaire.

En associant les liens identifiés entre l'hôte et le virus, une annotation taxonomique de contigs entiers a été utilisée. La comparaison avec la base de données NCBI-nr a été utilisée pour identifier l'éventuelle affiliation taxonomique des gènes dans les contigs viraux, puis comparée à l'affiliation taxonomique du contig et de l'hôte putatif liés (précédemment annotés à l'aide de l'outil Kaiju). Cependant, des erreurs d'annotation des contigs de l'hôte ont parfois été observées et des erreurs de prédiction des contigs viraux ont été causées par un manque d'homologues apparentés dans les bases de données ou par une forte proportion de gènes acquis

horizontalement qui sont partagés entre des espèces apparentées, ce qui peut conduire à de faux critères de validation (Bolotin et Hershberg 2017).

Dans l'ensemble, les résultats présentés dans cette thèse ont démontré des interactions dynamiques virus-hôtes à travers le gradient de pH du sol. Un nombre relativement plus important et une gamme de taille plus large de CRISPRs ont été détectés dans le sol à pH 4.5 par rapport à pH 7.5, ce qui suggère des fréquences d'infection virale contrastées à travers le gradient de pH (Bezuidt et al. 2020). Bien que les liens hôte-virus déterminés par les CRISPRs et les associations prévues par WIsH soient distincts entre les deux sols, ils impliquent en grande partie le même phyla. En outre, il est apparu que les mêmes contiguïtés d'hôtes étaient infectées par plusieurs virus différents, ce qui pourrait entraîner une compétition entre les virus. En particulier, la présence de gènes codant pour l'endonucléase de restriction (REase) dans la plupart des virus (1,842 VC, 20%) a mis en évidence la compétition entre virus (Chapitre III). La REase dans les génomes viraux peut conférer une résistance chez les hôtes à l'infection par d'autres phages qui peuvent également l'infecter (Lossouarn et al. 2019). De même, plusieurs REases ont été retrouvées dans les viromes du sol, ce qui suggère que la compétition virus-virus pourrait être courante dans les environnements du sol. En outre, un certain nombre de méthyltransférases (MTase) ont également été trouvées (446 VC, 5%), interférant potentiellement avec le système de défense antiviral de l'hôte RM, améliorant ainsi l'efficacité d'infection des virus (Labrie et al. 2010 ; Koonin et Krupovic 2020 ; Bezuidt et al. 2020). En général, les gènes codant pour la MTase se trouvent dans 20% de tous les génomes de bactériophages actuellement identifiés, ce qui suggère un rôle important dans l'interaction virus-hôte (Kaltz et Shykoff 1998 ; Murphy et al. 2013).

Les associations hôte-virus des procaryotes C et N-cyclant ont été identifiées en utilisant à la fois le CRISPR array et l'analyse d'ONF. Plus précisément, les virus interagissant avec les méthanotrophes (6 mVCs via le CRISPR ; 64 mVCs via le WIsH), les méthylotrophes (7 mVC via le WIsH) et les nitrifiants (4 mVCs associés aux AOA et 3 mVCs associés aux NOB via le WIsH) ont été identifiés dans les métagénomes CH_4- et CO_2-SIP, respectivement. L'identification des virus associés aux méthanotrophes obtenue par l'analyse de CRISPRs a démontré une interaction active dans le système du sol par la présence de séquences de spacers dans les MAG de Methylosinus et *Methylocystis* sp. En effet, deux des virus de la gamme d'hôtes infectent deux hôtes de Methylocystis, alors que d'autres virus sont spécifiques à un seul hôte. De plus, un virus semblait infecter plus fréquemment que les autres en raison du plus grand nombre de spacers trouvés dans les CRISPRs. Cependant, l'analyse des MAGs dérivés des métagénomes du sol natif (Chapitre II) ou du CO_2-SIP (Chapitre V) ne contenait pas de CRISPRs, et cette approche n'a donc pas pu être utilisée pour établir un lien avec les virus. Alors que l'analyse WIsH a révélé un grand nombre de virus associés aux méthanotrophes et aux méthylotrophes (64 et 7 mVCs, respectivement) à partir des métagénomes CH_4-SIP, seuls sept virus associés aux nitrifiants ont été identifiés.

La sélection de populations exclusivement enrichies en ^{13}C pour réduire la diversité microbienne et l'aide de l'outil DeepVirFinder ont permis de découvrir huit autres virus associés aux nitrifiants, mais ceux-ci étaient dérivés d'une très petite fraction des génomes viraux. L'abondance des mVCs associés aux méthanotrophes dans les métagénomes CH_4-SIP pourrait refléter l'abondance des communautés méthanotrophes enrichies dans ces métagénomes CH_4-SIP.

Le profilage de l'abondance a démontré la spécificité de l'habitat des hôtes et des virus le long d'un gradient de pH, comme l'ont montré d'autres études antérieures (Nicol et al. 2008 ; Adriaenssens et al. 2017). Toutefois, de manière inattendue, la comparaison des structures des communautés des virus et des hôtes procaryotes entre les sols de pH 4.5 et 7.5 a indiqué que la structure de la communauté virale était comparativement plus distincte, ce qui suggère que les virus n'ont pas les mêmes aires de répartition que leurs hôtes, et que, bien que limité par la structure de la communauté hôte, le pH du sol lui-même peut avoir un effet direct sur les populations virales.

2.4. Gammes d'hôtes du virus

Dans les analyses effectuées dans le cadre de cette thèse, la plupart des interactions sont apparues relativement spécifiques. Dans l'analyse des virus associés à une souche de Bacillus (Chapitre III), bien qu'il y ait eu une gamme diverse de virus infectants (génétique et morphologique), la comparaison avec les virus de référence a montré qu'ils étaient tous liés à d'autres virus Bacillus et représentaient donc probablement des virus spécifiques au *Bacillus*. Dans l'analyse des virus infectant deux souches étroitement apparentées de Methylocystis, trois virus étaient spécifiques à une souche, deux virus infectant les deux. Il se peut que la spécificité de l'hôte étroit soit courante dans le sol. Par exemple, dans une analyse des phages infectant le *Rhizobium* dans le sol de la rhizosphère (Santamaría et al. 2014), il a été démontré qu'ils présentaient une gamme étroite d'hôtes pour infecter 48 *Rhizobium* sp. testés. Cependant, la méthode de détermination de la gamme d'hôtes est limitée par le fait que tous les hôtes ne génèrent pas de plaques et qu'il n'existe pas de norme pour le nombre de souches ou d'espèces à tester (Ross et al. 2016). En général, on pense que les virus à gamme d'hôtes étroite sont répandus lorsque leurs hôtes sont abondants, alors que les virus à gamme d'hôtes large sont supposés infecter des hôtes peu abondants (Woolhouse et al. 2001 ; Sullivan et al. 2003 ; Elena et al. 2009 ; Dekel-Bird et al. 2015 ; Doron et al. 2016). Il pourrait en être de même pour les virus du sol. Cependant, comme les ensembles de données métagénomiques représentent souvent les organismes les plus abondants dans un échantillon (Rodriguez-R et Konstantinidis 2014), l'analyse des liaisons hôte-virus est effectuée en utilisant les organismes abondants, des virus à gamme d'hôtes étroite sont probablement sélectionnés. En outre, l'évolution rapide des séquences d'espacement au sein des génomes viraux peut entraîner l'absence de correspondance entre les spacers et les génomes viraux, ce qui peut

réduire la capacité à déterminer la liaison (Koonin et Krupovic 2020). Toutefois, l'analyse génomique des virus méthanotrophes liés au CRISPR a démontré que le nombre de gènes homologues entre le virus et l'hôte semble être lié à la spécificité de leur gamme d'hôtes. Les virus à spécificité étroite peuvent partager un plus grand nombre de gènes homologues à leurs cellules hôtes méthanotrophes ancestrales spécifiques, ce qui permet la spécificité étroite du virus alors que les virus à spécificité large semblent contenir une variété de gènes homologues à ceux trouvés dans d'autres méthanotrophes de souches ou de niveaux de genre différents. Néanmoins, comme les génomes du virus et de l'hôte n'étaient pas complets dans nos analyses et que la gamme d'hôtes a été définie par des preuves bioinformatiques (par exemple, la présence des spacers d'un virus dans différents hôtes), il n'est pas évident de constater dans quelle mesure cette hypothèse est étayée par les preuves présentées ici. Des travaux ultérieurs pourraient comporter des expériences en culture pour caractériser les virus à gamme d'hôtes large et étroite, ainsi que l'analyse des génomes complets de l'hôte et du virus pour mieux comprendre les caractéristiques génomiques définissant la gamme d'hôtes.

2.5. Gènes métaboliques auxiliaires

Les virus du sol contribuent au cycle biogéochimique en augmentant le potentiel métabolique de l'hôte après une infection virale par l'expression des AMG codées par le virus (Breitbart et al. 2007). Les AMG dérivées de l'hôte peuvent être acquises de leur hôte précédent immédiat ou d'hôtes plus lointains (Sharon et al. 2009 ; Sullivan et al. 2010 ; Kelly et al. 2013 ; Crummett et al. 2016). L'analyse des viromes et des virus dérivés du SIP ont permis d'identifier un grand nombre d'AMG impliquées dans le métabolisme du carbone, telles que les hydrolases et les peptidases de glycosides, qui ont été observées précédemment dans les virus du sol (Emerson et al. 2018 ; Trubl et al. 2018 ; Graham et al. 2019). Dans cette étude, les virus associés aux méthanotrophes contenaient des AMG codant pour des protéines impliquées dans l'oxydation du méthane (*pmoC* et gènes liés au cytochrome). Récemment, plusieurs gènes *pmoC* codés par des virus provenant de lacs d'eau douce ont été identifiés (Chen et al. 2020), des données transcriptomiques démontrant leur activité potentielle et un rôle possible des virus méthanotrophes modulant l'efflux de CH_4 en modifiant les taux d'oxydation du méthane des méthanotrophes infectés par des phages porteurs de *pmoC* (Chen et al. 2020). De même, une étude récente a signalé que les gènes *amoC* associés à Thaumarchaeota sont largement répandus dans les virus provenant de milieux marins (Roux et al. 2016 ; López-Pérez et al. 2019 ; Ahlgren et al. 2019). Bien que les gènes *pmoC* et *amoC* codés par des virus se soient avérés abondants dans les environnements aquatiques, un seul *pmoC* codé par un virus (et aucun *amoC* codé par un virus) a été trouvé dans cette étude. Une plus grande profondeur de séquençage ou des approches augmentant la proportion de virus associés aux nitrifiants (par exemple, le séquençage de fractions individuelles dans l'ADN-SIP

dont on sait qu'elles contiennent des quantités relativement élevées de gènes *Amo*) pourraient faciliter l'obtention de virus complets associés aux méthanotrophes ou aux nitrifiants. Bien que les virus infectant les méthanotrophes et les nitrifiants soient très distincts, il est intéressant de noter que les virus de chaque groupe fonctionnel acquièrent de préférence les gènes *pmoC* ou *amoC* (plutôt que *pmo*/*amoA* ou *pmo*/*amoB*) pour améliorer l'aptitude de leurs hôtes. Potentiellement, les virus associés aux méthanotrophes et aux nitrifiants pourraient influencer les taux d'oxydation du méthane et de nitrification dans le sol par la lyse des hôtes ou l'expression des AMG codés par le virus (par exemple les gènes *amoC* et *pmoC*), respectivement.

En général, l'identité de séquence entre donneur et receveur dans le transfert horizontal de gènes est initialement de 100%, mais elle diminue progressivement au fil du temps en raison des changements génétiques, tels que les mutations, les duplications de gènes et les réarrangements génomiques (Ochman et al. 2000 ; Raz et Tannenbaum 2010 ; Ku et Martin 2016). Ainsi, les transferts de gènes récents ont tendance à avoir un degré d'identité plus élevé avec les homologues de la lignée du donneur (Shoemaker et al. 2001 ; Smillie et al. 2011 ; Ku et Martin 2016). Les homologues partagés entre un hôte et un virus avaient généralement une identité plus élevée lorsqu'ils codaient pour des protéines impliquées dans la réplication du virus, ce qui suggère un transfert horizontal de gènes plus récent. Les homologues codant pour d'autres fonctions étaient généralement plus variables (30 à 90 %).

2.6. Perspectives des expériences

Entre les deux sols de pH contrastés, la diversité et la richesse virales étaient relativement plus importantes à un pH de 7.5. Cependant, il est important de considérer que les données présentées dans cette thèse ne reflètent pas la diversité virale complète. Les deux méthodes de récupération et de séquençage des virus utilisées ne détectent que les virus à ADN et excluent la détection des virus à ARN. Les procédures utilisées seraient aussi probablement biaisées en faveur des virus qui sont facilement extraits, et les interactions virus-matière organique varieront probablement avec un changement de pH du sol (Dowd et al. 1998 ; Lukasik et al. 2000 ; Chu et al. 2003 ; Zhao et al. 2008 ; Chen et al. 2014).

Une approche de test en plaque combinée à la métagénomique a permis de mettre en évidence les effets des interactions co-évolutives entre les populations d'hôtes et de virus, comme la présence de mécanismes antiviraux de l'hôte (par exemple RM et CRISPR-Cas-système) et de mécanismes de contre-défense du virus (par exemple mutation des spacers, présence des méthyltransférases). En outre, la diversité et l'infectiosité des populations de virus (par exemple le nombre d'UFP) se sont avérées plus importantes lorsque les populations de virus provenaient d'une niche différente du sol. Cependant, une seule souche bactérienne isolée du sol au pH 7.5 a été utilisée dans l'expérience. L'approche du test en plaque repose sur des souches bactériennes

cultivables, et sur celles qui sont capables de former des monocouches de cellules confluentes et qui sont sensibles aux infections virales. Cela conduit à une limitation de la capacité d'analyse de la grande majorité non cultivée. L'utilisation d'isolats bactériens supplémentaires de la même espèce et de ceux de différents genres pour tester s'il y a une observation réciproque pour une souche bactérienne isolée du sol à pH 4.5 est un objectif futur. L'enrichissement du virus et l'isolement bactérien ont utilisé un composite tamisé (0.2 g et 1 g, respectivement) de sols intacts (50 g). Bien que le tamisage du sol contribue à réduire l'hétérogénéité spatiale en éliminant les racines et les pierres et permette de produire des échantillons homogènes représentatifs, il peut réduire les multiples micro-habitats abritant des microbes qui sont associés aux agrégats du sol. Comme les agrégats de sol sont connus comme des micro-habitats qui favorisent une évolution microbienne parallèle, cela peut représenter un hot-spot pour les interactions virus-hôtes (Rilling et al. 2017 ; Pratama et al. 2018). Il serait donc plus réaliste de maintenir la structure naturelle du sol et de comparer les virus dans des agrégats individuels à différentes distances et à différentes échelles. Les agrégats de sol de la rhizosphère (par exemple, le sol d'adhésion des racines) seraient un modèle intéressant pour examiner les populations virales qui contrôlent les micro-organismes du sol par rapport à celles que l'on trouve dans les agrégats provenant du sol en vrac (par exemple, les agrégats de sol sans racines).

Le DNA-SIP combiné à un séquençage profond a permis de découvrir divers virus actifs enrichis en ^{13}C, et a permis une analyse relativement complète d'un réseau alimentaire microbien-viral en suivant l'oxidation du ^{13}C-methane. Même si ces analyses ont permis d'identifier une interaction récente à un moment précis, elles n'ont pas examiné les taux d'infection ou les changements réels dans les CRISPRs, ni les changements temporels dans la population et le nombre de virus associés. Les échantillons n'ont été séquencés qu'après 30 jours d'incubation dans le sol, et non périodiquement, ce qui ne permet pas de tirer des conclusions sur le rythme d'évolution du CRISPR. Ainsi, une série chronologique sur une période plus longue permettrait d'analyser les taux d'évolution du CRISPR, et les taux d'infection des populations virales associées à différents niveaux trophiques et à des sources croisées pourraient être examinés. Dans l'expérience CH_4-SIP, une concentration élevée de méthane (1%) a été utilisée en comparaison avec des concentrations atmosphériques de méthane de 0.00017%. Dans l'environnement complexe du sol, dans un sol bien drainé, les microsites avec de faibles concentrations d'oxygène ou d'anoxie (par exemple après une pluie) peuvent permettre des conditions qui entraînent une méthanogénèse et des concentrations élevées de méthane dans les microsites. Les conditions expérimentales utilisées ont probablement entraîné la croissance de méthanotrophes de faible affinité, et ont permis d'examiner les méthanotrophes de forte affinité qui sont importantes pour l'élimination de grandes quantités de méthane atmosphérique (Kneif et al. 2005). L'utilisation de différentes concentrations de méthane permettrait d'examiner d'autres membres de la

communauté des méthanotrophes. Dans l'expérience CO_2-SIP, la limitation évidente était l'absence de séparation de l'ADN entièrement enrichi en $^{13}CO_2$ en raison du chevauchement avec l'ADN à haut %GC non marqué des bactéries hétérotrophes. Le séquençage individuel de chaque fraction d'ADN et celui d'une extraction virale à partir des microcosmes de sol enrichis en ^{13}C permettrait de surmonter ces limitations. Une autre solution consisterait à étudier les virus associés aux nitrifiants en utilisant des cultures de nitrifiants (AOA, AOB et NOB) selon une approche basée sur la culture.

2.7. Conclusion

Ce travail a permis d'élargir nos connaissances sur les virus du sol. Les structures des communautés virales étaient étroitement limitées par leurs hôtes procaryotes et les virus semblaient avoir une gamme étroite d'hôtes. Le suivi du flux de carbone de l'hôte au virus a permis une analyse à haute résolution des communautés virales et a même permis d'identifier les interactions individuelles *in situ*. Cette approche peut donc convenir pour comprendre l'importance fonctionnelle et la dynamique des virus associés à de nombreux stades différents des cycles biogéochimiques dans le sol grâce à l'utilisation de substrats appropriés enrichis en isotopes.

References

Abby SS, Melcher M, Kerou M, et al (2018) *Candidatus* Nitrosocaldus cavascurensis, an ammonia oxidizing, extremely thermophilic archaeon with a highly mobile genome. Frontiers in Microbioloy 9:28.

Abdulrasheed M, Ibrahim HI, Maigari FU, et al (2018) Effect of soil pH on composition and abundance of nitrite-oxidizing bacteria. Journal of Biochemistry, Microbiology and Biotechnology 6:27–34.

Abedon ST (2012) Bacterial 'immunity' against bacteriophages. Bacteriophage 2:50–54.

Ackerman SH, Hofer MA, Weiner H (1978) Early maternal separation increases gastric ulcer risk in rats by producing a latent thermoregulatory disturbance. Science 201:373–376.

Ackermann H-W (1998) Tailed bacteriophages: The order caudovirales. In: Maramorosch K, Murphy FA, Shatkin AJ (eds) Advances in virus research volume 98. Elsevier Academic Press, Cambridge, pp. 135–201.

Adamsen AP, King GM (1993) Methane consumption in temperate and subarctic forest soils: rates, vertical zonation, and responses to water and nitrogen. Applied and Environmental Microbiology 59:485–490.

Adriaenssens EM, Brister JR (2017) How to name and classify your phage: an informal guide. Viruses 9:70.

Adriaenssens EM, Kramer R, Van Goethem MW, et al (2017) Environmental drivers of viral community composition in Antarctic soils identified by viromics. Microbiome 5:83.

Ahkami AH, Allen White R, Handakumbura PP, Jansson C (2017) Rhizosphere engineering: enhancing sustainable plant ecosystem productivity. Rhizosphere 3:233–243.

Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA (2019) Discovery of several novel, widespread, and ecologically distinct marine *Thaumarchaeota* viruses that encode amoC nitrification genes. The ISME Journal 13:618–631.

Ahlgren NA, Ren J, Lu YY, et al (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Research 45:39–53.

Aislabie J, Deslippe JR (2013) Soil microbes and their contribution to soil services. In: Dymond JR (eds) Ecosystem services in New Zealand: conditions and trends, Manaaki Whenua Press, New Zealand, pp 143–161.

Albers S-V (2016) Extremophiles: life at the deep end. Nature 538:457–457.

Albertsson P-åke, Frick G (1960) Partition of virus particles in a liquid two-phase system. Biochimica et Biophysica Acta 37:230–237.

Alneberg J, Bjarnason BS, de Bruijn I, et al (2014) Binning metagenomic contigs by coverage and composition. Nature Methods 11:1144–1146.

Andersson AF, Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320:1047–1050.

Angly FE, Felts B, Breitbart M, et al (2006) The marine viromes of four oceanic regions. PLOS Biology 4:e368.

Anthony C (1986) Bacterial oxidation of methane and methanol. In: Rose AH, Tempest DW (eds) Advances in microbial physiology volume 27. Elsevier Academic Press, Cambridge, pp 113–210

Aronson E, Allison S, Helliker BR (2013) Environmental impacts on the diversity of methane-cycling microbes and their resultant function. Frontiers in Microbiology 4:225.

Ashelford KE, Day MJ, Fry JC (2003) Elevated abundance of bacteriophage infecting *bacteria* in soil. Applied and Environmental Microbiology 69:285–289.

Aswad A, Katzourakis A (2018) Cell-derived viral genes evolve under stronger purifying selection in Rhadinoviruses. Journal of Virology 92:19.

Ayala-Castro C, Saini A, Outten FW (2008) Fe-S cluster assembly pathways in *bacteria*. Microbiology and Molecular Biology Reviews 72:110–125.

Baker M (2012) De novo genome assembly: what every biologist should know. Nature Methods 9:333–337.

Ballaud F, Dufresne A, Francez A-J, et al (2016) Dynamics of viral abundance and diversity in a sphagnum-dominated peatland: temporal fluctuations prevail over habitat. Frontiers in Microbiology 6:e1494.

Bankevich A, Nurk S, Antipov D, et al (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455–477.

Baquy MA-A, Li J-Y, Jiang J, et al (2018) Critical pH and exchangeable Al of four acidic soils derived from different parent materials for maize crops. Journal of Soils and Sediments 18:1490–1499.

Bartram AK, Jiang X, Lynch MDJ, et al (2014) Exploring links between pH and bacterial community composition in soils from the Craibstone Experimental Farm. FEMS Microbiology Ecology 87:403–415.

Baumann K, Dignac M-F, Rumpel C, et al (2013) Soil microbial diversity affects soil organic matter decomposition in a silty grassland soil. Biogeochemistry 114:201–212.

Benstead J, King GM (2001) The effect of soil acidification on atmospheric methane uptake by a maine forest soil. FEMS Microbiology Ecology 34:207–212.

Bertilsson S, Burgin A, Carey CC, et al (2013) The under-ice microbiome of seasonally frozen lakes. Limnology and Oceanography 58:1998–2012.

Best A, White A, Boots M (2009) The implications of coevolutionary dynamics to host-parasite interactions. The American Naturalist 173:779–791.

Bettstetter M, Peng X, Garrett RA, Prangishvili D (2003) AFV1, a novel virus infecting hyperthermophilic *archaea* of the genus *acidianus*. Virology 315:68–79.

Bezuidt OKI, Lebre PH, Pierneef R, et al (2020) Phages actively challenge niche communities in Antarctic soils. American Society for Microbiology 5:e234-20.

Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas systems in *bacteria* and *archaea*: versatile small RNAs for adaptive defense and regulation. Annual Review of Genetics 45:273–297.

Bi L, Yu D-T, Du S, et al (2020) Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. Environmental Microbiology 'in press'.

Bin Jang H, Bolduc B, Zablocki O, et al (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nature Biotechnology 37:632–639.

Bland C, Ramsey TL, Sabree F, et al (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.

Bock E, Koops H-P, Möller UC, Rudert M (1990) A new facultatively nitrite oxidizing bacterium, *Nitrobacter vulgaris* sp. nov. Archives of Microbiology 153:105–110.

Bock E, Sundermeyer-Klinger H, Stackebrandt E (1983) New facultative lithoautotrophic nitrite-oxidizing bacteria. Archives of Microbiology 136:281–284.

Bohannan BJM, Kerr B, Jessup CM, et al (2002) Trade-offs and coexistence in microbial microcosms. Antonie van Leeuwenhoek 81:107–115.

Bolduc B, Jang HB, Doulcier G, et al (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *archaea* and *bacteria*. PeerJ 5:e3243.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bolotin E, Hershberg R (2017) Horizontally acquired genes are often shared between closely related bacterial species. Frontiers in Microbiology 8:e1536

Bourne DG, McDonald IR, Murrell JC (2001) Comparison of pmoA PCR primer sets as tools for investigating methanotroph diversity in three Danish soils. Applied and Environmental Microbiology 67:3802–3809.

Bowman JP, Sly LI, Cox JM, Hayward AC (1990) *Methylomonas fodinarum* sp. nov. and *Methylomonas aurantiaca* sp.nov.: two closely related type I obligate methanotrophs. Systematic and Applied Microbiology 13:279–287.

Branton D, Deamer DW, Marziali A, et al (2008) The potential and challenges of nanopore sequencing. Nature Biotechnology 26:1146–1153.

Breitbart M, Hewson I, Felts B, et al (2003) Metagenomic analyses of an uncultured viral community from human feces. Journal of Bacteriology 185:6220–6223.

Breitbart M, Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? Trends in Microbiology 13:278–284.

Breitbart M, Salamon P, Andresen B, et al (2002) Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences 99:14250–14255.

Breitbart M, Thompson LR, Suttle CA, Sullivan MB (2007) Exploring the vast diversity of marine viruses. Oceanography 20:135–139

Brum JR, Ignacio-Espinoza JC, Roux S, et al (2015) Patterns and ecological drivers of ocean viral communities. Science 348:6237.

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

Buckling A, Rainey PB. (2002) Antagonistic coevolution between a bacterium and a bacteriophage. Proceedings of the Royal Society of London Series B, Biological Sciences 269:931–936.

Burz DS, Beckett D, Benson N, Ackers GK (1994) Self-assembly of bacteriophage lambda cI repressor: effects of single-site mutations on the monomer-dimer equilibrium. Biochemistry 33:8399–8405.

Canchaya C, Fournous G, Chibani-Chennoufi S, et al (2003a) Phage as agents of lateral gene transfer. Current Opinion in Microbiology 6:417–424.

Canchaya C, Proux C, Fournous G, et al (2003b) Prophage genomics. Microbiology and Molecular Biology Reviews 67:238–276.

Canfield DE, Glazer AN, Falkowski PG (2010) The evolution and future of earth's nitrogen cycle. Science 330:192–196.

Carbone A (2008) Codon bias is a major factor explaining phage evolution in translationally biased hosts. Journal of Molecular Evolution 66:210–223.

Casas V, Rohwer F (2007) Phage metagenomics. Methods in Enzymology. 421:259-68.

Chain P, Lamerdin J, Larimer F, et al (2003) Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. Journal of Bacteriology 185:2759–2773.

Chan Y, Nostrand JDV, Zhou J, et al (2013) Functional ecology of an Antarctic dry valley. Proceedings of the National Academy of Sciences 110:8990–8995.

Chaparro JM, Sheflin AM, Manter DK, Vivanco JM (2012) Manipulating the soil microbiome to increase soil health and plant fertility. Biology and Fertility of Soils 48:489–499.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36:1925–1927.

Chen L, Xun W, Sun L, et al (2014) Effect of different long-term fertilization regimes on the viral community in an agricultural soil of Southern China. European Journal of Soil Biology 62:121–126.

Chen Y, Dumont MG, Neufeld JD, et al (2008) Revealing the uncultivated majority: combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated *Methylocystis* in acidic peatlands. Environmental Microbiology 10:2609–2622.

Chen L-X, Méheust R, Crits-Christoph A, et al (2020) Large freshwater phages with the potential to augment aerobic methane oxidation. Nature Microbiology 1–12.

Chibani CM, Farr A, Klama S, et al (2019) Classifying the unclassified: a phage classification method. Viruses 11(2):195.

Chibani-Chennoufi S, Bruttin A, Dillmann M-L, Brüssow H (2004) Phage-host interaction: an ecological perspective. Journal of Bacteriology 186:3677–3686.

Choi J, Kotay SM, Goel R (2010) Various physico-chemical stress factors cause prophage induction in *Nitrosospira multiformis* 25196-an ammonia oxidizing bacteria. Water Research 44:4550–4558.

