Diversity and function of virus communities in hyperarid desert soils

By

Olivier D.J Zablocki

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor (Ph.D.) Microbiology

in the Faculty of Natural & Agricultural Sciences

University of Pretoria

November 2015

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Supervisor: Prof. Don Cowan

Co-supervisor: Dr. Evelien Adriaenssens

DECLARATION

I, Olivier Zablocki declare that this thesis, which I hereby submit for the degree *Philosophiae Doctor* Microbiology (Ph.D.) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: _____

Date: _____

ACKNOWLEDGMENTS

This research project would not have been possible without the support of many individuals and entities. I wish to express my sincere gratitudes to the following:

My family, for their encouragements, love and support,

The University of Pretoria, the Department of Microbiology, the Genomics Research Institute and the Centre for Microbial Ecology and Genomics for providing the necessary infrastructures and financial support,

The National Research Foundation, for their financial support throughout the duration of this project,

Prof. Don Cowan and Dr. Evelien Adriaenssens, for their invaluable mentorship, continuous support, patience, advices and opportunities,

Dr. Rolph Kramer, for testing the lysis assay,

Laboratory colleagues at the Centre for Microbial Ecology and Genomics, for their friendship and valuable technical advices,

Additional co-authors of publications, Prof. Ed Rybicki, Prof. Marla Tuffin, Prof. Craig Carry, Lonnie van Zyl, Dr. Enrico Rubagotti, Dr. Surendra Vikram, Dr. Mary Seely, Dr. Jean-Baptiste Ramond, Dr. Aline Frossard and Vincent Scola,

Prof. Eric Wommack, for his valuable recommendations on virus extraction optimization,

Prof. Gerhard Pietersen and Ronel Roberts, for producing transmission electron micrographs of the potyvirus isolate

THESIS SUMMARY

Viruses are the most numerous and genetically diverse biological entities in the biosphere. Yet, despite the demonstrated ubiquity of viruses, there remain substantial gaps in our understanding of their ecological roles, taxonomic diversity and distribution. In the oceans, where most research has been conducted on the subject, viruses have been demonstrated to perform key functions in global nutrient cycling and microbial diversification. However, it is still a matter of debate whether the same ecological importance holds true for viruses within terrestrial ecosystems. Thus, this study aimed to explore the taxonomic diversity and function of viruses in a major terrestrial biome: the desert ecosystem. Two hyperarid, thermally-contrasted desert biomes were studied: the Dry Valleys of Antarctica and the Namib Desert in Namibia. Prior to this research project, only very limited data were available on the presence and the genomes of viruses in these ecosystems.

In the Antarctic desert, virus diversity was compared between surface soil and rock-associated microbial communities (i.e., hypoliths). In the Namib Desert, the effects of a water availability gradient on virus diversity and distribution were assessed. Viruses were extracted from surface soils (Antarctic and Namib Desert samples) and rock-associated microbial communities (Antarctica only). Analysis of virus communities was conducted by the purification of viral metagenomic DNA, amplification by PCR and sequenced through NGS. Sequence data were used to determine the taxonomic composition of virus communities using homology-based searches against reference virus genomes in public sequence databases. For the Namib Desert study, the effects of soil microbial activity and physicochemical properties on the biogeography of soil viruses were determined. Bivariate correlation analyses were conducted to establish any putative links between virus distribution and edaphic factors.

Metagenomic analysis of soil virus communities in both desert biomes have shown that the virus composition is largely unknown (60-80%). Of the identified viruses, the dominant (90%) group were tailed, bacterial viruses (Order: *Caudovirales*). The remaining fraction was composed of eukaryotic viruses of algae, amoeba and insects. No archaeal viruses were found in this study. Furthermore, the taxonomic composition of viruses in all desert samples was strongly habitat- and sample- specific, with minimal overlap of virus species distribution. In the Namib Desert, significant correlations between virus abundance and microbial activity and several soil chemical constituents were found. It is suggested that virus communities in this hot hyperarid desert are not directly influenced by environmental conditions. Rather, the biogeographical patterns of the extracellular virus fraction are primarily determined by host distribution. In addition, a positive correlation between viral abundance and microbial activity is the first indirect indication of active viral propagation, and the possibility that soil viruses may contribute to nutrient cycling in a hot hyperarid desert soil environment. Collectively, the results in this thesis provide the first in-depth characterisation of a novel pool of viruses in hyperarid soils.

LIST OF ABBREVIATIONS

ALM	Ace Lake Mavirus
BLAST	Basic local alignment search tool
BSA	Bovine serum albumin
COG	Clusters of orthologous groups
CS	Coarse sand
CSV	Cocksfoot streak virus
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DS	Double-stranded
DUF	Domain of unknown function
EBI	European Bioinformatics institute
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EPS	Exopolysaccharide
FDA	Fluorescin diacetate assay
FS	Fine sand
HIV	Human immunodeficiency virus
HP	Hypothetical protein
HY	Hypolith
ICTV	International Committee on the Taxonomy of Viruses
IMG	Integrated Microbial Genomes
KB	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASL	Linker amplified shotgun library
MCP	Major coat protein
MDA	Multiple displacement amplification

MEGA	Molecular evolutionary genetics analysis
MS	Medium sand
MVSV	Miers Valley soil virophage
NCBI	National Centre for Biotechnology Information
NCLDV	Nucleocytoplasmic large DNA viruses
NGS	Next-generation sequencing
NIB	Nuclear inclusion gene B
NR	Non-redundant
OD	Optical density
OLV	Organic Lake virophage
OM	Organic matter
ORF	Open reading frame
OS	Open soil
OTU	Operational taxonomic unit
P	Precipitation
PATRIC	Pathosystems Resource Integration Center
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PET	Potential of evapotranspiration
PFGE	Pulse-field electrophoresis
pH	Potential of hydrogen
PHACTS	Phage classification tool set
PPCO	Polypropylene copolymer
RAPD	Randomly amplified polymorphic DNA
RH	Relative humidity
RNA	Ribonucleic acid
RNase	Ribonuclease
RP-SISPA	Random-priming, sequence independent, single primer amplification

RT-PCR	Reverse-transcription polymerase chain reaction
SDS	Sodium dodecyl sulphate
SPSS	Statistical Package for the Social Sciences
SS	Single stranded
TRFLP	Terminal restriction fragment length polymorphism
tRNA	Transfer RNA
UV	Ultra-violet
VBR	Virus-to-bacteria ratio
VFS	Very fine sand
VLP	Virus-like particle
WGA	Whole genome amplification
YSLV	Yellowstone Lake Virophage

LIST OF FIGURES

LIST OF TABLES

Contents

SECTION II: VIRUSES IN ANTARCTIC SOILS

SECTION III: VIRUSES IN NAMIB DESERT SOILS

SECTION I: THESIS BACKGROUND

Chapter 1:

An introduction to prokaryotic viruses and metaviromics

1.1 Chapter summary

This chapter consists of two main sections. The first section provides a general background on the biology, diversity and genomics of prokaryotic viruses. As this thesis deals with environmental viruses, the emphasis is on prokaryotic viruses, the dominant virus fraction in most natural ecosystems. Ecological aspects of viruses are not discussed in this chapter, and will be the focus of Chapter 2. In the second section, the basics of viral metagenomics are presented, as well as some of the most commonly used tools to analyse metavirome sequence data.

1.2 Prokaryotic viruses

1.2.1 Introduction

Viruses which infect cellular organisms belonging to the domain Bacteria and Archaea are the most abundant members of the virosphere (Comeau *et al*, 2008). Bacterial viruses were discovered twice, independently, at the beginning of the 20th century by Frederick Twort and Felix d'Herelle (Twort, 1915; d'Herelle, 1917). d'Herelle coined the term 'bacteriophage' or 'phage' to describe these "bacteria eaters". To this day, these terms are still widely used. The first report of a virus infecting Archaea came much later (Torsvik & Dundas, 1974), and was initially described as a bacteriophage of the hyperhalophilic *Halobacterium salinarium*. Recently, a naming convention for viruses of Archaea was proposed (Abedon & Murray, 2013), which encouraged the use of "archaeal virus" to avoid confusion when terms such as bacteriophages, archaeal phages and archeophage were used interchangeably. The name 'prokaryotic viruses' will be used here to refer to viruses which infect members of either of the prokaryotic domains.

1.2.2 Taxonomic diversity and genomics

The taxonomic assignment of prokaryotic viruses is governed by the International Committee on the Taxonomy of Viruses (ICTV) Bacterial and Archaeal viruses Subcommittee, and uses as many criteria as possible for the classification of viruses (Ackermann, 2011). The most important criteria include the type of the nucleic acid, virion morphology and physicochemical properties of the virion. As of 2015, the ICTV recognises 7 orders sub-divided in 26 families (Figure 1), in addition to 78 unassigned families (http://www.ictvonline.org/virustaxonomy). In the ICTV classification scheme, prokaryotic viruses are represented by two orders: the *Caudovirales* (tailed phages) and *Ligamenvirales*

(archaeal viruses). The order *Caudovirales* is sub-divided into three major tailed-virus families: *Siphoviridae, Myoviridae and Podoviridae*, and the *Ligamenvirales* divided into the *Lipothrixviridae* and *Rudiviridae* families.

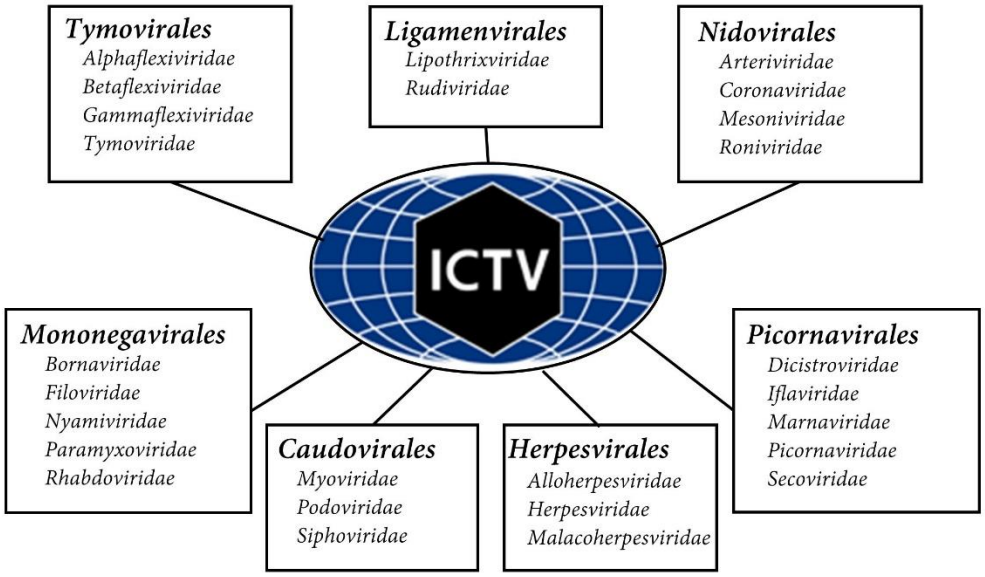

Figure 1. Overview of the 2014 ICTV virus taxonomy (latest official release). There are currently 7 orders sub-divided into 26 families. Many additional families are recognized, but have not been assigned into an order.

The virion morphology of prokaryotic viruses can vary greatly in terms of tail length, capsid shape, and whether the virus is enveloped or non-enveloped (Figure 2). Their genomes may either consist of dsDNA, ssDNA, ssRNA (+) or dsRNA and can be linear, circular or segmented. There also exists a great variation in genome size, ranging from the ~ 4 kb long ssDNA genome of levivirus GA (Inokuchi *et al*, 1986) to the large dsDNA, 497 kb-long myovirus G genome (Hendrix, 2009). About 96% of all known bacterial viruses are tailed, dsDNA viruses (Ackermann, 2007).

As of August 2015, the annotated, non-redundant full-length virus genome database RefSeq (URL: http://www.ncbi.nlm.nih.gov/genome/viruses/) contains 4877 genomes. Prokaryotic viruses constitute 34 % of the database, with 1670 genomes available (bacterial viruses: 1602; archaeal viruses: 68). A mere 138 bacterial genera are represented, with the top three hosts being *Mycobacterium* (15 %), *Pseudomonas* (8 %) and *Enterobacter* (7 %). *Acidianus* and *Sulfolobus* (family: *Sulfolobaceae*) are the most represented Archaeal genera. This wealth of virus genomic data has prompted their inclusion in taxonomic classification (Rohwer & Edwards, 2002). In contrast with the conserved universal genetic markers such as 16S ribosomal DNA used for the classification of prokaryotes and 18S for eukaryotes (Bik *et al*, 2012), comparative genomics has revealed that no single gene is shared amongst all known virus genomes (Rohwer & Edwards, 2002). Instead, nucleotide or amino acid sequences of viral structural components (e.g. capsid proteins) or enzymes (e.g. DNA polymerases) can be used for

classification of certain virus groups (Adriaenssens & Cowan, 2014). Classification of prokaryotic viruses based on the predicted phage proteome has also been attempted (Rohwer & Edwards, 2002) and recently used for a revised classification for the *Siphoviridae* family of tailed phages (Adriaenssens *et al*, 2014). The latter method consisted of a large collection of 105 representative phage genomes, which produced a universal "phage proteomic tree" (Figure 3). Since its introduction in 2002, the tree has grown in size and complexity, owing to the continuous inflow of new phage genomes. The latest iteration, produced in 2013, can be found at the following link: http://www.phantome.org/PhageProteomicTree/latest/.

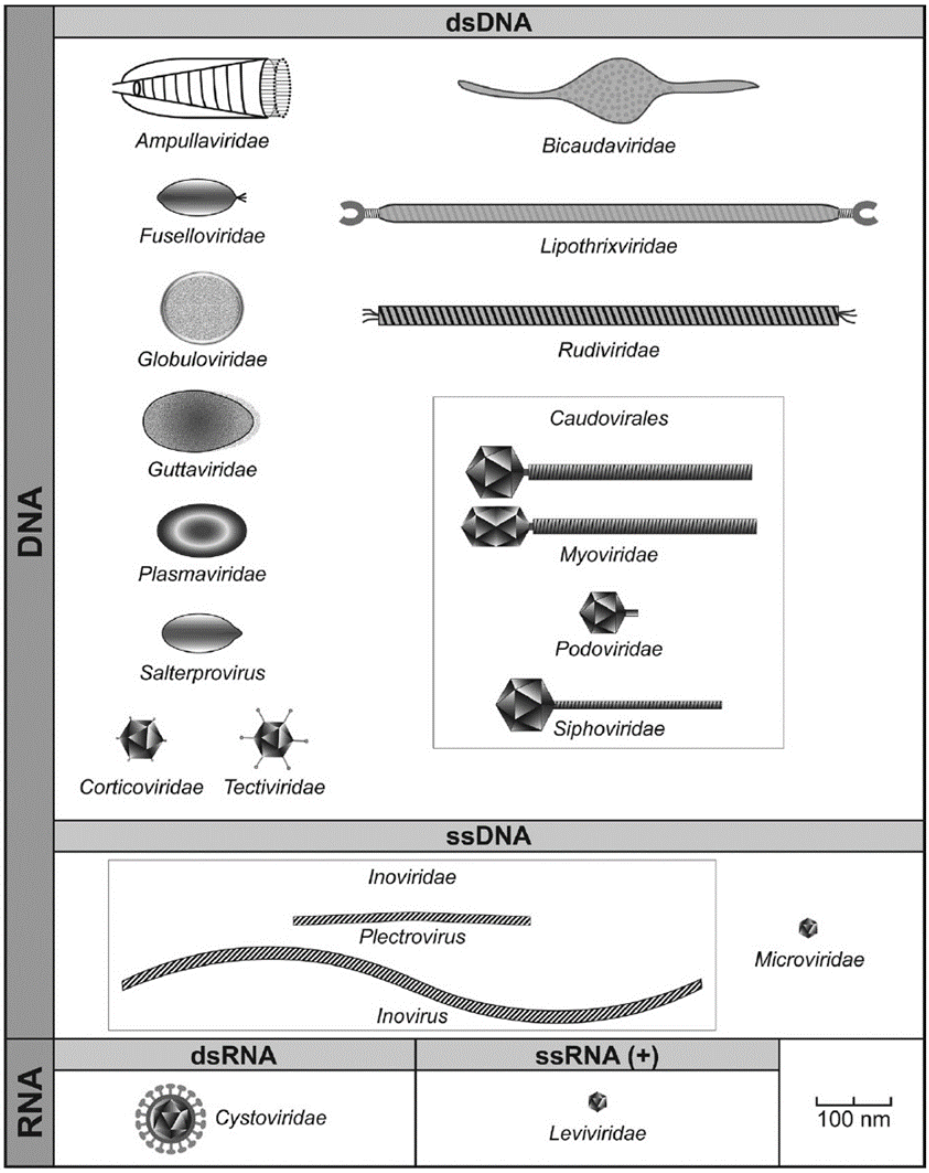

Figure 2. ICTV classification of prokaryotic viruses. Viruses are primarily classified according to nucleic acid type (DNA/RNA, single-stranded/double-stranded), and morphological features (Adapted from the IXth report of the ICTV, 2011).

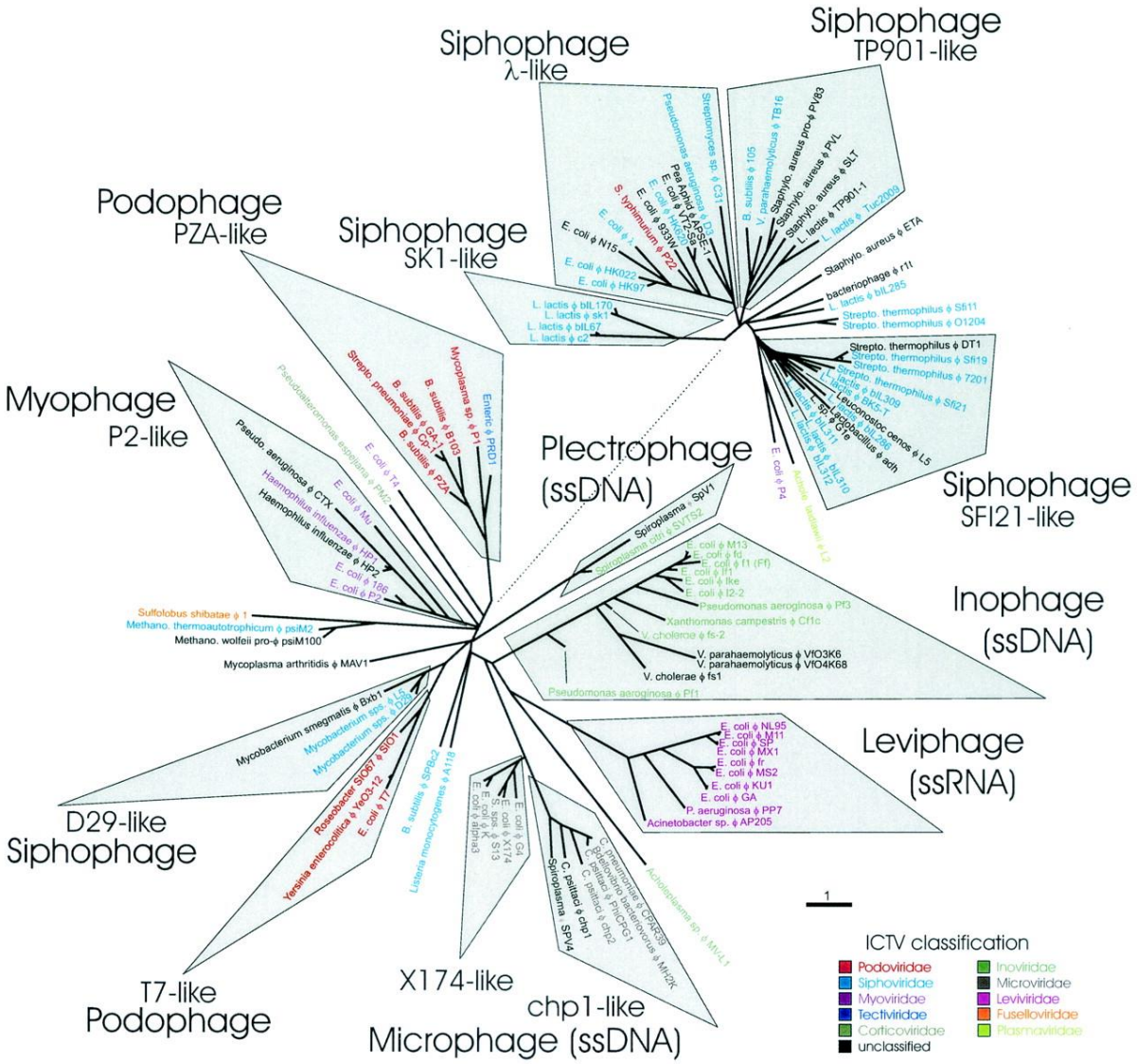

Figure 3. The universal phage proteomic tree. The dendrogram was constructed by using predicted phage proteome sequence data to classify bacterial viruses (adapted from Rohwer and Edwards, 2002).