Chu Y, Jin Y, Baumann T, Yates MV (2003) Effect of soil properties on saturated and unsaturated virus transport through columns. Journal of Environmental Quality 32:2017–2025.

Clokie MRJ, Millard AD, Letarov AV, Heaphy S (2011) Phages in nature. Bacteriophage 1:31–45.

Colombet J, Sime-Ngando T (2012) Use of PEG, polyethylene glycol, to characterize the diversity of environmental viruses. In: Méndez-Vilas A, (eds) Current microscopy contributions to advances in science and technology. Formatex Research Center, Spain, pp 316–322.

Conrad R (2009) The global methane cycle: recent advances in understanding the microbial processes involved. Environmental Microbiology Reports 1:285–292.

Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology 12:263–273.

Coulibaly ST, Rossolillo P, Winter F, et al (2015) Potent sensitisation of cancer cells to anticancer drugs by a quadruple mutant of the human deoxycytidine kinase. PLOS ONE 10:e140741.

Couvin D, Bernheim A, Toffano-Nioche C, et al (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research 46:246–251.

Cretu Stancu M, van Roosmalen MJ, Renkens I, et al (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Communications 8:1326.

Crummett LT, Puxty RJ, Weihe C, et al (2016) The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology 499:219–229.

Daims H, Lebedeva EV, Pjevac P, et al (2015) Complete nitrification by *Nitrospira* bacteria. Nature 528:504–509.

Dam B, Dam S, Kube M, et al (2012) Complete genome sequence of *Methylocystis* sp. Strain SC2, an aerobic methanotroph with high-affinity methane oxidation potential. Journal of Bacteriology 194:6008–6009.

Danovaro R, Dell'Anno A, Corinaldesi C, et al (2016) Virus-mediated archaeal hecatomb in the deep seafloor. Science Advances 2:e1600492.

de Gannes V, Eudoxie G, Hickey WJ (2014) Impacts of edaphic factors on communities of ammonia-oxidizing archaea, ammonia-oxidizing bacteria and nitrification in tropical soils. PLOS ONE 9:e89568

de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE (2019) Molecular and evolutionary determinants of bacteriophage host range. Trends in Microbiology 27:51–63.

Dedysh SN (2011) Cultivating uncultured bacteria from northern wetlands: knowledge gained and remaining gaps. Frontiers in Microbiology 2:184.

Dedysh SN, Knief C (2018) Diversity and phylogeny of described aerobic methanotrophs. In: Kalyuzhnaya MG, Xing X-H (eds) Methane biocatalysis: paving the way to sustainability. Springer, Cham, pp 17–42.

Dedysh SN, Panikov NS, Tiedje JM (1998) Acidophilic methanotrophic communities from sphagnum peat bogs. Applied and Environmental Microbiology 64:922–929

Dekel-Bird NP, Sabehi G, Mosevitzky B, Lindell D (2015) Host-dependent differences in abundance, composition and host range of cyanophages from the Red Sea. Environmental Microbiology 17:1286–1299.

Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annual Review of Microbiology 64:475–493.

Di HJ, Cameron KC, Shen JP, et al (2009) Nitrification driven by *bacteria* and not *archaea* in nitrogen-rich grassland soils. Nature Geoscience 2:621–624.

Dijkhuizen L, Levering PR, de Vries GE (1992) The physiology and biochemistry of aerobic methanol-utilizing gram-negative and gram-positive bacteria. In: Murrell JC, Dalton H (eds) Methane and Methanol Utilizers. Springer, New York, pp 149–181.

Dolinšek J, Lagkouvardos I, Wanek W, et al (2013) Interactions of nitrifying bacteria and heterotrophs: identification of a *Micavibrio*-like putative predator of *Nitrospira* spp. Applied and Environmental Microbiology 79:2027–2037.

Doron S, Fedida A, Hernández-Prieto MA, et al (2016) Transcriptome dynamics of a broad host-range cyanophage and its hosts. The ISME Journal 10:1437–1455.

Dowd SE, Pillai SD, Wang S, Corapcioglu MY (1998) Delineating the specific influence of virus isoelectric point and size on virus adsorption and transport through sandy soils. Applied and Environmental Microbiology 64:405–410.

Dryden DTF, Murray NE, Rao DN (2001) Nucleoside triphosphate-dependent restriction enzymes. Nucleic Acids Research 29:3728–3741.

Dupuis M-È, Villion M, Magadán AH, Moineau S (2013) CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nature Communications 4:2087.

Dutaur L, Verchot LV (2007) A global inventory of the soil CH4 sink. Global Biogeochemical Cycles 21:e4013.

Edwards RA, Rohwer F (2005) Viral metagenomics. Nature Reviews Microbiology 3:504–510.

Edwards U, Rogall T, Blöcker H, et al (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Research 17:7843–7853.

Elena SF, Agudelo-Romero P, Lalić J (2009) The evolution of viruses in multi-host fitness landscapes. Open Virology Journal 3:1–6.

Elsas JD van, Turner S, Bailey MJ (2003) Horizontal gene transfer in the phytosphere. New Phytologist 157:525–537.

Emerson JB (2019) Soil viruses: A new hope. Applied and Environmental Science 4(3):e120-19.

Emerson JB, Roux S, Brum JR, et al (2018) Host-linked soil viral ecology along a permafrost thaw gradient. Nature Microbiology 3:870–880.

Eren AM, Esen ÖC, Quince C, et al (2015) Anvi'o: an advanced analysis and visualization platform for omics data. PeerJ 3:e1319.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175–185.

Fauquet CM, Mayo MA, Maniloff J, et al (2005) Virus taxonomy: VIIIth report of the international committee on taxonomy of viruses. Virology division international union of microbiological societies, Elsevier Academic Press, Cambridge, 1162 pp.

Feiner R, Argov T, Rabinovich L, et al (2015) A new perspective on lysogeny: prophages as active regulatory switches of *bacteria*. Nature Reviews Microbiology 13:641–650.

Feng S, Tan CH, Constancias F, et al (2017) Predation by *Bdellovibrio bacteriovorus* significantly reduces viability and alters the microbial community composition of activated sludge flocs and granules. FEMS Microbiology Ecology 93:4.

Feng Y, Zhang Y, Ying C, et al (2015) Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13:4–16.

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.

Filippini M, Middelboe M (2007) Viral abundance and genome size distribution in the sediment and water column of marine and freshwater ecosystems. FEMS Microbiology Ecology 60:397–410.

Fineran PC, Charpentier E (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology 434:202–209.

Fogg PCM, Colloms S, Rosser S, et al (2014) New applications for phage integrases. Journal of Molecular Biology 426:2703–2716.

Freitag TE, Chang L, Clegg CD, Prosser JI (2005) Influence of inorganic nitrogen management regime on the diversity of nitrite-oxidizing bacteria in agricultural grassland soils. Applied and Environmental Microbiology 71:8323–8334.

Frias MJ, Melo-Cristino J, Ramirez M (2009) The autolysin LytA contributes to efficient bacteriophage progeny release in *Streptococcus pneumoniae*. Journal of Bacteriology 191:5428–5440.

Fu L, Niu B, Zhu Z, et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399:541–548.

Fuse H, Ohta M, Takimura O, et al (1998) Oxidation of trichloroethylene and dimethyl sulfide by a marine *Methylomicrobium* strain containing soluble methane monooxygenase. Bioscience, Biotechnology, and Biochemistry 62:1925–1931.

Galiez C, Siebert M, Enault F, et al (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics 33:3113–3114.

Gallet R, Kannoly S, Wang I-N (2011) Effects of bacteriophage traits on plaque formation. BMC Microbiology 11:181.

Gandon S, Buckling A, Decaestecker E, Day T (2008) Host–parasite coevolution and patterns of adaptation across time and space. Journal of Evolutionary Biology 21:1861–1866.

Garbeva P, Baggs EM, Prosser JI (2007) Phylogeny of nitrite reductase (nirK) and nitric oxide reductase (norB) genes from *Nitrosospira* species isolated from soil. FEMS Microbiology Letters 266:83–89.

García-López R, Pérez-Brocal V, Moya A (2019) Beyond cells – The virome in the human holobiont. Microbial Cell 6:373–396.

Gardner JG, Escalante-Semerena JC (2009) In *Bacillus subtilis*, the sirtuin protein deacetylase, encoded by the srtN gene (formerly yhdZ), and functions encoded by the acuABC genes control the activity of acetyl coenzyme A synthetase. Journal of Bacteriology 191:1749–1755.

Gelderblom HR (1996) Structure and classification of viruses. In: Baron S (eds) Medical microbiology 4th editions. The University of Texas Medical Branch at Galveston, Galveston, Chapter 41.

Golais F, Hollý J, Vítkovská J (2013) Coevolution of *bacteria* and their viruses. Folia Microbiologica (Praha) 58:177–186.

Goordial J, Davila A, Greer CW, et al (2017) Comparative activity and functional ecology of permafrost soils and lithic niches in a hyper-arid polar desert. Environmental Microbiology 19:443–458.

Gorter FA, Scanlan PD, Buckling A (2016) Adaptation to abiotic conditions drives local adaptation in *bacteria* and viruses coevolving in heterogeneous environments. Biology Letter 12:20150879.

Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. Molecular Biology and Evolution 4:426–444.

Graham EB, Paez-Espino D, Brislawn C, et al (2019) Untapped viral diversity in global soil metagenomes.BioRxiv 583997.

Graumann P, Wendrich TM, Weber MH, et al (1997) A family of cold shock proteins in *Bacillus subtilis* is essential for cellular growth and for efficient protein synthesis at optimal and low temperatures. Molecular Microbiology 25:741–756.

Greene J, Goldberg RB (1985) Isolation and preliminary characterization of lytic and lysogenic phages with wide host range within the *streptomycetes*. Journal of general microbiology 131:2459–2465.

Greiner T, Moroni A, Van Etten JL, Thiel G (2018) Genes for membrane transport proteins: not so rare in viruses. Viruses 10:456.

Gupta R, Prasad Y (2011) Efficacy of polyvalent bacteriophage P-27/HP to control multidrug resistant *Staphylococcus aureus* associated with human infections. Current Microbiology 62:255–260.

Gubry-Ragin C, Hai B, Quince C, et al. (2011). Niche specialization of terrestrial archaeal ammonia oxidizers. Proceedings of the National Academy of Sciences. 108:21206-21211.

Hagopian DS, Riley JG (1998) A closer look at the bacteriology of nitrification. Aquacultural Engineering 18:223–244.

Halary S, Temmam S, Raoult D, Desnues C (2016) Viral metagenomics: are we missing the giants? Current Opinion in Microbiology 31:34–43.

Han L-L, Yu D-T, Zhang L-M, et al (2017) Genetic and functional diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. Scientific Reports 7:45142.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews 68:669–685.

Hanson RS, Hanson TE (1996) Methanotrophic bacteria. Microbiology and Molecular Biology Reviews 60:439–471.

Hargreaves KR, Flores CO, Lawley TD, Clokie MRJ (2014) Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. American Society for Microbiology 5(5):e1045-13.

Harrison E, Brockhurst MA (2017) Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. BioEssays 39:1700112.

He J-Z, Hu H-W, Zhang L-M (2012) Current insights into the autotrophic thaumarchaeal ammonia oxidation in acidic soils. Soil Biology and Biochemistry 55:146–154.

Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends in Microbiology 8:504–508.

Hernandez-Doria JD, Sperandio V (2018) Bacteriophage transcription factor cro regulates virulence gene expression in enterohemorrhagic *Escherichia coli*. Cell Host Microbe 23:607-617.

Hill GT, Mitkowski NA, Aldrich-Wolfe L, et al (2000) Methods for assessing the composition and diversity of soil microbial communities. Applied Soil Ecology 15:25–36.

Ho A, Angel R, Veraart AJ, et al (2016) Biotic interactions in microbial communities as modulators of biogeochemical processes: methanotrophy as a model system. Frontiers in Microbiology 7:e1285.

Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of *bacteria* and *archaea*. Science 327:167–70.

Howe AC, Jansson JK, Malfatti SA, et al (2014) Tackling soil diversity with the assembly of large, complex metagenomes. Proceedings of the National Academy of Sciences 111:4904–4909.

Huntemann M, Ivanova NN, Mavromatis K, et al (2016) The standard operating procedure of the DOE-JGI metagenome annotation pipeline (MAP v.4). Standards in Genomic Sciences 11:17.

Hurwitz BL, Brum JR, Sullivan MB (2015) Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific ocean virome. The ISME Journal 9:472–484.

Huson DH, Beier S, Flade I, et al (2016) MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLOS Computational Biology 12:e1004957.

Hyatt D, Chen G-L, LoCascio PF, et al (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

Hyman P, Abedon ST (2010) Bacteriophage host range and bacterial resistance. In: Allen L, Sima S, Geoffrey G (eds) Advances in applied microbiology. Elsevier Academic Press, Cambridge, pp 217–248.

Hynes AP, Villion M, Moineau S (2014) Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nature Communications 5:4399.

IPCC (2013) Climate change 2013: the physical science basis, In: T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, et al. (eds) Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, New York, pp. 1–1535.

Iwasaki T (2010) Iron-sulfur world in aerobic and hyperthermoacidophilic *archaea Sulfolobus*. Hindawi Publishing Corporation, Archaea 2010:842639.

Jansen R, Embden JDA van, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology 43:1565–1575.

Jensen EC, Schrader HS, Rieland B, et al (1998) Prevalence of broad-host-range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. Applied and Environmental Microbiology 64:575–580.

Jin M, Guo X, Zhang R, et al (2019) Diversities and potential biogeochemical impacts of mangrove soil viruses. Microbiome 7:58.

Johnke J, Cohen Y, de Leeuw M, et al (2014) Multiple micro-predators controlling bacterial communities in the environment. Current Opinion in Biotechnology 27:185–190.

Jones P, Binns D, Chang H-Y, et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.

Jung M-Y, Park S-J, Kim S-J, et al (2014) A mesophilic, autotrophic, ammonia-oxidizing archaeon of thaumarchaeal group I.1a cultivated from a deep oligotrophic soil horizon. Applied and Environmental Microbiology 80:3645–3655.

Justino MC, Almeida CC, Gonçalves VL, et al (2006) *Escherichia coli* YtfE is a di-iron protein with an important function in assembly of iron-sulphur clusters. FEMS Microbiology Letters 257:278–284.

Kaltz, Shykoff (1998) Local adaptation in host–parasite systems. Heredity 81:361–470.

Kalyuzhnaya MG, Gomez OA, Murrell JC (2019) The methane-oxidizing bacteria (methanotrophs). In: McGenity TJ, Kenneth N. Timmis (eds) Taxonomy, genomics and ecophysiology of hydrocarbon-degrading microbes, Handbook of hydrocarbon and lipid microbiology. Springer, Berlin, pp 1–34.

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27–30.

Kanehisa M, Sato Y (2020) KEGG Mapper for inferring cellular functions from protein sequences. Protein Science 29:28–35.

Karakoç C, Radchuk V, Harms H, Chatzinotas A (2018) Interactions between predation and disturbances shape prey communities. Scientific Reports 8:2968.

Kauffman KM, Polz MF (2018) Streamlining standard bacteriophage methods for higher throughput. Elsevier, MethodsX 5:159–172.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. Ecology Letters 7:1225–1241.

Kelly L, Ding H, Huang KH, et al (2013) Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. The ISME Journal 7:1827–1841.

Khan Mirzaei M, Nilsson AS (2015) Isolation of phages for phage therapy: a comparison of spot tests and efficiency of plating analyses for determination of host range and efficacy. PLOS ONE 10:e118557.

Khot V, Strous M, Hawley AK (2020) Computational approaches in viral ecology. Computational and Structural Biotechnology Journal 18:1605–1612.

Kiedrowski MR, Crosby HA, Hernandez FJ, et al (2014) *Staphylococcus aureus* Nuc2 is a functional, surface-attached extracellular nuclease. PLOS ONE 9:e95574.

Kim J-G, Kim S-J, Cvirkaite-Krupovic V, et al (2019) Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. Proceedings of the National Academy of Sciences 116:15645–15650.

Kim K-H, Chang H-W, Nam Y-D, et al (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Applied and Environmental Microbiology 74:5975–5985.

Kimura M, Jia Z-J, Nakayama N, Asakawa S (2008) Ecology of viruses in soils: past, present and future perspectives. Soil Science and Plant Nutrition 54:1–32.

King AM, Lefkowitz E, Adams MJ, Carstens EB (2011) Virus taxonomy: ninth report of the international committee on taxonomy of viruses. Elsevier

Kits KD, Sedlacek CJ, Lebedeva EV, et al (2017) Kinetic analysis of a complete nitrifier reveals an oligotrophic lifestyle. Nature 549:269–272.

Kleiner M, Hooper LV, Duerkop BA (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics 16:7.

Knief C (2015) Diversity and habitat preferences of cultivated and uncultivated aerobic methanotrophic bacteria evaluated based on pmoA as molecular marker. Frontiers in Microbiology 6:1346

Knief C, Dunfield PF (2005) Response and adaptation of different methanotrophic bacteria to low methane mixing ratios. Environmental Microbiology Reports 7:1307–1317.

Knief C, Lipski A, Dunfield PF (2003). Diversity and activity of methanotrophic bacteria in different upland soils. Applied Environmental Microbiolgy 69:6703-6714.

Knowles B, Silveira CB, Bailey BA, et al (2016) Lytic to temperate switching of viral communities. Nature 531:466–470.

Koch H, van Kessel MAHJ, Lücker S (2019) Complete nitrification: insights into the ecophysiology of comammox *Nitrospira*. Applied Microbiology and Biotechnology 103:177–189.

Kolb S, Knief C, Stubner S, Conrad R (2003) Quantitative detection of methanotrophs in soil by novel pmoA-targeted real-time PCR assays. Applied and Environmental Microbiology 69:2423–2429.

Kono N, Arakawa K (2019) Nanopore sequencing: review of potential applications in functional genomics. Development, Growth & Differentiation 61:316–326.

Koonin EV, Krupovic M (2020) Phages build anti-defence barriers. Nature Microbiology 5:8–9.

Koskella B (2014) Bacteria-phage interactions across time and space: merging local adaptation and time-shift experiments to understand phage evolution. The American Naturalist 184 Suppl 1:S9-21.

Koskella B, Brockhurst MA (2014) Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiology Reviews 38:916–931.

Koskella B, Meaden S (2013) Understanding bacteriophage specificity in natural microbial communities. Viruses 5:806–823.

Krüger DH, Bickle TA (1983) Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. Microbiological Reviews 47:345–360

Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiology and Molecular Biology Reviews 75:610–635.

Krupovic M, Spang A, Gribaldo S, et al (2011) A thaumarchaeal provirus testifies for an ancient association of tailed viruses with *archaea*. Biochemical Society Transactions 39:82–88.

Ku C, Martin WF (2016) A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biology 14:89.

Kuhn JH, Becker S, Ebihara H, et al (2010) Proposal for a revised taxonomy of the family *Filoviridae*: classification, names of taxa and viruses, and virus abbreviations. Archives of Virology 155:2083–2103.

Kuno S, Yoshida T, Kaneko T, Sako Y (2012) Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. Applied and Environmental Microbiology 78:5353–5360.

Kuzyakov Y, Mason-Jones K (2018) Viruses in soil: nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. Soil Biology and Biochemistry 127:305–317.

Labonté JM, Swan BK, Poulos B, et al (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. The ISME Journal 9:2386–2399.

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nature Reviews Microbiology 8:317–327.

Lakay FM, Botha A, Prior BA (2007) Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. Journal of Applied Microbiology 102:265–273.

Lammel DR, Barth G, Ovaskainen O, et al (2018) Direct and indirect effects of a pH gradient bring insights into the mechanisms driving prokaryotic community structures. Microbiome 6:106.

Lance JC, Gerba CP (1984) Virus movement in soil during saturated and unsaturated flow. Applied and Environmental Microbiology 47:335–337

Langdon WB (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Mining 8:1.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Applied and Environmental Microbiology 75:5111–5120.

Lebedeva EV, Alawi M, Fiencke C, et al (2005) Moderately thermophilic nitrifying bacteria from a hot spring of the Baikal rift zone. FEMS Microbiology Ecology 54:297–306.

Lee CG, Watanabe T, Fujita Y, et al (2012) Heterotrophic growth of *cyanobacteria* and phage-mediated microbial loop in soil: examination by stable isotope probing (SIP) method. Soil Science and Plant Nutrition 58:161–168.

Lehtovirta-Morley LE, Stoecker K, Vilcinskas A, et al (2011) Cultivation of an obligate acidophilic ammonia oxidizer from a nitrifying acid soil. Proceedings of the National Academy of Sciences 108:15892–15897.

Leininger S, Urich T, Schloter M, et al (2006) *Archaea* predominate among ammonia-oxidizing prokaryotes in soils. Nature 442:806–809.

Lenski RE, Levin BR (1985) Constraints on the coevolution of *bacteria* and virulent phage: a model, some experiments, and predictions for natural communities. The American Naturalist 125:585–602.

Li C, Hu H, Chen Q-L, et al (2019) Comammox *Nitrospira* play an active role in nitrification of agricultural soils amended with nitrogen fertilizers. Soil Biology and Biochemistry 138:107609.

Li D, Luo R, Liu C-M, et al (2016a) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11.

Li Y, Ding K, Wen X, et al (2016b) A novel ammonia-oxidizing archaeon from wastewater treatment plant: its enrichment, physiological and genomic characteristics. Scientific Reports 6:23747.

Li Y, Sun H, Yang W, et al (2019) Dynamics of bacterial and viral communities in paddy soil with irrigation and urea application. Viruses 11(4):347.

Liang X, Zhang Y, Wommack KE, et al (2020) Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. Soil Biology and Biochemistry 144:107767.

Limpiyakorn T, Fürhacker M, Haberl R, et al (2013) amoA-encoding archaea in wastewater treatment plants: a review. Applied Microbiology and Biotechnology 97:1425–1439.

Lindell D, Sullivan MB, Johnson ZI, et al (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proceedings of the National Academy of Sciences101:11013–11018.

Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The genomes on line database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research 36:475–479.

Lodish H, Berk A, Zipursky SL, et al (2000) Viruses: structure, function, and uses. In: Freeman W H (eds) Molecular cell biology 4th edition. W. H. Freeman and Company, New York, Section 6.3.

López-Pérez M, Haro-Moreno JM, Torre JR de la, Rodriguez-Valera F (2019) Novel caudovirales associated with marine group I *thaumarchaeota* assembled from metagenomes. Environmental Microbiology 21:1980–1988.

Lossouarn J, Briet A, Moncaut E, et al (2019) *Enterococcus faecalis* countermeasures defeat a virulent *Picovirinae* bacteriophage. Viruses 11:48.

Loveland JP, Ryan JN, Amy GL, Harvey RW (1996) The reversibility of virus attachment to mineral surfaces. Colloids and Surfaces A: Physicochemical and Engineering Aspects 107:205–221.

Lukasik J, Scott TM, Andryshak D, Farrah SR (2000) Influence of salts on virus adsorption to microporous filters. Applied and Environmental Microbiology 66:2914–2920.

Madden TL, Tatusov RL, Zhang J (1996) Applications of network BLAST server. In: Methods in enzymology. Elsevier Academic Press, Cambridge, pp 131–141.

Maguire F, Jia B, Gray K, et al (2020) Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. BioRxiv 2020.03.31.997171.

Maher N, Dillon HK, Vermund SH, Unnasch TR (2001) Magnetic bead capture eliminates PCR inhibitors in samples collected from the airborne environment, permitting detection of *Pneumocystis carinii* DNA. Applied Environmental Microbiology 67:449–452.

Malone T, Blumenthal RM, Cheng X (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. Journal of Molecular Biology 253:618–632.

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.

Mann NH (2003) Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiology Reviews 27:17–34.

Maurier F, Beury D, Fléchon L, et al (2019) A complete protocol for whole-genome sequencing of virus from clinical samples: application to coronavirus OC43. Virology 531:141–148.

Martiny JBH, Riemann L, Marston MF, Middelboe M (2014) Antagonistic coevolution of marine planktonic viruses and their hosts. The Annual Review of Marine Science 6:393–414.

Martynov A, Severinov K, Ispolatov I (2017) Optimal number of spacers in CRISPR arrays. PLOS Computational Biology 13:e1005891.

Maruyama F, Ueki S (2016) Evolution and phylogeny of large DNA viruses, *Mimiviridae* and *Phycodnaviridae* including newly characterized *Heterosigma akashiwo* virus. Frontiers in Microbiology 7:120.

Mavromatis K, Ivanova N, Barry K, et al (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods 4:495–500.

Meijenfeldt FAB von, Arkhipova K, Cambuy DD, et al (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biology 20:127.

Mende DR, Waller AS, Sunagawa S, et al (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. PLOS ONE 7:e31386.

Menzel P, Ng KL, Krogh A (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications 7:11257.

Mg W, Ca S (1997) Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. Aquatic Microbial Ecology 13:225–232.

Michen B, Graule T (2010) Isoelectric points of viruses. Journal of Applied Microbiology 109:388–397.

Middelboe M, Jorgensen N, Kroer N (1996) Effects of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton. Applied and Environmental Microbiology 62:1991–1997.

Mihara T, Nishimura Y, Shimizu Y, et al (2016) Linking virus genomes with host taxonomy. Viruses 8(3):66.

Millard AD, Gierga G, Clokie MRJ, et al (2010) An antisense RNA in a lytic cyanophage links psbA to a gene encoding a homing endonuclease. The ISME Journal 4:1121–1135.

Miller ES, Heidelberg JF, Eisen JA, et al (2003) Complete genome sequence of the broad-host-range Vibriophage KVP40: comparative genomics of a T4-related bacteriophage. Journal of Bacteriology 185:5220–5233.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology, 155:733–740.

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

Monpoeho S, Maul A, Mignotte-Cadiergues B, et al (2001) Best viral elution method available for quantification of Enteroviruses in sludge by both cell culture and reverse transcription-PCR. Applied and Environmental Microbiology 67:2484–2488.

Morgan AD, MacLean RC, Hillesland KL, Velicer GJ (2010) Comparative analysis of *myxococcus* predation on soil *bacteria*. Applied and Environmental Microbiology 76:6920–6927.

Moser M, Weisse T (2011) Combined stress effect of pH and temperature narrows the niche width of flagellates in acid mining lakes. Journal of Plankton Research 33:1023–1032.

Müller K, Tidona CA, Bahr U, Darai G (1998) Identification of a thymidylate synthase gene within the genome of *Chilo iridescent* virus. Virus Genes 17:243–258.

Murase J, Frenzel P (2008) Selective grazing of methanotrophs by protozoa in a rice field soil. FEMS Microbiology Ecology 65:408–414.

Murphy J, Mahony J, Ainsworth S, et al (2013) Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. Applied and Environmental Microbiology 79:7547–7555.

Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59:695–700.

Naghili H, Tajik H, Mardani K, et al (2013) Validation of drop plate technique for bacterial enumeration by parametric and nonparametric tests. Veterinary Research Forum 4:179–183.

Narr A, Nawaz A, Wick LY, et al (2017) Soil viral communities vary temporally and along a land use transect as revealed by virus-like particle counting and a modified community fingerprinting approach (fRAPD). Frontiers in Microbiology 8:1975.

Nasko DJ, Ferrell BD, Moore RM, et al (2019) CRISPR spacers indicate preferential matching of specific Virioplankton genes. American Society for Microbiology 10:e2651-18.

Nasukawa T, Uchiyama J, Taharaguchi S, et al (2017) Virus purification by CsCl density gradient using general centrifugation. Archives of Virology 162:3523–3528.

Nazaries L, Murrell JC, Millard P, et al (2013) Methane, microbes and models: fundamental understanding of the soil methane cycle for future predictions. Environmental Microbiology 15:2395–2417.

Nazaries L, Tate KR, Ross DJ, et al (2011) Response of methanotrophic communities to afforestation and reforestation in New Zealand. The ISME Journal 5:1832–1836.

Nguyen N-L, Yu W-J, Gwak J-H, et al (2018) Genomic insights into the acid adaptation of novel methanotrophs enriched From acidic forest soils. Frontiers in Microbiology 9:1982.