1.2.3 Replication strategies

The replication cycles of prokaryotic viruses are diverse and may result in different infection outcomes, summarized in Figure 4. A common first step is adsorption of viral structural elements (e.g. tail fibers in tailed viruses) to extracellular components/receptors on its host surface, such as lipopolysaccharides (Adams, 1959). Once attached, nucleic acid injection into the cytoplasm is initiated. The ensuing infection cycle may then proceed in two ways, determined by the two main types of prokaryotic viruses: temperate and virulent. Temperate viruses (e.g. *Escherichia* phage λ) may either enter the lysogenic or the lytic cycle, whereas virulent viruses (e.g. *Escherichia* phage T4) can only follow the lytic cycle (Snyder *et al*, 2013).

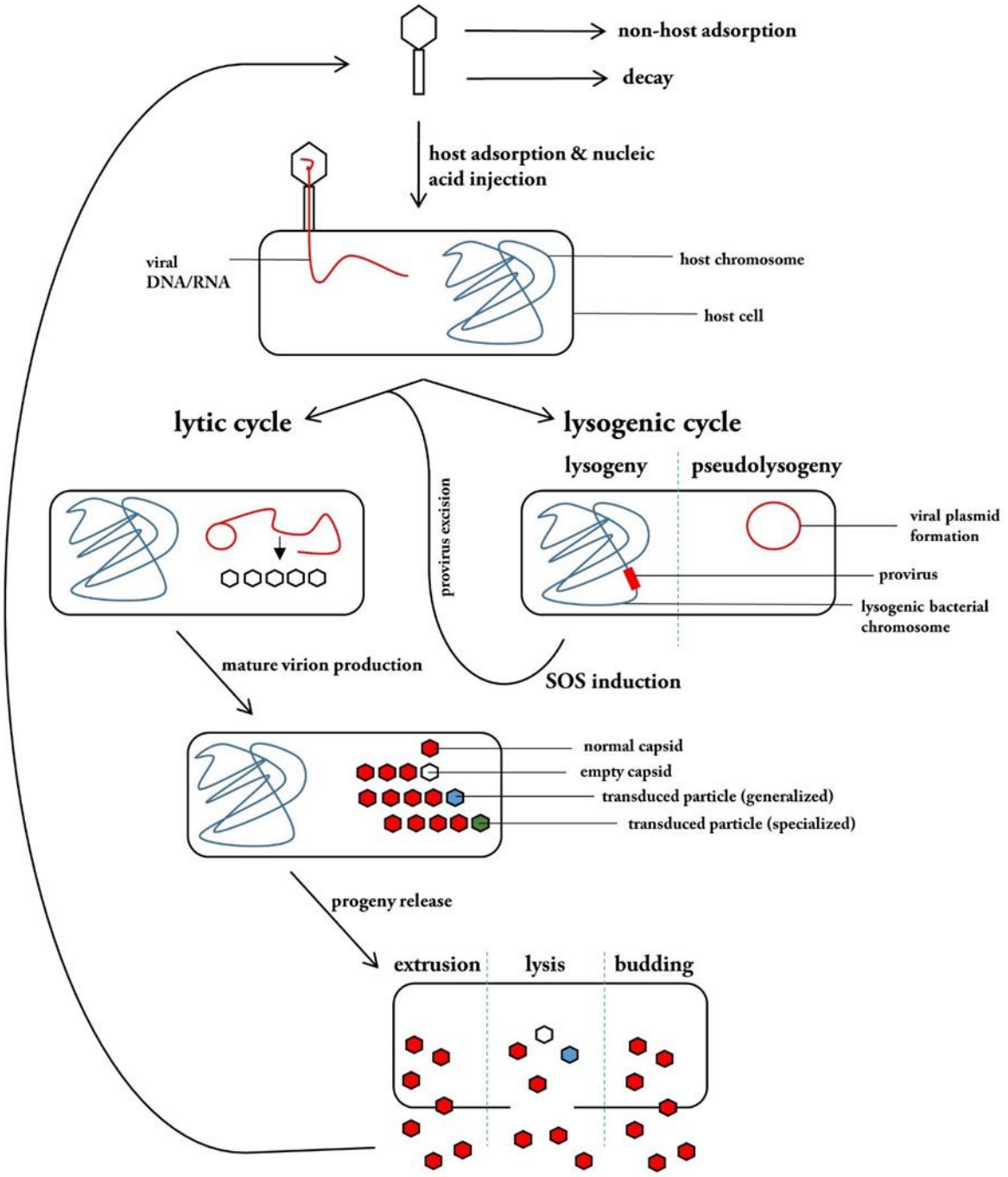

Figure 4. Generalized replication strategies of prokaryotic viruses. An infection cycle starts with attachment and adsorption to a host prokaryotic cell. Depending of virus type, either a lytic or lysogenic infection will occur. In virulent infections, only the lytic cycle is followed, where the virus replicates and bursts out of the host cell. Temperate viruses have the choice to follow either type of infection. During lysogeny, the viral genome is inserted into the host chromosome, and may also persist as an extra-chromosomal element (pseudolysogeny). If the bacterial SOS-response is triggered, the lytic cycle is induced. Four types of virion particles may result from either infection routes: 1) normal virion, which only contain viral DNA (red virions), 2) empty virions, without DNA (white virions), 3) transduced virions, with random host DNA (generalized particles, in blue) and 4) transduced virions, with specific host DNA following provirus excision (specialized particles, in green). Depending on the virus, particles may exit the cell by three mechanisms: extrusion (inoviruses and fuselloviruses), lysis (*Caudovirales* and lipothrixviruses) and budding (plasmaviruses). In a non-host environment, viruses may either adsorb irreversibly to certain molecules or surfaces (disabling infectivity) or degrade due to various factors.

Through the lysogenic cycle, a virus enters a latent infection state. In this cycle, the virus becomes integrated into the host chromosome, and can be vertically transmitted through subsequent rounds of host cell division (Lwoff, 1953). The integrated form of the virus is called a provirus (prophage is the commonly used term, but as some archaeal viruses are also temperate, this term may be more appropriate; the term provirus is usually used for integrated forms of retro-transcribing eukaryotic viruses; e.g. HIV) and the host cell is called a lysogen. An additional form of lysogeny occurs in certain viruses (e.g. phage P1, P22), whereby the virus achieves latency by genome circularization and host-dependent replication (plasmid-like formation). A prokaryotic host hosting such virus configuration is called a pseudolysogen (Abedon, 2009). The insertion of the virus genome into the host chromosome can be achieved either through site-specific recombination (e.g. phage λ; (Casjens & Hendrix, 2015)) or by transposition (e.g. phage Mu 1; (Montaño *et al*, 2012)). In site-specific recombination, circularization of the dsDNA virus genome is required for enzyme-mediated homologous recombination with the circular host chromosome.

A host harboring a provirus can acquire novel metabolic functions and immunity to superinfection by homologous viruses, termed lysogenic conversion (Menouni *et al*, 2013; Paul, 2008). The acquisition of virulence factors by non-pathogenic bacteria is a prime example of lysogenic conversion. Such factors include phage-encoded bacterial toxins such as the cholera toxin (Davis & Waldor, 2003), the Shiga toxin (Muniesa *et al*, 2004), and the neurotoxic botulism toxin from *Clostridium botulinum* (Sugiyama, 1980). Phage-borne genes have been observed to influence a range of other phenotypes in bacteria. For example, key cyanobacterial photosystem proteins (*psbA*) were present within phage genomes which infect cyanobacteria (cyanophages), which may be transferred to other bacterial species or members of community (Mann, 2003).

During lysogeny, virus genes involved in host lysis and virus genome replication are down-regulated (Hammerl *et al*, 2015). Environmental triggers such as ultraviolet radiation, nutrient deficiency or certain chemicals (e.g. antibiotics), or any form of stress which induces RecA production in the host cell (SOS response), will induce provirus excision and initiate the lytic cycle (Nanda *et al*, 2014). During the virion maturation phase, errors in capsid packaging can occur, resulting in mature virions containing host-derived DNA. This is the basis of transduction, which is the transfer of genes from one cell to another by a virus. Transduced virus particles (Figure 4) are generated by two mechanisms, either through specialized or generalized transduction (Snyder *et al*, 2013). Specialized transduction is directly related to lysogeny, as it results from errors in the excision of the provirus. Host genes adjacent to the virus insertion site may thus be packaged in viral capsids, along with the truncated virus genome. Generalized transduction results from erroneous packaging of random host DNA fragments, and therefore can occur in either virulent or temperate viruses. Lastly, empty capsid can also be produced and released during host lysis.

Bacterial conjugation, transformation and virus-mediated gene transfer (i.e. transduction) constitute the three horizontal gene transfer mechanisms in prokaryotes (Paul, 1999). Transduction

has been proposed as a major driver for prokaryote evolution and species diversification (Weinbauer & Rassoulzadegan, 2004). In marine ecosystems, high transduction rates (10^{13}-1.4×10^{14} transductants per year; (Jiang & Paul, 1998)) have led to suggestions that transduction might represent a significant, and understudied, microbial ecology process (Suttle, 2000; Weinbauer, 2004).

The lytic cycle ultimately leads to the release of progeny virions from the host cell. Host lysis is a programmed event, which must be carefully timed with the production of progeny virion. In dsDNA tailed viruses, temporal regulation of host lysis is mediated by a "holin-endolysin" system (Young & Bläsi, 1995). Holins are a genetically diverse family of virus-encoded transmembrane proteins, which serve as "biological timers" for the activation of their cognate endolysin (Wang *et al*, 2000). Endolysins (also called lysozymes) are cell-wall degrading enzymes, which are classified according to their enzymatic activity (e.g. muramidase, transglycosylase, amidase and endopeptidase) (Schmelcher *et al*, 2012; Nelson *et al*, 2012; Borysowski *et al*, 2006). Regardless of enzymatic specificity, all endolysins share a common function (i.e., they compromise the structural integrity of the cell wall by the cleavage of peptide or amide bonds). Depending on the virus, particles may exit the cell by three mechanisms (Figure 4): extrusion (inoviruses and fuselloviruses), lysis (*Caudovirales* and lipothrixviruses) and budding (plasmaviruses) (Calendar, 2012).

An alternative host lysis strategy to lysozyme-mediated lysis has been found only in small, single-stranded DNA/RNA viruses (Bernhardt *et al*, 2000), such as the *Microviridae* (ssDNA), *Leviviridae* ((+) ssRNA) and the proposed *Alleloviridae* family ((+) ssRNA). In this strategy, termed autolysis, the virus encodes an amurin, a protein with inhibitory activity towards the cell wall synthesis pathway (Bernhardt *et al.*, 2002). This mechanism of action is reminiscent of β-lactam antibiotics produced by *Penicillium* species. For this reason, amurins have sometimes been called "protein antibiotics".

The replication pathways described (i.e., lytic and lysogenic) are not the only possible infection outcomes in bacterial/archaeal viruses. Additional, lesser known and poorly understood replication cycles occur, such as non-integrative temperate infections and maintenance of chronic infections. However, as the molecular mechanisms for the lytic and lysogenic cycles are best understood, these cycles were put in emphasis.

1.3 Viral metagenomics

1.3.1 Introduction

Classical approaches to estimate virus abundance and diversity in a given environmental sample include virus counts (via transmission electron microscopy (EM) or fluorescent microscopy) and pulsed-field gel electrophoresis (PFGE) (Srinivasiah *et al*, 2008; Wommack *et al*, 2015, 1999). Direct EM observations of virus-like particles (VLPs) allow rudimentary virus classification based on morphological characteristics. PFGE is used for size estimation of viral genomes within a

metagenomic sample. By size-separating intact viral DNA, a range of distinct DNA bands varying in size and concentration is obtained (Carlson, 2005). A disadvantage of PFGE is a lack of gel resolution, where multiple virus genomes (of different species) may be visualized as single band of DNA. Both EM and PFGE are cost-effective and are useful in the partial assessment of virus richness and abundance in a given sample. However, these methods do not provide information about the genome/gene composition and taxonomic diversity of a virus population, particularly towards low titre virus species.

Since the introduction of next-generation sequencing (NGS) instruments over the last decade, sequencing costs per base have dramatically declined (Lam *et al*, 2012). This has permitted NGS to become routinely used for the genomic analysis of viruses on a community scale, termed viral metagenomics or metaviromics (Bzhalava & Dillner, 2013). The term metaviromics is formally defined as "culture-independent, functional and sequence-based analysis of an assemblage of phage genomes in an environment sample" (Riesenfeld *et al*, 2004). Applied to virus ecology, the metaviromics approach is a powerful tool for the determination of genotype abundance, community diversity and structure (Bzhalava & Dillner, 2013). Viral metagenomics was first applied for the analysis of two seawater samples (Breitbart *et al*, 2002). In this pioneering work, shotgun sequencing of two viral communities was achieved using clone libraries. At the time, 65% of sequences obtained did not show any similarity to database sequences, while the rest of the identified sequences belonged to the three major families of dsDNA tailed bacteriophages (i.e. *Myoviridae, Siphoviridae* and *Podoviridae*). Ecological insights gained from the use of metaviromics in aquatic habitats and in soils will be discussed in Chapter 2.

1.3.2 Generating a metavirome

An example of a standard, environmental metaviromics workflow is outlined in Figure 5. The first and most important step in metagenomics is efficient virus purification and recovery of high purity nucleic acids. No universal method exists for effective virus recovery in all sample types, although methods for some of the most commonly sampled environments (e.g. seawater) have been introduced (Thurber *et al*, 2009). For soil samples, this is a complex challenge, due to the combined effect of low viral DNA/RNA yields and the presence of chemical inhibitors (further discussed in Chapter 2).

The majority of metavirome projects have focused on the purification and analysis of dsDNA viruses, and have therefore omitted viruses with ssDNA, ssRNA and dsRNA genomes (Andrews-Pfannkoch *et al*, 2010). The taxonomic diversity of ssDNA viruses has been assessed using a metaviromics approach (Reavy *et al*, 2015; Hopkins *et al*, 2014; Roux *et al*, 2012; Kim *et al*, 2008; Bryson *et al*, 2015), although it has been argued that standard DNA purification methods used in metagenomic library construction are strongly biased towards the recovery of dsDNA replicative forms (Andrews-Pfannkoch *et al*, 2010). To reduce bias in virus diversity estimates, a

chromatography-based method (using hydroxyapatite) has been developed for the simultaneous recovery single-stranded and double-stranded forms of DNA and RNA virus genomes (Andrews-Pfannkoch *et al*, 2010).

An optional methodological pathway can be followed to recover proviruses (Figure 5, left path). Host culturing in nutrient broth (e.g. LB), antibiotic treatment and subsequent lysis allows for the recovery of proviruses (Mizuno *et al*, 2013). Virus particles can be concentrated by various means, but the most common first step is host removal by pore-size filtering and high speed centrifugation. Removal of co-occurring nucleic acids in the virus pellet is performed by adding nucleases (DNase and RNase), followed by the release of viral nucleic acids from their capsids by the addition of a proteolytic enzyme (Proteinase K). Viral DNA/RNA is then precipitated and can either be amplified randomly (using whole-genome amplification) or used for targeting of single genes by PCR amplification.

Prior to NGS, randomly amplified virus genome fragments were inserted into bacterial cloning vectors, and randomly selected for capillary sequencing (Breitbart *et al*, 2002). However, bias arising from cloning (e.g. preferential ligation of certain DNA fragments) and the limited number of clones typically selected for sequencing has made this technique restrictive (Mardis, 2007). Next-generation sequencers typically require DNA concentrations in the microgram range for optimal read output (Parkinson *et al*, 2012).

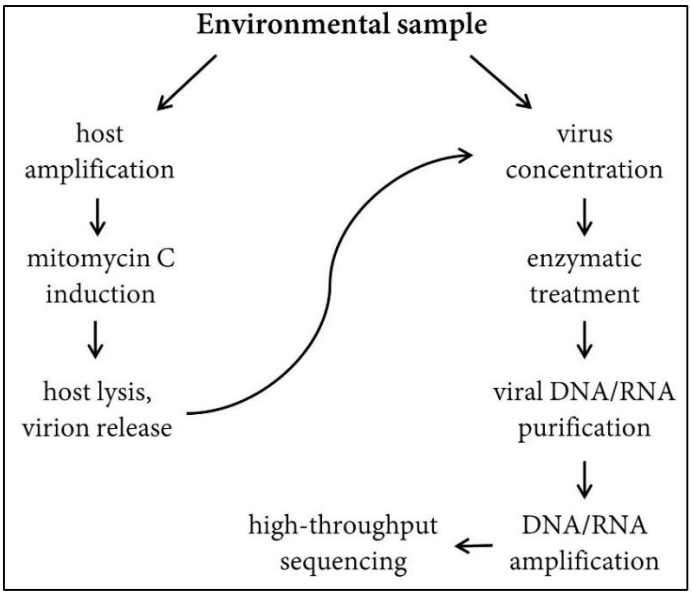

Figure 5. Environmental metagenomics workflow. Viruses may be recovered in two ways: either the provirus population is induced, or viruses are directly extracted from the sample.

1.3.3 Analysis of metavirome sequence data

NGS instruments can generate millions of short DNA sequences, called reads (Fumagalli *et al*, 2014). The analysis of such large datasets of virome reads is typically carried out in a bioinformatics

pipeline, divided in three consecutive main steps (Wooley *et al*, 2010): 1) read trimming/removal based on quality (Phred) scores (Ewing & Green, 1998), 2) *de novo* assembly of reads into contiguous sequences (contigs) and 3) taxonomic assignment of reads or contigs. A range of bioinformatics pipelines has been developed for a wide array of virus research applications, and have recently been reviewed (Sharma *et al*, 2015; Koboldt *et al*, 2013; Bzhalava & Dillner, 2013). Automated, server-based bioinformatics pipelines commonly used for virus ecology research include MG-RAST (Aziz *et al*, 2008; Meyer *et al*, 2008), MetaVir (Roux *et al*, 2014, 2011) and VIROME (Wommack *et al*, 2012). A brief overview of these pipelines is provided:

1. MG-RAST (http://metagenomics.anl.gov/): This pipeline was originally designed for the annotation of prokaryotic genomes using sub-systems, which reflect gene categories and functions (RAST, Rapid Annotation using Sub-systems Technology, (Aziz *et al*, 2008)). A metagenomic iteration of RAST was subsequently developed, called MG-RAST (MG, MetaGenomic). The pipeline permits read quality control, identification of rDNA sequences, taxonomic and functional assignment (sub-systems) of reads and/or contigs by BLAST comparison across multiple public databases. However, it is not recommended for the analysis of viromes due to high numbers of taxonomic misannotation of virus sequences identified as bacterial in origin.

2. MetaVir (http://metavir-meb.univ-bpclermont.fr): This pipeline is very similar to MG-RAST but is designed to identify virus sequences only. This is achieved by using BLASTp searches against the viral genome database RefSeq. Its main disadvantages are the limited number of reads which can be analyzed per metagenome (~ 2 million), and processing speed. Various statistical, phylogenetic and taxonomic analyses of viral reads/contigs can be performed. It offers the ability to compare a metavirome to others deposited on the server, and comparisons can be performed against datasets obtained from particular environments. Metagenome comparison is also available in MG-RAST and VIROME.

3. VIROME (Viral Informatics Resource for Metagenome Exploration, http://virome.dbi.udel.edu/): This pipeline, like MetaVir, is more focused on virus analysis. Unlike MetaVir, VIROME uses several additional databases for ORF classification. These include ribosomal databases, UniRef100 and MetaGenomesOnline. This permits the assessment and extent of bacterial/eukaryotic DNA contamination, as well as taxonomic assignment of virus reads/contigs.

1.4 Concluding remarks

Viruses of prokaryotes are the most diverse and abundant biological entities in the global biosphere, and have contributed immensely to our current understanding of molecular biology and

evolution (Suttle, 2005; Morange, 1998). In recognition of these facts, and to celebrate the hundredth anniversary of bacteriophage discovery (Twort, 1915), the year 2015 has been named the“ year of the phage” by viromics pioneer, Dr. Forest Rohwer (Koeris, 2015). Amazingly, it is only recently that the research community has begun to appreciate the diversity, abundance and crucial ecological roles prokaryotic viruses have in nature. Ecological perspectives of prokaryotic viruses will be discussed in chapter 2.

"The molecular biology of higher organisms does not stand on the shoulder of giants, but on the shoulder of dwarfs like phage T4 and lambda."

- Harald Brüssow

1.5 References

Abedon ST (2009) Disambiguating bacteriophage pseudolysogeny: an historical analysis of lysogeny, pseudolysogeny, and the phage carrier state. *Contemp. trends bacteriophage Res.***:** 285–307

Abedon ST & Murray KL (2013) Archaeal viruses, not archaeal phages: An archaeological dig. *Archaea* **2013:**

Ackermann HW (2007) 5500 Phages examined in the electron microscope. *Arch. Virol.* **152:** 227–243

Ackermann HW (2011) Bacteriophage taxonomy. *Microbiol. Aust.* **32:** 90–94

Adams MH (1959) Bacteriophages. *Bacteriophages*

Adriaenssens EM, A, Edwards R, B, Nash JHE, C, Mahadevan P, D, Seto D, E, Hans, -, Ackermann W, F, Lavigne R, G & and Andrew M. Kropinsk (2014) Integration of genomic and proteomic analyses in the classification of the Siphoviridae family.

Adriaenssens EM & Cowan DA (2014) Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.* **80:** 4470–4480

Andrews-Pfannkoch C, Fadrosh DW, Thorpe J & Williamson SJ (2010) Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl. Environ. Microbiol.* **76:** 5039–5045

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, et al (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75

Bernhardt TG, Roof WD & Young R (2000) Genetic evidence that the bacteriophage phi X174 lysis protein inhibits cell wall synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 4297–4302

Bernhardt TG, Wang IN, Struck DK & Young R (2002) Breaking free: 'Protein antibiotics' and phage lysis. *Res. Microbiol.* **153:** 493–501

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R & Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* **27:** 233–243 Available at: http://dx.doi.org/10.1016/j.tree.2011.11.010

Borysowski J, Weber-Dabrowska B & Górski A (2006) Bacteriophage endolysins as a novel class of antibacterial agents. *Exp. Biol. Med. (Maywood).* **231:** 366–377

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F & Rohwer F (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 14250–14255

Bryson SJ, Thurber AR, Correa AMS, Orphan VJ & Vega Thurber R (2015) A novel sister clade to the enterobacteria microviruses (family Microviridae) identified in methane seep sediments. *Environ. Microbiol.***:** n/a–n/a Available at: http://dx.doi.org/10.1111/1462-2920.12758

Bzhalava D & Dillner J (2013) Bioinformatics for Viral Metagenomics. *J Data Min. Genomics Proteomics* **4:** 602–2153

Calendar R (2012) The bacteriophages Springer Science & Business Media

Carlson K (2005) Appendix: working with bacteriophages: common techniques and methodological approaches. *Bacteriophages Biol. Appl.***:** 437–494

Casjens SR & Hendrix RW (2015) Bacteriophage lambda: Early pioneer and still relevant. *Virology* **479-480:** 310–330 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0042682215000537

Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH & Prangishvili D (2008) Exploring the prokaryotic virosphere. *Res. Microbiol.* **159:** 306–313

d'Herelle F (1917) Sur un microbe invisible antagoniste des bacilles dysentériques. *CR Acad. Sci. Paris* **165:** 373–375

Davis BM & Waldor MK (2003) Filamentous phages linked to virulence of Vibrio cholerae. *Curr. Opin. Microbiol.* **6:** 35–42

Ewing B & Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194

Fumagalli M, Vieira FG, Linderoth T & Nielsen R (2014) NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* **30:** 1486–1487

Hammerl JA, Roschanski N, Lurz R, Johne R, Lanka E & Hertwig S (2015) The Molecular Switch of Telomere Phages: High Binding Specificity of the PY54 Cro Lytic Repressor to a Single Operator Site. *Viruses* **7:** 2771–2793

Hendrix RW (2009) Jumbo bacteriophages. *Curr. Top. Microbiol. Immunol.* **328:** 229–240

Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M & Breitbart M (2014) Diversity of environmental single-stranded DNA phages revealed by PCR

amplification of the partial major capsid protein. *ISME J.*: 1–11 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24694711

Inokuchi Y, Takahashi R, Hirose T, Inayama S, Jacobson AB & Hirashima A (1986) The complete nucleotide sequence of the group II RNA coliphage GA. *J. Biochem.* **99:** 1169–1180

Jiang SC & Paul JH (1998) Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64:** 2780–2787

Kim K-H, Chang H-W, Nam Y-D, Roh SW, Kim M-S, Sung Y, Jeon CO, Oh H-M & Bae J-W (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **74:** 5975–5985

Koboldt DC, Steinberg KM, Larson DE, Wilson RK & Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. *Cell* **155:**

Koeris M (2015) Life in Our Phage World A Centennial Field Guide to the Earth's Most Diverse Inhabitants.

Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP & Snyder M (2012) Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30:** 562–562

Lwoff A (1953) Lysogeny. *Bacteriol. Rev.* **17:** 269–337 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=180777&tool=pmcentrez&rendertype=abstract

Mann NH (2003) Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol. Rev.* **27:** 17–34

Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods* **4:** 613–614

Menouni R, Champ S, Espinosa L, Boudvillain M & Ansaldi M (2013) Transcription termination controls prophage maintenance in Escherichia coli genomes. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 14414–9 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3761637&tool=pmcentrez&rendertype=abstract

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R & Wilke A (2008) The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386

Mizuno CM, Rodriguez-Valera F, Kimes NE & Ghai R (2013) Expanding the marine virosphere using metagenomics.

Montaño SP, Pigli YZ & Rice PA (2012) The μtranspososome structure sheds light on DDE recombinase evolution. *Nature* **491:** 413–7

Morange M (1998) History of molecular biology Wiley Online Library

Muniesa M, Serra-Moreno R & Jofre J (2004) Free Shiga toxin bacteriophages isolated from sewage

showed diversity although the stx genes appeared conserved. *Environ. Microbiol.* **6:** 716–725

Nanda AM, Heyer A, Krämer C, Grünberger A, Kohlheyer D & Frunzke J (2014) Analysis of SOS-induced spontaneous prophage induction in Corynebacterium glutamicum at the single-cell level. *J. Bacteriol.* **196:** 180–188

Nelson DC, Schmelcher M, Rodriguez-Rubio L, Klumpp J, Pritchard DG, Dong S & Donovan DM (2012) Endolysins as Antimicrobials. *Adv. Virus Res.* **83:** 299–365

Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP & Fischer MD (2012) Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res.* **22:** 125–133

Paul JH (1999) Microbial gene transfer: an ecological perspective. *J. Mol. Microbiol. Biotechnol.* **1:** 45–50

Paul JH (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* **2:** 579–589 Available at: http://www.nature.com/doifinder/10.1038/ismej.2008.35

Reavy B, Swanson MM, Cock P, Dawson L, Freitag TE, Singh BK, Torrance L, Mushegian AR & Taliansky M (2015) Distinct circular ssDNA Viruses Exist in Different Soil Types. *Appl. Environ. Microbiol.*: AEM.03878–14 Available at: http://aem.asm.org/lookup/doi/10.1128/AEM.03878-14

Riesenfeld CS, Schloss PD & Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38:** 525–552

Rohwer F & Edwards R (2002) The phage proteomic tree: A genome-based taxonomy for phage. *J. Bacteriol.* **184:** 4529–4535

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D & Enault F (2011) Metavir: A web server dedicated to virome analysis. *Bioinformatics* **27:** 3074–3075

Roux S, Krupovic M, Poulet A, Debroas D & Enault F (2012) Evolution and diversity of the microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* **7:** 1–12

Roux S, Tournayre J, Mahul A, Debroas D & Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15:** 76 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4002922&tool=pmcentrez&rendertype=abstract

Schmelcher M, Donovan DM & Loessner MJ (2012) Bacteriophage endolysins as novel antimicrobials. *Future Microbiol.* **7:** 1147–1171

Sharma D, Priyadarshini P & Vrati S (2015) Unraveling the web of viroinformatics: computational tools and databases in virus research. *J. Virol.* **89:** 1489–1501

Snyder L, Peters JE, Henkin TM & Champness W (2013) Molecular genetics of bacteria American Society of Microbiology

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T & Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* **159:** 349–357

Sugiyama H (1980) Clostridium botulinum neurotoxin. *Microbiol. Rev.* **44:** 419

Suttle CA (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. *Viral Ecol.* **1:** 247–296

Suttle CA (2005) Viruses in the sea. *Nature* **437:** 356–361

Thurber R V, Haynes M, Breitbart M, Wegley L & Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4:** 470–483

Torsvik T & Dundas ID (1974) Bacteriophage of Halobacterium salinarium. *Nature* **248:** 680–681

Twort FW (1915) An investigation on the nature of ultra-microscopic viruses. *Lancet* **186:** 1241–1243

Wang IN, Smith DL & Young R (2000) Holins: the protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* **54:** 799–825

Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28:** 127–181

Weinbauer MG & Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6:** 1–11

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S & Nasko DJ (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.* **6:** 427–439

Wommack KE, Nasko DJ, Chopyk J & Sakowski EG (2015) Counts and sequences, observations that continue to change our understanding of viruses in nature. *J. Microbiol.* **53:** 181–192 Available at: http://link.springer.com/10.1007/s12275-015-5068-6

Wommack KE, Ravel J, Hill RT, Chun J & Colwell RR (1999) Population dynamics of chesapeake bay virioplankton: Total-community analysis by pulsed-field gel electrophoresis. *Appl. Environ. Microbiol.* **65:** 231–240

Wooley JC, Godzik A & Friedberg I (2010) A primer on metagenomics. *PLoS Comput. Biol.* **6:**

Young R & Bläsi U (1995) Holins: Form and function in bacteriophage lysis. In *FEMS Microbiology Reviews* pp 191–205.

Chapter 2:

Diversity and ecology of viruses in hyperarid desert soils: a review

2.1 Chapter summary

In recent years, remarkable progress has been made in the field of virus environmental ecology. In marine ecosystems for example, viruses are now known to play pivotal roles in the biogeochemical cycling of nutrients and to be mediators of microbial evolution through horizontal gene transfer. In soils, the diversity and ecology of viruses is poorly understood, but evidence supports the view that these differ substantially from aquatic systems. Soils subjected to environmental extremes, as in hot hyperarid deserts, contain high virus titres, although even higher virion counts have been reported for cold desert soils. This observation, though based on a limited number of studies, has led to the hypothesis that temperature is the major determinant of the propensity for soil viruses to persist either in a lysogenic or lytic state. We suggest that viruses in desert soils have a limited capacity to shape system-wide community processes and population dynamics. Analyses of viral phylogenetic diversity in desert soils have shown a wide range of virus taxa, including families of the *Caudovirales* (bacteriophages), *Phycodnaviridae* and *Mimiviridae*. Many desert soil virus sequences show low identity values to virus genomes in public databases, suggesting the existence of distinct and as yet uncharacterised soil phylogenetic lineages. Preliminary analyses of viruses infecting key microbial ecosystem regulators (e.g., cyanobacteria) suggest that cyanophage differ substantially from their marine homologs. This review provides an overview of recent advances in the study of viruses in hyperarid soils, and the factors that contribute to viral abundance and diversity in these environments.

2.2 Introduction

Over recent decades, the critical roles that viruses play in the environment have become increasingly recognized by the research community (Rohwer *et al*, 2009). It has been estimated by direct counts of extracellular ('free floating') virus-like particles (VLPs) that the global "virosphere" may contain up to 10^{31} viral particles (Suttle, 2005), suggesting that viruses may be the most abundant biological entities on the planet and, potentially, the greatest reservoir of genetic diversity (Weinbauer & Rassoulzadegan, 2004; Frost *et al*, 2005; Dennehy, 2013). The ecological importance of viruses on a global scale has predominantly emerged from studies of marine and fresh water microbial communities (Fuhrman, 1999; Wommack & Colwell, 2000; Bettarel *et al*, 2004; Winget *et al*, 2005; Suttle, 2007; Rodriguez-Valera *et al*, 2009; Breitbart, 2012), where viruses have been demonstrated to be involved in core processes such as biogeochemical nutrient cycling (Fuhrman, 1999; Wommack & Colwell, 2000; Suttle, 2007), microbial population control through viral lysis (Wommack & Colwell, 2000; Bettarel *et al*, 2004) and microbial evolution via horizontal gene transfer (Rodriguez-Valera *et*

al, 2009).

Research on the virus ecology of soil environments has progressed more slowly and has received proportionally less attention (Breitbart, 2012; Kimura *et al*, 2008; Srinivasiah *et al*, 2008). However, enumeration of virus particles by electron microscopy (EM) on several soil types (Ashelford *et al*, 2003; Williamson *et al*, 2005, 2007) has shown high viral abundance values ranging from 1.5 x10^8 to 6.4 x10^8 per gram dry weight soils. Soil ecosystems are subject to unique abiotic ecological pressures, in part due to their wide compositional spectrum and spatial heterogeneity in terms of physicochemical properties (Schlesinger *et al*, 1990; Palmer, 2003). Environmental stresses are even greater in extremely arid soil systems, where soil organisms and communities may be simultaneously exposed to pulsed water events, and to the effects of desiccation-, solute- and UV-B radiation induced oxidative-stresses (Austin *et al*, 2004; Chen *et al*, 2009). Deserts represent the single largest terrestrial ecosystem type on Earth, covering ~33.6% of the global land mass, excluding Antarctica (Meigs, 1952), and are classified in terms of their aridity index, a ratio between precipitation (P) and potential of evapotranspiration (PET) (UNEP, 2013). This results in four desert categories, as dry-semiarid (0.5<P/PET<0.65), semiarid (0.2<P/PET<0.5), arid (0.05<P/PET<0.2) and hyperarid (P/PET<0.05). Hyperarid deserts generally receive annual precipitation of ≤ 70 mm and are often associated with intrinsic characteristics such as high pH (~7-9), high salinity levels, high surface radiation fluxes, long periods of desiccation and low water activity (Shmida & Wilson, 1985).

Desert soil microbial ecology research has primarily focused on bacterial communities, which have been shown to be largely responsible for primary production and the provision of key ecosystem services (Harel *et al*, 2004; Pointing *et al*, 2009; Tracy *et al*, 2010; Cowan *et al*, 2011). Soil virus populations and functions are seldom taken into consideration, thereby omitting a crucial variable within ecological models designed to predict microbial population dynamics. As a result, the ecological roles, survival mechanisms (against biotic and abiotic factors), the spatial and temporal changes in viral community structures (virus biogeography) and viral phylogenetic diversity, are still poorly understood in desert soils.

Since the recent re-emergence of soil virus ecology (Kimura *et al*, 2008), desert soils have possibly received the most attention (Table 1). With the advances in next generation sequencing (NGS) technologies, culture-independent methods have become the standard for sequence-based determination of viral diversity (Rosario & Breitbart, 2011). However, the rapidly growing volume of viral environmental sequence data has revealed that most sequences (~70%) have no homologs in public databases, and are typically labelled "viral dark matter" (Youle *et al*, 2012; Hatfull, 2015). Here, we discuss the current understanding of hot and cold desert soil virus diversity and function and identify key areas of future research.

Table 1. Mid-to high throughput soil-based studies pertaining to viral ecology since 2005.

Authors	Soil type	Location/country	Methods used
Pringent *et al.*, 2005	Hot desert surface sand	Sahara Desert in Morocco and Tunisia	EM[1], PFGE[2], Lytic induction
Williamson *et al.*, 2005	Agricultural, forest	Delaware,USA	Epifluorescence microscopy, EM[1]
Fierer *et al.*, 2007	Hot arid desert, tallgrass prairie, tropical rainforest	USA, Peru	Sanger sequencing of random viral metagenomics clones
Williamson *et al.*, 2007	Loamy and sandy soils, agricultural, forested wetlands	Antarctica (Tom and Obelisk pond); USA (Delaware)	Induction assays, Epifluorescence counting
Prestel *et al.*, 2008	Surface sand	Namib Desert	EM[1], PFGE[2], Sanger of cloned DNA fragments (LASL[3])
Swanson *et al.*, 2009	Dystric-Fluvic Cambisol soil	Dundee, Scotland	EM[1], epifluorescence counting
Meiring *et al.*, 2012	Soil underneath hypoliths	Miers Valley, Antarctica	Lytic induction, EM[1], phage isolation from culture
Pearce *et al.*, 2012	Surface soil	Alexander Island, Antarctica	Shotgun metagenome pyrosequencing
Swanson *et al.*, 2012	Surface soil (Antarctica)	Antarctica	EM[1], lytic induction, phage isolation
Prestel *et al.*, 2013	Dune surface sand	Mojave Desert, USA	EM[1], random amplification for viral DNA (Sanger)
Srinivasiah *et al.*, 2013	Surface soil	Antarctica (Tom and Obelisk pond)	RAPD[4] viral community fingerprinting
Adriaenssens *et al.*, 2014	Soil-associated rocks (hypoliths)	Namib Desert	Shotgun viral metagenome sequencing (Illumina)
Zablocki *et al.*, 2014	Antarctic surface soil and hypoliths	Miers Valley, Antarctica	Shotgun viral metagenome sequencing (Illumina)

[1]EM: electron microscopy; [2]PFGE: pulse field gel electrophoresis; [3]: LASL: linker amplified shotgun library; [4]RAPD: random amplified polymorphic DNA.

2.3 Diversity and abundance of viruses in desert soils

Hot deserts. Viral community analyses have been conducted on surface soil samples from three hot hyperarid deserts: the Sahara (Prigent *et al*, 2005), Namib (Prestel *et al*, 2008; Adriaenssens *et al*, 2015) and Mojave (Fierer *et al*, 2007; Prestel *et al*, 2013). In each of these studies, difficulties in detecting extracellular VLPs by electron microscopy (EM) or pulse field gel electrophoresis (PFGE) profiling were reported, suggesting a very low viral abundance within these soils. However, the inclusion of a lytic induction step (prophage excision stimulated by the addition of Mitomycin C (Ackermann & DuBow, 1987)) in the soil extraction protocol substantially increased the recovery of virus particles (Prigent *et al*, 2005; Prestel *et al*, 2008). For Sahara Desert surface sand samples, induced phage genomes were estimated to range in size from 45 to 270 kb. Electron microscopy (EM) of the induced phage fraction showed a majority of tailed virus morphotypes belonging to the *Myoviridae* family, some of which showed peculiar ribbon-like structures located at the tail tip of the virions

(Prestel *et al*, 2012). In the Namib Desert soil samples, twenty distinct morphotypes were identified, all members of the *Myoviridae* and *Siphoviridae* families with no apparent *Podoviridae*-like virions (Prestel *et al*, 2008). PFGE profiles from Namib soils indicated an average genome size of 55 to 65 kb, with several genomes of up to 350 kb in size (Prestel *et al*, 2008). EM visualisation of Mojave Desert sand samples showed eleven distinct tailed morphotypes, belonging to all three families of the *Caudovirales* (Prestel *et al*, 2013).