Nicol GW, Leininger S, Schleper C, Prosser JI (2008) The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. Environmental Microbiology 10:2966–2978.

Nisbet EG, Dlugokencky EJ, Manning MR, et al (2016) Rising atmospheric methane: 2007–2014 growth and isotopic shift. Global Biogeochemical Cycles 30:1356–1370.

Nishimura Y, Yoshida T, Kuronishi M, et al (2017) ViPTree: the viral proteomic tree server. Bioinformatics 33:2379–2380.

Norton JM, Klotz MG, Stein LY, et al (2008) Complete genome sequence of *Nitrosospira multiformis*, an ammonia-oxidizing bacterium from the soil environment. Applied and Environmental Microbiology 74:3559–3572.

Nurk S, Bankevich A, Antipov D, et al (2013) Assembling genomes and mini-metagenomes from highly chimeric reads. In: Deng M, Jiang R, Sun F, Zhang X (eds) Research in computational molecular biology. Springer, Berlin, pp 158–170.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Research 27:824–834.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Oliveira PH, Touchon M, Rocha EPC (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Research 42:10618–10631.

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. BMC Bioinformatics 12:385.

Op den Camp HJM, Islam T, Stott MB, et al (2009) Environmental, genomic and taxonomic perspectives on methanotrophic *Verrucomicrobia*. Environmental Microbiology Reports 1:293–306.

Orlowski J, Bujnicki JM (2008) Structural and evolutionary classification of type II restriction enzymes based on theoretical and experimental analyses. Nucleic Acids Research 36:3552–3569.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, et al (2016) Uncovering earth's virome. Nature 536:425–430.

Pal C, Maciá MD, Oliver A, et al (2007) Coevolution with viruses drives the evolution of bacterial mutation rates. Nature 450:1079–1081.

Pantastico-Caldas M, Duncan KE, Istock CA, Bell JA (1992) Population dynamics of bacteriophage and *Bacillus Subtilis* in soil. Ecology 73:1888–1902.

Papudeshi B, Haggerty JM, Doane M, et al (2017) Optimizing and evaluating the reconstruction of metagenome-assembled microbial genomes. BMC Genomics 18:915.

Park H-D, Wells GF, Bae H, et al (2006) Occurrence of ammonia-oxidizing archaea in wastewater treatment plant bioreactors. Applied and Environmental Microbiology 72:5643–5647.

Parks DH, Imelfort M, Skennerton CT, et al (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25:1043–1055.

Paterson S, Vogwill T, Buckling A, et al (2010) Antagonistic coevolution accelerates molecular evolution. Nature 464:275–278.

Patro R, Duggal G, Love MI, et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14:417–419.

Paul BG, Ding H, Bagby SC, et al (2017) Methane-oxidizing bacteria shunt carbon to microbial mats at a marine hydrocarbon seep. Frontiers in Microbiology 8:186.

Paul JH (2008) Prophages in marine *bacteria*: dangerous molecular time bombs or the key to survival in the seas? The ISME Journal 2:579–589.

Paul JH, Jiang SC, Rose JB (1991) Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. Applied and Environmental Microbiology 57:2197–2204.

Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013) Going back to the roots: the microbial ecology of the rhizosphere. Nature Reviews Microbiology 11:789–799.

Philipson L, Albertsson PÅ, Frick G (1960) The purification and concentration of viruses by aqueous polymer phase systems. Virology 11:553–571.

Philosof A, Yutin N, Flores-Uribe J, et al (2017) Novel abundant oceanic viruses of uncultured marine group II *Euryarchaeota*. Current Biology 27:1362–1368.

Pietilä MK, Demina TA, Atanasova NS, et al (2014) Archaeal viruses and bacteriophages: comparisons and contrasts. Trends in Microbiology 22:334–344.

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29:170–175.

Prangishvili D (2013) The wonderful world of archaeal viruses. Annual Review of Microbiology 67:565–585.

Prangishvili D, Bamford DH, Forterre P, et al (2017) The enigmatic archaeal virosphere. Nature Reviews Microbiology 15:724–739.

Pratama AA, van Elsas JD (2018) The "neglected" soil virome - potential role and impact. Trends in Microbiology 26:649–662.

Prosser JI, Hink L, Gubry-Rangin C, Nicol GW (2020) Nitrous oxide production by ammonia oxidizers: physiological diversity, niche differentiation and potential mitigation strategies. Global Change Biology 26:103–118.

Prosser JI, Nicol GW (2008) Relative contributions of *archaea* and *bacteria* to aerobic ammonia oxidation in the environment. Environmental Microbiology 10:2931–2941.

Prosser JI, Nicol GW (2012) Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation. Trends Microbiol 20:523–531.

Purkhold U, Pommerening-Röser A, Juretschko S, et al (2000) Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S rRNA and amoA sequence analysis: implications for molecular diversity surveys. Applied and Environmental Microbiology 66:5368–5382.

Quan M, Xie J, Liu X, et al (2016) Comparative analysis of genomics and proteomics in the new isolated *Bacillus thuringiensis* X022 revealed the metabolic regulation mechanism of carbon flux following Cu2+ treatment. Frontiers in Microbiology 7:792.

Quemin ERJ, Chlanda P, Sachse M, et al (2016) Eukaryotic-like virus budding in *archaea*. American Society for Microbiology 7(5):e1439-16.

Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. Nature 403:646–649.

Radajewski S, McDonald IR, Murrell JC (2003) Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. Current Opinion in Biotechnology 14:296–302.

Rajagopal I, Ahn BY, Moss B, Mathews CK (1995) Roles of *vaccinia* virus ribonucleotide reductase and glutaredoxin in DNA precursor biosynthesis. Journal of Biological Chemistry 270:27415–27418.

Raynaud X, Nunan N (2014) Spatial ecology of *Bacteria* at the microscale in soil. PLOS ONE 9:e87217.

Raz Y, Tannenbaum E (2010) The influence of horizontal gene transfer on the mean fitness of unicellular populations in static environments. Genetics 185:327–337.

Reavy B, Swanson MM, Cock PJA, et al (2015) Distinct circular single-stranded DNA viruses exist in different soil types. Applied and Environmental Microbiology 81:3934–3945.

Reeburgh WS (2003) Global methane biogeochemistry. Treatise on Geochemistry 4:347.

Ren B, Hu Y, Chen B, et al (2018a) Soil pH and plant diversity shape soil bacterial community structure in the active layer across the latitudinal gradients in continuous permafrost region of Northeastern China. Scientific Reports 8:5619.

Ren J, Ahlgren NA, Lu YY, et al (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5:69.

Ren J, Song K, Deng C, et al (2020) Identifying viruses from metagenomic data by deep learning. Quantitative Biology 8:64–77

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annual Review of Genetics 38:525–552.

Rillig MC, Muller LA, Lehmann A (2017) Soil aggregates as massively concurrent evolutionary incubators. The ISME Journal 11:1943–1948.

Roberts RJ, Belfort M, Bestor T, et al (2003) Survey and summary: a nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Research 31:1805–1812.

Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends in Genetics 18:291–294.

Rodriguez-Brito B, Li L, Wegley L, et al (2010) Viral and microbial community dynamics in four aquatic environments. The ISME Journal 4:739–751.

Rodriguez-Valera F, Martín-Cuadrado A-B, Rodriguez-Brito B, et al (2009) Explaining microbial population genomics through phage predation. Nature Precedings 1–1.

Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. Journal of Bacteriology 184:4529–4535.

Ross A, Ward S, Hyman P (2016) More is better: selecting for broad host range bacteriophages. Frontiers in Microbiology 7:1352.

Ross MO, Rosenzweig AC (2017) A tale of two methane monooxygenases. Journal of Biological Inorganic Chemistry 22:307–319.

Rotthauwe JH, Witzel KP, Liesack W (1997) The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. Applied and Environmental Microbiology 63:4704–4712.

Rouault TA (2012) Biogenesis of iron-sulfur clusters in mammalian cells: new insights and relevance to human disease. Disease Models & Mechanisms 5:155–164.

Rouault TA, Tong W-H (2005) Iron–sulphur cluster biogenesis and mitochondrial iron homeostasis. Nature Reviews Molecular Cell Biology 6:345–351.

Rousk J, Bååth E, Brookes PC, et al (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. The ISME Journal 4:1340–1351.

Roux S, Adriaenssens EM, Dutilh BE, et al (2019) Minimum information about an uncultivated virus genome (MIUViG). Nature Biotechnology 37:29–37.

Roux S, Brum JR, Dutilh BE, et al (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693.

Roux S, Enault F, Hurwitz BL, Sullivan MB (2015a) VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985.

Roux S, Hallam SJ, Woyke T, Sullivan MB (2015b) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. eLife 4:e8490.

Ruddiman WF, Thomson JS (2001) The case for human causes of increased atmospheric CH4 over the last 5000 years. Quaternary Science Reviews 20:1769–1777.

Rybniker J, Nowag A, Gumpel EV, et al (2010) Insights into the function of the WhiB-like protein of mycobacteriophage TM4 – a transcriptional inhibitor of WhiB2. Molecular Microbiology 77:642–657.

Sabree ZL, Rondon MR, Handelsman J (2009) Metagenomics. In: Schaechter M (eds) Encyclopedia of microbiology 3rd edition. Elsevier Academic Press, Cambridge, pp 622–632.

Saggar S, Tate KR, Giltrap DL, Singh J (2008) Soil-atmosphere exchange of nitrous oxide and methane in New Zealand terrestrial ecosystems and their mitigation options: a review. Plant Soil 309:25–42.

Sanguino L, Franqueville L, Vogel TM, Larose C (2015) Linking environmental prokaryotic viruses and their host through CRISPRs. FEMS Microbiology Ecology 91:5.

Santamaría RI, Bustos P, Sepúlveda-Robles O, et al (2014) Narrow-host-range bacteriophages that infect *Rhizobium etli* associate with distinct genomic types. Applied and Environmental Microbiology 80:446–454.

Santoro AE, Dupont CL, Richter RA, et al (2015) Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": an ammonia-oxidizing archaeon from the open ocean. Proceedings of the National Academy of Sciences USA 112:1173–1178.

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Research 20:1165–1173.

Schmerer M, Molineux IJ, Bull JJ (2014) Synergy as a rationale for phage therapy using phage cocktails. PeerJ 2:e590.

Schoenfeld T, Patterson M, Richardson PM, et al (2008) Assembly of viral metagenomes from Yellowstone hot springs. Applied and Environmental Microbiology 74:4164–4174.

Schulz F, Alteio L, Goudeau D, et al (2018) Hidden diversity of soil giant viruses. Nature Communications 9:4881.

Schwietzke S, Sherwood OA, Bruhwiler LMP, et al (2016) Upward revision of global fossil fuel methane emissions based on isotope database. Nature 538:88–91.

Scott DR, Marcus EA, Wen Y, et al (2010) Cytoplasmic histidine kinase (HP0244)-regulated assembly of urease with UreI, a channel for urea and its metabolites, CO2, NH3, and NH4(+), is necessary for acid survival of Helicobacter pylori. Journal of Bacteriology 192:94–103.

Seccareccia I, Kost C, Nett M (2015) Quantitative analysis of *Lysobacte*r predation. Applied and Environmental Microbiology 81:7098–7105.

Segobola J, Adriaenssens EM, Tsekoa T, et al (2018) Exploring viral diversity in a unique South African soil habitat. Scientific Reports 8:111.

Sharon I, Alperovitch A, Rohwer F, et al (2009) Photosystem I gene cassettes are present in marine virus genomes. Nature 461:258–262.

Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science 5:209.

Shen J, Zhang L, Zhu Y, et al (2008) Abundance and composition of ammonia-oxidizing bacteria and ammonia-oxidizing archaea communities of an alkaline sandy loam. Environmental Microbiology Reports 10:1601–1611.

Shen W, Le S, Li Y, Hu F (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file Manipulation. PLOS ONE 11:e163962.

Shiau Y-J, Cai Y, Jia Z, et al (2018) Phylogenetically distinct methanotrophs modulate methane oxidation in rice paddies across Taiwan. Soil Biology and Biochemistry 124:59–69.

Shindell DT, Faluvegi G, Koch DM, et al (2009) Improved attribution of climate forcing to emissions. Science 326:716–718.

Shoemaker NB, Vlamakis H, Hayes K, Salyers AA (2001) Evidence for extensive resistance gene transfer among *Bacteroides* spp. and among *Bacteroides* and other genera in the human colon. Applied and Environmental Microbiology 67:561–568.

Silveira CB, Rohwer FL (2016) Piggyback-the-winner in host-associated microbial communities. npj Biofilms and Microbiomes 2:1–5.

Sirajuddin S, Rosenzweig AC (2015) Enzymatic oxidation of methane. Biochemistry 54:2283–2294.

Skennerton CT, Imelfort M, Tyson GW (2013) Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Resarch 41:e105.

Smillie CS, Smith MB, Friedman J, et al (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244.

Snyder JC, Bateson MM, Lavin M, Young MJ (2010) Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Applied and Environmental Microbiology 76:7251–7258.

Snyder JC, Bolduc B, Young MJ (2015) 40 years of archaeal virology: expanding viral diversity. Virology 479–480:369–378.

Snyder JC, Brumfield SK, Kerchner KM, et al (2013) Insights into a viral lytic pathway from an archaeal virus-host system. Journal of Virology 87:2186–2192.

Sorokin DY, Lücker S, Vejmelkova D, et al (2012) Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum *Chloroflexi*. The ISME Journal 6:2245–2256.

Sorokin DYu, Jones BE, Gijs Kuenen J (2000) An obligate methylotrophic, methane-oxidizing *Methylomicrobium* species from a highly alkaline environment. Extremophiles 4:145–155.

Srinivasiah S, Bhavsar J, Thapar K, et al (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Research in Microbiology 159:349–357.

Srinivasiah S, Lovett J, Ghosh D, et al (2015) Dynamics of autochthonous soil viral communities parallels dynamics of host communities under nutrient stimulation. FEMS Microbiology Ecology 91:.

Stahl DA, de la Torre JR (2012) Physiology and diversity of ammonia-oxidizing archaea. Annual Review of Microbiology 66:83–101.

Starr EP, Nuccio EE, Pett-Ridge J, et al (2019) Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. Proceedings of the National Academy of Sciences 116:25900–25908.

Stein LY, Arp DJ, Berube PM, et al (2007) Whole-genome analysis of the ammonia-oxidizing bacterium, *Nitrosomonas eutropha* C91: implications for niche adaptation. Environmental Microbiology Reports 9:2993–3007.

Stephen JR, Kowalchuk GA, Bruns M-A, et al (1998) Analysis of beta-subgroup proteobacterial ammonia oxidizer populations in soil by denaturing gradient gel electrophoresis analysis and hierarchical phylogentic probing. Applied and Environmental Microbiolgy 64:2958-2965.

Stern A, Sorek R (2011) The phage-host arms-race: shaping the evolution of microbes. Bioessays 33:43–51.

Steudler PA, Bowden RD, Melillo JM, Aber JD (1989) Influence of nitrogen fertilization on methane uptake in temperate forest soils. Nature 341:314–316.

Steward GF, Culley AI, Mueller JA, et al (2013) Are we missing half of the viruses in the ocean? The ISME Journal 7:672–679.

Stingl K, Altendorf K, Bakker EP (2002) Acid survival of *Helicobacter pylori*: how does urease activity trigger cytoplasmic pH homeostasis? Trends Microbiol 10:70–74.

Strich JR, Chertow DS (2019) CRISPR-Cas biology and its application to infectious diseases. Journal of Clinical Microbiology 57:e1307-18.

Sullivan MB, Coleman ML, Weigele P, et al (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. PLOS Biology 3:e144.

Sullivan MB, Huang KH, Ignacio-Espinoza JC, et al (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environmental Microbiology 12:3035–3056.

Sullivan MB, Waterbury JB, Chisholm SW (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature 424:1047–1051.

Suttle CA (2005) Viruses in the sea. Nature 437:356–361.

Swanson MM, Fraser G, Daniell TJ, et al (2009) Viruses in soils: morphological diversity and abundance in the rhizosphere. Annals of Applied Biology 155:51–60.

Szafranek-Nakonieczna A, Wolińska A, Zielenkiewicz U, et al (2019) Activity and identification of methanotrophic bacteria in arable and no-tillage soils from Lublin region (Poland). Microbial Ecology 77:701–712.

Tamreihao K, Salam N, Ningthoujam DS (2018) Use of acidophilic or acidotolerant *Actinobacteria* for austainable agricultural production in acidic soils. In: Egamberdieva D, Birkeland N-K, Panosyan H, Li W-J (eds) Extremophiles in Eurasian ecosystems: ecology, diversity, and applications. Springer, Singapore, pp 453–464.

Tan MH, Austin CM, Hammer MP, et al (2018) Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. Gigascience 7:3.

Tate KR (2015) Soil methane oxidation and land-use change – from process to mitigation. Soil Biology and Biochemistry 80:260–272.

Taylor AE, Vajrala N, Giguere AT, et al (2013) Use of aliphatic n-alkynes to discriminate soil nitrification activities of ammonia-oxidizing thaumarchaea and bacteria. Applied and Environmental Microbiology 79:6544–6551.

tenOever BR (2016) The evolution of antiviral defense systems. Cell Host Microbe 19:142–149.

Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. Current Opinion in Microbiology 14:321–327.

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology 11:472–477.

Theisen AR, Ali MH, Radajewski S, et al (2005) Regulation of methane oxidation in the facultative methanotroph *Methylocella silvestris* BL2. Molecular Microbiology 58:682–692.

Tiutikov FM, Beliaeva NN, Smirnova ZS, et al (1976) Isolation of bacteriophages of methane oxidizing bacteria and study of their properties. Mikrobiologiia 45:1056–1062

Tock MR, Dryden DTF (2005) The biology of restriction and anti-restriction. Current Opinion in Microbiology 8(4):466-72.

Tourna M, Freitag TE, Nicol GW, Prosser JI (2008) Growth, activity and temperature responses of ammonia-oxidizing archaea and bacteria in soil microcosms. Environmental Microbiology Reports 10:1357–1364.

Tourna M, Stieglmeier M, Spang A, et al (2011) *Nitrososphaera viennensis*, an ammonia oxidizing archaeon from soil. Proceedings of the National Academy of Sciences 108:8420–8425.

Trubl G, Hyman P, Roux S, Abedon ST (2020) Coming-of-age characterization of soil viruses: a user's guide to virus isolation, detection within metagenomes, and viromics. Soil Systems 4:23.

Trubl G, Jang HB, Roux S, et al (2018) Soil viruses are underexplored players in ecosystem carbon processing. Applied and Environmental Science 3(5):e76-18.

Trubl G, Roux S, Solonenko N, et al (2019) Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. PeerJ 7:e7265.

Trubl G, Solonenko N, Chittick L, et al (2016) Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. PeerJ 4:e1999.

Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. American Journal of Human Genetics 85:142–154.

Tyutikov FM, Bespalova IA, Rebentish BA, et al (1980) Bacteriophages of methanotrophic bacteria. Journal of Bacteriology 144:375–381.

Tyutikov FM, Yesipova VV, Rebentish BA, et al (1983) Bacteriophages of methanotrophs isolated from fish. Applied and Environmental Microbiology 46:917–924.

Uchiyama J, Rashel M, Maeda Y, et al (2008) Isolation and characterization of a novel *Enterococcus faecalis* bacteriophage phiEF24C as a therapeutic candidate. FEMS Microbiology Letters 278:200–206.

Uritskiy GV, DiRuggiero J, Taylor J (2018) MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6:158.

Valen LV (1973) Body size and numbers of plants and animals. Evolution 27:27–35.

van Houte S, Ekroth AKE, Broniewski JM, et al (2016a) The diversity-generating benefits of a prokaryotic adaptive immune system. Nature 532:385–388.

van Houte S, Ekroth AKE, Broniewski JM, et al (2016b) The diversity-generating benefits of a prokaryotic adaptive immune system. Nature 532:385–388.

van Regenmortel MHV, Mayo MA, Fauquet CM, Maniloff J (2000) Virus nomenclature: consensus versus chaos. Archives of Virology 145:2227–2232.

van Teeseling MCF, Pol A, Harhangi HR, et al (2014) Expanding the verrucomicrobial methanotrophic world: description of three novel species of *Methylacidimicrobium* gen. nov. Applied and Environmental Microbiology 80:6782–6791.

Vasu K, Nagaraja V (2013) Diverse functions of restriction-modification systems in addition to cellular defense. Microbiology and Molecular Biology Reviews 77:53–72.

Vermassen A, Leroy S, Talon R, et al (2019) Cell wall hydrolases in *bacteria*: insight on the diversity of cell wall amidases, glycosidases and peptidases toward peptidoglycan. Frontiers in Microbiology 10:331.

Vinod MG, Shivu MM, Umesha KR, et al (2006) Isolation of *Vibrio harveyi* bacteriophage with a potential for biocontrol of luminous vibriosis in hatchery environments. Aquaculture 255:117–124.

Vivant A-L, Garmyn D, Maron P-A, et al (2013) Microbial diversity and structure are drivers of the biological barrier effect against *Listeria monocytogenes* in soil. PLOS ONE 8:e76991.

Vollmer W, Blanot D, De Pedro MA (2008) Peptidoglycan structure and architecture. FEMS Microbiology Reviews 32:149–167.

Volossiouk T, Robb EJ, Nazar RN (1995) Direct DNA extraction for PCR-mediated assays of soil organisms. Applied and Environmental Microbiology 61:3972–3976.

Walker CB, de la Torre JR, Klotz MG, et al (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine *crenarchaea*. Proceedings of the National Academy of Sciences 107:8818–8823.

Wang B, Lerdau M, He Y (2017) Widespread production of nonmicrobial greenhouse gases in soils. Global Change Biology 23:4472–4482.

Wang G, Liu Q, Pei Z, et al (2020a) The diversity of the CRISPR-as system and prophages present in the genome reveals the co-evolution of *Bifidobacterium pseudocatenulatum* and phages. Frontiers in Microbiology 11:e1088.

Wang X, Wang S, Jiang Y, et al (2020b) Comammox bacterial abundance, activity, and contribution in agricultural rhizosphere soils. Science of The Total Environment 727:138563.

Wang Y, Fu L, Ren J, et al (2018) Identifying group-specific sequences for microbial communities using long k-mer sequence signatures. Frontiers in Microbiology 9:e872.

Wang Z, Zong H, Zheng H, et al (2015) Reduced nitrification and abundance of ammonia-oxidizing bacteria in acidic soil amended with biochar. Chemosphere 138:576–583.

Wartiainen I, Hestnes AG, McDonald IR, Svenning MM (2006) *Methylobacter tundripaludum* sp. nov., a methane-oxidizing bacterium from Arctic wetland soil on the Svalbard islands, Norway (78 degrees N). International Journal of Systematic and Evolutionary Microbiology 56:109–113.

Watanabe K, Kodama Y, Harayama S (2001) Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. Journal of Microbiological Methods 44:253–262.

Watt M, Hugenholtz P, White R, Vinall K (2006) Numbers and locations of native *bacteria* on field-grown wheat roots quantified by fluorescence in situ hybridization (FISH). Environmental Microbiology 8:871–884.

Weber-Dąbrowska B, Jończyk-Matysiak E, Żaczek M, et al (2016) Bacteriophage procurement for therapeutic purposes. Frontiers in Microbiology 7:1177.

Wei STS, Lacap-Bugler DC, Lau MCY, et al (2016) Taxonomic and functional diversity of soil and hypolithic microbial communities in Miers Valley, McMurdo Dry Valleys, Antarctica. Frontiers in Microbiology 7:e1642.

Weinbauer MG, Brettar I, Höfle MG (2003) Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. Limnology and Oceanography 48:1457–1465.

Weinbauer MG, Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? Environmental Microbiology 6:1–11.

Weinberger AD, Sun CL, Pluciński MM, et al (2012) Persisting viral sequences shape microbial CRISPR-based immunity. PLOS Computational Biology 8:e1002475.

Weisse T, Moser M, Scheffel U, et al (2013) Systematics and species-specific response to pH of *Oxytricha acidotolerans* sp. nov. and *Urosomoida* sp. (Ciliophora, Hypotricha) from acid mining lakes. European Journal of Protistology 49:255–271.

Wichels A, Biel SS, Gelderblom HR, et al (1998) Bacteriophage diversity in the North Sea. Applied and Environmental Microbiology 64:4128–4133

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13:e1005595.

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiology Reviews 35:957–976.

Wilhelm SW, Suttle CA (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 KE, Corzo KA, Drissi CL, et al (2013) Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biology and Fertility of Soils 49:857–869.

Williamson KE, Fuhrmann JJ, Wommack KE, Radosevich M (2017) Viruses in soil ecosystems: an unknown quantity within an unexplored territory. Annual Review of Virology 4:201–219.

Williamson KE, Radosevich M, Wommack KE (2005) Abundance and diversity of viruses in Six Delaware soils. Applied and Environmental Microbiology 71:3119–3125.

Williamson KE, Wommack KE, Radosevich M (2003) Sampling natural viral communities from soil for culture-independent analyses. Applied and Environmental Microbiology 69:6628–6633.

Williamson SJ, Houchin LA, McDaniel L, Paul JH (2002) Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. Applied and Environmental Microbiology 68:4307–4314.

Wilson GG, Murray NE (1991) Restriction and modification systems. Annual Review of Genetics 25:585–627.

Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews 64:69–114.

Wommack KE, Williamson KE, Helton RR, et al (2009) Methods for the isolation of viruses from environmental samples. In: Clokie MRJ, Kropinski AM (eds) Bacteriophages: methods and protocols volume 1: isolation, characterization, and interactions. Humana Press, Totowa, pp 3–14.

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLOS Computational Biology 6:e1000667.

Woolhouse ME, Taylor LH, Haydon DT (2001) Population biology of multihost pathogens. Science 292:1109–1112.

Wu Y-W, Simmons BA, Singer SW (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607.

Xiang X, Chen L, Huang X, et al (2005) *Sulfolobus tengchongensis* spindle-shaped virus STSV1: virus-host interactions and genomic features. Journal of Virology 79:8677–8686.

Xu S, Wang B, Li Y, et al (2020) Ubiquity, diversity, and activity of comammox *Nitrospira* in agricultural soils. Science of the Total Environment 706:135684.

Young CC, Burghoff RL, Keim LG, et al (1993) Polyvinylpyrrolidone-agarose gel electrophoresis purification of polymerase chain reaction-amplifiable DNA from soils. Applied and Environmental Microbiology 59:1972–1974.

Yu P, Mathieu J, Li M, et al (2016) Isolation of polyvalent bacteriophages by sequential multiple-host approaches. Applied and Environmental Microbiology 82:808–815.

Zablocki O, Zyl L van, Adriaenssens EM, et al (2014) High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. Applied and Environmental Microbiology 80:6888–6897.

Zeng Q, Dong Y, An S (2016) Bacterial community responses to soils along a latitudinal and vegetation gradient on the Loess Plateau, China. PLOS ONE 11:e152894.

Zhalnina KV, Dias R, Leonard MT, et al (2014) Genome sequence of *Candidatus* Nitrososphaera evergladensis from group I.1b enriched from Everglades soil reveals novel genomic features of the ammonia-oxidizing archaea. PLOS ONE 9:e101648.

Zhao B, Zhang H, Zhang J, Jin Y (2008) Virus adsorption and inactivation in soil as influenced by autochthonous microorganisms and water content. Soil Biology and Biochemistry 40:649–659.

Zhao J, Meng Y, Drewer J, et al (2020) Differential ecosystem function stability of ammonia-oxidizing archaea and bacteria following short-term environmental perturbation. Applied and Environmental Science 5:e309-20.

Appendix

Supplementary Table 2.1. Summary statistics of the metagenomic assembled genomes (MAG).