Sanger sequencing of randomly selected cloned phage fragments from the Mojave Desert soil virus communities showed that 36% of sequenced clones had no homologs in public sequence databases (Prestel *et al*, 2013). Within the identified virus sequences, the majority were homologous to bacteriophages infecting common soil bacteria such as members of the Proteobacteria, including *Bacillus* and *Rhizobium*. From the same samples, 38 bacterial isolates were grown in pure culture and 84% were shown to harbour at least one SOS-inducible phage. A similar study on loamy sand from a different area of the Mojave Desert showed that a large majority of randomly selected metaviral clone sequences had no database homologs (Fierer *et al*, 2007). Of those clones with significant sequence identity (tBLASTx search with a 10^{-3} threshold), phages associated with *Actinoplanes*, *Mycobacterium*, *Myxococcus* and *Streptomyces* were the most common. Other virus signals detected included archaeal (*Haloarcula* phage) and herpes-like viruses. Using a similar methodology, 50% of the viral sequences from three Namib Desert surface sand samples had no homologs in public sequence databases, with most positive hits showing homology to *Siphoviridae* phages linked to Gram-positive bacteria (Prestel *et al*, 2008). Most recently, a shotgun NGS approach was used to investigate the metavirome of Namib Desert hypoliths (Adriaenssens *et al*, 2015), cyanobacteria-dominated microbial niche communities on the ventral surfaces of translucent rocks (Makhalanyane *et al*, 2013b). The most abundant sequences belonged to *Geobacillus*- and *Bacillus*-infecting phages, while cyanophage markers were unexpectedly found only in low numbers. The distinct phylogenetic clustering of assembled *phoH* genes (a cyanophage marker (Goldsmith *et al*, 2011)) suggested that desert soil cyanophages were only distantly related to their well-studied marine counterparts (Adriaenssens *et al*, 2015), and that the dominance of marine cyanophage sequences in sequence databases might account for the low cyanophage hit rate of sequences in the Namib Desert hypolithon metavirome. This observation has wider implications for studies of soil metaviromics, where an underestimation of cyanophage abundance and diversity may skew estimates of the functional importance (and population dynamics) of soil cyanobacteria, arguably the most important taxonomic group in most desert soil microbial communities (Tracy *et al*, 2010; Cowan *et al*, 2011).

Cold deserts. Studies of viral communities in cold hyperarid desert soils have been almost exclusively conducted in the major ice-free regions of Antarctica (e.g., the East Antarctic McMurdo Dry Valleys). Direct viral counts by epifluorescence microscopy (Williamson *et al*, 2007) showed high VLP densities, in the range of 2.3 - 6.4 × 10^8 extracellular VLPs per gram of dry soil. The prevalence of bacterial lysogens within these soils was between 4.6 and 21.1%, a much lower occurrence level than

estimated for bacteria in hot desert soils (84% (Fierer *et al*, 2007)). Using epifluorescence direct counts of extractable bacteria and extracellular virus particles, virus-to-bacteria ratios (VBR) ranging from 170 to 8200 were calculated, the highest recorded for any soil ecosystem (Williamson *et al*, 2007).

Antarctic soil bacterial isolates have yielded several unique virus genomic structures. The distinct temperate siphophages (SpaA1 and BceA1) isolated from *Staphylococcus pasteuri* and *Bacillus cereus* both contained almost complete additional phage genomes (MZTP02) (Swanson *et al*, 2012). This "Russian doll" gene arrangement had not been previously described for soil bacteriophages, and has led to speculation that it may represent a 'fast-track' route for virus evolution and horizontal gene transfer, with a possible role in host range expansion.

Pyrosequencing of Antarctic soil metagenomic DNA has identified a wide diversity of bacteria, archaea, microeukaryotes and viruses (Pearce *et al*, 2012). From the total sequence dataset, 494 phage-related hits (0.18% of the total number of sequences) were identified. Top BLAST hits against public databases were related to phages known to infect to *Mycobacteria*, *Burkholderia*, *Bordetella*, *Pseudomonas*, *Enterobacteria*, *Flavobacterium*, *Myxococcus*, *Synechococcus*, *Prochlorococcus* and *Sinorhizobium*. However, viral DNA was not specifically enriched in this study, and this may have resulted in an underestimation of viral diversity.

The spatial composition and dynamics of viral communities along an Antarctic soil transect have been recently reported (Srinivasiah *et al*, 2013). Using random PCR amplification of polymorphic DNA (RAPD-PCR) assays, viral community fingerprints were used to assess short-term changes in the composition of viral communities. To maximize the number of viruses sequences amplified, RAPD-PCR primer design was based on the identification of recurring dodecamer sequences (G+C content ≥70%) within 22 selected viral metagenomes. Qualitative comparisons of the Antarctic fingerprint patterns demonstrated that heterogeneous soil conditions and associated environmental factors (e.g., carbon levels, moisture content, pH and light exposure frequency) impacted the composition of viral assemblages across geographic distances as short as 20 metres. The RAPD-PCR fingerprint data also suggested that virus assemblages were not present as inactive, inert particles, but were dynamically involved in infection of co-existing microbial hosts. Furthermore, environmental pressures known to influence bacterial community structures in the Antarctic desert (Yergeau & Kowalchuk, 2008) were shown to have a similarly influential role on virus community dynamics.

Abundance estimates (Williamson *et al*, 2007) suggest that Antarctic desert soils contain a substantially higher proportion of free extracellular VLPs than hot hyperarid desert soils, where a lysogenic lifestyle appears to be prevalent (Prigent *et al*, 2005; Prestel *et al*, 2008, 2013). A sequence-based metagenomic comparison of viral assemblages (single- and double-stranded DNA viruses only) in surface soils and hypolithic communities in the Antarctic McMurdo Dry Valleys (Zablocki *et al*, 2014a) demonstrated that bacteriophages constituted the majority of the identified viruses, representing all *Caudovirales* families. *Mycobacterium* phage sequences were the most highly represented in the viral fraction (Pearce *et al*, 2012). No archaeal virus sequences were recorded, in

line with previous observations that archaea are either absent or present in very low numbers in this environment (Pointing *et al*, 2009; Makhalanyane *et al*, 2013a). Within the hypolith metavirome dataset, the fraction of cyanophage sequences was under-represented, with low sequence similarities to known cyanophages. Dry Valley surface soils also contained a number of other virus signatures, including phycodnaviruses, mimiviruses and virophage capsid protein genes (Zablocki *et al*, 2014a), many of which are most commonly identified in aquatic systems.

2.4 Factors shaping viral community structures in desert soils

Soil virus populations display different dynamics from marine and freshwater systems (Srinivasiah *et al*, 2013) (Figure 6). In marine systems, two major factors influence viral abundance: the biological productivity of the system and microbial diversity and abundance (Weinbauer & Rassoulzadegan, 2004; Dennehy, 2013). Viral abundance has been shown to increase as the productivity of a system increases. Co-occurring virus host communities also influence viral abundance, as in microbial bloom events, which increase the number of lytic infections thereby releasing additional phage particles. Marine-associated abiotic parameters such as temperature, salinity and pH are stable on relatively large spatial scales (Kimura *et al*, 2008), and do not appear to significantly affect viral abundance.

Soil systems, particularly desert soils, are inhomogeneous, in that soil particles are semi-discrete. Under dry conditions, such as in typical desert soils, microbial populations are restricted to discrete biofilms adsorbed to particle surfaces and often embedded in EPS matrices (Davey & O'toole, 2000; Donlan, 2002; Gorbushina, 2007). While the biofilms on each of these particles may be considered to be independently homogenous systems, the system would only be considered as homogeneous during times of inundation (i.e., during significant rainfall events, which may be extremely infrequent (Reynolds *et al*, 2004)). We therefore suggest that the system-wide dispersal/transport of viruses in soil environments may be an extremely rare event. The corollary to this suggestion is that the capacity of viruses to impose large-scale controls on the population dynamics of their hosts in desert soils may be very limited. If the limited capacity for the transport of virions is combined with the observed propensity of phages in desert soil communities to exist in a lysogenic state (Prestel *et al*, 2013; Thurber, 2009), the implication is that desert soil viruses may play a less significant role in system-wide community processes and population dynamics. However, it is also clear from this consideration that there is a need for more targeted experimental analysis of the functional role of viruses and phage in soil systems, such as the potential beneficial character of these

viruses regarding lysogenic conversion and presence of auxiliary metabolic genes.

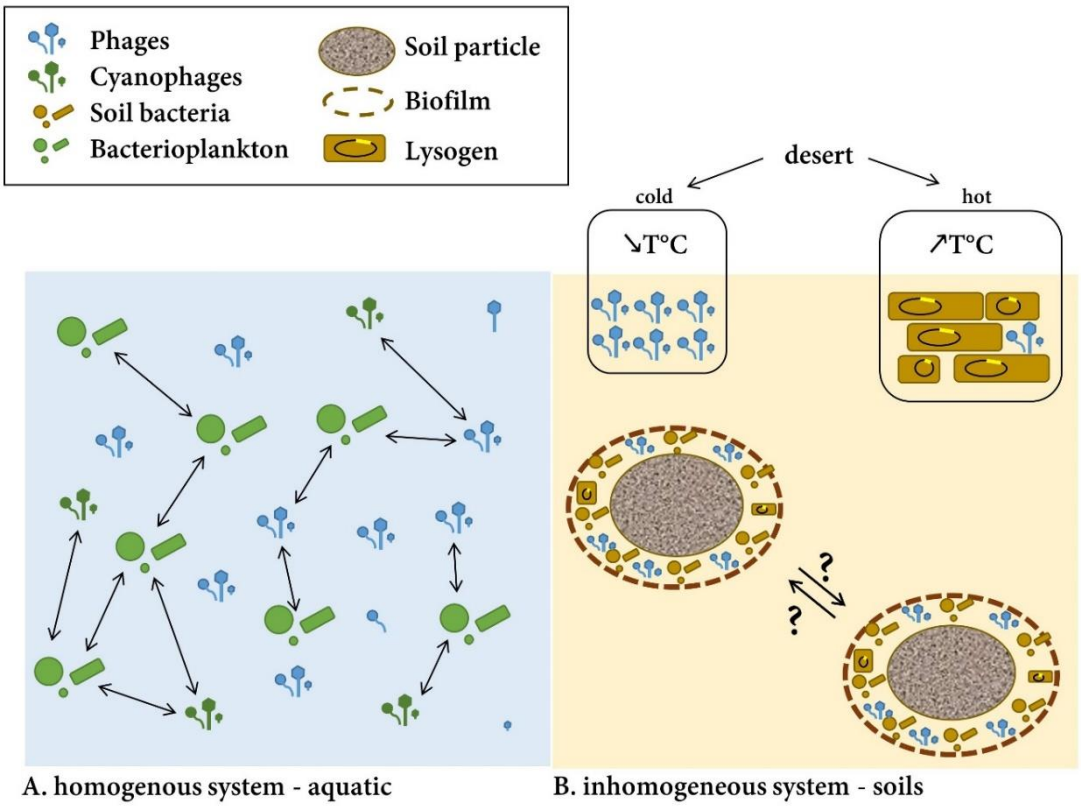

Figure 6. Virus community dynamics in aquatic (A) and a soil (B) ecosystems. Marine and freshwater systems can be regarded as homogenous systems, where the distribution of virus particles (e.g. phages) and host organisms (e.g. bacterioplankton) is relatively even, and high rates of infection, gene transfer and dispersal occur in the continuous medium. Abiotic properties such as temperature, pH and salinity remain relatively constant, and may have minimal impact on viral abundance. In contrast, hyperarid soil microbial communities exist as discrete systems, with soil microbial communities existing primarily in the form of biofilms around sand particles, protected by layers of extracellular polymeric substances. The level of virus-host interactions within biofilms and between individual biofilm communities remains an open question, but diffusion rates are expected to be low on both small and large spatial scales. Soil abiotic factors such as pH, clay content and organic matter vary through soil matrices, and contribute to the overall heterogeneity of the system. The mean temperature regime is suggested to influence the incidence of lysogeny in soil communities, affecting the number of extracellular virus particles.

On the basis of the arguments above, we argue that abiotic factors may be more important drivers of soil virus and host processes than biotic factors (c.f., aquatic and marine environments). Temperature has been shown to be one of the major factors controlling viral survival rates in soils (Hurst *et al*, 1980; Nasser & Oman, 1999), with lower temperatures enabling longer survival rates, extended latent periods and reduced burst sizes while warmer temperatures have been associated with reduced virus proliferation and greater inactivation rates (Straub *et al*, 1992; Leonardopoulos *et al*,

1996). Soil physicochemical variables such as pH, clay type, organic matter content, heavy metal concentration, salinity and moisture content have also been shown to significantly contribute to the regulation of virus communities (see reviews: (Kimura *et al*, 2008; Jończyk *et al*, 2011)). However, in Antarctic soils, no correlation between soil physical properties (% organic matter, % water content and pH) and the incidence of lysogeny has been established (Williamson *et al*, 2007), suggesting that the long-standing hypothesis that the high spatial heterogeneity of the soil environment selects for lysogenic replication (Stewart & Levin, 1984; Marsh & Wellington, 1994) does not hold for all soil habitats.

It has been suggested that for hot deserts, where viruses and their hosts are exposed to a range of environmental extremes (Prestel *et al*, 2013), lysogeny might constitute a survival mechanism for long-term preservation of phage genomes (Thurber, 2009). In Antarctic soils, organic matter and water content have shown a negative correlation with viral abundance, despite high extracellular VLP abundance (Williamson *et al*, 2007). It is hypothesized that lower temperatures may favour the preservation of extracellular virus particles, making them more abundant and detectable (Williamson, 2011). In contrast, high surface soil temperatures (maxima of 50°C) of the Namib Desert (Makhalanyane *et al*, 2013b) may explain why no detectable extracellular VLP were observed with TEM. Additionally, though not reported for desert soils, adsorption to clays has been shown to protect viruses against inactivation, thus enabling viral persistence for long periods of time (Stotzky, 1986).

Viral operational taxonomic unit (OTU) abundance estimates from metagenome data have provided some insights into the factors that shape the diversity of viral communities in desert soils (Fierer *et al*, 2007). Comparisons of viral community compositions across three contrasting soil ecosystems (prairie, desert and rainforest) have demonstrated that microbial communities were both locally and globally diverse. Comparative phylogenetic analyses showed little taxonomic overlap between soils sampled from the three different habitats, as well as low identity values to annotated sequences in public databases. However, the factors that may be responsible for the observed diversity distributions are, as yet, unknown.

Similar habitat-specific viral community compositions have been reported through the use of hierarchical clustering of metaviromes, based on dinucleotide frequencies (Willner *et al*, 2009). This method is especially useful for gaining ecological insights from metagenomic datasets containing a majority of unaffiliated reads to public databases. Dinucleotide frequencies within metaviromes have showed distinct virus community clustering within single habitat types such as desert soils. Cluster analysis of hypolith and open soil metaviromes from Antarctic and Namib Desert soil has shown that both hypolith metaviromes clustered at a single node while, conversely, both open soil metaviromes displayed an identical pattern (Zablocki *et al*, 2014b). Despite the great geographic distances or differing environmental conditions, similar habitat types harboured more closely related viral communities. The most obvious common factor between the two contrasting deserts is aridity, which may be a key driver of community speciation and recruitment in these soils.

2.5 Conclusions and future research avenues

Research on desert soil viruses is technically challenging, partly due to the physical properties of soil. Desert soils frequently produce sub-optimal viral DNA yields (≤10 ng/µl) (Kim & Bae, 2011), forcing the inclusion of a random PCR amplification step for NGS library construction. The use of whole genome amplification (WGA) by multiple displacement amplification (MDA) almost certainly results in biased amplification of certain virus groups (Kim *et al*, 2008) and prevents the accurate determination of viral abundances. While this amplification step is widely accepted as a necessity in metaviromic studies (Delwart, 2007), the development of more efficient and effective metaviromic DNA extraction technologies, so as to obviate the need for WGA, would represent a substantial advance in the field.

Alternatively, two current approaches could be used to minimize amplification-related obstacles. Firstly, WGA on low DNA metavirome samples can be carried out by random-priming, sequence independent, single primer amplification (RP-SISPA) (Weynberg *et al*, 2014). Advantages of the method are a reduced number of sequence chimeras generated during amplification and a decreased preferential amplification of ssDNA templates (Polson *et al*, 2011). However, RP-SISPA amplifies in a template-dependent manner which results in low sequence coverage of non-dominant virus sequences within a metavirome (Karlsson *et al*, 2013; Rosseel *et al*, 2013). Secondly, high-throughput sequencing of low DNA samples without prior WGA is becoming accessible (ThruPLEX, Rubicon genomics). Technical developments in Illumina paired-end sequencing library construction kits have reduced the minimum genomic DNA requirement to around 50 picograms, but this method has not yet been applied to the sequencing of viral metagenomes.

Enumeration of virus particles is a valuable approach for monitoring viral population turnover over space and time (Wommack *et al*, 2015). Soil particulates and/or extraction buffers may produce background obstructions to epifluorescence microscopy (Williamson *et al*, 2003) and prevent accurate virus counts. However, EM visualization of virus morphology remains an important criterion for viral taxonomic classification (Buechen-Osmond & Dallwitz, 1996), and the inability to distinguish virus particles under EM in some soils (e.g. desert soils (Prigent *et al*, 2005; Prestel *et al*, 2008, 2013)) remains a major technical hindrance. It is currently assumed that this limitation is due to low numbers of extracellular particles (Prestel *et al*, 2008), although this has not been empirically determined. The further development of optimised soil-specific extraction methods would provide more accurate virus counts as well as less biased analyses of viral populations.

Sequence-based identification of viral communities, using either multiple gene markers (Adriaenssens & Cowan, 2014) or full virome sequencing (Sullivan, 2015), is becoming more routine. In marine virus ecology, the use of conserved viral marker genes such as DNA polymerases (Chen *et al*, 1996), ribonucleotide reductases (Sakowski *et al*, 2014) and T4-related structural proteins (Hambly *et al*, 2001; Short & Suttle, 2005) has provided detailed data on viral biodiversity, on intra- and inter-

viral evolutionary relationships and on oceanic viral turnover rates. The use of these methods to study virus diversity and biogeography in desert soils is relatively new and most commonly involves the sequencing of whole metaviromes (Adriaenssens *et al*, 2015; Srinivasiah *et al*, 2013). However, as both sequence-based and metaviromic approaches generally result in a large number of unknown sequences (Hatfull, 2015), it can be argued that sequencing efforts should also be focused on functionally essential genes, an approach which can provide both informative data on viral richness and insights into the functional roles of viruses in soil ecosystems.

The incidence of lysogeny in hyperarid desert soils appears to be correlated with abiotic factors such as temperature and aridity, although the role of the hosts in the balance of lytic and lysogenic viruses is still poorly understood. Recently, the application of single cell genomics (Lasken, 2012) to a pool of single amplified marine genomes allowed the analysis of *in situ* microbial interactions in a culture-independent manner (Labonté *et al*, 2015), making it possible to identify infectious stages (e.g., late lytic) within bacterial genomes and predict the genetic capacity for lysogeny in uncultured viruses. This approach, applied to desert soil bacterial communities, has the potential to identify host-associated factors contributing to the abundance and diversity of lytic and lysogenic viruses.

A common feature of many desert ecosystems is the occurrence of hypolithic niches (Thomas, 2011). These rock-associated cryptic microbial communities are usually dominated by photosynthetic cyanobacteria, but contain a wide diversity of members of the phyla Actinobacteria, Acidobacteria, Bacteroidetes and Proteobacteria (Schlesinger *et al*, 2003; Warren-Rhodes *et al*, 2006, 2007; Pointing *et al*, 2009). Cyanobacteria are of particular importance, due to their key roles in primary productivity and nitrogen input in depauperate ecosystems (Tracy *et al*, 2010; Cowan *et al*, 2011). To date, no fully characterized desert soil-associated cyanophage isolates have been reported. Preliminary metagenomic data on Namib hypoliths (Adriaenssens *et al*, 2015) have shown evidence of novel soil cyanophage lineages, the sequences of which have low identities to characterized marine cyanophage genomes. As cyanobacteria are readily amenable to culturing (Garcia-Pichel *et al*, 2001), this provides opportunities for the isolation of their phages and access to full-length soil cyanophage genomes. Such data would support downstream applications such as primer design for targeted amplification of related taxa, and monitoring of these assemblages within desert soil ecosystems.

2.6 Hypotheses and aims

Aims

The goal of this study broadly consists of the sampling and analysis of metaviromics sequence data in order to characterize viral populations from soils of the Namib Desert and the Antarctic Dry Valleys. In current literature, a certain bias towards aquatic environments has been observed, and this project aims to bridge that gap by not only looking at the soil ecosystem, but specifically arid ones. We aim to bypass the host culturing step, specifically when it comes to phage enrichment through chemical lytic induction, usually performed on bacteria that manage to grow under laboratory

conditions. This will reduce the host culturing bias and allow an enhanced estimation of viral diversity. The characterization, assembly and annotation of the viral metagenomes obtained using next-generation sequencing technologies from these ecosystems/microbial communities are expected to contribute to the knowledge of viral diversity present in these soils.

As such, the main research questions are the following:

1) What is the viral diversity in hypolith communities found underneath rocks and surrounding surface soil in the Dry Valleys of Antarctica?

2) What is the viral diversity associated with surface soils of the Namib Desert, and how is community structure influenced by environmental gradients across the desert ecosystem?

Hypotheses

1) From both desert ecosystems, bacteriophages from at least one the *Caudovirales* families (i.e. *Myoviridae*, *Podoviridae* and *Siphoviridae*) will be present.

2) A large fraction of virus-specific sequences will not be matched against deposited sequences in public databases.

3) Novel, previously un-characterized viruses from both desert ecosystems will be assembled from obtained sequence data.

2.6 References

Ackermann H-W & DuBow MS (1987) Viruses of prokaryotes CRC press

Adriaenssens EM & Cowan DA (2014) Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.* **80:** 4470–4480

Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M & Cowan DA (2015) Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* **17:** 480–495 Available at: http://europepmc.org/abstract/MED/24912085

Ashelford KE, Day MJ & Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. *Appl. Environ. Microbiol.* **69:** 285–289

Austin AT, Yahdjian L, Stark JM, Belnap J, Porporato A, Norton U, Ravetta DA & Schaeffer SM (2004) Water pulses and biogeochemical cycles in arid and semiarid ecosystems. *Oecologia* **141:** 221–235

Bettarel Y, Sime-Ngando T, Amblard C & Dolan J (2004) Viral activity in two contrasting lake ecosystems. *Appl. Environ. Microbiol.* **70:** 2941–2951

Breitbart M (2012) Marine Viruses: Truth or Dare. *Ann. Rev. Mar. Sci.* **4:** 425–448

Buechen-Osmond C & Dallwitz M (1996) Towards a universal virus database—progress in the

ICTVdB. *Arch. Virol.* **141:** 392–399

Chen F, Suttle CA & Short SM (1996) Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl. Environ. Microbiol.* **62:** 2869–2874

Chen LZ, Wang GH, Hong S, Liu A, Li C & Liu YD (2009) UV-B-induced oxidative damage and protective role of exopolysaccharides in desert cyanobacterium Microcoleus vaginatus. *J. Integr. Plant Biol.* **51:** 194–200

Cowan DA, Pointing SB, Stevens MI, Craig Cary S, Stomeo F & Tuffin IM (2011) Distribution and abiotic influences on hypolithic microbial communities in an Antarctic Dry Valley. *Polar Biol.* **34:** 307–311

Davey ME & O'toole GA (2000) Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.* **64:** 847–867

Delwart EL (2007) Viral metagenomics. *Rev. Med. Virol.* **17:** 115–131

Dennehy JJ (2013) What Ecologists Can Tell Virologists. *Annu. Rev. Microbiol.*: 117–135 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24847957

Donlan RM (2002) Biofilms: Microbial life on surfaces. *Emerg. Infect. Dis.* **8:** 881–890

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards R a., Felts B, Rayhawk S, Knight R, Rohwer F & Jackson RB (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73:** 7059–7066

Frost LS, Leplae R, Summers AO & Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3:** 722–732

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

Garcia-Pichel F, López-Cortés A & Nübel U (2001) Phylogenetic and Morphological Diversity of Cyanobacteria in Soil Desert Crusts from the Colorado Plateau. *Appl. Environ. Microbiol.* **67:** 1902–1910

Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa RA & Breitbart M (2011) Development of phoH as a novel signature gene for assessing marine phage diversity. *Appl. Environ. Microbiol.* **77:** 7730–7739

Gorbushina AA (2007) Life on the rocks. *Environ. Microbiol.* **9:** 1613–1631

Hambly E, Tétart F, Desplats C, Wilson WH, Krisch HM & Mann NH (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 11411–11416

Harel Y, Ohad I & Kaplan A (2004) Activation of photosynthesis and resistance to photoinhibition in cyanobacteria within biological desert crust. *Plant Physiol.* **136:** 3070–3079

Hatfull GF (2015) Dark matter of the biosphere: The amazing world of bacteriophage diversity. *J. Virol.*: JVI.01340–15 Available at: http://jvi.asm.org/lookup/doi/10.1128/JVI.01340-15

Hurst CJ, Gerba CP & Cech I (1980) Effects of environmental variables and soil characteristics on virus survival in soil. *Appl. Environ. Microbiol.* **40:** 1067–1079

Jończyk E, Kłak M, Międzybrodzki R & Górski A (2011) The influence of external factors on bacteriophages--review. *Folia Microbiol. (Praha).* **56:** 191–200

Karlsson OE, Belák S & Granberg F (2013) The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecur. Bioterror.* **11 Suppl 1:** S227–34 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23971810

Kim KH & Bae JW (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77:** 7663–7668

Kim K-H, Chang H-W, Nam Y-D, Roh SW, Kim M-S, Sung Y, Jeon CO, Oh H-M & Bae J-W (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **74:** 5975–5985

Kimura M, Jia ZJ, Nakayama N & Asakawa S (2008) Ecology of viruses in soils: Past, present and future perspectives. *Soil Sci. Plant Nutr.* **54:** 1–32

Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Wommack K & Stepanauskas R (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.*: 1–14 Available at: http://www.nature.com/doifinder/10.1038/ismej.2015.48

Lasken RS (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10:** 631–640

Leonardopoulos J, Papaconstantinou A & Georgakopoulou-Papandreou E (1996) The meaning of soil characteristics and temperature for the survival of bacteriophages in naturally contaminated soil samples. *ACTA Microbiol. Hell.* **41:** 309–316

Makhalanyane T, Valverde A, Birkeland N-K, Cary SC, Tuffin IM & Cowan DA (2013a) Evidence for successional development in Antarctic hypolithic bacterial communities. *ISME J.* **7:** 2080–90 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23765099

Makhalanyane T, Valverde A, Lacap D, Pointing S, Tuffin M & Cowan D (2013b) Evidence of species recruitment and development of hot desert hypolithic communities. *Environ. Microbiol. Rep.* **5:** 219–224

Marsh P & Wellington EMH (1994) Phage-host interactions in soil. *FEMS Microbiol. Ecol* **15:** 99–108

Meigs P (1952) Arid and semiarid climatic types of the world. In *Proceedings, VIII General Assembly and XVII International Congress, International Geographical Union, Washington DC* pp 135–138.

Nasser AM & Oman SD (1999) Quantitative assessment of the inactivation of pathogenic and indicator viruses in natural water sources. *Water Res.* **33:** 1748–1752

Palmer TM (2003) Spatial habitat heterogeneity influences competition and coexistence in an African acacia ant guild. *Ecology* **84:** 2843–2855

Pearce DA, Newsham KK, Thorne MAS, Calvo-Bado L, Krsek M, Laskaris P, Hodson A & Wellington EM (2012) Metagenomic analysis of a southern maritime Antarctic soil. *Front. Microbiol.* **3:** 1–13

Pointing SB, Chan Y, Lacap DC, Lau MCY, Jurgens JA & Farrell RL (2009) Highly specialized microbial diversity in hyper-arid polar desert. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 19964–19969

Polson SW, Wilhelm SW & Wommack KE (2011) Unraveling the viral tapestry (from inside the capsid out). *ISME J.* **5:** 165

Prestel E, Regeard C, Andrews J, Oger P & DuBow MS (2012) A Novel Bacteriophage Morphotype with a Ribbon-like Structure at the Tail Extremity. *Res. J. Microbiol.* **7:** 75–81

Prestel E, Regeard C, Salamitou S, Neveu J & Dubow MS (2013) The bacteria and bacteriophages from a Mesquite Flats site of the Death Valley desert. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **103:** 1329–1341

Prestel E, Salamitou S & Dubow MS (2008) An examination of the bacteriophages and bacteria of the Namib desert. *J. Microbiol.* **46:** 364–372

Prigent M, Leroy M, Confalonieri F, Dutertre M & DuBow MS (2005) A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles* **9:** 289–296

Reynolds JF, Kemp PR, Ogle K & Fernández RJ (2004) Modifying the 'pulse-reserve' paradigm for deserts of North America: Precipitation pulses, soil water, and plant responses. *Oecologia* **141:** 194–210

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F & Mira A (2009) Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7:** 828–836

Rohwer F, Prangishvili D & Lindell D (2009) Roles of viruses in the environment. *Environ. Microbiol.* **11:** 2771–2774

Rosario K & Breitbart M (2011) Exploring the viral world through metagenomics. *Curr. Opin. Virol.* **1:** 289–297

Rosseel T, Van Borm S, Vandenbussche F, Hoffmann B, van den Berg T, Beer M & Höper D (2013) The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. *PLoS One* **8:**

Sakowski EG, Munsell E V, Hyatt M, Kress W, Williamson SJ, Nasko DJ, Polson SW & Wommack KE (2014) Ribonucleotide reductases reveal novel viral diversity and predict biological and

ecological features of unknown marine viruses. *Proc. Natl. Acad. Sci.* **111:** 15786–15791

Schlesinger WH, Pippen JS, Wallenstein MD, Hofmockel KS, Klepeis DM & Mahall BE (2003) Community composition and photosynthesis by photoautotrophs under quartz pebbles, southern Mojave Desert. *Ecology* **84:** 3222–3231

Schlesinger WH, Reynolds JF, Cunningham GL, Huenneke LF, Jarrell WM, Virginia RA & Whitford WG (1990) Biological feedbacks in global desertification. *Science* **247:** 1043–1048

Shmida A & Wilson M (1985) Biological Determinants of Species Diversity. *J. Biogeogr.* **12:** 1–20 Available at: http://www.jstor.org/stable/2845026

Short CM & Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* **71:** 480–486

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T & Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* **159:** 349–357

Srinivasiah S, Lovett J, Polson S, Bhavsar J, Ghosh D, Roy K, Fuhrmann JJ, Radosevich M & Wommack KE (2013) Direct assessment of viral diversity in soils by random PCR amplification of polymorphic DNA. *Appl. Environ. Microbiol.* **79:** 5450–5457

Stewart FM & Levin BR (1984) The population biology of bacterial viruses: why be temperate. *Theor. Popul. Biol.* **26:** 93–117

Stotzky G (1986) Influence of soil mineral collids on metabolic processes, growth, adhesion, and ecology of microbes and viruses In: Huang PM, Schnitzer Meds. *Interact. Soil. Miner. with Nat. Oganics Microbies Madision Soil Sci. Soc. Am.* **428:**

Straub TM, Pepper IL & Gerba CP (1992) Persistence of viruses in desert soils amended with anaerobically digested sewage sludge. *Appl. Environ. Microbiol.* **58:** 636–641

Sullivan MB (2015) Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *J. Virol.* **89:** 2459–2461

Suttle CA (2005) Viruses in the sea. *Nature* **437:** 356–361

Suttle CA (2007) Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5:** 801–812

Swanson MM, Reavy B, Makarova KS, Cock PJ, Hopkins DW, Torrance L, Koonin E V. & Taliansky M (2012) Novel bacteriophages containing a genome of another bacteriophage within their genomes. *PLoS One* **7:**

Thomas DSG (2011) Arid environments: their nature and extent. *Arid Zo. Geomorphol. Process. Form Chang. Drylands, Third Ed.:* 1–16

Thurber RV (2009) Current insights into phage biodiversity and biogeography. *Curr. Opin. Microbiol.* **12:** 582–587

Tracy CR, Streten-Joyce C, Dalton R, Nussear KE, Gibb KS & Christian KA (2010) Microclimate and limits to photosynthesis in a diverse community of hypolithic cyanobacteria in northern

Australia. *Environ. Microbiol.* **12:** 592–607

UNEP (2013) Global environment outlook 2000 Routledge

Warren-Rhodes KA, Rhodes KL, Boyle LN, Pointing SB, Chen Y, Liu S, Zhuo P & McKay CP (2007) Cyanobacterial ecology across environmental gradients and spatial scales in China's hot and cold deserts. *FEMS Microbiol. Ecol.* **61:** 470–482

Warren-Rhodes KA, Rhodes KL, Pointing SB, Ewing SA, Lacap DC, Gómez-Silva B, Amundson R, Friedmann EI & McKay CP (2006) Hypolithic cyanobacteria, dry limit of photosynthesis, and microbial ecology in the hyperarid Atacama Desert. *Microb. Ecol.* **52:** 389–398

Weinbauer MG & Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6:** 1–11

Weynberg KD, Wood-Charlson EM, Suttle C a. & van Oppen MJH (2014) Generating viral metagenomes from the coral holobiont. *Front. Microbiol.* **5:** 1–11

Williamson KE (2011) Soil phage ecology: abundance, distribution, and interactions with bacterial hosts. In *Biocommunication in Soil Microorganisms* pp 113–136. Springer

Williamson KE, Radosevich M, Smith DW & Wommack KE (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9:** 2563–2574

Williamson KE, Radosevich M & Wommack KE (2005) Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol.* **71:** 3119–3125

Williamson KE, Wommack KE & Radosevich M (2003) Sampling Natural Viral Communities from Soil for Culture-Independent Analyses. *Appl. Environ. Microbiol.* **69:** 6628–6633

Willner D, Thurber RV & Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.* **11:** 1752–1766

Winget DM, Williamson KE, Helton RR & Wommack KE (2005) Tangential flow diafiltration: An improved technique for estimation of virioplankton production. *Aquat. Microb. Ecol.* **41:** 221–232

Wommack KE & Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64:** 69–114

Wommack KE, Nasko DJ, Chopyk J & Sakowski EG (2015) Counts and sequences, observations that continue to change our understanding of viruses in nature. *J. Microbiol.* **53:** 181–192 Available at: http://link.springer.com/10.1007/s12275-015-5068-6

Yergeau E & Kowalchuk GA (2008) Responses of Antarctic soil microbial communities and associated functions to temperature and freeze-thaw cycle frequency. *Environ. Microbiol.* **10:** 2223–2235

Youle M, Haynes M & Rohwer F (2012) Scratching the surface of biology's dark matter. In *Viruses: Essential agents of life* pp 61–81. Springer

Zablocki O, van Zyl L, Adriaenssens E, Rubagotti E, Tuffin M, Cary S & Cowan D (2014a) High-level

diversity of tailed Phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. *Appl. Environ. Microbiol.* **80:** 6888–6897

Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC & Cowan D (2014b) Niche-dependent genetic diversity in Antarctic metaviromes. *Bacteriophage* **4:** e980125 Available at: http://www.tandfonline.com/doi/abs/10.4161/21597081.2014.980125

SECTION II: VIRUSES IN ANTARCTIC SOILS

Chapter 3:

High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils

3.1 Chapter summary

The metaviromes of two distinct Antarctic hyperarid desert soil communities have been characterized. Hypolithic communities, cyanobacterium-dominated assemblages situated on the ventral surfaces of quartz pebbles embedded in the desert pavement, showed higher virus diversity than surface soils, which correlated with previous bacterial community studies. Prokaryotic viruses (i.e., phages) represented the largest viral component (particularly *Mycobacterium* phages) in both habitats, with an identical hierarchical sequence abundance of families of tailed phages (*Siphoviridae* > *Myoviridae* > *Podoviridae*). No archaeal viruses were found. Unexpectedly, cyanophages were poorly represented in both metaviromes and were phylogenetically distant from currently characterized cyanophages. Putative phage genomes were assembled and showed a high level of unaffiliated genes, mostly from hypolithic viruses. Moreover, unusual gene arrangements in which eukaryotic and prokaryotic virus-derived genes were found within identical genome segments were observed. *Phycodnaviridae* and *Mimiviridae* viruses were the second-most abundant taxa and more numerous within open soil. Novel virophage-like sequences (within the Sputnik clade) were identified. These findings highlight high-level virus diversity and novel species discovery potential within Antarctic hyperarid soils and may serve as a starting point for future studies targeting specific viral groups.

3.2 Introduction

Antarctica is the coldest, driest place on Earth (Fortuin & Oerlemans, 1990). Exposed soil areas comprise approximately 0.4% of the continent's surface and are mainly located in coastal areas, particularly on the Antarctic Peninsula and in the McMurdo Dry Valleys (Bockheim & McLeod, 2008). These mineral soils are exposed to a range of "extreme" abiotic factors, including very low temperatures, high soil salinity, low water availability and nutrient levels, high levels of UV radiation, and strong, cold winds descending from glaciers or mountain tops (katabatic). Due to these conditions, the most morphologically distinct soil communities (i.e., type I, II, and III hypoliths) are associated with lithic surfaces (Cary *et al*, 2010; Wierzchos *et al*, 2012). Hypolithic communities, occurring on the ventral surfaces of translucent quartz rocks, have been shown to be mostly composed of phototrophic cyanobacterial species (Wood *et al*, 2008). These photoautotroph- dominated communities have crucial roles within the Antarctic soil ecosystem, such as primary productivity and nitrogen input (Tracy *et al*, 2010; Cowan *et al*, 2011b). While the composition of these communities

is now reasonably well understood (Smith *et al*, 2000; Pointing *et al*, 2009; Cowan *et al*, 2010, 2011a; Khan *et al*, 2011), the associated viruses, with their potential to influence microbial population dynamics and nutrient cycling via viral lysis (Laybourn-Parry, 2009), have yet to be characterized.

No comprehensive analyses of the collective viral genomic content (i.e., the metavirome) of Antarctic soils have yet been published, with the limited number of reported Antarctic viral metagenomic studies focusing on aquatic systems (Säwström *et al*, 2008; López-Bueno *et al*, 2009; Yau *et al*, 2011) and Antarctic megafauna such as seals (Kennedy *et al*, 2000) and penguins (Wallensten *et al*, 2006). Metaviromic surveys of saline meromictic lakes have shown a high level of diversity of virus-like particles (mostly phages) and several virophages (López-Bueno *et al*, 2009; Yau *et al*, 2011). To date, the few studies of viruses in Antarctic soils have focused on classical phage isolation, lytic induction experiments from culturable bacterial species and virus abundance counts (Williamson *et al*, 2007; Meiring *et al*, 2012; Swanson *et al*, 2012). Here we report a comprehensive characterization of virus diversity using a metagenomic approach in Antarctic desert soils, with a focus on the double-stranded DNA (dsDNA) virus composition of two common microhabitats: open surface soils and hypolithic communities.

3.3 Materials and methods

Sampling location. Samples were collected from the Miers Valley, Ross Dependency in eastern Antarctica (GPS coordinates, 78°05.6'S, 163°48.6'E) during the austral summer period of 2011. For the open soil sample, 1.5 kg of surface soil (0- to 2-cm depth) was collected from an approximately 1-m^2 area at a single location. The hypolith sample consisted of 0.5 kg of hypolith scrapings gathered aseptically from a collection of cyanobacterial-type hypoliths (n > 50) from an area of approximately 50 m^2. The open-soil sample was recovered from within this area. Samples were transferred and stored in sterile Whirl-Pak bags (product no. B01445WA; Nasco) at below 0°C in the field and during transport and at – 80°C in the laboratory.