MAG Id	Completeness (%)	Contamination (%)	GC (%)	N50	Size (bp)	Phylum
bin.1	55.59	4.77	65	14,227	2,015,931	*Proteobacteria*
bin.2	52.79	9.18	69	7,989	1,917,286	*Actinobacteriota*
bin.3	65.09	0.85	68	11,970	1,716,518	*Actinobacteriota*
bin.4	97.28	2.53	62	360,210	2,718,555	*Proteobacteria*
bin.5	52.58	3.44	68	6,599	3,122,607	*Actinobacteriota*
bin.6	51.46	9.95	60	6,372	3,827,690	*Acidobacteriota*
bin.7	68.37	4.34	64	68,064	4,188,629	*Acidobacteriota*
bin.8	57.20	8.40	67	9,926	2,642,992	*Actinobacteriota*
bin.9	51.72	9.48	69	15,335	1,656,273	*Actinobacteriota*

Supplementary Table 2.2. Auxiliary metabolic genes (AMGs) of viral contigs (VCs) encoding for the glycoside hydrolase (GH) families and peptidases.

Function	VCs	pH 4.5 VCs	pH 7.5 VCs
Glycoside hydrolase superfamily	203	54	67
Glycoside hydrolase, family 1	1	0	0
Glycoside hydrolase, family 5	12	1	4
Glycoside hydrolase, family 8	1	0	0
Glycoside hydrolase, family 10	1	0	0
Glycoside hydrolase, family 12	2	0	0
Glycoside hydrolase, family 16	14	0	9
Glycoside hydrolase, family 18	0	0	1
Glycoside hydrolase, family 19	14	1	6
Glycoside hydrolase, family 22	1	0	0
Glycoside hydrolase, family 24	38	5	5
Glycoside hydrolase, family 25	53	23	16
Glycoside hydrolase, family 25	3	0	0
Glycoside hydrolase, family 26	15	1	2
Glycoside hydrolase, family 37	3	1	0
Glycoside hydrolase, family 42	3	0	0
Glycoside hydrolase, family 45	1	0	0
Glycoside hydrolase, family 81	0	0	2
Glycoside hydrolase, family 71	7	0	1
Peptidase A24A, N-terminal	1	0	0
Peptidase A8, signal peptidase II	4	0	1
Peptidase C1A, papain C-terminal	29	5	6
Peptidase C26	8	3	4
Peptidase C39-like	5	1	2
Peptidase C39, bacteriocin processing	1	0	1

Peptidase G1 superfamily	2	2	0
Peptidase G2, IMC cleavage domain	1	0	0
Peptidase HybD-like domain superfamily	1	0	0
Peptidase M10 serralysin, C-terminal	1	0	0
Peptidase M10, metallopeptidase	8	0	4
Peptidase M11, gametolysin	2	0	0
Peptidase M13	1	0	4
Peptidase M14, carboxypeptidase A	1	0	0
Peptidase M15A, C-terminal	30	2	20
Peptidase M15B	16	0	11
Peptidase M15C	34	0	14
Peptidase M16, C-terminal	1	0	0
Peptidase M17, C-terminal	1	0	0
Peptidase M2, peptidyl-dipeptidase A	1	0	1
Peptidase M20	1	0	0
Peptidase M20, dimerisation domain	1	0	0
Peptidase M23	61	19	31
Peptidase M24, methionine aminopeptidase	2	0	0
Peptidase M28	1	0	0
Peptidase M4 domain	1	0	1
Peptidase M41-like	4	2	1
Peptidase M48	0	0	1
Peptidase M48, protease HtpX, putative	2	0	0
Peptidase M50	1	0	0
Peptidase M54, archaemetzincin	2	0	0
Peptidase S1, PA clan	16	1	3
Peptidase S11, C-terminal	1	0	0
Peptidase S11, N-terminal	1	0	0
Peptidase S1B	2	0	2
Peptidase S1C	3	3	1
Peptidase S24/S26A/S26B/S26C	7	1	0
Peptidase S26	0	0	4
Peptidase S26A, conserved site	1	0	0
Peptidase S26A, serine active site	1	0	0
Peptidase S49	16	2	7
Peptidase S53, activation domain	2	0	0
Peptidase S8, subtilisin-related	0	1	0
Peptidase S8, subtilisin, Asp-active site	5	0	0
Peptidase S8, subtilisin, Ser-active site	2	0	1
Peptidase S8/S53 domain	3	0	1
Peptidase S8/S53 domain superfamily	11	0	4
Peptidase U9, T4 prohead protease	14	3	0

Supplementary Table 2.3. Auxiliary metabolic genes (AMGs) of viral contigs (VCs) encoding for ATPase, ABC transporter and other membrane transporters.

Function	VCs	pH 4.5 VCs	pH 7.5 VCs
ATP synthase subunit alpha, N-terminal	41	7	12
ATP synthase, F0 complex, subunit A superfamily	1	0	0
ATP synthase, F0 complex, subunit C, DCCD-binding site	1	0	0
ATP synthase, F1 complex, delta/epsilon subunit	1	0	0
ATP synthase, F1 complex, delta/epsilon subunit, N-terminal	1	0	0
ATP synthase, F1 complex, gamma subunit	1	0	0
ATP synthase, F1 complex, gamma subunit conserved site	1	0	0
ATP-cone domain	10	1	0
ATPase RavA-like, AAA lid domain	2	1	1
ATPase terminase subunit, putative	7	2	1
ATPase, AAA-3	1	3	0
ATPase, AAA-type, conserved site	2	1	1
ATPase, AAA-type, core	23	4	9
ATPase, dynein-related, AAA domain	32	7	17
ATPase, F1/V1/A1 complex, nucleotide-binding domain	1	0	0
ATPase, OSCP/delta subunit	2	0	0
ABC transporter type 1, transmembrane domain superfamily	3	0	5
ABC transporter-like	6	1	0
ABC transporter, BtuC-like	1	0	0
ABC transporter, conserved site	2	0	0
ABC-2 transporter	5	0	0
Mechanosensitive channel	2	0	0
Potassium channel domain	1	0	0
Aquaporin-like	1	0	0
Ammonium transporter	1	0	0
Sodium/solute symporter superfamily	1	0	0
FAD/NAD(P)-binding domain superfamily	31	3	5
NAD-dependent DNA ligase	2	0	0
NAD-dependent DNA ligase, adenylation	2	1	0
NAD-dependent DNA ligase, N-terminal	1	0	0
NAD-dependent epimerase/dehydratase	36	10	16
NAD(+) synthetase	1	0	0
NAD(P)-binding domain superfamily	60	9	28
NADH-ubiquinone oxidoreductase, iron-sulphur binding domain	1	0	2
NADH:ubiquinone oxidoreductase, conserved site	1	0	0
NADP-dependent oxidoreductase domain	1	0	0
NADPH-dependent 7-cyano-7-deazaguanine reductase QueF	3	0	1
Ferredoxin-NADP reductase (FNR), nucleotide-binding domain	3	0	1
Ferrodoxin-fold anticodon-binding domain	1	0	0
YubB, ferredoxin-like domain	10	0	8
2Fe-2S ferredoxin-like superfamily	1	0	3
2Fe-2S ferredoxin, iron-sulphur binding site	1	0	2
2OGFeDO, oxygenase domain	9	1	7

Cupredoxin	6	0	0
Periplasmic copper-binding protein NosD, beta helix domain	1	0	0
Superoxide dismutase-like, copper/zinc binding domain superfamily	1	0	0
Tyrosinase copper-binding domain	1	0	0

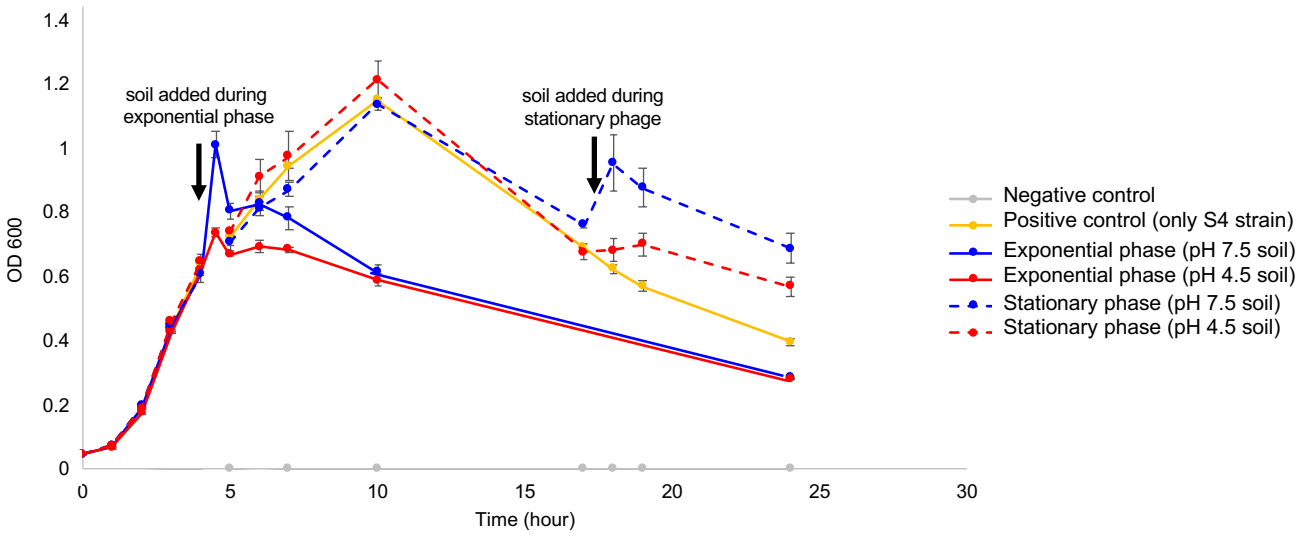
Supplementary Table 3.1. Soil pH and moisture content across the soil pH gradient.

Sample ID	Soil pH	Wet soil (g)	Dry soil (g)	Water content (%)
pH4.5_1	4.17	3.02	2.69	12.26
pH4.5_2	4.31	3.05	2.7	12.96
pH4.5_3	4.18	3.01	2.66	13.15
pH5.5_1	4.8	3.08	2.81	9.60
pH5.5_2	4.78	3.02	2.74	10.21
pH5.5_3	4.78	3.03	2.75	10.18
pH6.5_1	6.27	3.03	2.77	9.38
pH6.5_2	6.55	3.09	2.77	11.55
pH6.5_3	6.51	3.05	2.87	6.27
pH7.5_1	7.36	3	2.75	9.09
pH7.5_2	7.28	3	2.73	9.89
pH7.5_3	7.38	3.03	2.77	9.38

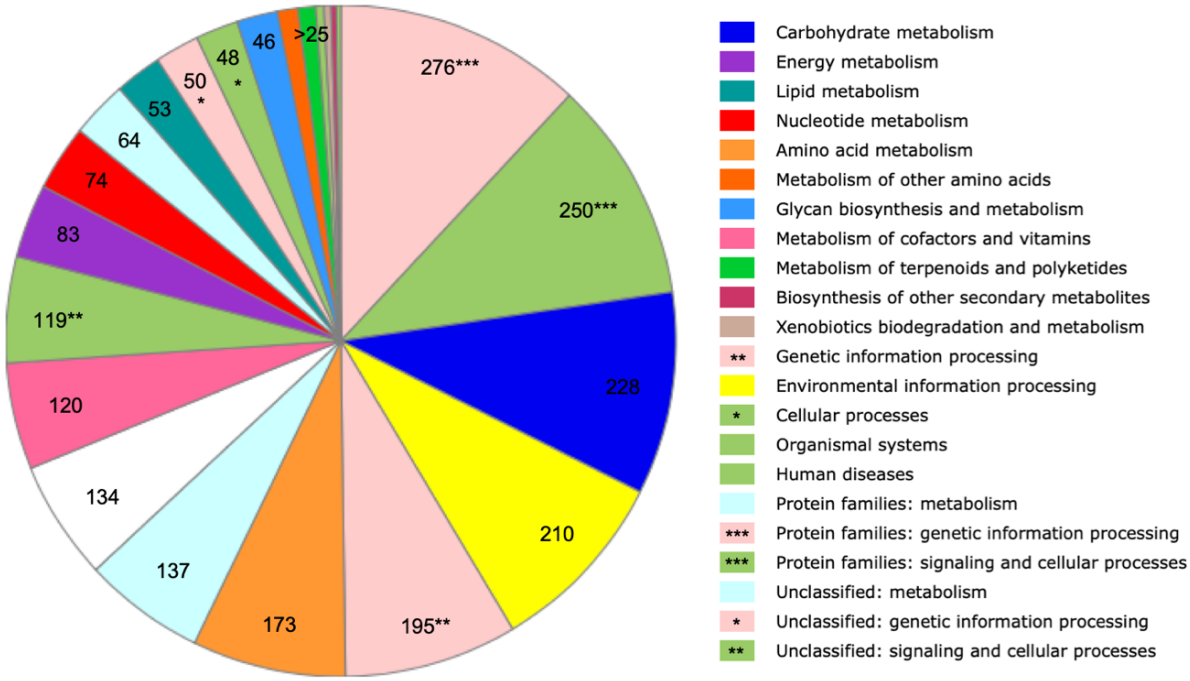
Supplementary Table 3.2. Gene annotation of the metagenomic viral contigs (mVCs) and prophages using the NCBI nr database with VIPTree. Genes annotated as uncharacterized proteins are not shown.

Viral gene ID	Function	Identity (%)	E value	bit score
mVC_1_1	Terminase large subunit	97.9	1.7e-163	583
mVC_1_77	Terminase large subunit	98.6	7.6e-161	574
mVC_2_16	Putative tail protein	31.4	5.4e-13	82
mVC_2_17	Putative tail tape measure protein	57.5	0.0e+0	1253
mVC_2_26	Putative major capsid protein	58.7	1.1e-107	398
mVC_2_32	Putative minor capsid protein	40.1	1.1e-69	273
mVC_2_33	Capsid portal protein	53.0	1.9e-149	538
mVC_2_34	Terminase large subunit	61.1	2.9e-142	513
mVC_2_40	Putative terminase small subunit	50.3	4.9e-38	164
mVC_3_1	Portal protein	78.5	3.8e-223	782
mVC_3_2	Terminase large subunit	84.8	2.4e-220	773
mVC_3_4	Putative terminase small subunit	54.4	3.6e-70	273
mVC_3_44	Tail length tape measure protein	56.3	0.0e+0	1281
mVC_3_47	Major capsid protein	76.7	8.7e-59	233
mVC_3_49	Minor capsid protein	65.8	4.1e-40	171
mVC_3_53	Major capsid protein	79.1	4.3e-123	449
mVC_3_55	Major capsid protein	80.3	2.8e-25	122
mVC_3_56	Minor capsid protein	74.7	1.5e-150	540
mVC_4_1	Terminase small subunit	61.1	2.7e-68	266
mVC_4_35	Tail length tape measure protein	65.9	0.0e+0	1498
mVC_4_38	Major capsid protein	89.6	3.3e-71	275
mVC_4_40	Minor capsid protein	83.8	5.0e-51	208
mVC_4_41	Putative minor capsid protein 1	73.2	6.6e-43	181
mVC_4_44	Major capsid protein	88.6	5.9e-141	508
mVC_4_45	Major capsid protein	73.8	1.0e-70	273

mVC_4_46	Minor capsid protein	88.5	3.3e-168	599
mVC_4_47	Portal protein	83.1	1.6e-245	857
mVC_4_49	Terminase large subunit	90.7	2.9e-234	819
mVC_5_37	Tail fiber protein	77.1	0.0e+0	5664
mVC_5_39	Putative adsorbtion tail protein	93.1	0.0e+0	2138
mVC_5_41	Baseplate J family protein	93.4	3.7e-183	649
mVC_5_42	Baseplate protein	86.9	2.8e-96	358
mVC_5_45	Putative tail fiber	93.7	2.5e-210	739
mVC_5_48	Tail assembly chaperone	74.1	1.5e-18	100
mVC_5_49	Tail assembly chaperone	61.8	5.1e-43	181
mVC_5_50	Tail tube subunit	97.8	4.3e-71	274
mVC_5_51	Tail sheath protein	95.8	0.0e+0	1093
mVC_5_59	Tail fiber protein	81.2	0.0e+0	2292
mVC_5_61	Butative adsorbtion tail protein	93.6	0.0e+0	2149
mVC_5_63	Baseplate J family protein	93.1	1.9e-182	646
mVC_5_64	Baseplate protein	86.4	3.7e-96	358
mVC_5_67	Putative fiber protein	93.9	4.6e-209	735
mVC_5_68	Putative tail lysin	79.6	0.0e+0	1150
S4_Prophage_1_21	Phage integrase	97.0	2.1e-28	132
S4_Prophage_1_22	Phage integrase	100.0	4.7e-33	148
S4_Prophage_1_31	Phage portal protein, HK97 family	99.7	1.5e-215	757
S4_Prophage_1_38	Phi13 family phage major tail protein	99.5	1.2e-106	393
S4_Prophage_2_1	Integrase	98.7	4.4e-131	475
S4_Prophage_4_4	Phage minor structural protein	100.0	6.6e-43	181
S4_Prophage_4_5	Phage minor structural protein	100.0	2.2e-234	819
S4_Prophage_4_6	Phage minor structural protein	98.9	6.6e-254	884
S4_Prophage_4_7	Phage minor structural protein	94.7	6.5e-51	207
S4_Prophage_4_8	Phage minor structural protein	94.6	1.1e-147	530
S4_Prophage_4_9	Phage minor structural protein	100.0	2.4e-8	66
S4_Prophage_4_10	Phage minor structural protein	97.4	1.4e-56	226
S4_Prophage_4_18	Phi13 family phage major tail protein	99.5	1.2e-106	393
S4_Prophage_4_36	Integrase	100.0	1.0e-96	359



Supplementary Figure 3.1. Impact of soil pH (4.5 and 7.5) and growth stages (exponential phase and overnight grown stage) on the growth of *Bacillus* sp. S4.



Supplementary Figure 3.2. KEGG pathway reconstructions of the *Bacillus* sp. S4 strain. Metabolic functions are visualized in the pie chart. Number of genes matched to KEGG function is shown on the pie chart.

Supplementary Table 4.1. Spacer sequences matching to VirSorter predicted metagenomic viral contigs (mVCs). Positive and negative strands are presented by + and -, respectively.

Virus ID	CRISPR spacer ID	Matched spacer sequences	Strand	Start[1]	End[2]
mVC_07548-cat_2	pH4.5_1_CRISPR_03-S_13	TCCTCGCACCATTCGGCAATGTCGCGCTGGCGC	+	26343	26375
mVC_00108-cat_2	pH4.5_1_CRISPR_03-S_15	AACGCCCGCGCCTCAGCATCCGTAAGAGCCATAA	-	329	362
mVC_12213-cat_2	pH4.5_1_CRISPR_03-S_16	CGGCATTTTGTCGGCTCCCCAGCCCGTTCCATC	+	121	153
mVC_00108-cat_2	pH4.5_1_CRISPR_03-S_17	TCCTCCGGCGGCCTCGCCGGCGCGCTGGCCGCCGC	+	32059	32093
mVC_00108-cat_2	pH4.5_1_CRISPR_03-S_18	GCCAGATCGTCGCCAGCTACAAAAAGCTGACGGC	+	31044	31077
mVC_07548-cat_2	pH4.5_1_CRISPR_03-S_28	GCCTTCAGCTCGTCGGCCGTCTCGGTCGTGAGG	+	27345	27377
mVC_00234-cat_2	pH4.5_1_CRISPR_03-S_58	ATCTACAAGGACGGCGAAGACCCGATCCTGATTG	+	3493	3526
mVC_00234-cat_2	pH4.5_1_CRISPR_03-S_61	GCGCGATCCATATTCAAGCGGCCGAGCGGGCGGC	-	12661	12694
mVC_08211-cat_3	pH4.5_1_CRISPR_03-S_7	CTTCAGATCGCCAAGCTGCGACCGGCGGGCTCCC	-	2288	2321
mVC_12213-cat_2	pH4.5_1_CRISPR_03-S_73	ACGTCGGAGGTCGCATGACGTCTTCCGCCGAAAC	-	45583	45616
mVC_08211-cat_3	pH4.5_1_CRISPR_03-S_74	CGGTCGAGGCTACGGAGAAGATGTTGCGCGCGGC	-	4136	4169
mVC_07548-cat_2	pH4.5_1_CRISPR_03-S_79	GTTCGTGGCGCCTGGGTGATATTTCCCGAGCCGG	+	24894	24927
mVC_07548-cat_2	pH4.5_1_CRISPR_03-S_8	TCGCGCACCTCGGCCTGCTTGACTCGCAAGCCCA	+	24439	24472
mVC_12249-cat_2	pH4.5_1_CRISPR_03-S_83	CAATTTGTCGTCGGCTCCTCGGTGATTTGCAAGG	-	19663	19696
mVC_00108-cat_2	pH4.5_1_CRISPR_03-S_83	CAATTTGTCGTCGGCTCCTCGGTGATTTGCAAGG	+	32594	32627
mVC_07580-cat_2	pH4.5_1_CRISPR_03-S_86	TCGGCGCAGACGCGCAAGGCGACGAGCTTTAAAG	+	8905	8938
mVC_08211-cat_3	pH4.5_1_CRISPR_03-S_87	AAATGAGCGGCAACAAGGTTCCTGCGACAACCAG	+	1670	1703
mVC_07548-cat_2	pH4.5_1_CRISPR_03-S_9	ACGAGGAGCTGCGCTATCTCGAAAATCTCGGCGC	-	28003	28036
mVC_12249-cat_2	pH4.5_1_CRISPR_03-S_93	AGCAGCGACGAAGGACAGCGCGCCGTCGCGCAT	-	760	792
mVC_00234-cat_2	pH4.5_1_CRISPR_03-S_93	AGCAGCGACGAAGGACAGCGCGCCGTCGCGCAT	+	10287	10319
mVC_12213-cat_2	pH4.5_1_CRISPR_03-S_95	TGCATCCTGGTGCAGAATTCTGGGGTAGTTCGG	-	3274	3306
mVC_07580-cat_2	pH4.5_1_CRISPR_05-S_7	TCATCCTGCAGGGGCTGGAAGCTTGGAGAG	-	8304	8333
mVC_12249-cat_2	pH4.5_1_CRISPR_09-S_20	CTCTGTGAGGTCTGCGCGCCGCGCCATTTCTGGCC	-	2361	2395
mVC_07548-cat_2	pH4.5_1_CRISPR_09-S_23	GCTCGCCAATCTTCTCGATGTAGCCCCCCTTTGCGA	+	37979	38014
mVC_07548-cat_2	pH4.5_1_CRISPR_09-S_24	CGCGGCAGCATCTCGGCTCCGCACAGCCACATCTG	+	42992	43026
mVC_07548-cat_2	pH4.5_1_CRISPR_09-S_27	TGTACGTCCTGCTGCAGGATGACGTCCTGCGCGT	+	43298	43331
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_33	TTTCGCGCGCCATCATTTCCCCCGCACGGAGTC	-	56428	56460
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_33	TTTCGCGCGCCATCATTTCCCCCGCACGGAGTC	+	8668	8700
mVC_12249-cat_2	pH4.5_1_CRISPR_09-S_35	GCGGCCTCTTTCGCAAGGCTGCGCACCTCGGAATT	+	13766	13800
mVC_00108-cat_2	pH4.5_1_CRISPR_09-S_35	GCGGCCTCTTTCGCAAGGCTGCGCACCTCGGAATT	-	38502	38536

mVC_07548-cat_2	pH4.5_1_CRISPR_09-S_40	GAAAGGCCGCCCAATCCTTGATCGAATAGTTTTT	+	24950	24983
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_43	CAGCGGCTACCACGACACGGATGAGGAAGACGAA	-	13900	13933
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_44	AAACTACCGGAAATTTCTCGATCGCTTCGGGCTC	+	54463	54496
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_45	AGGCTGGAAAGACGCCGATCAAGCGCGAGGATTG	+	53958	53991
mVC_00108-cat_2	pH4.5_1_CRISPR_09-S_47	CGGCCGGCACGGTCCATACGCCCGCCTCGAGGTC	-	1069	1102
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_51	CCGGATCAGGCTTACGAGACAGCATCCAAGGCAG	+	2236	2269
mVC_12249-cat_2	pH4.5_1_CRISPR_09-S_52	TGTCGCCGGCGCTGTGCGGGCGCATGTCGCGCGT	+	19061	19094
mVC_00108-cat_2	pH4.5_1_CRISPR_09-S_52	TGTCGCCGGCGCTGTGCGGGCGCATGTCGCGCGT	-	33196	33229
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_60	CCGAGAGCCTCAAAAATTTGCTCGGCGCTAACCCC	-	55969	56003
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_60	CCGAGAGCCTCAAAAATTTGCTCGGCGCTAACCCC	+	9125	9159
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_68	CGCTTTGCTGTCACCGCGCTCTCCATCCTGCAAC	-	57133	57166
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_70	CGCCAGGGCTGATCGGTATCTGTGACGGGGTGG	+	9391	9423
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_74	CCGGATGGAATGGTTGCGTTGATGTTCGAATCATA	+	8417	8451
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_75	ATCGCCGCCTGGATCAGGCTCTGGTAATTGATCG	+	10727	10760
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_76	TTTTTGAGGGCTGCGGACGTGTCGCGATCAGGACA	+	6968	7002
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_77	ACGCATTCGGAAAAGCCAGCGGGAATGTGATGGT	+	8680	8713
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_78	CTGATAAATGCCCGTCAGCGTTCCGGCAGGGCAAG	+	11221	11255
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_79	GTAGTTTCCCATCCCCACCGAGCTGGTTGATGAC	+	5978	6011
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_80	GCGAAAATGGATGGCGCTTCCTACACCAGATACCA	+	56568	56602
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_80	GCGAAAATGGATGGCGCTTCCTACACCAGATACCA	-	8526	8560
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_81	GATCTCGATTGGAAAGAGAAACGCGGGACTGAG	-	7008	7040
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_83	CTGGCGCCGGCAGAACGGCGTCTGCCCTATCAC	-	6683	6715
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_85	GTTGATCTTCCGCCCCGCGCATCATTGCCGGATG	-	49482	49515
mVC_08211-cat_3	pH4.5_1_CRISPR_09-S_85	GTTGATCTTCCGCCCCGCGCATCATTGCCGGATG	+	14180	14213
mVC_12213-cat_2	pH4.5_1_CRISPR_09-S_87	CTGCACGATGCTGTAGGCGACGCCAGCGGGCGTGG	+	9527	9561
mVC_07548-cat_2	pH4.5_1_CRISPR_09-S_9	TATGAGCTGGAGTTCTTCGCCCATGACGATGGC	-	5361	5393
mVC_12494-cat_3	pH4.5_1_CRISPR_10-S_10	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	-	23220	23254
mVC_01376-cat_3	pH4.5_1_CRISPR_10-S_10	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	+	5771	5805
mVC_01376-cat_3	pH4.5_1_CRISPR_10-S_5	TCTTCTCATCGGTGACGATGGATCGGA	-	1204	1230
mVC_13605-cat_3	pH4.5_1_CRISPR_10-S_5	TCTTCTCATCGGTGACGATGGATCGGA	+	3664	3690
mVC_08964-cat_3	pH4.5_1_CRISPR_10-S_5	TCTTCTCATCGGTGACGATGGATCGGA	-	7067	7093

mVC_00397-cat_3	pH4.5_1_CRISPR_21-S_10	GAGCGCCTCTGCGACCGGCCCGAGCTCTCCGGTCGAGTCGAGACATC	-	2108	2154
mVC_12213-cat_2	pH4.5_1_CRISPR_86-S_3	GGTCTGCCCGCCGTCCGCGATGACCTCGCCGTAGAAT	+	40157	40193
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_1	GGGCGGCGGCGGCGAAGCGGACCTGGGAAAA	+	7866	7896
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_2	GGGTGGCGGCGCTGAACAG	+	7917	7935
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_3	GGGCGGCGGCGCTGAACAG	+	7956	7974
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_4	TGACGGCGGCGGCGAAGCGGACCTGGGAAAA	+	7995	8025
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_5	TCGTGGCGGCGCGGAACAG	+	8046	8064
mVC_07580-cat_2	pH4.5_2_CRISPR_00-S_6	TGGCGGCGGCGGCGAAGCTGACCTGGGAAAA	+	8085	8115
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_10	TGCAAGGTCACACCGACACAGCACCAGTCCGGC	+	19478	19510
mVC_07548-cat_2	pH4.5_2_CRISPR_03-S_107	GTCGTCGAGATGCGCAACCTGCCTTTCGACAGC	-	26174	26206
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_12	CGCCGTTGGCGACAGCCAGCTGTGTGCCGCTCCAA	+	5234	5268
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_13	TCCATCTCTTCGCCCGCCTTGTCGAGGATCTCGG	-	31630	31663
mVC_07548-cat_2	pH4.5_2_CRISPR_03-S_13	TCCATCTCTTCGCCCGCCTTGTCGAGGATCTCGG	-	44177	44210
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_18	AACCTCCCTGACAGCTACGTCGCGGGCCAGAACAT	-	25746	25780
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_20	ACTCGCACGCCGGGAAGCATCCGCGCCCGTCGC	-	26489	26521
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_21	AGGACAATTTCTCCGCCATCCGCGACGCCGAGCG	-	23609	23642
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_23	AATTCGAAAACAATGGCGTCAACGTCAAGTTGT	-	12116	12148
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_24	ATGTCGCTCAGGGTCATCCCGTAGATGGCGATGTC	+	12424	12458
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_25	GCTCGCGCAGATGTATTTCAATGAGGCGGCGCT	-	20624	20656
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_26	ACCGAGAAGATGCGCGCCGGCCCGGTCATCCCCG	-	21282	21315
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_27	CATCGCCGGCACGGTGCTTACGGTCACGGGGGCTG	-	18170	18204
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_28	CCCGTAAGATAGCCGTCAACGAGACCGGATCAT	-	12596	12628
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_3	AGGAAAACGTCGCGCTGATCAAGAGCATCCCGC	-	26732	26764
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_31	ATCGCCGCCTCGCCAAGCGGCGTGTCCTTGAACT	+	43488	43521
mVC_07548-cat_2	pH4.5_2_CRISPR_03-S_33	TGGGCGATCCTCCTGCGGCGGATGAGCTGGCCTG	-	2865	2898
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_34	TTTTTCTCGGCGTGGTAGCGAAGGATTGGGTCAA	+	41419	41452
mVC_08211-cat_3	pH4.5_2_CRISPR_03-S_36	CTTCAGATCGCCAAGCTGCGACCGGCGGGCTCCC	-	2288	2321
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_37	GTTGAACATGCGAGTGACGATAATGCGCGAGCCC	+	40454	40487
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_4	TTCCCGACACTCTCAAGGTAGCTCTTCGACTGC	-	49989	50021
mVC_08211-cat_3	pH4.5_2_CRISPR_03-S_4	TTCCCGACACTCTCAAGGTAGCTCTTCGACTGC	+	13671	13703
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_48	ACCAAAAAGCGCATCAACGTCACGTTCGAAATG	-	12543	12575
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_5	TGGGATCTACGCCATCCGTAACGTTTATGGGGAA	+	46515	46548

mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_51	CCTCCAAAGAGAACGAGACCCGCTGGCTTCGCGC	+	42559	42592
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_54	CTACCAGCCGGGACCGAATGCGAAGAAGGTCCTG	-	16384	16417
mVC_08211-cat_3	pH4.5_2_CRISPR_03-S_57	GGCCCCGAAGGGGTCGAGCCTCTCCTCGACGAG	-	11842	11874
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_64	TATCTCGGGCAGGCGGCCCCCACCTCGACGACGA	-	9874	9907
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_65	GTCATCAACCAGCTCGGTGGGGATGGGAAACTAC	-	5978	6011
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_7	ACAAAACAGCGGCGTCACGCCGGGCGCCGTCGC	+	18900	18932
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_72	ACGTCGGAAGGCGCTATCGATCCGATCCTCAAGC	+	49867	49900
mVC_08211-cat_3	pH4.5_2_CRISPR_03-S_72	ACGTCGGAAGGCGCTATCGATCCGATCCTCAAGC	-	13792	13825
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_74	ATCACCAGCCTGCTCGACGAGTTCGGCCCCGAA	+	51747	51779
mVC_07548-cat_2	pH4.5_2_CRISPR_03-S_77	ATGCGCCAACAGATGCTCGTCGACGGCAAGCCGG	-	2503	2536
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_8	GTCCAATAGCCGCGTGTCGTCATATCGAGTGCGG	+	4348	4381
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_86	CCCTGCCACTCCTTCACGCTTCGGCCCTTGGCCTT	-	57189	57223
mVC_08211-cat_3	pH4.5_2_CRISPR_03-S_89	CCTGAAGACGTTTGCGCTTTGCCTCACCCATGG	+	2417	2449
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_9	CACGCGACGGCGCCCGGCGTGACGCCGCTGTTT	-	18903	18935
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_91	CTCTCCTACGGGTCTGCGACATCGAGACTACCGG	+	54600	54633
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_93	CGCGAACTTCACCAACAGCAAGCGTGGCGCAGAG	+	49548	49581
mVC_12213-cat_2	pH4.5_2_CRISPR_03-S_94	TAGCCCTCGAACTCGTTGACTACATGCGCCGCGA	+	51369	51402
mVC_12213-cat_2	pH4.5_2_CRISPR_08-S_1	GAGCGCCAATGGTTCACCGAGGAAGTGACCGGCGG	+	54911	54945
mVC_08211-cat_3	pH4.5_2_CRISPR_08-S_1	GAGCGCCAATGGTTCACCGAGGAAGTGACCGGCGG	-	10183	10217
mVC_07548-cat_2	pH4.5_2_CRISPR_08-S_12	TCGTTTTGCAACAGCGCCTGACCGACCGTGGCCA	+	22119	22152
mVC_08211-cat_3	pH4.5_2_CRISPR_08-S_2	CGTTCCATCCAGCCATGATCCCGAGGCCGTCGCG	+	213	246
mVC_00108-cat_2	pH4.5_2_CRISPR_08-S_23	GGCAAGGAAGGTTCGATTTGGGCCGATGACGCCG	-	5760	5793
mVC_08211-cat_3	pH4.5_2_CRISPR_08-S_9	ACCGCAAGGGCTGCACCGGCGCGAGCACGATTCAC	-	7728	7762
mVC_00108-cat_2	pH4.5_2_CRISPR_09-S_13	GGCAAGGAAGGTTCGATTTGGGCCGATGACGCCG	-	5760	5793
mVC_07548-cat_2	pH4.5_2_CRISPR_09-S_7	CGGCAGGAGCTGACGCCGATCCTGCAGAAGAAGG	-	23758	23791
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_1	GATCAGTCGGGTGGCGGTCGATGAGTCGAGACATCC	-	2555	2590
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_10	TAACAGATCGCGATTGGGCGAGCGGCCGGAGCGACTCCTGGTCGAGACATC	-	1966	2016
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_11	CGACACACGTGGACGGATGACAGACCAGGCGACGAGGTCGTCGAGACATC	-	1895	1944
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_12	CGACCGTTGGTGCTGCGCTCGCCGCGCTCTTCTTCTACATGTCGAGACATC	-	1823	1873
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_13	CGACGCGCGTGTGCGCCACTCGCGGGTGTCGAGACATC	-	1764	1801
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_14	CGAACGTGAAGATCGAGCACTACTCGCTGGGAGTCGAGACATC	-	1700	1742
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_15	CGACTCGACTTGCCGCGTGCACTCCTTCGCGGCCGCTACATGTCGAGACATC	-	1627	1678

mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_16	CGACGCGGGAACGTCTTCAATAGGTCATCCTCCGTAGTCTGGTCGAGACATC	-	1554	1605
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_17	CGACTCGGACGACGCCTGCTCGAAGTGGAACTCGAGCTTCCACGTCGAGACATC	-	1479	1532
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_2	CCTAGATCCGCACCACGTCGAGACATC	-	2507	2533
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_3	CGACGGCTCCTTGCGCTCATTCGCCTCTCGTCGAGACATT	-	2446	2485
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_4	GTTCTGCGACGACCGCGACGACCAGCGCGGCGTCCTAGGGTCGAGACATC	-	2375	2424
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_5	ACCATTCGCATCGAAACCGCGCTGTCGAGACACT	-	2320	2353
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_6	CGACCCGCTGTCGTAGCCGCTCCGCGCGGCGTTGCACATTGTCGAGACACC	-	2248	2298
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_7	AGACTCCAGGTGCGCCACGAATTGCGTCCCGTCGCGGTCGAGACATC	-	2180	2226
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_8	CGACGAGCGCCTCTGCGACCGGCCCGAGCTCTCCGGTCGAGTCGAGACATC	-	2108	2158
mVC_00397-cat_3	pH4.5_2_CRISPR_10-S_9	TGACGCTAAATTGTGACACGTCCGTCCTGGTAGATGTAGTCGAGACACT	-	2038	2086
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_2	GGTTCTGCGACAACCGGCTCGCGACTTATGCGCGGGCGAGTGAGGAT	-	1128	1174
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_3	GGTTCTGTGACTTGGCGTATCGTAGGAACTCGTCGCCATCACTGTCGG	-	1055	1102
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_4	AGTTCTGCGACCTTCCCCAAGATAGAGGGGACGGGTCGCTAGGCAT	-	984	1029
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_5	AGTTCTGCGACCCGAGTACGCAACCTCGGCCTGGAGGACGAGCTTGAT	-	911	958
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_6	AGTTCTGCGACCTTCCGTGCAAGACTTCCTGGGCGGAGGACGAACCAT	-	838	885
mVC_13626-cat_3	pH4.5_2_CRISPR_11-S_6	AGTTCTGCGACCTTCCGTGCAAGACTTCCTGGGCGGAGGACGAACCAT	+	85	132
mVC_00397-cat_3	pH4.5_2_CRISPR_11-S_7	AGTTCTGCGACTCGTAGGGACGAAAAAGTCGCACTCGAAATGCAGGAG	-	765	812
mVC_13626-cat_3	pH4.5_2_CRISPR_11-S_7	AGTTCTGCGACTCGTAGGGACGAAAAAGTCGCACTCGAAATGCAGGAG	+	158	205
mVC_12494-cat_3	pH4.5_2_CRISPR_26-S_3	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	-	23220	23254
mVC_01376-cat_3	pH4.5_2_CRISPR_26-S_3	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	+	5771	5805
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_1	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7869	7899
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_1	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7998	8028
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_2	TGGCGGCGCTGAACAGCCC	+	7920	7938
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_3	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7869	7899
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_3	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7998	8028
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_4	TGGCGGCGCTGAACAGCCC	+	7920	7938
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_5	CGGCGGCGCTGAACAGCCC	+	7959	7977
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_6	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7869	7899
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_6	CGGCGGCGGCGAAGCGGACCTGGGAAAACCC	+	7998	8028
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_7	TGGCGGCGCGGAACAGCCC	+	8049	8067
mVC_07580-cat_2	pH4.5_3_CRISPR_00-S_8	CGGCGGCGGCGAAGCTGACCTGGGAAAACCC	+	8088	8118
mVC_12249-cat_2	pH4.5_3_CRISPR_01-S_14	TGCAATGGCGCGCGCGCCAAATCCGCACGCTCGCC	+	41317	41351

mVC_00108-cat_2	pH4.5_3_CRISPR_01-S_14	TGCAATGGCGCGCGCCAAATCCGCACGCTCGCC	-	10297	10331
mVC_07548-cat_2	pH4.5_3_CRISPR_01-S_21	TCGAGCGTGCCGACGAGCGCGGTCAGCTTCGCCTT	+	24082	24116
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_49	ATTCTACGGCGAGGTCATCGCGGACGGCGGGCAGACC	-	40157	40193
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_51	CCGAGCGCGGGATGGGACATGCGCGCCAGACTTT	-	40299	40332
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_54	TCCTCTCCGCGATCAGTTCCGACCGTGTGGTGAT	-	37286	37319
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_57	CCTGCGCGTCGCGCATGGCCGCCACCATGCAGTG	+	38903	38936
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_58	GCGGGGAAATCATGGCCCTGAGATGGGCTGCAACA	-	37115	37149
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_61	CTGACAGCTCTCTGGACCTCGAGCGGCCCTTCGC	+	29639	29672
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_64	CCCCTGGAGACCGCAAGCGCAACAAGGCAAAAAAGG	-	38807	38842
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_66	TGCTGGCCGTTCTGATGCCGATGAGGGTCTAAGC	-	37196	37229
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_68	GGCGTCATGAGCGCCCTCCCAGACCTCCGCGAA	+	48273	48305
mVC_12213-cat_2	pH4.5_3_CRISPR_01-S_69	ACAGCCATTGCCTGATGGTCGAGAGATGGTCGA	+	42494	42526
mVC_00397-cat_3	pH4.5_3_CRISPR_05-S_1	GCGGATACGCTCGTCAACATGTTCCACGGCCGCATGTCGAGACAT	-	1190	1234
mVC_13626-cat_3	pH4.5_3_CRISPR_05-S_1	GCGGATACGCTCGTCAACATGTTCCACGGCCGCATGTCGAGACAT	+	25	69
mVC_00397-cat_3	pH4.5_3_CRISPR_05-S_2	CTTCCGTGCAAGACTTCCTGGGCGGAGGACGAACCATGTCGAGACAC	-	828	874
mVC_13626-cat_3	pH4.5_3_CRISPR_05-S_2	CTTCCGTGCAAGACTTCCTGGGCGGAGGACGAACCATGTCGAGACAC	+	96	142
mVC_00397-cat_3	pH4.5_3_CRISPR_05-S_3	TCGTAGGGACGAAAAAGTCGCACTCGAAATGCAGGAGGTCGAGACAC	-	755	801
mVC_13626-cat_3	pH4.5_3_CRISPR_05-S_3	TCGTAGGGACGAAAAAGTCGCACTCGAAATGCAGGAGGTCGAGACAC	+	169	215
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_1	GATCAGTCGGGTGGCGGTCGATGAGTCGAGACATCC	-	2555	2590
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_10	TAACAGATCGCGATTGGGCGAGCGGCCGGAGCGACTCCTGGTCGAGACATC	-	1966	2016
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_11	CGACACACGTGGACGGATGACAGACCAGGCGACGAGGTCGTCGAGACATC	-	1895	1944
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_12	CGACCGTTGGTGCTGCGCTCGCCGCGCTCTTCTTCTACATGTCGAGACATC	-	1823	1873
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_13	CGACGCGCGTGTGCGCCACTCGCGGGTGTCGAGACATC	-	1764	1801
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_14	CGAACGTGAAGATCGAGCACTACTCGCTGGGAGTCGAGACATC	-	1700	1742
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_15	CGACTCGACTTGCCGCGTGCACTCCTTCGCGGCCGCTACATGTCGAGACATC	-	1627	1678
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_2	CCTAGATCCGCACCACGTCGAGACATC	-	2507	2533
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_3	CGACGGCTCCTTGCGCTCATTCGCCTCTCGTCGAGACATT	-	2446	2485
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_4	GTTCTGCGACGACCGCGACGACCAGCGCGGCGTCCTAGGGTCGAGACATC	-	2375	2424
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_5	ACCATTCGCATCGAAACCGCGCTGTCGAGACACT	-	2320	2353
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_6	CGACCCGCTGTCGTAGCCGCTCCGCGCGGCGTTGCACATTGTCGAGACACC	-	2248	2298
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_7	AGACTCCAGGTGCGCCACGAATTGCGTCCCGTCGCGGTCGAGACATC	-	2180	2226
mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_8	CGACGAGCGCCTCTGCGACCGGCCCGAGCTCTCCGGTCGAGTCGAGACATC	-	2108	2158

mVC_00397-cat_3	pH4.5_3_CRISPR_08-S_9	TGACGCTAAATTGTGACACGTCCGTCCTGGTAGATGTAGTCGAGACACT	-	2038	2086
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_1	ATCAAGCTCGTCCTCCAGGCCGAGGTTGCGTACTCGGGTCGCAGAACT	+	911	958
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_10	CAGACTACGGAGGATGACCTATTGAAGACGTTCCCGCGTCGCAGAACC	+	1565	1612
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_2	ATGCCTAGCGACCCGTCCCCTCTATCTTGGGGAAGGTCGCAGAACT	+	984	1029
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_3	CCGACAGTGATGGCGACGAGTTCCTACGATACGCCAAGTCACAGAACC	+	1055	1102
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_4	ATCCTCACTCGCCCGCGCATAAGTCGCGAGCCGGTTGTCGCAGAACC	+	1128	1174
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_5	ATGCGGCCGTGGAACATGTTGACGAGCGTATCCGCGTCGCAGAACC	+	1200	1245
mVC_13626-cat_3	pH4.5_3_CRISPR_16-S_5	ATGCGGCCGTGGAACATGTTGACGAGCGTATCCGCGTCGCAGAACC	-	14	59
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_6	CCACTACTGCGCCGACGGCGGCTCCTTGTCGCGCAGCGGTCGCAGAACC	+	1271	1319
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_7	ATGGCGACCCTGTTCGATCGCCTCAGCGGCGACATGTCGCAGAACC	+	1345	1390
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_8	CTGCCGTGCGCGGAAGACTTCCGGCTGAGCAACGCCGAGTCGCAGAACC	+	1416	1464
mVC_00397-cat_3	pH4.5_3_CRISPR_16-S_9	GTGGAAGCTCGAGTTCCACTTCGAGCAGGCGTCGTCCGAGTCGCAGAACC	+	1490	1539
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_1	TGAGTCCGCGGCGCTTGAGTTCGGTTCGGACGGT	-	55289	55322
mVC_08211-cat_3	pH4.5_3_CRISPR_17-S_1	TGAGTCCGCGGCGCTTGAGTTCGGTTCGGACGGT	+	9806	9839
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_2	CTTTTTGAAGTTCTCGGGTCGTTTGGTGACGAGAAT	-	58261	58296
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_3	CTTTTCGCGCGCCATCATTTCCCCCGCACGGAGTC	-	56428	56462
mVC_08211-cat_3	pH4.5_3_CRISPR_17-S_3	CTTTTCGCGCGCCATCATTTCCCCCGCACGGAGTC	+	8666	8700
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_4	CCCCGAAGCATGGGGAGAGAGACGACCCTCTAGCGC	+	56763	56798
mVC_08211-cat_3	pH4.5_3_CRISPR_17-S_4	CCCCGAAGCATGGGGAGAGAGACGACCCTCTAGCGC	-	8330	8365
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_5	CGGCGCGGTATTGTCGCCGCGCTCGTTGCATCC	+	59130	59162
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_8	GTCCCTCAAGAACGAACTTGTGAACGGCGCGGCC	-	53840	53873
mVC_12213-cat_2	pH4.5_3_CRISPR_17-S_9	CCGCCAGACTGGCGCGACCATCCTGAAGGTTTC	+	57639	57671
mVC_08211-cat_3	pH4.5_3_CRISPR_17-S_9	CCGCCAGACTGGCGCGACCATCCTGAAGGTTTC	-	6308	6340
mVC_12213-cat_2	pH4.5_3_CRISPR_22-S_4	CTCATCGACAGCATCAACAGACGAGCCGCCTGACCAT	-	38318	38354
mVC_12213-cat_2	pH4.5_3_CRISPR_27-S_7	TCTGCGGGCCGTATGGCATCCGTTGTGAGTCAGC	-	42763	42796
mVC_01376-cat_3	pH4.5_3_CRISPR_28-S_1	CCTGGCGGCGCCGTAACCATGTGAATTTTGCCGGCA	-	3297	3332
mVC_13605-cat_3	pH4.5_3_CRISPR_28-S_1	CCTGGCGGCGCCGTAACCATGTGAATTTTGCCGGCA	+	1562	1597
mVC_08964-cat_3	pH4.5_3_CRISPR_28-S_1	CCTGGCGGCGCCGTAACCATGTGAATTTTGCCGGCA	-	9160	9195
mVC_12494-cat_3	pH4.5_3_CRISPR_32-S_3	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	-	23220	23254
mVC_01376-cat_3	pH4.5_3_CRISPR_32-S_3	GAGTTCGTGCGCGGTAAGACGGTCAAGATATACGC	+	5771	5805
mVC_07580-cat_2	pH4.5_3_CRISPR_40-S_2	GAAAGAGTGACACAAGATACTGATTACTCACAA	+	14772	14804

mVC_07580-cat_2	pH4.5_3_CRISPR_40-S_4	GCGGCCAATACGCCGTCGAGGCGCCCAAGGTTCC	+	21335	21368
mVC_12213-cat_2	pH4.5_3_CRISPR_40-S_5	CTACAACCGAGTCATATGCACGGACTTCCTTTCGC	-	39814	39848
mVC_07580-cat_2	pH4.5_3_CRISPR_48-S_2	GACGGCCAGCGGGCGCGGGGCCATATCACGCAGC	-	13675	13708
mVC_07580-cat_2	pH4.5_3_CRISPR_48-S_4	GAGCGTGGCGGCAAGATTTTGCAGGTGCGGCTG	-	12458	12490
mVC_12213-cat_2	pH4.5_3_CRISPR_53-S_1	GACGCTCCCCCATTCCAGGCGGCCGTATTCGTT	+	40424	40456
mVC_12213-cat_2	pH4.5_3_CRISPR_53-S_4	CCTGCGCGTCGCGCATGGCCGCCACCATGCAGTG	+	38903	38936
mVC_01376-cat_3	pH4.5_3_CRISPR_82-S_2	CCGAAGGAACTCCGATCCATCGTCACCGATGAGAAGA	+	1194	1230
mVC_13605-cat_3	pH4.5_3_CRISPR_82-S_2	CCGAAGGAACTCCGATCCATCGTCACCGATGAGAAGA	-	3664	3700
mVC_08964-cat_3	pH4.5_3_CRISPR_82-S_2	CCGAAGGAACTCCGATCCATCGTCACCGATGAGAAGA	+	7057	7093
mVC_29706-cat_2	pH7.5_1_CRISPR_03-S_4	CAGCTTGCGATGAGATCGTCGCGCGCGGCGTG	-	38327	38358
mVC_20191-cat_2	pH7.5_1_CRISPR_03-S_4	CAGCTTGCGATGAGATCGTCGCGCGCGGCGTG	+	15656	15687
mVC_20158-cat_2	pH7.5_1_CRISPR_03-S_7	AGCCGAGGATGTAGCAGACCTCTCGCACATCG	-	15441	15472
mVC_20187-cat_2	pH7.5_1_CRISPR_06-S_2	GCGAGACGGTCTTCGCGGCGCTCGCCTTGTTGAGCGT	-	36878	36914
mVC_29706-cat_2	pH7.5_1_CRISPR_12-S_19	CCGGCAATCGAGGGCATTTACGCCGCTGCTGG	+	49129	49160
mVC_20191-cat_2	pH7.5_1_CRISPR_12-S_19	CCGGCAATCGAGGGCATTTACGCCGCTGCTGG	-	4854	4885
mVC_20187-cat_2	pH7.5_2_CRISPR_02-S_5	GCGAGACGGTCTTCGCGGCGCTCGCCTTGTTGAGCGT	-	36878	36914
mVC_20457-cat_2	pH7.5_2_CRISPR_03-S_44	GCTGCCGACACGCTGCTCCAGGCGGCGATCGACGCC	-	162	197
mVC_20457-cat_2	pH7.5_2_CRISPR_03-S_44	GCTGCCGACACGCTGCTCCAGGCGGCGATCGACGCC	-	54	89
mVC_29706-cat_2	pH7.5_2_CRISPR_05-S_1	CCAGCAGCGGCGTAAATGCCCTCGATTGCCGG	-	49129	49160
mVC_20191-cat_2	pH7.5_2_CRISPR_05-S_1	CCAGCAGCGGCGTAAATGCCCTCGATTGCCGG	+	4854	4885
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_1	TCGACTGCAAGGACTGCCGGCACTATCGGC	+	3760	3789
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_2	CCCATCTCGAATAGCCGCGACACGCCCTCA	+	3826	3855
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_3	CGTGGTCGAAACGATGGGACATCATCGCCA	+	3892	3921
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_4	GCCTGCCTGGAAATAGCCCTTCCACACGAC	+	3958	3987
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_5	AATCGCTCCGATATCAGGAACGCGCTCAAA	+	4024	4053
mVC_20844-cat_3	pH7.5_2_CRISPR_08-S_6	GGGATGACCAAGGGCGAGTTGATCCAGTTC	+	4090	4119
mVC_20457-cat_2	pH7.5_2_CRISPR_103-S_3	CCGTGATTGGTCGGGCGGGACATCACCATGCCTGCATAGT	+	11277	11316
mVC_21051-cat_3	pH7.5_2_CRISPR_138-S_2	TCGCCGACGATCTTCGCCGCGCACAATGCGAA	+	8197	8228
mVC_20158-cat_2	pH7.5_2_CRISPR_15-S_13	CGATGTGCGAGAGGTCTGCTACATCCTCGGCT	+	15441	15472
mVC_29706-cat_2	pH7.5_2_CRISPR_15-S_16	CACGCCGCGCGCGACGATCTCATCGCAAGCTG	+	38327	38358

mVC_20191-cat_2	pH7.5_2_CRISPR_15-S_16	CACGCCGCGCGCGACGATCTCATCGCAAGCTG	-	15656	15687
mVC_20457-cat_2	pH7.5_2_CRISPR_152-S_1	GTGGGCGGATGCATGCTCGTGATGTTCAAGGAGATCGAT	+	12293	12331
mVC_20180-cat_2	pH7.5_2_CRISPR_21-S_44	TGTAACCGGTCGACCCAGACGTTTTTGTCCAA	-	42238	42269
mVC_20457-cat_2	pH7.5_2_CRISPR_63-S_2	CGCCTCCGCCGATATCGACAAGATGGCGGCGCTCGCGAA	-	9816	9854
mVC_16898-cat_2	pH7.5_2_CRISPR_87-S_2	TTTCGCGATCCGTTCAGCGAGGACGTTAACGA	+	18457	18488
mVC_20126-cat_2	pH7.5_2_CRISPR_98-S_6	GCGCATCGCCGATCTCACGGCCGAGGTCCACAG	-	6525	6557
mVC_29706-cat_2	pH7.5_3_CRISPR_01-S_4	CAGCTTGCGATGAGATCGTCGCGCGCGGCGTG	-	38327	38358
mVC_20191-cat_2	pH7.5_3_CRISPR_01-S_4	CAGCTTGCGATGAGATCGTCGCGCGCGGCGTG	+	15656	15687
mVC_20158-cat_2	pH7.5_3_CRISPR_01-S_7	AGCCGAGGATGTAGCAGACCTCTCGCACATCG	-	15441	15472
mVC_20457-cat_2	pH7.5_3_CRISPR_02-S_49	GCTGCCGACACGCTGCTCCAGGCGGCGATCGACGCC	-	162	197
mVC_20457-cat_2	pH7.5_3_CRISPR_02-S_49	GCTGCCGACACGCTGCTCCAGGCGGCGATCGACGCC	-	54	89
mVC_20187-cat_2	pH7.5_3_CRISPR_03-S_5	GCGAGACGGTCTTCGCGGCGCTCGCCTTGTTGAGCGT	-	36878	36914
mVC_29706-cat_2	pH7.5_3_CRISPR_05-S_40	CCGGCAATCGAGGGCATTTACGCCGCTGCTGG	+	49129	49160
mVC_20191-cat_2	pH7.5_3_CRISPR_05-S_40	CCGGCAATCGAGGGCATTTACGCCGCTGCTGG	-	4854	4885
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_1	GAACTGGATCAACTCGCCCTTGGTCATCCC	-	4090	4119
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_2	TTTGAGCGCGTTCCTGATATCGGAGCGATT	-	4024	4053
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_3	GTCGTGTGGAAGGGCTATTTCCAGGCAGGC	-	3958	3987
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_4	TGGCGATGATGTCCCATCGTTTCGACCACG	-	3892	3921
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_5	TGAGGGCGTGTCGCGGCTATTCGAGATGGG	-	3826	3855
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_6	GCCGATAGTGCCGGCAGTCCTTGCAGTCGA	-	3760	3789
mVC_20844-cat_3	pH7.5_3_CRISPR_06-S_7	ATGCGTTCAGGGGAGGCCGCA	-	3703	3723
mVC_20180-cat_2	pH7.5_3_CRISPR_48-S_7	TTGGACAAAAACGTCTGGGTCGACCGGTTACA	+	42238	42269

Start[1] and End[2] position of spacer sequences within the mVCs.