Sample processing, DNA extraction, and sequencing. Processing of both types of samples was performed similar to the methods in Adriaenssens *et al.* (2014). Both the open-soil sample and pooled hypolithic samples were suspended in 3 litres of deionized water and shaken vigorously. The solids were allowed to settle, and the supernatant was decanted. The process was repeated, and both supernatants were mixed. The supernatant was centrifuged at 1,593 × *g* for 10 min (Beckman JA10 rotor), decanted, and passed through a 0.22-µm filter (Stericup [500 ml, 0.22 µm]; catalog no. SCGPU05RE; Millipore). Virus particles were collected from the filtrate by centrifugation in a Beckman JA20 rotor at 43,667 × *g* for 6 h in autoclaved 30-ml Nalgene polypropylene copolymer (PPCO) tubes (catalog no. 3119-0030). The 6 litres was spun down by discarding the supernatant from each 30-ml tube (8 tubes in a JA20 rotor) after a round of centrifugation and then adding another 30 ml of the extract to the tube. The individual pellets were resuspended in 3 ml successively: the first pellet was resuspended in 3 ml Tris-EDTA (TE) buffer, the liquid was then transferred to the next tube, the pellet was resuspended properly and then transferred to the next tube and so on until all pellets were resuspended. The pellets were treated with DNase I (catalog no. EN0521; Fermentas) and RNase

A (catalog no. EN0531; Fermentas) to a final concentration of 0.1 μg/ml at 37°C for 1 h. The presence of bacterial DNA was checked by amplifying the 16S rRNA gene (primers E9F and U1510R (Hansen *et al*, 1998; Reysenbach *et al*, 1995)) as follows: 1 μl of genomic DNA was mixed with 2.5 μl of each primer (10 mM), 2.5 μl of 2 mM deoxynucleoside triphosphates (dNTPs), 2.5 μl of 10 × DreamTaq buffer (ThermoFisher Scientific, MA, USA), 1 μl of 10- mg/ml bovine serum albumin (BSA), 0.125 μl DreamTaq polymerase (ThermoFisher Scientific, MA, USA), and Milli-Q water to a total volume of 25 μl. PCR was conducted under the following thermal regime: (i) 5 min at 95°C; (ii) 30 cycles, with 1 cycle consisting of 30 s at 95°C, 30 s at 52°C, and 85 s at 72°C; and (iii) 10 min at 72°C. The virus suspension was treated with proteinase K (Fermentas) at a final concentration of 1 μg/ml at 55°C for 2 h. Seventy microliters of SDS (20%) was added and incubated at 37°C for 1 h. Nucleic acids were purified by performing two rounds of phenol-chloroform-isoamyl alcohol (25:24:1) extraction followed by chloroform-isoamyl alcohol (24:1) phase separation. DNA was precipitated by the addition of 1/10 volume of sodium acetate (3 M; pH 5.2) and 2 volumes of 100% ethanol and left overnight at 4°C. Samples were centrifuged at 29,000 × *g* for 10 min to pellet the DNA, which was resuspended in 30 μl of TE buffer. The DNA was further cleaned using the Qiagen gel extraction kit (Qiaex II; catalog no. 20021; Qiagen). Ten nanograms of each sample was then used to perform Phi29 amplification (GenomiPhi HY DNA amplification kit; catalog no. 25-6600-20; GE Healthcare) using the manufacturer's recommendations. Library preparation included a 10% phiX V3 spike per the manufacturer's instructions with the Illumina Nextera XT library prep kit/MiSeq reagent kit V2. The amplified DNA was sequenced (2 × [forward and reverse sequencing] 250-bp reads, ~250-bp average insert size) on the Illumina MiSeq sequencer platform located at the University of the Western Cape, Cape Town, South Africa.

Sequence data analysis. Sequence reads were curated for quality control and adapter trimmed using CLC Genomics version 6.0.1 (CLC, Denmark), using the default parameters. Unpaired reads were aligned against each other using Bowtie under default parameters. *De novo* assembly for each read data set was performed with both CLC Genomics and DNASTAR Lasergene SeqMan assembler suite using the default parameters. Reads and contigs were uploaded to the MetaVir (Roux *et al*, 2011) server (http://metavir-meb.univ-bpclermont.fr/) and MG-RAST (Meyer *et al*, 2008) (http://metagenomics.anl.gov/) server for virus diversity estimations (data available from these webservers). Taxonomic composition by MetaVir was computed from a BLAST comparison with the RefSeq complete viral genome protein sequence database from NCBI (1 May 2013 release) using BLASTp with a threshold of 10^{-5} for the E-value. Assembled reads were searched for open reading frames (ORFs) and compared to the RefSeq complete viral database (through the MetaVir pipeline) and MG-RAST, which include annotations using the following databases, for functional and organism assignment: GenBank, Integrated Microbial Genomes (IMG), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pathosystems Resource Integration Center (PATRIC), RefSeq, SEED, Swiss-Prot, tremble, and eggnog. The subset of affiliated (i.e., predicted genes with a database match) contigs generated by CLC Genomics were compared to the contigs generated by LaserGene using BLASTx under standard parameters. For the assignment of functional hierarchy, COG (clusters of orthologous

groups), KEGG Orthology (KO), and NOG databases were used. Guanine-plus-cytosine (G+C) content was determined by importing .fasta files into BioEdit (Hall, 1999). The presence of tRNAs in annotated contigs was assessed with the tRNAscanSE software accessible through http://lowelab.ucsc.edu/tRNAscan-SE/ (Schattner *et al*, 2005). For prediction of phage lifestyle and host Gram stain reaction, whole-genome protein sequences of candidate phage genomes were submitted to the online version of PHACTS (http://www.phantome.org/PHACTS/) (McNair *et al*, 2012). Aligned marker genes showing sufficient homology (> 150 bp; MetaVir) against the contigs were recovered, and phylogenetic analysis was performed using MEGA5 (http://www.megasoftware.net/). Rooted dendrograms were inferred using the maximum likelihood method with a bootstrap test of 1,000 pseudoreplicates. Phylogenetic analysis for virophage sequences was performed independently from MetaVir. Metavirome virophage amino acid sequences, as well as 9 virophage major capsid protein (MCP) sequences obtained from the NCBI GenBank database were aligned with the online version of MAFFT version 7 (http://mafft.cbrc.jp/alignment/software/). Tree construction was conducted as outlined by Zhou *et al.* (2013).

Accession numbers. These sequence data have been submitted to the DDBJ/EMBL/GenBank databases (sequence read archive [SRA]) under study accession no. SRP038018 (hypolith library) and no. SRP035457 (open-soil library).

3.4 Results

Viral diversity estimations. The presence of bacterial contamination was deemed negligible in both metavirome libraries (using the 16S gene fragment), as no discernible bands of amplified products were obtained. MiSeq reads ranged from 236 to 241 bp, with an overall higher G+C content within reads obtained from the open-soil library. Sequencing metadata, assembly metrics, and BLASTp searches are summarized in Table 2. BLASTx comparison of contig data sets from both habitat libraries (generated by two separate assemblers) revealed that 99.41% (hypolith library) and 99.5% (open-soil library) of affiliated contigs were shared between the two assembled read data sets. Contigs from CLC Genomics were used for the remainder of the analysis. Aligned against each other, libraries contained 66.01% of reads that were unique to each habitat, while 33.99% were shared (a read aligned at least once). In both read data sets, bacteria were the most represented hits (80.7 to 94.5%).

Table 2. Next-generation sequencing metadata, including assembly, annotation, and diversity statistics produced by CLC Genomics and MG-RAST server.

Parameter[a]	Value for parameter	
	Open-soil library	Hypolith library
Pre-QC no. of reads	1,622,598	3,771,948
Post-QC no. of reads	1,597,524	3,729,606
Average read length (post- QC)	236.95	241.25
Mean G+C content ± SD (%)	52 ± 12	47 ± 8
No. of reads (%) not assembled into contigs	111,385 (6.97)	274,272 (7.35)
No. of contigs generated	22,237	53,695
Minimum length (nt)	200	200
Maximum length (nt)	177,571	50,044
$N_{25}/N_{50}/N_{75}$	865/446/267	945/553/354
% unknown proteins	58.5	81.3
% annotated proteins	39.2	14.5
Species richness	344.508	1,058.398

[a] QC, quality control; nt, nucleotides.

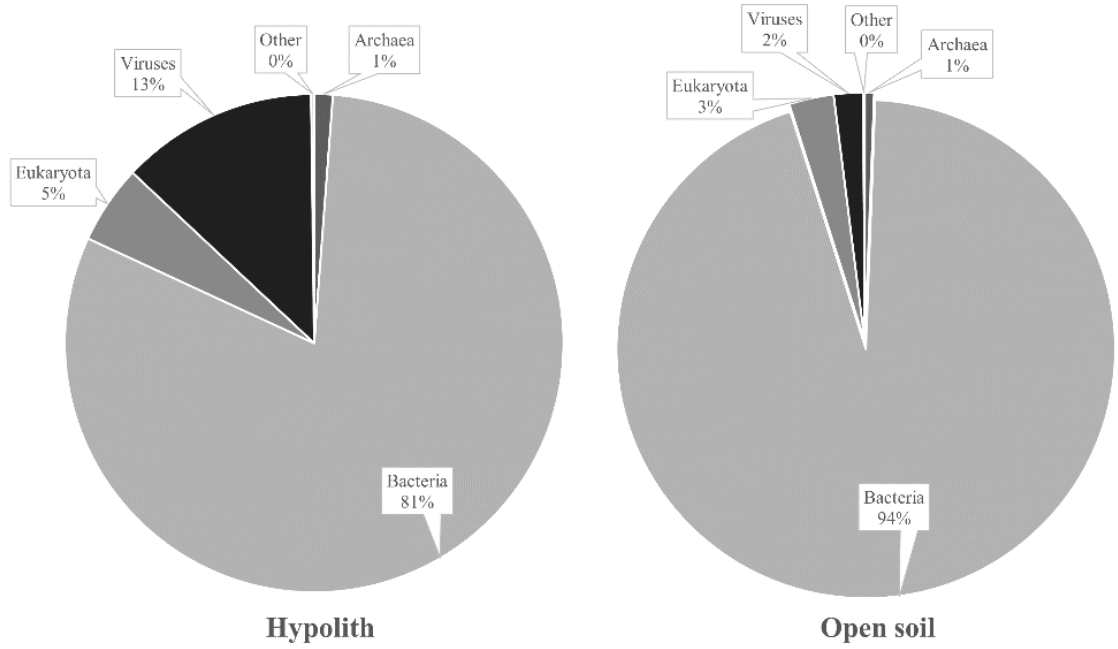

Figure 7. Taxonomic affiliation of the predicted open reading frames (ORFs) in both habitat libraries, assigned by the MG-RAST server.

However, these estimations varied depending on the metagenomic platform used and whether reads or contigs were submitted. BLASTp searches of the MetaVir server using contigs produced significantly more virus-related hits compared to searches of the MG-RAST server. For example, 1.9% of the open-soil contigs were predicted to be viral in origin in MG-RAST, while MetaVir with the same data set predicted 18.8%. The same was true for the hypolith library, where MG-RAST predicted 12.8% for viruses, while MetaVir predicted 19.2%. Archaea, Eukaryota, and "other" represented the smallest fraction, while viruses (particularly in hypolith) were second in terms of contig affiliations (Figure 7). Species richness was computed by MG-RAST using normalized values, since unequal distribution of reads between open-soil and hypolith libraries were obtained. The hypolith library was ~3-fold more diverse than the open-soil library (344.5 versus 1,058.4 species). In contrast, the open soil showed higher taxonomic abundance (species evenness, γ-diversity) compared to the hypolith (15,663 versus 11,480; Table 3). *Proteobacteria* and *Firmicutes* were the most abundant in both libraries, with viruses in hypolith the third-most-abundant organisms. Rarefaction curves generated by MG-RAST (see Figure 12) showed that the hypolith library was sampled more comprehensively compared to the open-soil library.

Nineteen virus families were identified by MetaVir (Table 4), in which prokaryotic viruses were the most abundant in both habitats (76.0% in open soil and 82.3% in hypolith). Identified phages were dominated by the order *Caudovirales* in the following abundance ranking (identical for both habitats): *Siphoviridae* > *Myoviridae* > *Podoviridae*. The next most highly represented virus families were *Mimiviridae* and *Phycodnaviridae*, both more numerous in the open-soil sample. Viral parasites of large dsDNA viruses, i.e., virophages (La Scola *et al*, 2008), were exclusively identified in the open-soil habitat. Signatures from *Adenoviridae*, *Bicaudaviridae*, *Hytrosaviridae*, *Retroviridae*, and *Rudiviridae* were found in low numbers in the hypolith habitat only. Both habitats contained 13.5 to 15.1% of sequences identified as unclassified viruses. Due to the lack of universal markers for viruses (such as the 16S rRNA gene marker used for bacteria or the 18S rRNA gene marker for eukaryotes), markers targeting virus families/species were used instead as an alternative to improve taxonomic affiliation of the annotated ORFs from both assembled reads (contigs) and reads alone. Sequences with significant homology to reference markers are shown in Table 5 in the supplemental material. The large terminase subunit (*terL*) marker, required for packaging initiation in members of the *Caudovirales* (Black, 1995), was the most common match in both habitats. This was consistent with the taxonomic affiliations of contigs in the virus families shown in Table 4. Nonbacterial viruses (such as *Paramecium bursaria chlorella* virus and *Emiliana huxleyi* virus, which belong to the family *Phycodnaviridae*, and invertebrate viruses belonging to the family *Ascoviridae*), identified with the major capsid protein (*mcp*) and DNA polymerase family B (*polB*) gene markers, were found exclusively within the open-soil community. For virophage-related sequences, 5 candidate ORFs were submitted to a tBLASTn query and showed closest similarity to the Zamilon (Gaia *et al*, 2014) and Sputnik (La Scola *et al*, 2008) virophages, both isolated from soil and aquatic environments, respectively. We attempted to determine its phylogenetic relationship, as among the very few currently recognized virophages, one has been isolated from Organic Lake, Antarctica. Among these ORFs, a

partial *mcp* sequence (344 amino acids long) was identified and aligned with other known virophage MCP sequences (Zhou *et al*, 2013). The *mcp* tree in Figure 13 (end of chapter) shows a clustering pattern identical to the tree (Zhou *et al*, 2013) and indicates that the virophage sequence from open soil (Miers Valley soil virophage [MVSV]) belonged to the Sputnik virophage group (cluster 1) and was not more closely related to its Antarctic counterpart. Its position within the tree suggests that MVSV shares a genetically distant ancestor with Sputnik and Zamilon virophages. No reads with significant homology to the *psbA* gene (a marine cyanophage photosynthesis-related gene) were identified.

Table 3. Relative abundance of the most represented phyla in both biotopes identified by MG-RAST (based on reads).

Top phylum	Relative abundance[a]		% total abundance[a]	
	OS	HY	OS	HY
Proteobacteria	9,493	3,492	60.6	30.4
Firmicutes	2,722	2,647	17.4	23
Viruses	302	1,469	1.9	12.8
Actinobacteria	521	1,039	3.3	9
Bacteroidetes	1,133	876	7.2	7.6
Cyanobacteria	173	443	1.1	3.9
Chloroflexi	60	121	0.4	1.1
Verrucomicrobia	165	110	1.1	1
Chordata	91	109	0.6	0.9
Unclassified eukaryotes	85	106	0.5	0.9
Planctomycetes	103	87	0.7	0.8

[a] OS, open soil; HY, hypolith

However, other cyanophage sequences were detected within the *g20* and *phoH* phylogenies from the hypolith data set alone (Figure 8), present as highly divergent sequences at the root of cyanophages sequence clusters. A summary of marker-identified phage species for each marker is shown in Table 6.

Functional composition of hypoliths and open soil. The hypolith data set was highly uncharacterized (predicted proteins with no significant homologs), with 81.3% compared to the open soil with 58.5%. Twenty-six functional categories were assigned to both libraries (Figure 9), each subdivided into distinct subsystems. Apart from the phage category, functional abundance in all categories was greater in open soil. Highest abundance variations between both biotopes included several metabolic pathways involving phosphorus, nitrogen, aromatic compounds, and iron

metabolism. Dormancy and sporulation-related functions were also notably higher in the open soil. A similar trend was found for stress-related functions, including oxidative, osmotic, and acid stress. However, found almost exclusively in the hypolith library were desiccation stress-related protein functions. Virus-specific functional components were retrieved manually from the MetaVir server, counted, and classified into several virus component categories, shown in Table 7. Both habitat samples contained genes encoding numerous virus structural proteins (portal, tape measure, and capsid) and enzymes (terminases, DNA/RNA polymerases, helicases, and lysins), consistent with an abundance of tailed-phage-related components.

Table 4. Taxonomic abundance of identified viral ORFs (contig BLASTp with a threshold of 10^{-5} for the E value) identified by MetaVir in both Antarctic biotopes.

		Taxonomic abundance of viral ORFs (%) (no. of sequence hits)	
Virus order and family	Host(s)	Hypolith	Open soil
Caudovirales			
Myoviridae	Bacteria, archaea	20.9 (1,305)	26.1 (415)
Podoviridae	Bacteria	9.04 (565)	11.0 (175)
Siphoviridae	Bacteria, archaea	52.6 (3,287)	38.3 (610)
Herpesvirales			
Herpesviridae	Vertebrates	0.11 (7)	0.13 (2)
Virus families not assigned into an order			
Adenoviridae	Vertebrates	0.02 (1)	0.00 (0)
Ascoviridae	Invertebrates	0.03 (2)	0.37 (6)
Asfarviridae	Swine, arthropod borne	0.03 (2)	0.06 (1)
Baculoviridae	Invertebrates	0.08 (5)	0.25 (4)
Bicaudaviridae	Archaea	0.05 (3)	0.00 (0)
Hytrosaviridae	Diptera (flies)	0.02 (1)	0.00 (0)
Inoviridae	Bacteria	0.16 (10)	0.06 (1)
Iridoviridae	Amphibians, fishes, invertebrates	0.11 (7)	0.44 (7)
Microviridae	Bacteria	0.30 (19)	0.50 (8)

Mimiviridae	Amoebae	0.88 (55)	2.32 (37)
Phycodnaviridae	Algae	1.74 (109)	4.33 (69)
Polydnaviridae	Parasitoid wasps	0.02 (1)	0.00 (0)
Poxviridae	Humans, arthropods, vertebrates	0.06 (4)	0.56 (9)
Retroviridae	Vertebrates	0.02 (1)	0.00 (0)
Rudiviridae	Thermophilic archaea	0.03 (2)	0.00 (0)
Viruses not assigned into families			
Unclassified viruses	N/A	13.5 (844)	15.1 (241)
Sputnik virophage	Mimivirus-infected amoebae	0.00 (0)	0.37 (6)

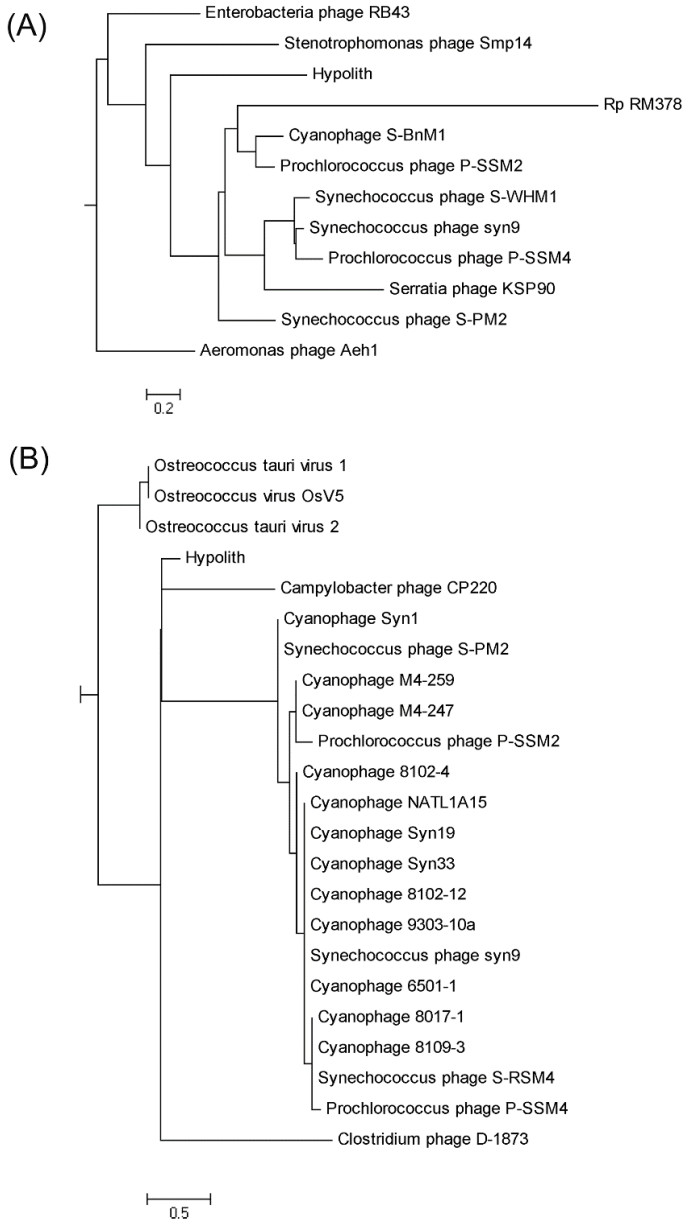

Figure 8. Selected cyanophages subtree phylogenies from *g20* (A) and *phoH* (B) marker genes based on protein alignments retrieved from the MetaVir 2.0 analysis server (metagenomic read selection and tree construction methods outlines by Roux *et al.*, 2011 . Scale bars indicate the number of base substitutions per site. Rp, *Rhodothermus* phage.