Supplementary Table 4.2. Taxonomic annotation of contigs that contained CRISPR arrays.

Contig ID	CRISPR ID	Class	Family	Genus
c_01020	pH4.5_1-CRISPR-03	*Alphaproteobacteria*	*Rhodobacteraceae*	*Pannonibacter*
c_01322	pH4.5_1-CRISPR-05	*Alphaproteobacteria*	*Bradyrhizobiaceae*	
c_02851	pH4.5_1-CRISPR-09			
c_04960	pH4.5_1-CRISPR-10			
c_07580	pH4.5_2-CRISPR-00	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_12123	pH4.5_2-CRISPR-00	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_09682	pH4.5_2-CRISPR-03			
c_00397	pH4.5_2-CRISPR-10			
c_12123	pH4.5_3-CRISPR-00	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_12195	pH4.5_3-CRISPR-01	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_00397	pH4.5_3-CRISPR-08			
c_04960	pH4.5_3-CRISPR-32			
c_18269	pH7.5_1-CRISPR-03	*Alphaproteobacteria*	*Methylobacteriaceae*	*Methylobacterium*
c_29866	pH7.5_1-CRISPR-12	*Alphaproteobacteria*	*Xanthobacteraceae*	*Xanthobacter*
c_20392	pH7.5_2-CRISPR-02	*Alphaproteobacteria*	*Beijerinckiaceae*	*Beijerinckia*
c_20455	pH7.5_2-CRISPR-03	*Alphaproteobacteria*	*Rhodobacteraceae*	*Pannonibacter*
c_29866	pH7.5_2-CRISPR-05	*Alphaproteobacteria*	*Xanthobacteraceae*	*Xanthobacter*
c_20844	pH7.5_2-CRISPR-08	*Gammaproteobacteria*	*Pseudomonadaceae*	*Pseudomonas*
c_25299	pH7.5_2-CRISPR-15	*Betaproteobacteria*	*Burkholderiaceae*	*Burkholderia*
c_29713	pH7.5_2-CRISPR-15	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_25299	pH7.5_3-CRISPR-01	*Betaproteobacteria*	*Burkholderiaceae*	*Burkholderia*
c_29713	pH7.5_3-CRISPR-01	*Alphaproteobacteria*	*Methylocystaceae*	*Methylocystis*
c_29737	pH7.5_3-CRISPR-02	*Alphaproteobacteria*	*Chelatococcaceae*	*Chelatococcus*
c_20392	pH7.5_3-CRISPR-03	*Alphaproteobacteria*	*Beijerinckiaceae*	*Beijerinckia*
c_29781	pH7.5_3-CRISPR-03	*Deltaproteobacteria*	*Desulfovibrionaceae*	*Desulfovibrio*
c_29866	pH7.5_3-CRISPR-05	*Alphaproteobacteria*	*Xanthobacteraceae*	*Xanthobacter*
c_20844	pH7.5_3-CRISPR-06	*Gammaproteobacteria*	*Pseudomonadaceae*	*Pseudomonas*

Supplementary Table 4.3. BLASTp alignment between protein sequences of metagenomic assembled genomes (MAGs) and protein sequences from methanotroph-associated metagenomic viral contigs (mVCs) with identity > 30%, E value < $10e^{-5}$, bit score > 50 and query cover > 70% cutoff. The length of query and subject are presented in bp.

Query ID (Viral protein ID)	VP AA length[1]	Subject ID (Host protein ID)	HP AA length[2]	Identity (%)	E value	bit score	Query cover
mVC_000108-cat_2_1	379	bin.21_02902	404	48.29	6.87e-110	325	98
mVC_000108-cat_2_1	379	bin.14_01987	405	45.57	4.84e-105	313	98
mVC_000108-cat_2_1	379	bin.14_01627	405	35.82	1.15e-48	167	96
mVC_000108-cat_2_3	152	bin.21_01140	139	58.33	2.44e-50	155	99
mVC_000108-cat_2_3	152	bin.2_02473	150	48.66	7.88e-35	116	97
mVC_000108-cat_2_4	79	bin.2_02479	82	53.24	7.38e-20	73.9	94
mVC_000108-cat_2_43	213	bin.21_02323	176	30.81	5.18e-13	62.8	79
mVC_000108-cat_2_43	213	bin.14_00936	176	31.55	3.96e-12	60.5	79
mVC_007548-cat_2_29	298	bin.21_00796	301	42.24	3.09e-85	256	99
mVC_007548-cat_2_31	347	bin.21_00795	316	51.52	3.07e-94	281	94
mVC_007548-cat_2_49	170	bin.21_01541	169	63.69	4.21e-69	205	99
mVC_007548-cat_2_49	170	bin.2_02674	243	43.29	4.40e-37	125	96
mVC_007548-cat_2_49	170	bin.14_03309	183	43.75	1.66e-32	112	85
mVC_007548-cat_2_49	170	bin.2_02140	180	47.91	9.58e-31	107	85
mVC_007548-cat_2_49	170	bin.21_02786	180	45.13	1.82e-29	104	85
mVC_007548-cat_2_49	170	bin.14_03464	241	42.44	2.37e-27	100	82
mVC_007548-cat_2_49	170	bin.14_00002	393	30.65	1.10e-06	45.4	78
mVC_007548-cat_2_9	81	bin.21_01224	77	36.36	2.10e-08	44.7	91
mVC_008211-cat_3_1	370	bin.21_01138	370	99.45	0.0	746	100

mVC_008211-cat_3_17	379	bin.21_00915	365	54.69	1.25e-139	400	98
mVC_008211-cat_3_17	379	bin.14_02736	382	55.08	1.20e-138	398	99
mVC_008211-cat_3_2	70	bin.21_01139	70	98.57	1.86e-46	140	100
mVC_008211-cat_3_3	146	bin.21_01140	139	93.47	6.08e-96	271	95
mVC_008211-cat_3_3	146	bin.2_02473	150	50.00	8.28e-39	126	95
mVC_012213-cat_2_50	373	Bin.14_00929	372	50.93	2.59e-123	358	100
mVC_012213-cat_2_50	373	bin.21_00585	372	50.40	6.15e-122	355	100
mVC_012249-cat_2_12	150	bin.21_00988	145	34.00	8.98e-12	57.4	95
mVC_012249-cat_2_12	150	bin.2_02315	144	34.18	1.21e-09	51.6	73
mVC_012249-cat_2_37	213	bin.21_02323	176	30.81	5.18e-13	62.8	79
mVC_012249-cat_2_37	213	bin.14_00936	176	31.55	3.96e-12	60.5	79
mVC_012249-cat_2_43	182	bin.21_00974	211	32.86	3.11e-15	68.6	73
mVC_012249-cat_2_9	216	bin.21_01239	276	85.22	1.08e-110	316	81
mVC_012249-cat_2_9	216	bin.21_00984	272	83.05	1.83e-109	313	81
mVC_020187-cat_2_15	95	bin.2_01433	86	72.00	9.22e-33	107	79
mVC_020187-cat_2_15	95	bin.14_00738	80	71.23	5.91e-31	102	77
mVC_020187-cat_2_15	95	bin.21_01207	84	67.12	2.02e-30	101	77
mVC_020187-cat_2_15	95	bin.21_02304	85	68.49	8.33e-26	89.7	77
mVC_020187-cat_2_15	95	bin.21_02731	89	41.02	4.15e-16	65.1	82
mVC_020187-cat_2_25	594	bin.2_00624	532	31.22	4.69e-70	233	87
mVC_020187-cat_2_3	380	bin.2_02481	380	97.14	0.0	693	92
mVC_020187-cat_2_39	212	bin.14_00273	212	61.50	6.92e-88	256	99
mVC_020187-cat_2_39	212	bin.2_02313	213	62.25	4.33e-83	244	94

mVC_020187-cat_2_39	212	bin.21_00990	212	57.67	1.23e-81	240	100
mVC_020187-cat_2_40	297	bin.2_02314	292	42.08	7.27e-77	234	99
mVC_020187-cat_2_40	297	bin.21_00989	294	40.54	5.04e-66	206	99
mVC_020187-cat_2_40	297	bin.14_00274	296	39.73	2.25e-64	202	99
mVC_020187-cat_2_41	148	bin.21_00988	145	54.54	7.73e-52	159	97
mVC_020187-cat_2_41	148	bin.14_00275	142	54.54	1.13e-51	159	97
mVC_020187-cat_2_41	148	bin.2_02315	144	53.84	4.80e-48	150	97
mVC_020187-cat_2_42	1368	bin.2_02316	1285	38.75	0.0	831	100
mVC_020187-cat_2_42	1368	bin.14_00276	1292	38.08	0.0	796	97
mVC_020187-cat_2_42	1368	bin.21_00987	1296	37.13	0.0	768	97
mVC_020187-cat_2_43	526	bin.2_02317	526	70.39	1.64e-127	380	100
mVC_020187-cat_2_44	183	bin.2_02301	188	85.95	1.22e-115	324	97
mVC_020187-cat_2_47	82	bin.21_01240	80	59.45	3.53e-27	92.4	90
mVC_020187-cat_2_47	82	bin.21_00983	84	59.45	1.45e-26	90.9	90
mVC_020187-cat_2_48	63	bin.2_02669	69	46.15	2.26e-10	48.9	83
mVC_020187-cat_2_48	63	bin.2_00434	69	46.15	2.26e-10	48.9	83

[1]VP AA= Viral Protein, Amino Acid ; [2]HP AA= Host Protein, Amino Acid

Supplementary Table 4.4. Gene annotation of the metagenomic viral contigs (mVCs) that were linked to metagenomic assembled genomes (MAGs) via CRISPR arrays. Genes were annotated with VIPTree using the NCBI-nr database.

Gene ID	Hit gene	Organism	Identity (%)	E value	bit score
mVC_00108-cat_2_1	Integrase	*Methylocystis*	67.1	1.1e-141	511
mVC_00108-cat_2_2	hypothetical protein	*Methylosinus*	38.6	5.5e-13	81
mVC_00108-cat_2_3	Uncharacterized protein	*Methylovirgula*	66	7.0e-13	81
mVC_00108-cat_2_4	Uncharacterized protein	*Methylosinus*	49.4	2.2e-9	69
mVC_00108-cat_2_5	ASCH domain-containing protein	*Nitrosospira*	67.7	1.2e-39	170
mVC_00108-cat_2_7	Uncharacterized protein	Marine sediment metagenome	46.6	1.2e-7	64
mVC_00108-cat_2_9	Uncharacterized protein	*Alphaproteobacteria bacterium*	62	3.8e-54	218
mVC_00108-cat_2_11	Uncharacterized protein	Candidatus *Propionivibrio*	40.7	1.2e-15	90
mVC_00108-cat_2_12	Uncharacterized protein	*Sphingomonas*	46.2	3.6e-49	203
mVC_00108-cat_2_14	Uncharacterized protein	*Oryza sativa*	41.8	3.6e-4	52
mVC_00108-cat_2_15	Uncharacterized protein	*Methylocystis*	49.3	3.4e-7	62
mVC_00108-cat_2_16	ParB domain protein nuclease	*Methylocystis*	45	1.2e-124	456
mVC_00108-cat_2_24	Uncharacterized protein	*Nitrospira*	43.8	2.0e-15	89
mVC_00108-cat_2_26	helix-turn-helix domain-containing protein	*Azospirillum*	47.5	4.4e-7	62
mVC_00108-cat_2_30	Uncharacterized protein	*Methylobacterium*	34.2	1.8e-11	76
mVC_00108-cat_2_32	Bacteriophage protein (Modular protein)	*uncultured*	61.7	3.6e-12	79
mVC_00108-cat_2_36	MT-A70 family protein	*Methylosinus*	52	2.1e-86	327
mVC_00108-cat_2_37	Uncharacterized protein	*Rhodoblastus*	37.1	4.3e-18	98
mVC_00108-cat_2_38	phosphohydrolase	*Bosea*	58.2	8.2e-46	190
mVC_00108-cat_2_39	DEAD/DEAH box helicase	*Methylosinus*	60.6	4.4e-200	706
mVC_00108-cat_2_40	Uncharacterized protein	*Rhodoblastus*	46.3	4.7e-9	68
mVC_00108-cat_2_42	hypothetical protein	*Mesorhizobium*	42	3.5e-4	52
mVC_00108-cat_2_43	hypothetical protein	*Pseudomonas*	30.9	2.8e-10	74
mVC_00108-cat_2_44	Uncharacterized protein	*Methylocystis*	48.3	7.0e-38	165

mVC_00108-cat_2_45	HNH endonuclease	*Methylosinus*	57.3	6.1e-25	121
mVC_00108-cat_2_46	phage terminase small subunit P27 family	*Pleomorphomonas*	33.3	4.3e-10	72
mVC_00108-cat_2_47	Uncharacterized protein	*Methylosinus*	61.7	5.6e-152	546
mVC_00108-cat_2_48	phage portal protein	*Rhizobium*	57.4	9.6e-132	478
mVC_00108-cat_2_49	hypothetical protein	*Rhizobium*	40.7	1.1e-77	299
mVC_00108-cat_2_50	phage major capsid protein	*Rhizobium*	70.7	3.2e-184	653
mVC_00108-cat_2_51	hypothetical protein	*Rhizobium*	51.9	2.4e-37	162
mVC_00108-cat_2_52	hypothetical protein	*Rhizobium*	64	4.1e-85	322
mVC_00108-cat_2_53	Uncharacterized protein	*Azospirillum*	47.2	2.6e-2	46
mVC_00108-cat_2_54	Uncharacterized protein	*Devosia*	45.3	9.2e-13	80
mVC_00108-cat_2_57	hypothetical protein	*Rhizobiales*	43.7	1.4e-43	184
mVC_00108-cat_2_58	Uncharacterized protein	*Methylovirgula*	52.1	2.5e-29	136
mVC_00108-cat_2_59	Uncharacterized protein	*Methylovirgula*	45.1	1.3e-19	103
mVC_00108-cat_2_60	Uncharacterized protein	*Methylovirgula*	27.8	7.8e-4	51
mVC_00108-cat_2_61	Uncharacterized protein	*Roseiarcus*	40.9	1.1e-18	100
mVC_00108-cat_2_63	hypothetical protein	*Methyloferula*	53.3	2.5e-58	233
mVC_00108-cat_2_64	Uncharacterized protein	*Beijerinckia*	34.3	5.5e-5	55
mVC_07548-cat_2_1	Cell wall hydrolase	*Methylocystis*	97.4	1.0e-14	87
mVC_07548-cat_2_2	Uncharacterized protein	*Ruminococcaceae*	29.8	1.1e-17	98
mVC_07548-cat_2_3	Uncharacterized protein	*Chloroflexi*	33.6	1.1e-19	104
mVC_07548-cat_2_4	hypothetical protein	*Salinispora*	25.8	2.7e-4	54
mVC_07548-cat_2_6	hypothetical protein	*Pseudomonas*	54.9	1.5e-52	213
mVC_07548-cat_2_7	Uncharacterized protein	*Rhodoblastus*	45.8	2.9e-68	267
mVC_07548-cat_2_8	phage tail protein	*Alphaproteobacteria*	54.7	1.0e-11	77
mVC_07548-cat_2_9	Uncharacterized protein	*Bosea*	36.7	5.5e-5	55
mVC_07548-cat_2_10	Uncharacterized protein	*Rhodoblastus*	44.9	1.8e-24	119
mVC_07548-cat_2_11	phage tail tape measure protein	*Methylocystis*	47.3	7.2e-124	453
mVC_07548-cat_2_13	Uncharacterized protein	*Hartmannibacter*	27.6	7.5e-2	44

mVC_07548-cat_2_14	Uncharacterized protein	*Rhodoblastus*	52.2	6.7e-43	181
mVC_07548-cat_2_15	Uncharacterized protein	*Rhodoblastus*	44.3	3.2e-94	354
mVC_07548-cat_2_16	hypothetical protein	*Methylobacter*	36.2	4.1e-8	65
mVC_07548-cat_2_17	Uncharacterized protein	*Rhodospirillales*	43.3	4.2e-49	204
mVC_07548-cat_2_18	Uncharacterized protein	*Methylocystis*	66.1	3.8e-295	1022
mVC_07548-cat_2_19	phage tail protein I	*Methylosinus*	61.8	8.2e-66	257
mVC_07548-cat_2_20	baseplate assembly protein	*Methylocystis*	56	1.2e-86	328
mVC_07548-cat_2_21	baseplate assembly protein	*Bosea*	52.2	2.4e-24	119
mVC_07548-cat_2_23	Uncharacterized protein	*Blastochloris*	29.6	6.3e-9	68
mVC_07548-cat_2_24	Uncharacterized protein	*Rhodoblastus*	36	2.8e-28	132
mVC_07548-cat_2_25	Uncharacterized protein	*Rhodoblastus*	47.2	2.6e-31	142
mVC_07548-cat_2_26	DUF1320 domain-containing protein	*Rhodoblastus*	45.5	3.1e-24	119
mVC_07548-cat_2_28	Uncharacterized protein	*Ensifer*	54.2	5.9e-88	332
mVC_07548-cat_2_29	hypothetical protein	*Blastochloris*	49.6	2.6e-23	116
mVC_07548-cat_2_30	Uncharacterized protein	*Rhodoblastus*	41.9	4.1e-65	257
mVC_07548-cat_2_32	Uncharacterized protein	*Rhodoblastus*	53.4	5.7e-100	374
mVC_07548-cat_2_33	Uncharacterized protein	*Rhodoblastus*	52.8	1.7e-153	551
mVC_07548-cat_2_34	Uncharacterized protein	*Methylocystis*	70.6	3.2e-207	730
mVC_07548-cat_2_35	Uncharacterized protein	*Rhodoblastus*	48.1	6.9e-43	181
mVC_07548-cat_2_36	hypothetical protein	*Methylocystis*	78.9	6.5e-35	154
mVC_07548-cat_2_37	DUF2730 family protein	*Methylocystis*	71.2	8.0e-49	200
mVC_07548-cat_2_38	hypothetical protein	*Methylocystis*	68.7	4.6e-20	105
mVC_07548-cat_2_39	hypothetical protein	*Methylocystis*	42.9	4.3e-2	45
mVC_07548-cat_2_44	D-alanyl-D-alanine carboxypeptidase-like protein	*Bosea*	59.7	4.8e-73	282
mVC_07548-cat_2_45	hypothetical protein	*Methylocystis*	46.8	1.3e-22	113
mVC_07548-cat_2_46	hypothetical protein	*Sphingomonas*	53	1.6e-47	196
mVC_07548-cat_2_48	Phosphatidylserine synthase	*Methylocystis*	70.3	7.6e-55	220

mVC_07548-cat_2_49	Uncharacterized protein	*Aurantimonas*	46.3	6.9e-24	117
mVC_07548-cat_2_50	hypothetical protein	*Methylocystis*	52.9	1.1e-10	74
mVC_07548-cat_2_51	hypothetical protein	*Methylocystis*	57	1.0e-27	130
mVC_07548-cat_2_53	regulatory protein GemA	*Methylocystis*	53.8	3.8e-55	222
mVC_07548-cat_2_55	hypothetical protein	*Methylocystis*	63.9	9.7e-18	97
mVC_07548-cat_2_56	Uncharacterized protein	*Afifella*	68.7	4.4e-77	295
mVC_07548-cat_2_57	hypothetical protein	*Methylocystis*	48.9	2.3e-60	240
mVC_07548-cat_2_58	hypothetical protein	*Methylocystis*	35	1.7e-6	60
mVC_07548-cat_2_59	hypothetical protein	*Rhodopseudomonas*	52.1	6.7e-27	127
mVC_07548-cat_2_60	hypothetical protein	*Methylocystis*	36.5	2.7e-32	146
mVC_07548-cat_2_61	Transposase	*Methylobacterium*	34.9	2.3e-35	157
mVC_07548-cat_2_62	hypothetical protein	*Methylocystis*	71	3.9e-283	982
mVC_07548-cat_2_63	hypothetical protein	*Methylobacterium*	53.4	1.3e-9	70
mVC_07548-cat_2_64	hypothetical protein	*Methylocystis*	61.2	2.1e-84	320
mVC_07548-cat_2_65	hypothetical protein	*Methylocystis*	56.7	9.4e-21	107
mVC_07548-cat_2_66	Uncharacterized protein	*Methylocystis*	55.6	1.2e-64	254
mVC_07548-cat_2_69	Uncharacterized protein	*Ensifer*	59.7	2.2e-14	86
mVC_07548-cat_2_71	Aminotransferase class IV	*Methylocystis*	73.7	1.2e-4	54
mVC_08211-cat_3_1	Integrase	*Beijerinckiaceae*	51.3	9.8e-92	345
mVC_08211-cat_3_3	Uncharacterized protein	*Methylovirgula*	66.1	5.7e-15	88
mVC_08211-cat_3_5	Uncharacterized protein	*Methylocystis*	78.9	3.2e-34	152
mVC_08211-cat_3_7	Centrosomal protein of 164 kDa	*Hondaea*	32.3	2.2e-5	57
mVC_08211-cat_3_8	hypothetical protein	*Sulfitobacter*	67.7	3.7e-17	95
mVC_08211-cat_3_12	Uncharacterized protein	*Sphingomonas*	45.8	8.1e-49	202
mVC_08211-cat_3_14	Restriction endonuclease	*Mesorhizobium*	49.5	7.1e-21	107
mVC_08211-cat_3_15	ATP-binding protein	*Enterovirga*	61.8	2.0e-124	454
mVC_08211-cat_3_16	Uncharacterized protein	*Methylocapsa*	58.6	2.3e-11	76
mVC_08211-cat_3_19	Uncharacterized protein	*Kaistia*	51.4	8.8e-93	349

mVC_08211-cat_3_20	hypothetical protein	*Methylobacterium*	53.4	1.4e-61	244
mVC_08211-cat_3_21	Uncharacterized protein	*Methylobacterium*	51.4	2.8e-78	300
mVC_08211-cat_3_26	Uncharacterized protein	*Xanthobacter*	55.6	5.7e-39	170
mVC_08211-cat_3_29	hypothetical protein	*Pelagibacterium*	53.1	1.5e-26	126
mVC_12213-cat_2_6	acetylmuramidase	*Sinorhizobium*	47.2	1.7e-35	156
mVC_12213-cat_2_7	hypothetical protein	*Burkholderia*	32.4	1.3e-28	135
mVC_12213-cat_2_8	hypothetical protein	*Rhizobiales*	31.1	2.7e-57	232
mVC_12213-cat_2_9	Uncharacterized protein	*Shewanella*	62.6	2.0e-31	143
mVC_12213-cat_2_10	Uncharacterized protein	*Mesorhizobium*	41.8	4.2e-36	159
mVC_12213-cat_2_11	hypothetical protein	*Rhizobium*	38.5	8.5e-22	111
mVC_12213-cat_2_12	DUF4376 domain-containing protein	*Camelimonas*	37.5	6.6e-11	74
mVC_12213-cat_2_13	Uncharacterized protein	*Rhizobium*	37.2	5.0e-49	203
mVC_12213-cat_2_14	DUF2612 domain-containing protein	*Mesorhizobium*	54.1	1.3e-65	257
mVC_12213-cat_2_15	hypothetical protein	*Rhizobium*	56.4	1.0e-148	535
mVC_12213-cat_2_16	hypothetical protein	*Methylosinus*	45.3	1.9e-64	254
mVC_12213-cat_2_18	hypothetical protein	*Mesorhizobium*	52.9	6.9e-24	117
mVC_12213-cat_2_19	Uncharacterized protein		52.3	6.6e-45	188
mVC_12213-cat_2_20	M23 family metallopeptidase	*Methylobacterium*	30.8	2.2e-47	199
mVC_12213-cat_2_21	hypothetical protein	*Curvibacter*	53.2	5.9e-7	61
mVC_12213-cat_2_22	Uncharacterized protein	*Pseudomonas*	58.5	5.3e-37	161
mVC_12213-cat_2_23	hypothetical protein	*Burkholderia*	53.9	5.9e-36	158
mVC_12213-cat_2_24	Uncharacterized protein	*Shewanella*	55.9	6.2e-195	689
mVC_12213-cat_2_25	Uncharacterized protein	*Rhizobium*	62.4	1.0e-82	314
mVC_12213-cat_2_26	Uncharacterized protein	*Rhizobium*	42.7	6.0e-20	104
mVC_12213-cat_2_27	Uncharacterized protein	*Shewanella*	68.8	3.1e-48	198
mVC_12213-cat_2_28	DUF4054 domain-containing protein	*Methylobacterium*	55.9	3.4e-31	142
mVC_12213-cat_2_29	hypothetical protein	*Acetobacter*	39.8	2.3e-16	92
mVC_12213-cat_2_30	DUF2184 domain-containing protein	*Yersinia*	51.5	2.5e-94	354

mVC_12213-cat_2_31	hypothetical protein	*Rhizobium*	46.6	1.2e-62	248
mVC_12213-cat_2_32	DUF2213 domain-containing protein	*Rhizobium*	46.2	1.7e-129	472
mVC_12213-cat_2_33	Uncharacterized protein	*Brevundimonas*	26.5	1.1e-4	55
mVC_12213-cat_2_34	phage head morphogenesis protein	*Rhizobium*	60.8	3.3e-92	346
mVC_12213-cat_2_35	Uncharacterized protein	*Methylobacterium*	55.5	4.0e-138	500
mVC_12213-cat_2_37	hypothetical protein	*Bradyrhizobium*	63.9	1.5e-155	557
mVC_12213-cat_2_38	hypothetical protein	*Neorhizobium*	52.1	4.1e-45	188
mVC_12213-cat_2_40	Uncharacterized protein	*Ensifer*	59.7	2.2e-14	86
mVC_12213-cat_2_41	D-3-phosphoglycerate dehydrogenase	*bacterium*	51	1.8e-3	50
mVC_12213-cat_2_44	hypothetical protein	*Rhizobium*	49.4	1.9e-101	378
mVC_12213-cat_2_45	hypothetical protein	*Nitratireductor*	40.9	1.7e-6	60
mVC_12213-cat_2_46	DNA cytosine methyltransferase	*Methylosinus*	51.8	2.3e-65	257
mVC_12213-cat_2_47	Uncharacterized protein	*Microvirga*	55	2.5e-70	273
mVC_12213-cat_2_48	Vitamin B12-dependent ribonucleotide reductase	*Rhodospirillaceae*	46.4	4.2e-13	82
mVC_12213-cat_2_49	RecName: Full=Beta sliding clamp	*Methylocystis*	50.7	5.9e-92	346
mVC_12213-cat_2_50	Uncharacterized protein	*Chelatococcus*	41.1	6.2e-66	260
mVC_12213-cat_2_51	DUF4942 domain-containing protein	*Rhizobiales*	60.6	1.4e-172	614
mVC_12213-cat_2_54	hypothetical protein	*Bosea*	74.6	2.0e-80	305
mVC_12213-cat_2_56	hypothetical protein	*Methylobacterium*	39.9	6.9e-24	117
mVC_12213-cat_2_58	Uncharacterized protein	*Methylobacterium*	28.1	3.1e-8	65
mVC_12213-cat_2_61	hypothetical protein	*Labrenzia*	30.6	8.6e-3	47
mVC_12213-cat_2_62	hypothetical protein	*Pseudoruegeria*	47.4	4.2e-34	151
mVC_12213-cat_2_64	Uncharacterized protein	*Methylocystis*	77.6	3.5e-28	132
mVC_12213-cat_2_65	Putative phage repressor	*Rhodopseudomonas*	32.4	1.2e-21	111
mVC_12213-cat_2_67	Fis family transcriptional regulator	*Mesorhizobium*	44.4	6.6e-32	144
mVC_12213-cat_2_73	Uncharacterized protein	*Xanthobacter*	55	2.0e-38	168
mVC_12213-cat_2_78	Uncharacterized protein	*uncultured*	57.5	1.0e-51	210