Due to the large number of assembled contigs, a subset of 26 were selected for further analysis (see Table 8) based on a combination of criteria: size (≥10,000 bp), percentage of annotated ORFs within a contig (11 to 100%), and predicted circularity of the putative genome. On average, the percentage of homologous genes from public databases was 40.2 ± 21.2 in open soil and 31.9 ± 12.5 in the hypolith. The average values for G+C content in open-soil and hypolith contigs were 55.6% ± 7.3%

and 44.3% ± 3.2%, respectively. Phage genomes were submitted to PHACTS (McNair *et al*, 2012) for lifestyle (temperate or virulent) and host Gram reaction prediction. As a general trend for both habitats, putative temperate phages dominated (61.5%), while the predicted host range was 88.5% Gram negative. These predictions are supported by a recent study (Makhalanyane *et al*, 2013), which reported that Gram-negative *Proteobacteria* were the dominant phylum in hypolithic and open-soil habitats within the McMurdo Dry Valleys.

For the open-soil habitat alone, contigs contained genes from two virus families infecting algae (*Phycodnaviridae*-like) and amoeba (*Mimiviridae*-like), positioned between phage-related genes. The largest contig (AntarOS_1 [Antar stands for Antarctic, and OS stands for open soil], 177,571 bp) contained one gene from *Acanthamoeba polyphaga* mimivirus and one from *Paramecium bursaria* chlorella virus A1, while the rest of the genes were phage related. Several core genes (Iyer *et al*, 2001) from the nucleocytoplasmic large DNA viruses (NCLDVs) were identified in the open soil (also to a lesser extent in the hypolith library) contig data set. These core genes included topoisomerase II, RNA polymerase subunit 2, guanylyltransferase, RuvC, dUTPase 2, thymidylate kinase, MutT/ NUDIX motif, and ankyrin repeat genes. A hybrid gene arrangement from different viruses was found in another contig, AntarOS_17 (Figure 10). This 24,870-bp contig was divided into 30 predicted ORFs, 21 of which showed significant homology to virus genes in the RefSeq database (detailed BLAST results for individual ORFs can be found in Table 9 in the supplemental material). Of the 30 predicted ORFs from gene 11 to gene 30 but excluding gene 18, 67% showed significant homology at the amino acid level with a single microalga-infecting virus species (unclassified *Tetraselmis viridis* virus S1, GenBank accession no. NC_020869.1). Genes 4 to 8, 10, and 18 showed similarity to several phages infecting *Burkholderia*, *Rhodobacter*, and *Azospirillum* species. The genome size was too short to be considered phycodnavirus-like (Van Etten *et al*, 2002) and possessed phage genes that were in usual functional synteny toward phage head maturation (such as *terS*, *terL*, and the capsid protein gene [within ORFs 4 to 8]). A tBLASTn search using the protein sequence of the *terL* gene against the NCBI nr database showed highest similarity (34 to 35% identity, E value of 6×10^{-76}) to various *Streptococcus* phi phages, including SsUD1, m46.1, and D12. To verify that this contig did not result from read misassembly (i.e., chimeric), two sets of primers were designed to amplify fragments from the overlapping region between ORF 10 and ORF 11, which based on the contig annotations, appeared to delineate two gene sets from different viruses. Amplicons of expected sizes were obtained and sequenced bidirectionally by Sanger technology. The sequenced fragments aligned with their respective regions, which indicated that this contig region was correctly assembled.

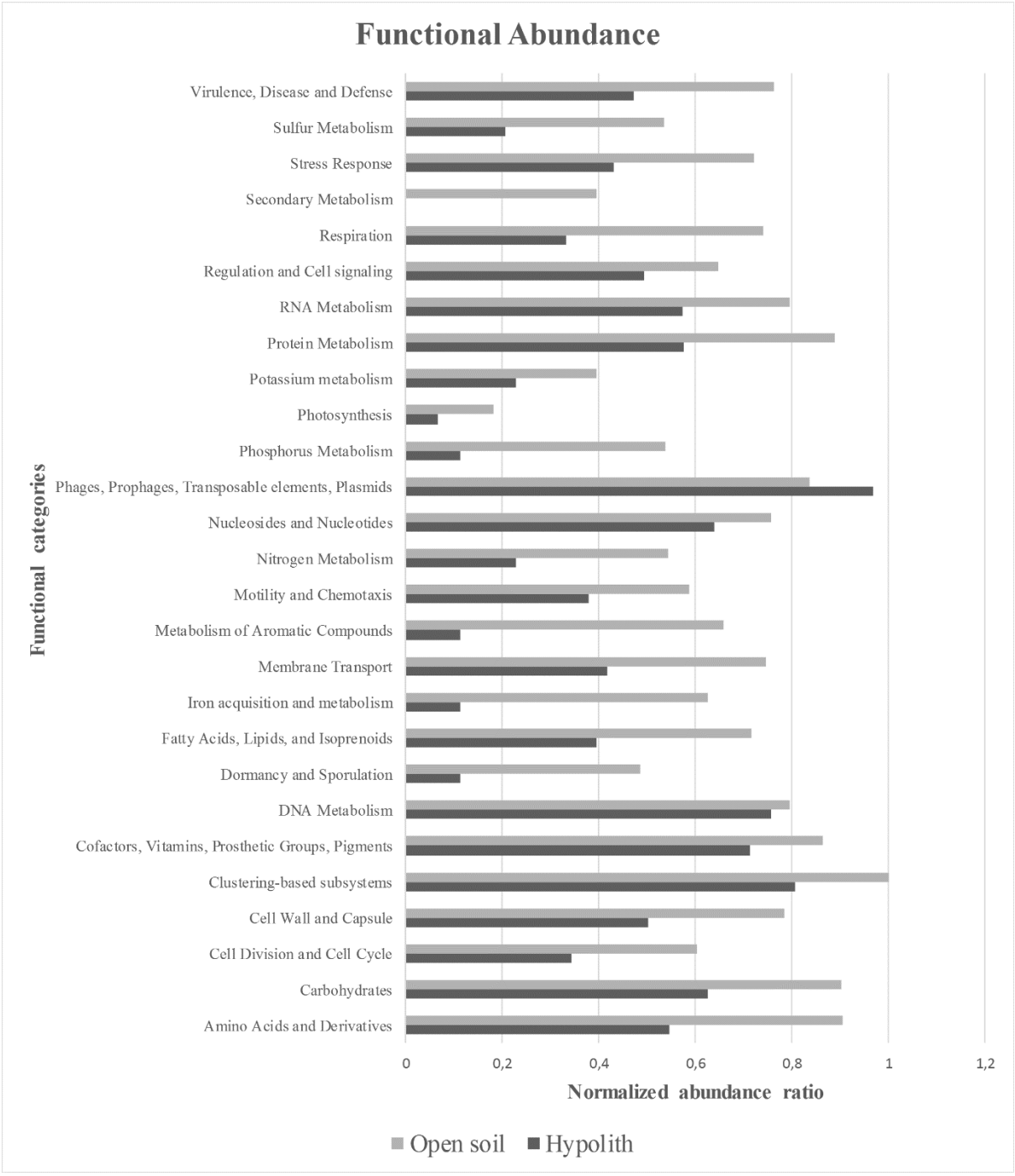

Figure 9. Functional assignment of predicted ORFs compared in both soil habitats. (Functional annotation was performed by MG-RAST using a 60% similarity cut-off).

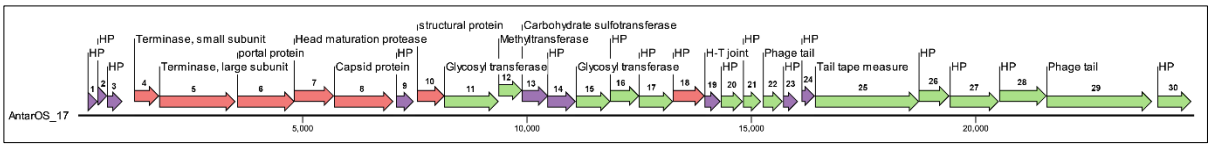

Figure 10. Predicted genome organization of phage AntarOS_17, assembled from open-soil reads. Arrows represent open reading frames (ORFs) and their orientation. ORFs without taxonomic affiliations or ORFs that were predicted without a known function (ORFan) (purple), ORFs that were more closely related to bacteriophage genes (red), and ORFs that were more closely related to *Tetraselmis viridis* virus S1 (a phycodnavirus) (green) are indicated. Numbers below the continuous black line (whole contig length) represent the nucleotide number at a given point. The numbers above each ORF arrow indicate the gene number. HP denotes a hypothetical protein.

Phage-host associations. As both habitats showed a high level of diversity of phage-related sequences, taxonomic affiliation of the reads (marker gene independent) were categorized according to host and relative sequence abundances in both habitat samples (Figure 11). Phage sequences identified most closely to host species spanning 5 bacterial phyla: *Firmicutes* (7 bacterial genera), *Proteobacteria* (8 bacterial genera), *Cyanobacteria* (3 bacterial genera), *Bacteroidetes* (*Flavobacterium*), and *Actinobacteria* (4 bacterial genera). By comparing the bacterial operational taxonomic unit (OTU) distributions in the same soil environments generated by Makhalanyane *et al.* (2013), we attempted to correlate presence/absence of bacterial OTUs based on the phage sequences obtained. Additionally, we included in our comparison 454 sequencing-based soil metagenomic data (Pearce *et al*, 2012), obtained from moraine soil collected from the margins of a permanent melt water pond located at Mars Oasis on Alexander Island, west of the Antarctic Peninsula. On the basis of identified phage species, only members of *Firmicutes* were found in both soil habitats in this study, but not in the 16S/terminal restriction fragment length polymorphism (TRFLP) bacterial data (Makhalanyane *et al*, 2013). This discrepancy between phage and bacterial data was also observed for hypoliths in hot desert soils (Adriaenssens *et al*, 2015). The other major bacterial phyla (*Proteobacteria*, *Cyanobacteria*, *Bacteroidetes*, and *Actinobacteria*) were present in both the metavirome data and 16S/TRFLP sequence data. However, bacterial genera indirectly identified by their phages from this study, were all found in the survey by Pearce *et al.* (2012).

At the level of individual phages, *Lactococcus* and *Mycobacterium* phage sequences were most common in the hypolith sample (> 10%), whereas in open soil, the largest fraction (> 6%) was composed of *Bacillus*, *Pseudomonas*, and *Mycobacterium* phage sequences. Few phage host species could be linked to the 16S/TRFLP data, but at the phylum level, *Proteobacteria* and *Actinobacteria* were present in both data sets. *Caulobacter* and *Flavobacterium* were both found in this study (identified by their phage) and 16S/TRFLP data. However, the top 10 virus isolate BLAST hits obtained by Pearce et al. (2012) were similar to that found in this study, where *Mycobacterium* phages ranked first (in the case of the hypolith sample), but also included *Pseudomonas*, *Enterobacteria*, *Flavobacterium*, and *Synechococcus* phages.

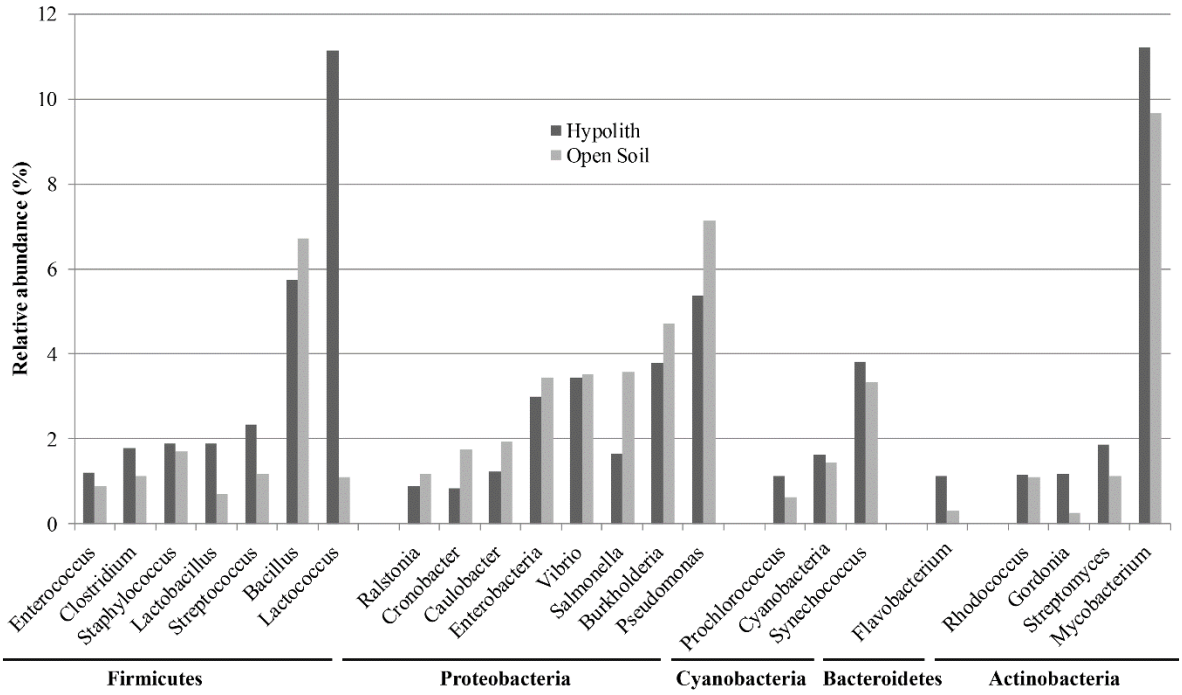

Figure 11. Relative abundance of identified reference phage genome sequences based on predicted ORFs identified by MetaVir (E-value cut-off, 10^{-5}) against the RefSeq database. Reference phage genome fragments were clustered and counted per their bacterial hosts in their respective habitats (hypolith or open soil). Black lines below a subset of phage-infecting bacterial species indicate the bacterial host phylum.

3.5 Discussion

Unlike aquatic ecosystems which have received considerable attention since the advent of viral metagenomics (Srinivasiah *et al*, 2008), virus diversity in many soil habitats has not been characterized extensively. In studies of Antarctic continental microbiology, only freshwater lake metaviromes have been reported thus far (Säwström *et al*, 2008; López-Bueno *et al*, 2009). Recent phylogenetic studies of Antarctic Dry Valley soils have shown that hypolithic communities represent the most biodiverse and complex biological assemblages in this hyperarid soil biome (Cary *et al*, 2010). Diversity estimations generated from our data (including viruses) suggest a similar pattern (3-fold-higher diversity compared to open soil). Furthermore, read libraries shared a 33% similarity overlap, indicating that hypolithic communities are distinct and differ from their surroundings. This uniqueness, coupled with its higher microbial diversity, may make hypolithic communities biodiversity "micro-hot spots" in this hyperarid desert. A large fraction of ORFs (62.5 to 84.5%) from both soil habitat samples had no significant homologs in public sequence databases, also observed in another published soil metavirome (Fierer *et al*, 2007). Rarefaction curves indicated that the open-soil biotope has not been sampled sufficiently, and therefore, a greater sequencing depth would be advisable in future metagenome experiments for this habitat.

Taxonomic/functional affiliation of predicted ORFs and gene marker analyses (e.g., using *terL*) were consistent with the conclusion that tailed bacteriophages were the primary virus component

in both soil habitats. Furthermore, an identical family-specific hierarchical abundance was observed for both habitats (*Siphoviridae* > *Myoviridae* > *Podoviridae*) but with a higher sequence diversity in hypolith communities. Compared to other virus groups, phage sequences were predictably overrepresented in both metaviromes, given that the biotic component in these soils is dominated by prokaryotes (Pointing *et al*, 2009). However, we note that our nucleic acid extraction method would exclude RNA viruses (either single or double stranded), and therefore, we do not claim that our data reflect the complete viral diversity in these soil habitats. As a sampling bias, viruses in a prophage state may constitute a large (and unsurveyed) proportion of the dsDNA phage diversity, given that it has been reported (Williamson *et al*, 2007) that many soil borne bacteria appear to contain prophages, including those from Antarctic soils. Conversely, our data suggest that ~ 61% of phage assemblages are temperate. Very few archaeal virus signatures were found in either soil habitat, consistent with previous prokaryote diversity studies (Pointing *et al*, 2009; Makhalanyane *et al*, 2013).

Unexpectedly, cyanophages were poorly represented in the hypolith sample (in terms of sequence abundance and diversity). Given the dominance of cyanobacteria in type I hypoliths (Cowan *et al*, 2010; Khan *et al*, 2011), it was reasonably predicted that cyanophages would represent a major clade. The apparent success of cyanobacteria as dominant elements of the hypolithic community might possibly be linked to the low abundance of associated viruses (where the levels of phage infection of other bacterial groups such as mycobacteria, *Bacillus*, *Flavobacterium*, and pseudomonads were higher and their host populations were under tighter predation control). However, this is in contradiction to the general understanding that the most abundant phage groups in any given environment reflect the abundance of microbial community members found in that environment (Breitbart & Rohwer, 2005). Thus, while it is possible that cyanophages genuinely represent a minor component of phage diversity, we suggest that this result is an artifact of the substantial underrepresentation of soil-associated cyanophage genome sequences in public databases, further accentuated by the fact that most characterized cyanophages are of aquatic origin (Hurst, 2011). To our knowledge, only a small fraction of cyanophages of soil origin have been described thus far (Wang *et al*, 2011), which also reported a high phylogenetic distance from "common" marine cyanophages, emphasizing the fact that little is known about these cyanophages. In support of this, cyanophages communities in paddy field soils have been shown to be different from those in freshwater, marine water, and even paddy floodwater, identifying unique *g20* subclusters specific to soil-derived cyanophages (Wang *et al*, 2011). In the current study, marker gene analysis successfully identified several metavirome sequences at the root of cyanophage clusters, suggesting that these represent novel phage phylotypes with a high genetic distance from currently characterized cyanophages. The high abundance of phages infecting certain bacterial genera such as *Mycobacterium*, *Lactococcus*, *Bacillus*, and *Pseudomonas* phages (~6 to 10% of all identified phage sequences) in both Antarctic desert soil habitats has been reported (Pearce *et al*, 2012).