mVC_12213-cat_2_79	Uncharacterized protein	*Devosia*	40	8.2e-22	111
mVC_12213-cat_2_80	Uncharacterized protein	*Methylobacterium*	53	1.6e-86	327
mVC_12213-cat_2_81	hypothetical protein	*Methylobacterium*	54.3	9.1e-61	241
mVC_12213-cat_2_82	Uncharacterized protein	*Kaistia*	51.6	3.3e-92	347
mVC_12213-cat_2_85	Uncharacterized protein	*Methylocapsa*	56.9	6.7e-11	74
mVC_12213-cat_2_86	Restriction endonuclease-like	*Paracoccus*	52.6	2.3e-26	127
mVC_12213-cat_2_87	Uncharacterized protein	*Rhodospirillaceae bacterium*	61.6	3.4e-50	205
mVC_12213-cat_2_93	ASCH domain-containing protein	*Fodinicurvata*	87.1	6.8e-8	64
mVC_12249-cat_2_4	Uncharacterized protein	*Methylovirgula*	82.7	6.5e-27	127
mVC_12249-cat_2_5	Uncharacterized protein	*Rhodoblastus*	40.4	4.4e-15	88
mVC_12249-cat_2_8	hypothetical protein	*Ancalomicrobiaceae bacterium*	52.3	1.1e-18	100
mVC_12249-cat_2_10	lysozyme	*Methylosinus*	60.4	1.5e-54	220
mVC_12249-cat_2_11	Uncharacterized protein	Marine sediment metagenome	22.1	5.3e-8	69
mVC_12249-cat_2_12	Uncharacterized protein	*Methylocystis*	46.9	1.2e-183	652
mVC_12249-cat_2_13	Uncharacterized protein	*Methylocystis*	65.2	2.3e-48	199
mVC_12249-cat_2_14	DUF2163 domain-containing protein	*Methylocystis*	51.7	7.5e-77	295
mVC_12249-cat_2_15	conserved protein of unknown function	*Methylocella*	46	8.7e-43	181
mVC_12249-cat_2_16	Uncharacterized protein	*Rhodoblastus*	36.2	4.0e-62	249
mVC_12249-cat_2_18	Uncharacterized protein	*Beijerinckia*	34.3	7.2e-5	54
mVC_12249-cat_2_19	hypothetical protein	*Methyloferula*	53.3	2.5e-58	233
mVC_12249-cat_2_21	Uncharacterized protein	*Roseiarcus*	40.9	1.1e-18	100
mVC_12249-cat_2_22	Uncharacterized protein	*Methylovirgula*	27.8	7.8e-4	51
mVC_12249-cat_2_23	Uncharacterized protein	*Methylovirgula*	45.1	1.3e-19	103
mVC_12249-cat_2_24	Uncharacterized protein	*Methylovirgula*	52.9	4.9e-30	138
mVC_12249-cat_2_25	hypothetical protein	*Rhizobiales bacterium*	42.3	1.3e-41	177
mVC_12249-cat_2_28	Uncharacterized protein	*Devosia*	45.3	9.2e-13	80
mVC_12249-cat_2_29	Uncharacterized protein	*Azospirillum*	47.2	2.6e-2	46
mVC_12249-cat_2_30	hypothetical protein	*Rhizobium*	64.4	8.2e-86	324

mVC_12249-cat_2_31	hypothetical protein	*Rhizobium*	51.9	2.4e-37	162
mVC_12249-cat_2_32	phage major capsid protein	*Rhizobium*	69	6.9e-179	635
mVC_12249-cat_2_33	hypothetical protein	*Rhizobium*	40.7	8.6e-78	299
mVC_12249-cat_2_34	phage portal protein	*Rhizobium*	57.4	1.6e-131	478
mVC_12249-cat_2_35	Terminase	*Roseomonas*	51.2	1.9e-152	548
mVC_12249-cat_2_36	phage terminase small subunit P27 family	*Pleomorphomonas*	33.3	4.3e-10	72
mVC_12249-cat_2_37	HNH endonuclease	*Methylosinus*	57.3	6.1e-25	121
mVC_12249-cat_2_38	Uncharacterized protein	*Methylocystis*	48.3	7.0e-38	165
mVC_12249-cat_2_39	hypothetical protein	*Pseudomonas*	30.9	2.8e-10	74
mVC_12249-cat_2_40	hypothetical protein	*Devosia*	35.9	1.2e-4	54
mVC_12249-cat_2_42	Uncharacterized protein	Marine sediment metagenome	46.3	4.7e-9	68
mVC_12249-cat_2_43	DEAD/DEAH box helicase	*Methylosinus*	62.8	1.7e-199	704
mVC_12249-cat_2_44	phosphohydrolase	*Bosea*	57.6	5.3e-45	188
mVC_12249-cat_2_45	Uncharacterized protein	*Rhodoblastus*	37.8	1.5e-18	100
mVC_12249-cat_2_46	MULTISPECIES: MT-A70 family protein	*Methylosinus*	52	2.1e-86	327
mVC_12249-cat_2_50	Bacteriophage protein (Modular protein)	*uncultured*	61.7	3.6e-12	79
mVC_12249-cat_2_52	Uncharacterized protein	*Methylobacterium*	34.2	1.8e-11	76
mVC_12249-cat_2_56	helix-turn-helix domain-containing protein	*Azospirillum*	47.5	4.4e-7	62
mVC_12249-cat_2_58	Uncharacterized protein	*Nitrospira*	43.8	2.0e-15	89
mVC_12249-cat_2_65	ParB domain protein nuclease	*Methylocystis*	45.6	2.2e-126	461
mVC_12249-cat_2_66	Uncharacterized protein	*Methylocystis*	45.5	2.6e-7	62
mVC_12249-cat_2_67	Uncharacterized protein	*Oryza*	41.8	3.6e-4	52
mVC_12249-cat_2_69	Uncharacterized protein	*Sphingomonas*	45.8	8.1e-49	202
mVC_12249-cat_2_73	Uncharacterized protein	Marine sediment metagenome	46.6	1.2e-7	64
mVC_12249-cat_2_75	ASCH domain-containing protein	*Fodinicurvata*	87.1	6.8e-8	64
mVC_20187-cat_2_1	hypothetical protein	*Pantoea*	53.8	1.8e-16	93
mVC_20187-cat_2_2	SIR2 family protein	*Salinisphaera*	61.6	3.1e-298	1033
mVC_20187-cat_2_5	MULTISPECIES: site-specific integrase	*Methylosinus*	69.5	5.4e-141	509

mVC_20187-cat_2_6	hypothetical protein	*Methylosinus*	44.1	5.0e-27	128
mVC_20187-cat_2_7	MULTISPECIES: chromosome partitioning protein ParB	*Methylosinus*	57.4	4.8e-210	740
mVC_20187-cat_2_8	MULTISPECIES: hypothetical protein	*Methylosinus*	67.7	1.0e-14	87
mVC_20187-cat_2_9	hypothetical protein	*Methylosinus*	43	4.3e-11	75
mVC_20187-cat_2_10	Uncharacterized protein	*Methylocystis*	48.1	7.3e-2	44
mVC_20187-cat_2_11	MULTISPECIES: hypothetical protein	*Methylosinus*	56.9	4.6e-54	218
mVC_20187-cat_2_12	XRE family transcriptional regulator	*Methylosinus*	55.7	1.7e-14	86
mVC_20187-cat_2_15	hypothetical protein	*Methylosinus*	64.6	1.1e-42	180
mVC_20187-cat_2_16	hypothetical protein	*Pseudomonas*	49.2	9.8e-23	114
mVC_20187-cat_2_17	MULTISPECIES: DUF2312 domain-containing protein	*Methylosinus*	77.9	7.3e-34	151
mVC_20187-cat_2_18	MULTISPECIES: hypothetical protein	*Methylosinus*	74.5	6.5e-14	84
mVC_20187-cat_2_20	MULTISPECIES: hypothetical protein	*Methylosinus*	66.7	5.7e-63	247
mVC_20187-cat_2_21	MULTISPECIES: MT-A70 family protein	*Methylosinus*	67.2	2.8e-119	436
mVC_20187-cat_2_22	MULTISPECIES: hypothetical protein	*Methylosinus*	58.4	1.4e-45	190
mVC_20187-cat_2_23	helix-turn-helix domain-containing protein	*Methylocystis*	50.5	6.5e-79	302
mVC_20187-cat_2_24	Uncharacterized protein	*Methylosinus*	76.3	1.5e-88	333
mVC_20187-cat_2_25	MULTISPECIES: hypothetical protein	*Methylosinus*	82.9	5.1e-11	75
mVC_20187-cat_2_26	P27 family phage terminase small subunit	*Methylocystis*	81.2	7.7e-90	337
mVC_20187-cat_2_27	terminase large subunit	*Rhizobiales*	69.7	3.1e-238	833
mVC_20187-cat_2_28	Phage portal protein	*Methylocella*	55.3	4.3e-133	483
mVC_20187-cat_2_29	Clp protease ClpP	*Methylocystis*	68.7	5.9e-99	368
mVC_20187-cat_2_30	phage major capsid protein	*Methylocystis*	82.8	1.3e-202	714
mVC_20187-cat_2_31	Uncharacterized protein	*Methylobacterium*	37.3	5.5e-5	55
mVC_20187-cat_2_32	MULTISPECIES: phage gp6-like head-tail connector protein	*Methylosinus*	50.2	1.2e-48	200
mVC_20187-cat_2_33	MULTISPECIES: head-tail adaptor protein	*Methylosinus*	77.5	6.7e-43	181
mVC_20187-cat_2_34	Uncharacterized protein	*Methylocella*	46.8	1.1e-35	157

mVC_20187-cat_2_36	DUF3168 domain-containing protein	*Methylocystis*	81.1	8.2e-62	243
mVC_20187-cat_2_37	MULTISPECIES: hypothetical protein	*Methylosinus*	71.2	9.7e-55	220
mVC_20187-cat_2_38	gene transfer agent family protein	*Methylocystis*	72.6	1.8e-40	173
mVC_20187-cat_2_40	MULTISPECIES: hypothetical protein	*Methylosinus*	41.7	1.8e-118	435
mVC_20187-cat_2_41	TIGR02217 family protein	*Blastochloris*	65.9	3.5e-72	278
mVC_20187-cat_2_42	Beta tubulin	*Methylocystis*	78.1	6.3e-126	458
mVC_20187-cat_2_43	Peptidase P60	*Methylocystis*	73.9	5.2e-56	224
mVC_20187-cat_2_44	Uncharacterized protein	*Methylocystis*	70.3	0.0e+0	2028
mVC_20187-cat_2_45	Uncharacterized protein	*Methylocystis*	40.5	9.7e-84	319
mVC_20187-cat_2_46	Uncharacterized protein	*bacterium*	55.4	7.2e-43	181
mVC_20187-cat_2_47	MULTISPECIES: DNA adenine methylase	*Methylosinus*	93.2	2.9e-147	529
mVC_20187-cat_2_48	MULTISPECIES: hypothetical protein	*Methylosinus*	76.3	3.2e-42	178
mVC_20187-cat_2_49	MULTISPECIES: hypothetical protein	*Methylosinus*	59.1	1.9e-13	83
mVC_20187-cat_2_50	MULTISPECIES: hypothetical protein	*Methylosinus*	48.1	2.1e-4	53

Supplementary Table 4.5. Host-virus linkage using the WIsH tool. Gene homology analysis between viral metagenomic viral contigs (mVCs) and host contigs, bins and NCBI-nr database prokaryotes.

VirSorter mVCs	Host contig	Bin	Genus of database prokayotes	Number of homologs with host contig	Number of homologs with bin	Number of homologs with genus	Matrix	*P*-value
mVC_00010-cat_5	c_18004		*Methylomonas*	4		4	-1.37	0
mVC_00059-cat_3	c_00052		*Methylocystis*	15		2	-1.35	1.22e-8
mVC_00100-cat_6	c_07034			5			-1.36	0
mVC_00114-cat_3	c_27207		*Myxococcus*	0		0	-1.38	0.04
mVC_00150-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.02
mVC_00204-cat_6	c_07946		*Methyloceanibacter*	8		3	-1.32	0
mVC_00307-cat_2	c_01373		*Stigmatella*	0		0	-1.38	2.52e-3
mVC_00374-cat_6	c_07155		Candidatus *Puniceispirillum*	1		0	-1.37	0
mVC_00412-cat_3	c_14963			3			-1.35	0
mVC_00453-cat_3	c_30908		*Methylomonas*	0		1	-1.38	0.01
mVC_00633-cat_6	c_13428		*Bosea*	5		0	-1.34	0
mVC_00743-cat_3	c_10505			3			-1.34	0
mVC_00769-cat_2	c_15791		*Methylobacterium*	6		2	-1.33	0
mVC_00884-cat_3	c_14425	bin.17		0	0		-1.38	0.01
mVC_00907-cat_3	c_10684		*Caulobacter*	1		0	-1.38	1.56e-3
mVC_00976-cat_6	c_06108			2			-1.37	0
mVC_01419-cat_3	c_20605		*Sorangium*	0		0	-1.38	4.59e-3
mVC_01490-cat_3	c_25479		*Massilia*	0		0	-1.38	1.23e-3
mVC_01586-cat_2	c_12119	bin.5	*Chelatococcus*	0	13	0	-1.30	0.02
mVC_01595-cat_3	c_10118		*Hyphomicrobium*	11		1	-1.29	0
mVC_01664-cat_3	c_20947		*Methylomicrobium*	0		0	-1.38	1.22e-3
mVC_01801-cat_3	c_03446		*Ochrobactrum*	6		0	-1.36	0

mVC_01839-cat_3	c_16557		*Oleispira*	0		0	-1.38	5.01e-14
mVC_01905-cat_6	c_00778		*Marivivens*	1		0	-1.36	3.06e-10
mVC_02020-cat_3	c_21045		*Methylomicrobium*	0		0	-1.38	2.69e-3
mVC_02077-cat_3	c_13280		*Ralstonia*	10		0	-1.33	0
mVC_02279-cat_3	c_15211			4			-1.27	0
mVC_02315-cat_3	c_01322			1			-1.37	3.31e-7
mVC_02377-cat_3	c_09879		*Methylocella*	0		1	-1.38	4.31e-9
mVC_02484-cat_3	c_12119	bin.5	*Chelatococcus*	1	2	0	-1.28	5.58e-3
mVC_02519-cat_3	c_12332		*Brevundimonas*	6		0	-1.36	0
mVC_02569-cat_3	c_07752		*Methylocystis*	4		2	-1.31	0
mVC_02890-cat_3	c_12869		*Methylobacterium*	7		1	-1.31	0
mVC_02956-cat_3	c_00825		*Devosia*	8		0	-1.32	0
mVC_03011-cat_3	c_13103		*Nitrosomonas*	1		0	-1.37	1.16e-9
mVC_03086-cat_3	c_08685		*Novosphingobium*	8		0	-1.25	0
mVC_03352-cat_3	c_26026	bin.7	*Bdellovibrio*	0	0	0	-1.38	7.33e-7
mVC_03487-cat_2	c_12696		*Methylocystis*	9		0	-1.28	0
mVC_03505-cat_3	c_08766		*Methylocystis*	1		0	-1.37	0
mVC_03533-cat_2	c_09040		*Methylocella*	3		0	-1.30	0
mVC_03614-cat_3	c_12119	bin.5	*Chelatococcus*	0	1	0	-1.28	6.43e-3
mVC_03682-cat_3	c_09201			12			-1.22	0
mVC_03702-cat_2	c_29870		*Methylomonas*	0		0	-1.38	0.02
mVC_04118-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	2	-1.30	0.02
mVC_04134-cat_3	c_02240		*Methanosarcina*	0		0	-1.38	3.80e-9
mVC_04884-cat_3	c_13580		*Methylocella*	0		0	-1.35	0
mVC_05086-cat_3	c_06532		*Rhodoplanes*	2		0	-1.37	0
mVC_05201-cat_3	c_09773	bin.4		7	9		-1.30	0
mVC_05285-cat_3	c_12119	bin.5	*Chelatococcus*	0	2	0	-1.30	0.03
mVC_05859-cat_1	c_09111		*Rhizobium*	3		1	-1.35	0

mVC_06341-cat_2	c_00661	bin.4	*Beijerinckia*	5	5	1	-1.28	0
mVC_06925-cat_3	c_09275		*Sinorhizobium*	4		0	-1.36	0
mVC_06956-cat_2	c_10049		*Campylobacter*	3		0	-1.35	0
mVC_07402-cat_5	c_30619		*Halothiobacillus*	10		10	-1.37	0
mVC_07499-cat_2	c_12119	bin.5	*Chelatococcus*	0	3	0	-1.30	0.03
mVC_07548-cat_2	c_12119	bin.5	*Chelatococcus*	0	6	0	-1.30	0.03
mVC_07580-cat_2	c_12119	bin.5	*Chelatococcus*	0	1	0	-1.30	0.04
mVC_07725-cat_6	c_12257		*Beijerinckia*	2		0	-1.35	0
mVC_07905-cat_3	c_07386		*Methylobacterium*	3		0	-1.36	0
mVC_07925-cat_6	c_24246		*Janthinobacterium*	2		0	-1.37	0
mVC_08004-cat_3	c_08366		*Nitrosomonas*	2		0	-1.38	1.85e-3
mVC_08048-cat_3	c_00202		*Methylocella*	3		1	-1.37	2.38e-8
mVC_08272-cat_3	c_12710		*Methylocella*	1		0	-1.38	6.10e-3
mVC_08336-cat_3	c_07343		*Planktomarina*	0		0	-1.37	5.55e-17
mVC_08415-cat_3	c_03446		*Ochrobactrum*	4		0	-1.36	0
mVC_08608-cat_3	c_14956		*Hyphomicrobium*	11		1	-1.30	0
mVC_08659-cat_5	c_21189	bin.7	*Bdellovibrio*	0	0	0	-1.38	5.47e-6
mVC_08708-cat_3	c_03440	bin.4	*Mycobacterium*	3	3	0	-1.32	0
mVC_08712-cat_3	c_01101		*Blastomonas*	3		0	-1.34	0
mVC_08785-cat_2	c_09673		*Methylobacterium*	5		14	-1.36	0
mVC_08826-cat_3	c_01742		*Yangia*	3		0	-1.35	0
mVC_08851-cat_2	c_07439		Candidatus *Methylopumilus*	0		0	-1.38	1.22e-3
mVC_08939-cat_3	c_08513		*Beijerinckia*	1		0	-1.38	0.02
mVC_08954-cat_3	c_04091			2			-1.37	0
mVC_08991-cat_3	c_13464			10			-1.23	0
mVC_09003-cat_6	c_09201			10			-1.33	1.80e-14
mVC_09060-cat_3	c_12642		*Acidiphilium*	3		0	-1.36	0
mVC_09064-cat_2	c_11052		*Paraburkholderia*	0		0	-1.38	0

mVC_09233-cat_3	c_25501		*Methylomicrobium*	0		0	-1.38	9.56e-3
mVC_09255-cat_2	c_12119	bin.5	*Chelatococcus*	1	2	0	-1.27	4.14e-3
mVC_09256-cat_3	c_01573			6			-1.25	0
mVC_09268-cat_3	c_02286		*Hyphomonas*	4		1	-1.32	0
mVC_09301-cat_3	c_08129		*Methylomonas*	2		2	-1.37	0
mVC_09343-cat_3	c_08623			5			-1.36	0
mVC_09403-cat_3	c_12125		*Collimonas*	0		0	-1.38	0.04
mVC_09424-cat_3	c_02227		*Rhizobium*	4		1	-1.34	0
mVC_09442-cat_3	c_12124		*Janthinobacterium*	0		0	-1.38	0.04
mVC_09460-cat_3	c_12119	bin.5	*Chelatococcus*	0	3	0	-1.30	0.02
mVC_09474-cat_3	c_11709			1			-1.36	0
mVC_09608-cat_3	c_07642	bin.4	*Methylocystis*	5	5	0	-1.35	3.48e-10
mVC_09622-cat_3	c_12264		*Methylocystis*	3		2	-1.35	1.13e-12
mVC_09650-cat_3	c_12529		*Conexibacter*	2		0	-1.32	0
mVC_09678-cat_2	c_09360		*Bradyrhizobium*	1		0	-1.36	0
mVC_09700-cat_3	c_01742		*Yangia*	2		0	-1.37	0
mVC_09760-cat_3	c_12651		*Methylocella*	11		0	-1.27	0
mVC_09941-cat_3	c_10865			1			-1.37	0
mVC_09979-cat_3	c_05572		*Lysobacter*	2		0	-1.36	0
mVC_10028-cat_3	c_08116		*Acidiphilium*	3		0	-1.37	3.15e-11
mVC_10423-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.26	0.00
mVC_10504-cat_3	c_15010		*Bradyrhizobium*	1		0	-1.37	0
mVC_10571-cat_3	c_03446		*Ochrobactrum*	4		0	-1.36	0
mVC_11036-cat_3	c_16224		*Methylocystis*	6		5	-1.30	0
mVC_11479-cat_3	c_14680		*Methylocystis*	1		1	-1.36	0
mVC_11501-cat_2	c_00969		*Cupriavidus*	1		0	-1.38	2.95e-13
mVC_11663-cat_3	c_11486		*Mesorhizobium*	4		2	-1.36	0
mVC_11787-cat_2	c_10077		*Nitrosomonas*	2		0	-1.36	0

mVC_12120-cat_6	c_12125		*Collimonas*	34		0	-1.38	0.04
mVC_12123-cat_5	c_12119	bin.5	*Chelatococcus*	95	254	0	-1.30	0.03
mVC_12131-cat_5	c_12119	bin.5	*Chelatococcus*	27	0	6	-1.28	8.00e-3
mVC_12139-cat_6	c_24976		*Yersinia*	5		0	-1.38	2.51e-3
mVC_12147-cat_5	c_30823		*Methylomicrobium*	3		0	-1.37	0
mVC_12177-cat_6	c_22473		*Methylomonas*	2		1	-1.38	6.09e-3
mVC_12188-cat_5	c_12119	bin.5	*Chelatococcus*	9	0	0	-1.26	1.62e-3
mVC_12236-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_12288-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_12294-cat_6	c_21910	bin.7	*Bdellovibrio*	0	0	1	-1.38	1.82e-3
mVC_12336-cat_3	c_23699		*Methylomonas*	1		0	-1.38	2.43e-8
mVC_12448-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.03
mVC_12456-cat_5	c_09202		*Methylocystis*	10		0	-1.32	0
mVC_12555-cat_3	c_10582			2			-1.36	0
mVC_12577-cat_6	c_00075	bin.4	*Methylocystis*	10	0	0	-1.32	1.68e-13
mVC_12726-cat_6	c_08685		*Novosphingobium*	15		2	-1.26	0
mVC_12855-cat_3	c_19044		*Methylomicrobium*	0		0	-1.38	2.30e-3
mVC_12904-cat_3	c_10240			2			-1.33	0
mVC_12959-cat_3	c_07343		*Planktomarina*	2		0	-1.36	0
mVC_12995-cat_6	c_00654		*Polymorphum*	7		0	-1.32	0
mVC_13098-cat_2	c_08410		*Rhodoplanes*	8		0	-1.28	0
mVC_13121-cat_2	c_07944		*Sphingobium*	9		2	-1.23	0
mVC_13134-cat_3	c_08430		*Clostridium*	10		0	-1.31	0
mVC_13198-cat_3	c_02233		*Methylovorus*	0		0	-1.38	9.71e-3
mVC_13236-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.03
mVC_13278-cat_3	c_00202		*Methylocella*	2		5	-1.36	1.14e-12
mVC_13387-cat_6	c_08410		*Rhodoplanes*	2		0	-1.35	5.55e-17
mVC_13458-cat_3	c_07034			3			-1.36	0

mVC_13486-cat_3	c_07034			3			-1.33	0
mVC_13544-cat_3	c_01768		*Magnetospira*	4		0	-1.33	0
mVC_13586-cat_3	c_11733		*Hyphomicrobium*	1		1	-1.37	0
mVC_13808-cat_3	c_04960			4			-1.29	0
mVC_13938-cat_3	c_01768		*Magnetospira*	2		0	-1.33	0
mVC_14028-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.02
mVC_14108-cat_2	c_04357		*Methylocella*	4		0	-1.28	0
mVC_14130-cat_3	c_07646		*Methylocystis*	10		0	-1.21	0
mVC_14233-cat_3	c_12642		*Acidiphilium*	0		0	-1.36	0
mVC_14718-cat_3	c_07596	bin.22	*Methylocella*	3	0	0	-1.32	0
mVC_15978-cat_1	c_00778		*Marivivens*	2		0	-1.30	0
mVC_16632-cat_3	c_08303		*Methylobacterium*	7		0	-1.30	0
mVC_16880-cat_2	c_22800			3			-1.37	0
mVC_16891-cat_5	c_20753		*Bradyrhizobium*	13		0	-1.31	0
mVC_16898-cat_2	c_00005		*Methylomicrobium*	0		0	-1.38	2.50e-3
mVC_16915-cat_3	c_31667		*Thiolapillus*	5		0	-1.37	0
mVC_16933-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_17002-cat_5	c_20091		*Thermodesulfatator*	0		0	-1.37	1.29e-3
mVC_17043-cat_2	c_18547			6			-1.38	1.12e-3
mVC_17048-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_17106-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.31	0.04
mVC_17199-cat_6	c_23822			1			-1.38	0
mVC_17282-cat_6	c_23330		*Methylomicrobium*	1		0	-1.38	0
mVC_17297-cat_2	c_12119	bin.5	*Chelatococcus*	1	0	0	-1.31	0.04
mVC_17367-cat_2	c_20090		*Trichodesmium*	0		0	-1.37	2.53e-3
mVC_17484-cat_3	c_32891			5			-1.35	0
mVC_17916-cat_2	c_33597			5			-1.34	0
mVC_18575-cat_2	c_32020			7			-1.24	0

mVC_18819-cat_1	c_30137		*Steroidobacter*	0		0	-1.38	8.68e-6
mVC_18915-cat_2	c_20986		*Corynebacterium*	3		0	-1.28	0
mVC_18921-cat_2	c_12119	bin.5	*Chelatococcus*	5	0	0	-1.28	5.36e-3
mVC_19055-cat_1	c_07785		*Methylophilus*	0		0	-1.38	7.88e-6
mVC_19156-cat_3	c_19844		*Herbaspirillum*	7		8	-1.36	0
mVC_19312-cat_3	c_21845		*Methylomonas*	0		3	-1.38	7.45e-9
mVC_19342-cat_3	c_30617		*Hyphomicrobium*	1		0	-1.37	0
mVC_19378-cat_3	c_32520		*Filomicrobium*	2		0	-1.35	0
mVC_19769-cat_3	c_24347		*Desulfovibrio*	3		0	-1.36	0
mVC_20167-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.26	1.74e-3
mVC_20170-cat_2	c_24962			5			-1.38	8.15e-3
mVC_20180-cat_2	c_12128	bin.18	*Methylomonas*	0	0	0	-1.38	6.38e-3
mVC_20187-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	1	-1.28	7.72e-3
mVC_20207-cat_5	c_21149	bin.7	*Bdellovibrio*	0	0	16	-1.37	0
mVC_20210-cat_2	c_21787	bin.7	*Bdellovibrio*	0	0	12	-1.38	1.41e-3
mVC_20234-cat_2	c_18547			6			-1.37	0
mVC_20303-cat_3	c_31667		*Thiolapillus*	6		0	-1.37	0
mVC_20461-cat_2	c_17037		*Yersinia*	18		0	-1.29	0
mVC_20573-cat_5	c_20090		*Trichodesmium*	0		0	-1.38	0.02
mVC_20627-cat_3	c_33128			4			-1.35	0
mVC_20636-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_20710-cat_3	c_21184		*Hyphomonas*	1		0	-1.38	1.10e-7
mVC_20749-cat_2	c_20090		*Trichodesmium*	0		0	-1.37	1.05e-3
mVC_20886-cat_3	c_22688		*Geobacter*	1		0	-1.38	0
mVC_21051-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.31	0.04
mVC_21116-cat_3	c_29019			0			-1.38	1.06e-11
mVC_21153-cat_3	c_18205			3			-1.32	0
mVC_21185-cat_3	c_26429		*Burkholderia*	0		0	-1.38	1.06e-10

mVC_21233-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.03
mVC_21418-cat_2	c_12457	bin.18	*Methylomonas*	0	0	0	-1.38	4.47e-3
mVC_22100-cat_2	c_22588		*Gallionella*	0		0	-1.38	8.49e-3
mVC_22606-cat_3	c_17712			2			-1.35	0
mVC_22720-cat_3	c_18023		*Nitrospira*	1		0	-1.34	0
mVC_23012-cat_2	c_31074		*Desulfobulbus*	11		0	-1.23	0
mVC_23185-cat_3	c_32489		*Comamonas*	0		0	-1.38	5.32e-3
mVC_23226-cat_3	c_12146	bin.18	*Methylomicrobium*	0	0	0	-1.38	7.69e-6
mVC_23241-cat_3	c_29019			0			-1.38	3.52e-10
mVC_23589-cat_2	c_00005		*Methylomicrobium*	0		0	-1.38	2.88e-3
mVC_23849-cat_2	c_30755		*Magnetospirillum*	8		0	-1.25	0
mVC_24032-cat_2	c_30736			3			-1.34	0
mVC_24042-cat_2	c_31333			0			-1.38	0
mVC_24162-cat_3	c_33537			3			-1.34	0
mVC_24164-cat_3	c_00003	bin.1	*Sideroxydans*	0	0	0	-1.38	2.10e-3
mVC_24622-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.27	2.00e-3
mVC_24690-cat_3	c_21670		*Methylomonas*	2		3	-1.37	0
mVC_24780-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.03
mVC_25059-cat_3	c_33537			6			-1.33	0
mVC_25087-cat_2	c_07420	bin.19	*Methylomonas*	4	0	9	-1.38	5.20e-5
mVC_25108-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.28	6.65e-3
mVC_25342-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_25450-cat_3	c_20299		*Methylomonas*	1		1	-1.38	1.69e-6
mVC_25723-cat_3	c_00005		*Methylomicrobium*	0		0	-1.38	2.48e-3
mVC_26213-cat_2	c_31966		*Methylomonas*	2		0	-1.31	0
mVC_26577-cat_3	c_07420	bin.19	*Methylomonas*	5	0	4	-1.37	5.13e-7
mVC_26784-cat_3	c_20406		*Methylomicrobium*	0		0	-1.38	1.63e-3
mVC_27735-cat_1	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.02

mVC_27875-cat_2	c_12160	bin.18	*Methylomicrobium*	0	0	1	-1.38	2.17e-5
mVC_28150-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.28	8.97e-3
mVC_28568-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_28899-cat_2	c_20560		*Methylomicrobium*	0		5	-1.38	1.33e-7
mVC_29706-cat_2	c_12119	bin.5	*Chelatococcus*	1	0	0	-1.30	0.02
mVC_29722-cat_5	c_12119	bin.5	*Chelatococcus*	1	0	0	-1.27	3.11e-3
mVC_29731-cat_5	c_12119	bin.5	*Chelatococcus*	3	0	0	-1.26	9.07e-3
mVC_29901-cat_3	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	4.36e-3
mVC_29943-cat_3	c_31667		*Thiolapillus*	5		0	-1.36	0
mVC_30023-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.31	0.04
mVC_30098-cat_2	c_07420	bin.19	*Methylomonas*	6	0	1	-1.37	6.40e-7
mVC_30189-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.28	5.86e-3
mVC_30392-cat_6	c_25516	bin.23		4	0		-1.33	0
mVC_30410-cat_3	c_18379		*Methylomonas*	8		0	-1.30	0
mVC_30689-cat_3	c_12119	bin.5	*Chelatococcus*	1	0	0	-1.27	2.67e-3
mVC_30698-cat_6	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_30773-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.30	0.03
mVC_30899-cat_3	c_00001	bin.1	*Gallionella*	0	0	0	-1.38	2.63e-3
mVC_30959-cat_3	c_17286		*Hyphomicrobium*	3		0	-1.33	0
mVC_31068-cat_2	c_12119	bin.5	*Chelatococcus*	0	0	0	-1.29	0.01
mVC_31183-cat_3	c_20429		*Bradyrhizobium*	3		5	-1.36	2.61e-5
mVC_31281-cat_2	c_21557		*Pseudomonas*	4		0	-1.28	0
mVC_31458-cat_2	c_25375			4			-1.27	0
mVC_31794-cat_2	c_31327		*Methylomicrobium*	0		0	-1.38	4.99e-16
mVC_31857-cat_3	c_21232			4			-1.31	0
mVC_32350-cat_3	c_17836		*Polaromonas*	6		0	-1.31	0
mVC_32993-cat_2	c_18958			7			-1.24	0

Supplementary Table 5.1. Summary information for the metagenomic viral contigs (mVCs) and their predicted host contigs using the WIsH tool

Virus ID	Host ID	Phylum	Genus	Specie	Matrix	*p*-value
mVC_44311-cat_3	c_000249	*Actinobacteria*	*Streptomyces*	*Streptomyces griseochromogenes*	-1.266	0
mVC_35019-cat_2	c_000258	*Actinobacteria*	*Conexibacter*	*Conexibacter woesei*	-1.245	0
mVC_10150-cat_2	c_000849	*Proteobacteria*	*Methylobacterium*	*Methylobacterium aquaticum*	-1.313	0
mVC_45267-cat_6	c_001222	*Acidobacteria*	*Granulicella*	*Granulicella tundricola*	-1.272	0
mVC_35071-cat_3	c_001340	*Proteobacteria*	*Methylomicrobium*	*Methylomicrobium alcaliphilum*	-1.268	0
mVC_15429-cat_3	c_002245	*Proteobacteria*	*Myxococcus*	*Myxococcus fulvus*	-1.276	0
mVC_15043-cat_3	c_003864	*Actinobacteria*	*Nakamurella*	*Nakamurella multipartita*	-1.331	0
mVC_55446-cat_2	c_004273	*Firmicutes*	*Desulfotomaculum*	*Desulfotomaculum reducens*	-1.361	0
mVC_14206-cat_3	c_004522	*Actinobacteria*	*Nakamurella*	*Nakamurella multipartita*	-1.319	0
mVC_19191-cat_3	c_004928				-1.302	0
mVC_38254-cat_3	c_007426	*Proteobacteria*	*Steroidobacter*	*Steroidobacter denitrificans*	-1.316	0
mVC_17060-cat_3	c_007426	*Proteobacteria*	*Steroidobacter*	*Steroidobacter denitrificans*	-1.327	0
mVC_19307-cat_3	c_008003				-1.307	0
mVC_00634-cat_3	c_009897	*Proteobacteria*	*Bradyrhizobium*	*Bradyrhizobium japonicum*	-1.309	0
mVC_42471-cat_2	c_010602	*Actinobacteria*	*Nakamurella*	*Nakamurella multipartita*	-1.234	0
mVC_48952-cat_2	c_011761				-1.297	0
mVC_36328-cat_2	c_014497	*Spirochaetes*	*Turneriella*	*Turneriella parva*	-1.257	0
mVC_35869-cat_3	c_015570	*Actinobacteria*	*Kitasatospora*	*Kitasatospora setae*	-1.376	2.391e-11
mVC_53588-cat_3	c_015570	*Actinobacteria*	*Kitasatospora*	*Kitasatospora setae*	-1.375	8.93e-15
mVC_40940-cat_2	c_016065	*Actinobacteria*	*Acidimicrobium*	*Acidimicrobium ferrooxidans*	-1.279	0
mVC_45031-cat_3	c_016791				-1.318	0
mVC_34267-cat_3	c_016791				-1.315	0
mVC_48727-cat_3	c_016883				-1.334	0
mVC_60020-cat_3	c_017720	*Proteobacteria*	*Methylobacterium*	*Methylobacterium* sp. 4-46	-1.255	0
mVC_35372-cat_2	c_017744	*Actinobacteria*	*Actinoalloteichus*		-1.282	0

mVC_49362-cat_2	c_018694	*Actinobacteria*	*Ilumatobacter*	*Ilumatobacter coccineus*	-1.257	0
mVC_10443-cat_3	c_019426	*Proteobacteria*	*Bradyrhizobium*	*Bradyrhizobium icense*	-1.307	0
mVC_35074-cat_3	c_021587				-1.359	0
mVC_03365-cat_3	c_022238				-1.277	0
mVC_12050-cat_3	c_022397	*Actinobacteria*	*Micromonospora*		-1.257	0
mVC_61586-cat_3	c_023169	*Proteobacteria*	*Rhodoplanes*	*Rhodoplanes* sp. Z2-YC6860	-1.284	0
mVC_67341-cat_3	c_024170				-1.264	0
mVC_76192-cat_2	c_025473	Actinobacteria	*Frankia*	*Frankia* symbiont of *Datisca glomerata*	-1.267	0
mVC_24309-cat_3	c_026067	Proteobacteria	*Mesorhizobium*	*Mesorhizobium loti*	-1.235	0
mVC_67345-cat_3	c_026195	Proteobacteria	*Grimontia*	*Grimontia hollisae*	-1.249	0
mVC_73506-cat_2	c_030526	Actinobacteria	*Streptomyces*	*Streptomyces leeuwenhoekii*	-1.256	0
mVC_73822-cat_3	c_030917				-1.258	0
mVC_52152-cat_2	c_034615	Actinobacteria	*Conexibacter*	*Conexibacter woesei*	-1.238	0
mVC_14236-cat_3	c_035124	Actinobacteria	*Frankia*	*Frankia* sp. EAN1pec	-1.242	0
mVC_59030-cat_3	c_035981				-1.327	0
mVC_17466-cat_3	c_038997				-1.244	0
mVC_34234-cat_6	c_042120	*Firmicutes*	*Desulfosporosinus*	*Desulfosporosinus acidiphilus*	-1.381	1.015e-12
mVC_17296-cat_2	c_042479	*Proteobacteria*	*Nitrobacter*	*Nitrobacter hamburgensis*	-1.292	0
mVC_02817-cat_3	c_042848	*Proteobacteria*	*Bradyrhizobium*	*Bradyrhizobium diazoefficiens*	-1.298	0
mVC_11165-cat_2	c_043209	*Actinobacteria*	*Actinoplanes*	*Actinoplanes friuliensis*	-1.281	0
mVC_33890-cat_5	c_044276	*Firmicutes*	*Syntrophobotulus*	*Syntrophobotulus glycolicus*	-1.346	1.163e-12
mVC_36338-cat_3	c_046536				-1.299	0
mVC_41852-cat_3	c_046676	*Actinobacteria*	*Nonomuraea*	*Nonomuraea* sp. ATCC 55076	-1.350	0
mVC_03841-cat_3	c_046898				-1.307	0
mVC_39726-cat_3	c_047056	*Proteobacteria*	*Bradyrhizobium*	*Bradyrhizobium* sp.	-1.276	0
mVC_02745-cat_2	c_048003	*Actinobacteria*	*Conexibacter*	*Conexibacter woesei*	-1.289	0
mVC_55871-cat_3	c_049823				-1.320	0
mVC_58079-cat_2	c_050213	*Actinobacteria*	*Conexibacter*	*Conexibacter woesei*	-1.306	0.035

mVC_13892-cat_3	c_050571	*Actinobacteria*	*Conexibacter*	*Conexibacter woesei*	-1.333	0
mVC_34364-cat_3	c_050571	*Actinobacteria*	*Conexibacter*	*Conexibacter woesei*	-1.362	1.260e-07
mVC_01602-cat_3	c_050604	*Deinococcus-Thermus*	*Deinococcus*	*Deinococcus swuensis*	-1.261	0
mVC_14986-cat_3	c_050663	*Actinobacteria*	*Intrasporangium*	*Intrasporangium calvum*	-1.294	0
mVC_35969-cat_3	c_050733	*Proteobacteria*	*Blastochloris*	*Blastochloris viridis*	-1.307	0
mVC_01872-cat_3	c_051227	*Firmicutes*	*Mahella*	*Mahella australiensis*	-1.325	6.724e-13
mVC_45715-cat_6	c_051227	*Firmicutes*	*Mahella*	*Mahella australiensis*	-1.235	0
mVC_38396-cat_3	c_051836	*Actinobacteria*	*Rubrobacter*	*Rubrobacter xylanophilus*	-1.238	0
mVC_03109-cat_3	c_052032	*Actinobacteria*	*Arsenicicoccus*	*Arsenicicoccus* sp. oral taxon 190	-1.236	0
mVC_46078-cat_2	c_052032	*Actinobacteria*	*Arsenicicoccus*	*Arsenicicoccus* sp. oral taxon 190	-1.227	0
mVC_47357-cat_3	c_052289	*Actinobacteria*	*Streptomyces*	*Streptomyces gilvosporeus*	-1.364	0
mVC_36522-cat_3	c_052420	*Proteobacteria*	*Burkholderia*	*Burkholderia* sp. PAMC 28687	-1.335	0
mVC_22913-cat_3	c_052568	*Actinobacteria*	*Acidimicrobium*	*Acidimicrobium ferrooxidans*	-1.328	0
mVC_02500-cat_3	c_052830				-1.260	0
mVC_46006-cat_3	c_053343				-1.262	0
mVC_20140-cat_3	c_054037	*Proteobacteria*	*Anaeromyxobacter*	*Anaeromyxobacter* sp. Fw109-5	-1.258	0
mVC_47791-cat_2	c_054643	*Actinobacteria*	*Pimelobacter*	*Pimelobacter simplex*	-1.245	0
mVC_47657-cat_1	c_055705	*Chloroflexi*	*Sphaerobacter*	*Sphaerobacter thermophilus*	-1.292	0
mVC_38346-cat_3	c_056690				-1.333	0
mVC_12581-cat_3	c_057817	*Actinobacteria*	*Mycobacterium*	*Mycobacterium avium*	-1.27Z	0
mVC_15853-cat_3	c_058899	*Synergistetes*	*Thermovirga*	*Thermovirga lienii*	-1.293	0
mVC_10028-cat_2	c_059111	*Proteobacteria*	*Oligotropha*	*Oligotropha carboxidovorans*	-1.299	0
mVC_46520-cat_3	c_060282	*Actinobacteria*	*Rhodococcus*		-1.332	0
mVC_66365-cat_3	c_060671	*Acidobacteria*	*Granulicella*	*Granulicella tundricola*	-1.270	0.004
mVC_06649-cat_3	c_060671	*Acidobacteria*	*Granulicella*	*Granulicella tundricola*	-1.273	0.006
mVC_21215-cat_2	c_060671	*Acidobacteria*	*Granulicella*	*Granulicella tundricola*	-1.277	0.010
mVC_73334-cat_6	c_060671	*Acidobacteria*	*Granulicella*	*Granulicella tundricola*	-1.284	0.019

mVC_32298-cat_2	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.289	0.029
mVC_28653-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.280	0.012
mVC_66531-cat_2	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.278	0.011
mVC_24086-cat_2	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.294	0.046
mVC_08590-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.256	8.457 e-04
mVC_08495-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.275	0.007
mVC_57171-cat_2	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.286	0.022
mVC_26841-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.294	0.045
mVC_11398-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.293	0.040
mVC_51863-cat_3	c_060671	Acidobacteria	*Granulicella*	*Granulicella tundricola*	-1.290	0.032
mVC_31723-cat_3	c_062598	Proteobacteria	*Rhodoplanes*	*Rhodoplanes* sp. Z2-YC6860	-1.325	0
mVC_24232-cat_3	c_062598	Proteobacteria	*Rhodoplanes*	*Rhodoplanes* sp. Z2-YC6860	-1.318	0
mVC_28437-cat_3	c_064088				-1.262	0
mVC_31725-cat_3	c_064771	Proteobacteria	*Rhodoplanes*	*Rhodoplanes* sp. Z2-YC6860	-1.274	0
mVC_71803-cat_3	c_065314	Proteobacteria	*Nitrobacter*	*Nitrobacter winogradskyi*	-1.380	4.264e-09
mVC_67361-cat_3	c_065610	Thaumarchaeota	*Nitrososphaera*	*Nitrososphaera viennensis*	-1.383	1.365e-14
mVC_73139-cat_3	c_065704				-1.346	0
mVC_28057-cat_1	c_070900	Proteobacteria	*Rhodopseudomonas*	*Rhodopseudomonas palustris*	-1.289	0
mVC_23811-cat_2	c_071629	Actinobacteria	*Cryobacterium*	*Cryobacterium arcticum*	-1.293	0
mVC_58257-cat_2	c_073106	Thaumarchaeota	Candidatus *Nitrosotenuis*	*Candidatus Nitrosotenuis cloacae*	-1.384	8.186e-13
mVC_62847-circular-cat_3	c_073203	Thaumarchaeota	Candidatus *Nitrosotenuis*	*Candidatus Nitrosotenuis cloacae*	-1.384	1.300e-09
mVC_26441-cat_2	c_073589				-1.240	0
mVC_64625-cat_5	c_074223	Proteobacteria	*Methylobacterium*	*Methylobacterium extorquens*	-1.254	0
mVC_71296-cat_3	c_074499	Proteobacteria	*Nitrobacter*	*Nitrobacter winogradskyi*	-1.279	0
mVC_61037-cat_3	c_074537	Thaumarchaeota	Candidatus *Nitrosopelagicus*	*Candidatus Nitrosopelagicus brevis*	-1.285	0

Supplementary Table 5.2. Gene homology of nitrifier-associated metagenomic viral contigs (mVCs) against the NCBI nr and Interproscan 5 database.

Gene ID	Identity (%)	E value	bit score	Host	Function (nr)	Function (Interproscan 5)
mVC_58257-cat_2-AOA_1	88.8	2.0e-52	212	Candidatus *Nitrosotalea devanaterra*	Uncharacterized protein	
mVC_58257-cat_2-AOA_2	96.2	5.1e-35	154	Candidatus *Nitrosotalea devanaterra*	Uncharacterized protein	
mVC_58257-cat_2-AOA_3	42	3.3e-18	99	*Nitrosopumilales* archaeon	Uncharacterized protein	
mVC_58257-cat_2-AOA_4	65.5	1.8e-238	834	*Thaumarchaeota* archaeon	Uncharacterized protein	Terminase large subunit, T4 like virus-type
mVC_58257-cat_2-AOA_5	96.7	4.3e-112	411	Candidatus *Nitrosotalea devanaterra*	Metal-sulfur cluster biosynthetic enzyme	Coenzyme PQQ synthesis protein D (PqqD)
mVC_58257-cat_2-AOA_6	93.9	1.0e-171	611	Candidatus *Nitrosotalea okcheonensis*	Putative methylthioribose-1-phosphate isomerase	
mVC_61037-cat_3-AOA_2	74.5	7.1e-37	161	*Thaumarchaeota* archaeon N4	hypothetical protein	
mVC_61037-cat_3-AOA_3	80.9	1.2e-39	170	Candidatus *Nitrosotalea bavarica*	hypothetical protein	SNase-like, OB-fold superfamily
mVC_61037-cat_3-AOA_4	59.5	1.2e-20	107	Marine Group I *thaumarchaeote* SCGC AAA799-N04	hypothetical protein	
mVC_61037-cat_3-AOA_6	57.6	2.1e-4	53	*Nitrosopumilus*	MULTISPECIES: hypothetical protein	
mVC_61037-cat_3-AOA_8	79.5	2.1e-25	122	*Thaumarchaeota* archaeon N4	Lrp/AsnC family transcriptional regulator	
mVC_61037-cat_3-AOA_9	71.7	1.1e-113	418	*Candidatus Nitrosotenuis uzonensis*	Transcription initiation factor IIB 2	Transcription factor TFIIB, conserved site
mVC_61037-cat_3-AOA_10	45.7	2.2e-250	874	Marine Group I *thaumarchaeote* SCGC RSA3	hypothetical protein	
mVC_61037-cat_3-AOA_11	55.6	9.1e-49	202	*Nitrosopumilus* sp. CG10	Uncharacterized protein	
mVC_61037-cat_3-AOA_12	28.3	2.4e-36	161	*Nitrosopumilales* archaeon CG	Uncharacterized protein	

mVC_61037-cat_3-AOA_14	81	7.7e-103	381	Candidatus *Nitrosotenuis chungbukensis*	prepilin peptidase	Prepilin type IV endopeptidase, peptidase domain
mVC_62847-circular-cat_3-AOA_2	48.6	9.4e-13	80	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_3	47.3	7.1e-13	81	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_4	26.1	5.7e-9	72	Candidatus *Wolfebacteria bacterium* RIFCSPLOWO2	Uncharacterized protein	Immunoglobulin-like fold
mVC_62847-circular-cat_3-AOA_10	29.5	1.1e-2	47	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_11	28.2	4.6e-6	63	*Phialocephala subalpina*	Structural maintenance of chromosomes protein	
mVC_62847-circular-cat_3-AOA_13	31.6	1.5e-28	136	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_18	25	2.6e-2	46	*Bifidobacterium breve*	phage N-6-adenine-methyltransferase	DNA N-6-adenine-methyltransferase
mVC_62847-circular-cat_3-AOA_20	28.2	3.0e-3	50	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_21	27.1	4.8e-16	94	*Thaumarchaeota archaeon*	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_27	48.8	4.4e-2	45	*Uncultured marine crenarchaeote* SAT1000-21-C11	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_29	28.8	1.2e-28	136	marine metagenome	Uncharacterized protein	
mVC_62847-circular-cat_3-AOA_32	36.9	8.7e-3	47	archaeon	Uncharacterized protein	
mVC_67361-cat_3-AOA_1	62	1.1e-8	67	*Nitrososphaera gargensis* (strain Ga9.2)	Putative UbiA prenyltransferase	
mVC_67361-cat_3-AOA_2	35.8	3.7e-17	95	*Nitrosopumilales archaeon* CG11	Uncharacterized protein	
mVC_67361-cat_3-AOA_4	56	1.7e-6	60	Candidatus *Nitrosotenuis aquarius*	AAA family ATPase	Winged helix DNA-binding domain superfamily

mVC_67361-cat_3-AOA_10	72.6	7.9e-20	104	*Nitrososphaera gargensis* (strain Ga9.2)	Uncharacterized protein	
mVC_67361-cat_3-AOA_11	81.5	1.3e-27	130	*Thaumarchaeota archaeon*	Deoxyribonuclease	
mVC_67361-cat_3-AOA_12	70.8	3.5e-20	105	*Thaumarchaeota archaeon*	Uncharacterized protein	
mVC_67361-cat_3-AOA_13	57	7.6e-23	114	*Nitrososphaera viennensis*	hypothetical protein	Winged helix DNA-binding domain superfamily
mVC_473879_LOW_GC_2	97.3	1.2e-12	80	Candidatus *Nitrosotalea devanaterra*	Uncharacterized protein	Serralysin-like metalloprotease, C-terminal
mVC_1363578_LOW_GC_1	93.1	7.5e-63	247	Candidatus *Nitrosotalea okcheonensis*	Uncharacterized protein	
mVC_51245_LOW_GC_2	41.3	1.3e-59	240	*Nitrosopumilales archaeon*	6-bladed beta-propeller	NHL repeat, subgroup
mVC_92579_LOW_GC_1	27.7	6.6e-3	48	*Nitrosopumilales archaeon*	Uncharacterized protein	
mVC_92579_LOW_GC_2	33.1	9.2e-14	87	*Clostridium* sp.	Uncharacterized protein	Purple acid phosphatase-like
mVC_179812_LOW_GC_1	46.7	7.4e-23	116	*Planctomycetia bacterium*	Uncharacterized protein	
mVC_186378_LOW_GC_1	28.1	2.5e-16	95	*Spirosoma radiotolerans*	hypothetical protein	Pectin lyase fold/virulence factor
mVC_17296-cat_2-NOB_1	42.9	4.5e-15	88	*Bradyrhizobium* sp. OK095	Outer membrane immunogenic protein	
mVC_17296-cat_2-NOB_2	39.1	1.4e-33	151	*Pseudorhodoplanes sinuspersici*	Uncharacterized protein	Outer membrane protein beta-barrel domain
mVC_17296-cat_2-NOB_3	76.2	1.3e-59	236	*Pseudolabrys* sp. Root1462	Peptidase P60	Endopeptidase, NLPC/P60 domain
mVC_17296-cat_2-NOB_4	70.8	5.9e-120	438	*Pseudolabrys* sp. Root1462	Beta tubulin	Bacteriophage phiJL001, Gp84, N-terminal
mVC_17296-cat_2-NOB_5	62.5	1.2e-20	107	*Bradyrhizobium icense*	type II toxin-antitoxin system RelE/ParE family toxin	Toxin-antitoxin system, RelE/ParE toxin domain superfamily
mVC_17296-cat_2-NOB_6	71.2	1.9e-42	179	*Bradyrhizobium icense*	XRE family transcriptional regulator	Cro/C1-type helix-turn-helix domain

mVC_17296-cat_2-NOB_7	81.2	6.5e-95	354	*Pseudolabrys* sp. Root1462	TIGR02217 family protein	Protein of unknown function DUF2460
mVC_17296-cat_2-NOB_8	64.2	8.1e-49	200	*Nitrobacter* sp. 62-13	Phage tail protein	
mVC_17296-cat_2-NOB_9	74.2	1.2e-18	100	*Pseudolabrys* sp. GY_H	Phage tail assembly chaperone	
mVC_17296-cat_2-NOB_10	65.7	6.7e-27	127	*Pseudolabrys* sp. GY_H	Gene transfer agent family protein	Phage tail tube protein, GTA-gp10
mVC_17296-cat_2-NOB_11	87.9	2.6e-60	238	*Pseudolabrys* sp. GY_H	Phage major tail protein, TP901-1 family	Gene transfer agent, major tail protein
mVC_17296-cat_2-NOB_12	74.8	2.6e-52	212	*Pseudolabrys* sp. Root1462	Uncharacterized protein	Tail completion protein
mVC_17296-cat_2-NOB_13	60.8	1.5e-26	126	*Pseudolabrys* sp. Root1462	Uncharacterized protein	head-tail adaptor superfamily
mVC_17296-cat_2-NOB_14	69.4	9.6e-50	203	*Pseudolabrys* sp. Root1462	Uncharacterized protein	
mVC_71296-cat_3-NOB_1	66.4	2.9e-82	312	*Bradyrhizobium* sp. LVM 105	hypothetical protein	
mVC_71296-cat_3-NOB_4	68.4	1.4e-20	108	*Nitrobacter hamburgensis*	Uncharacterized protein	
mVC_71296-cat_3-NOB_5	93 ;8	7.2e-8	64	*Nitrobacter winogradskyi*	Uncharacterized protein	
mVC_71296-cat_3-NOB_6	49	3.4e-15	89	*Bradyrhizobium* sp. CCBAU 51781	hypothetical protein	
mVC_71296-cat_3-NOB_7	91.1	1.2e-41	176	*Nitrobacter winogradskyi*	Phage integrase	
mVC_71296-cat_3-NOB_8	87.4	1.0e-120	440	*Nitrobacter* sp. Nb-311A	Phage integrase	DNA breaking-rejoining enzyme, catalytic core
mVC_71803-cat_3-NOB_1	66.1	8.4e-14	84	*Bradyrhizobium* sp. AC87j1	Uncharacterized protein	
mVC_71803-cat_3-NOB_2	87	5.0e-27	128	*Nitrobacter vulgaris*	Cold-shock protein	
mVC_71803-cat_3-NOB_3	87.7	3.7e-25	122	*Nitrobacter vulgaris*	Uncharacterized protein	
mVC_71803-cat_3-NOB_4	65.6	1.8e-27	129	*Nitrobacter* sp. Nb-311A	Uncharacterized protein	
mVC_71803-cat_3-NOB_5	63.9	1.2e-10	74	*Nitrobacter vulgaris*	hypothetical protein	
mVC_71803-cat_3-NOB_6	83.9	2.2e-59	235	*Bradyrhizobium lablabi*	Molecular chaperone IbpA	
mVC_71803-cat_3-NOB_7	66.3	3.6e-20	105	*Nitrobacter vulgaris*	Uncharacterized protein	
mVC_71803-cat_3-NOB_8	78.5	9.6e-26	124	*Nitrobacter vulgaris*	Uncharacterized protein	

mVC_71803-cat_3-NOB_9	71.6	9.2e-21	107	*Nitrobacter vulgaris*	Uncharacterized protein
mVC_71803-cat_3-NOB_10	56.4	1.2e-4	54	*Bradyrhizobium erythrophlei*	Uncharacterized protein