Sequences with close homology to large dsDNA eukaryotic virus families such as *Mimiviridae* and *Phycodnaviridae*-like genomic elements were found as the second largest virus component (0.88% to 4.33%) in both habitats (excluding the unclassifiable virus fraction [13.5% to 15.1%]). Mimivirus-

related sequences were unexpected, as a 0.22-µm filter size should have excluded large virus particles (~0.7 µm (Claverie *et al*, 2009)), as well as repeated centrifugation steps. Phycodnaviruses, at ~0.16 ± 0.06 µm (Dunigan *et al*, 2006), would be expected to be recovered in the filtrate. However, detection of *mcp* components from a novel Sputnik-like virophage, a parasite of large dsDNA viruses (La Scola *et al*, 2008), provided further indirect evidence for the presence of mimivirus-like populations in the open-soil habitat. In addition, the identified virophage sequence was more closely related to geographically distant isolates (France and Tunisia) compared to the other virophage isolate from Organic Lake in Antarctica. A recent study (Pearce *et al*, 2012) showed sequences belonging to both host genera (*Paramecium*, *Chlorella*, and *Acanthamoeba*) and their associated viruses (chlorovirus and mimivirus) in moraine Antarctic soil. La Scola *et al.* (2010) first demonstrated the presence of mimiviruses in soil (previously only isolated from aquatic habitats). The present metavirome sequences, combined with pyrosequencing data of metagenomic libraries from Pearce *et al.* (2012), provide additional evidence for the presence of mimivirus-like genome elements in Antarctic soils. Further sampling to isolate virophages from Antarctic soils would provide further understanding into the ecology and function of these infectious agents, given that their contributions into the regulation of viral populations are starting to become apparent in other habitats (Gaia *et al*, 2014). The unusual gene configuration observed within contig AntarOS_17 (where phage and eukaryotic viruses were predicted) was confirmed by PCR on the original DNA sample, therefore ruling out misassembly of the reads for this region. Most likely, this was caused by misannotation of the predicted ORFs, caused by a lack of closer homologs in databases. While read misassembly is still a possibility in other generated contigs, confidence level in assembly accuracy was high, as the Illumina sequencing control used in both runs was phiX174 (a ~ 5,000-bp single-stranded DNA [ssDNA] virus) which was reassembled almost completely (99.7%) and correctly annotated by the MetaVir pipeline.

Virus families representing less than 0.5% of sequence abundance (Table 4) included those infecting infected Diptera, arthropods, and other invertebrates and were mostly found in the open soil habitat. As these hosts have been shown to occur on the Antarctic peninsula (Convey *et al*, 2008; Treonis *et al*, 1999), this may represent an additional pool of uncharacterized viruses within the Antarctic invertebrate fauna.

A positive correlation between phage genera from this study and their associated hosts identified in other bacterial diversity studies was established (Makhalanyane *et al*, 2013; Pearce *et al*, 2012; Friedmann & Thistle, 1993; Smith *et al*, 2006). As in previous hypolith/ open-soil community diversity (α- diversity) comparisons (Makhalanyane *et al*, 2013), where hypoliths showed a higher degree of diversity than open soil, the same was demonstrated to be true for their associated viruses.

This study represents an initial broad survey of virus diversity in Antarctic hyperarid desert soils and has demonstrated that these local virus assemblages are highly diverse and largely uncharacterized. Due to a huge gap in terms of homologous sequences in databases at this time, the generation of additional metagenomic sequence data is not likely to yield usable information. This emphasizes the need for more "traditional" studies, performed in parallel on identical sample sources. These include morphological data from microscopy, lytic induction (e.g., mitomycin C) upon raw soil,

and Sanger sequencing of clones targeting specific virus families. Unfortunately, a large fraction will most likely remain uncharacterizable *in vitro*, as the majority of their hosts (bacteria in particular) remain unculturable. Larger eukaryotic viruses infecting algae, amoebae, and invertebrates have not previously been characterized in this environment, and our data demonstrate that these viruses represent an unknown virus population that awaits characterization. Such data would further advance our understanding of the trophic structure and function of communities inhabiting this cold, hyperarid desert biome.

3.6 Additional figures and tables (included here for purposes of text continuity)

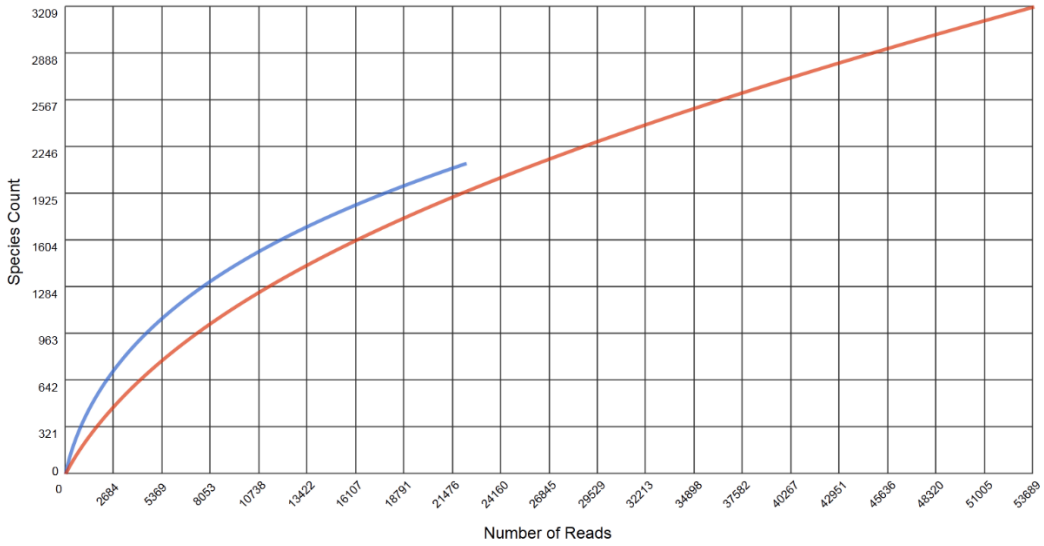

Figure 12. Rarefaction curves generated from both Antarctic read libraries through the MG-RAST pipeline (maximum E-value of 10^{-5} and minimum identity of 60 %) encompassing total abundance (i.e. not limited to viruses). Blue curve: open soil; orange curve: hypolith.

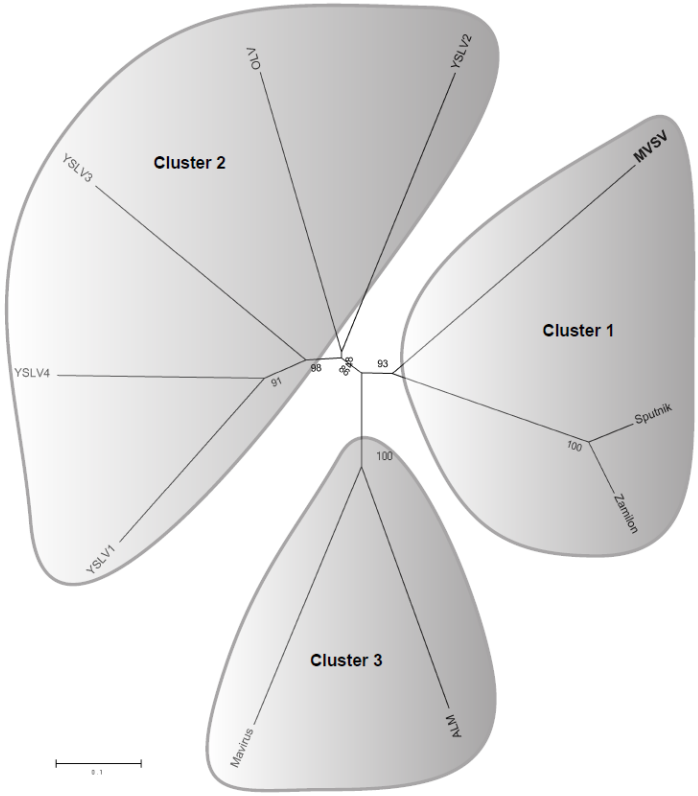

Figure 13. Phylogenetic analysis of the major capsid protein (*mcp*) showing the three current virophage groups (clusters 1-3). The open soil metavirome sequence, MVSV, is indicated in bold within cluster 1. Numbers present at tree nodes indicate bootstrap confidence values. ALM: Ace Lake Mavirus; YSLV: Yellowstone lake virophage; OLV: Organic Lake; MVSV: Miers Valley soil virophage.

Table 5. Marker gene matches against reads (>150bp homology) in both soil type metaviromes.

Marker gene	**Viral target**	**No. of sequences found**	
		Open soil	**Hypolith**
g20	T4-like phages	1	1
gp23	T4-like phages	4	0
mcp	Large eukaryotic dsDNA viruses	2	0
phoH	dsDNA phages/ *Phycodnaviridae*	5	2
polB	dsDNA viruses	3	3
polB2	*Adenoviridae*, Salterproviruses, *Caudovirales* and *Ampullaviridae*	1	1
terL	*Caudovirales*	164	138
vp1	*Microviridae*	11	55

Table 6. Phage species most closely related to the metavirome reads by phylogenetic analysis of the *terL*, *polB2*, *phoH*, *g20* marker genes. VP1 species were not included as it was probably residual contamination of the phiX V3 used during sequencing.

Marker gene	Closest phage isolates	Hypolith	Open soil
terL	*Aeromonas* phage phi018P		X
	Burkholderia phage Bcep 6B	X	X
	Burkholderia phage Bcep Mu	X	
	Clostridium phage 39-O		X
	Flavobacterium phage 11b	X	
	Haemophilus phage Aaphi23		X
	Haemophilus phage HP1	X	
	Iodobacterium phage phiPLPE	X	
	Listeria phage 006		X
	Listeria phage B054	X	X
	Methanobacterium phage psi M1	X	
	Methanothermobacter page psiM100	X	
	Mycobacterium phage Barnyard		X
	Mycobacterium phage chah	X	
	Mycobacterium phage orion	X	
	Mycobacterium phage PG1	X	
	Mycobacterim phage TM4	X	X
	Natrialba phage PhiCh1		X
	Pseudomonas phage 116	X	
	Pseudomonas phage B3	X	
	Pseudomonas phage LUZ24	X	
	Pseudomonas phage phiCTX		X
	Roseobacter phage SI01	X	
	Staphylococcus phage G1	X	X
	Staphylococcus phage K		X
	Staphylococcus phage X88		X
	Streptococcus phage 858	X	
	Streptococcus phage MM1	X	X

	Streptomyces phage mu 1/6	X	X
	Temperate phage phiN1H 1.1		X
	Sulfitobacter phage EE36 phi1	X	X
	Xanthomonas phage Xp15	X	
g20	Cyanophage cluster[1]	X	
	Rhodothermus phage RM378		X
gp23	*Rhodothermus* phage RM378		X
mcp	*Paramecium bursaria* chlorella virus		X
	Ascoviridae virus cluster[1]		X
phoH	*Campylobacter* phage CP220	X	
	Staphylococcus phage G1/Twort	X	
polB	*Emiliana huxleyi* virus 86		X
polB2	Actinomyces phage AV-1	X	
	Lactococcus phage asccphi28		X

[1]Reads could not be associated with a specific phage but was in an ancestry state of a sequence cluster.

Table 7. Major viral components found in the hypolith and open soil metavirome libraries.

Main functional class	**Phage component**	**Hypolith matches**		**Open Soil matches**	
		Count[1]	**%[2]**	**Count[1]**	**%[2]**
Phage structural components/ packaging	Portal	199	0.37	48	0.34
	Baseplate	20	0.037	11	0.078
	Tape measure	160	0.30	25	0.17
	Tail sheath	19	0.036	10	0.072
	Major capsid/head/structural	99	0.188	18	0.13
	Minor capsid/head/structural	7	0.013	4	0.03
	Tail-associated	189	0.36	50	0.357
	Head-tail connector	24	0.05	9	0.064
	Terminase	407	0.77	76	0.54
Host integration/restriction	Recombinase	2	0.004	4	0.03

	Integrase	65	0.123	38	0.272
	Resolvase	15	0.03	13	0.093
	Transposase	35	0.066	40	0.286
	Exonuclease	25	0.05	14	0.100
	Endonuclease	119	0.23	15	0.11
	DNA helicase	268	0.51	16	0.115
	RNA helicase	11	0.021	2	0.0143
	DNA polymerase (A)	155	0.294	22	0.15
	DNA polymerase (B)	62	0.12	22	0.15
	RNA polymerase	4	0.0075	6	0.043
	Reverse transcriptase	2	0.004	3	0.021
	DNA Ligase	25	0.05	16	0.115
	DNA Primase	25	0.05	11	0.0787
	RNA Ligase	6	0.011	3	0.0215
	DNA Methylase	454	0.863	97	0.69
	RNase H/T	12	0.023	9	0.064
	RNA Methylase	0	0	1	0.0071
Host lysis/maturation	Peptidase	215	0.41	93	0.665
	Amidase	74	0.141	11	0.0787
Uncharacterized	ORFans[3]	44 443	84.5	8 733	62.5
	Semi-ORFans class I[4]	958	1.82	291	2.01
	Semi-ORFans class II[5]	785	1.49	440	3.15

[1]total amount of sequences in the dataset bearing homology

[2]percentage of sequences in the whole dataset

[3]predicted ORFs with no significant homology in any database

[4]predicted ORFs with a bit score but no protein homologs in pfam

[5]predicted ORF with no bitscore but a protein homolog of unknown function

Table 8. Predicted genomic characteristics for a selected subset of annotated contigs from open soil and hypolith habitats.

Library	Contig name	Length (bp)	No. of predicted ORFs	% of affiliated ORFs	G+C %	tRNAs found	Predicted lifestyle	Predicted host Gram stain
Open Soil	AntarOS_1	177 571	244	27.86	53.15	19	Lytic	(+)
	AntarOS_3	57 271	94	21.27	65.33	0	Temperate	(-)
	AntarOS_9	48 915	72	29.16	55.52	1	Temperate	(+)
	AntarOS_8	42 417	62	37.09	65.39	0	Lytic	(-)
	AntarOS_4	40 059	57	19.29	44.31	0	Lytic	(-)
	AntarOS_20	39 073	57	42.12	61.03	0	Temperate	(-)
	AntarOS_80	27 419	40	72.5	62.58	0	Lytic	(-)
	AntarOS_17	24 870	30	76.6	54.12	0	Undetermined	Undetermined
	AntarOS_617	16 838	24	58.33	47.41	0	Temperate	(-)
	AntarOS_7	36 701	52	23.1	55.80	0	Temperate	(-)
	AntarOS_544	19 355	28	25	43.63	0	Lytic	(-)
	AntarOS_146	13 230	24	33.3	58.36	0	Temperate	(-)
	AntarOS_125	11 794	18	50	54.07	0	Temperate	(-)
	AntarOS_55	11 341	18	17	59.99	0	Lytic	(-)
	AntarOS_548	11 305	14	21.42	52.61	0	Lytic	(-)
	AntarOS_504	10 845	16	37.5	63.72	0	Temperate	(-)

	AntarOS_380	10 705	8	87.5	61.78	0	Temperate	(-)
	AntarOS_6	10 599	9	44.44	43.30	0	Temperate	(-)
Hypolith	AntarHYP_182	50 044	73	28.76	44.97	1	Lytic	(-)
	AntarHYP_58	42 093	55	54.55	40.62	0	Temperate	(+)
	AntarHYP_47	28 660	27	25.93	45.68	4	Temperate	(-)
	AntarHYP_284	17 600	27	11.11	41.63	0	Temperate	(-)
	AntarHYP_845	17 085	31	35.48	45.38	0	Lytic	(-)
	AntarHYP_323	16 563	35	31.43	43.36	0	Temperate	(-)
	AntarHYP_219_4	16 377	27	40.74	42.03	1	Temperate	(-)
	AntarHYP_170_2	13 162	11	27.3	50.73	0	Lytic	(-)

Table 9. BLAST search results against the RefSeq database (generated by MetaVir) for each predicted ORF identified in contig AntarOS_17. All predicted ORFs within this table were in a forward orientation. Please refer to Figure 10 for the contig map associated with this data.

ORF #	Length (in nucleotides)	Best BLAST hit accession	Predicted taxonomy	Predicted function	PFAM affiliation
1	224	-	-	Hypothetical protein	-
2	212	-	-	Hypothetical protein	-
3	344	-	-	Hypothetical protein	-
4	560	NP_543087.1	*Enterobacteria* phage phiP27	Terminase, small subunit	PF05119.7
5	1703	YP_003090178.1	*Burkholderia* phage KS9	Terminase, large subunit	PF03354.10
6	1283	NP_945033.1	Burkholderia phage phi1026b	Phage portal	PF04860.7
7	896	YP_006585.1	*Klebsiella* phage phiKO2	Peptidase S49	PF01343.13
8	1325	YP_001686874.1	*Azospirillum* phage Cd	Phage capsid	PF05065.8
9	395	-	-	Hypothetical protein	-
10	611	YP_002284344.1	*Pseudomonas* phage PAJU2	Hypothetical protein	-
11	1217	YP_007675761.1	*Tetraselmis viridis* virus S1	Domain of unknown function	PF08759.6
12	527	YP_007675759.1	*Tetraselmis viridis* virus S1	Methyltransferase domain	PF13489.1
13	581	-	-	Sulfotransferase	PF03567.9
14	632	-	-	Hypothetical protein	-
15	767	YP_007675758.1	*Tetraselmis viridis* virus S1	Glycosyl transferase family 2	PF00535.21
16	644	YP_007675756.1	*Tetraselmis viridis* virus S1	Hypothetical protein	-
17	758	YP_007675755.1	*Tetraselmis viridis* virus S1	-	PB001172
18	710	YP_007518392.1	*Rhodobacter* phage RcapNL	-	PB001622

19	356	-	-	Head-tail joining protein	PF05521.6
20	488	YP_007675753.1	*Tetraselmis viridis* virus S1	Domain of unknown function	PF04883.7
21	401	YP_007675752.1	*Tetraselmis viridis* virus S1	Domain of unknown function	PF11367.3
22	440	YP_007675751.1	*Tetraselmis viridis* virus S1	Major tail protein 2	PF06199.6
23	329	-	-	Hypothetical protein	-
24	296	-	-	Hypothetical protein	-
25	2321	YP_007675748.1	*Tetraselmis viridis* virus S1	Hypothetical protein	-
26	683	YP_007675747.1	*Tetraselmis viridis* virus S1	Domain of unknown function	PF09343.5
27	1088	YP_007675746.1	*Tetraselmis viridis* virus S1	Hypothetical protein	-
28	1055	YP_007675745.1	*Tetraselmis viridis* virus S1	Domain of unknown function	PF09931.4
29	2351	YP_007675743.1	*Tetraselmis viridis* virus S1	Tail protein	PF13550.1
30	761	YP_007675742.1	*Tetraselmis viridis* virus S1	Protein of unknown function	PF10983.3

3. 6 References

Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M & Cowan DA (2015) Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* **17:** 480–495 Available at: http://europepmc.org/abstract/MED/24912085

Black LW (1995) DNA packaging and cutting by phage terminases: control in phage T4 by a synaptic mechanism. *Bioessays* **17:** 1025–1030

Bockheim JG & McLeod M (2008) Soil distribution in the McMurdo Dry Valleys, Antarctica. *Geoderma* **144:** 43–49

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

Cary SC, McDonald IR, Barrett JE & Cowan D a (2010) On the rocks: the microbiology of Antarctic

Dry Valley soils. *Nat. Rev. Microbiol.* **8:** 129–138 Available at: http://dx.doi.org/10.1038/nrmicro2281

Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H & Abergel C (2009) Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. *J. Invertebr. Pathol.* **101:** 172–180

Convey P, Gibson JAE, Hillenbrand CD, Hodgson DA, Pugh PJA, Smellie JL & Stevens MI (2008) Antarctic terrestrial life - Challenging the history of the frozen continent? *Biol. Rev.* **83:** 103–117

Cowan DA, Khan N, Pointing SB & Cary SC (2010) Diverse hypolithic refuge communities in the McMurdo Dry Valleys. *Antarct. Sci.* **22:** 714–720

Cowan DA, Pointing SB, Stevens MI, Craig Cary S, Stomeo F & Tuffin IM (2011a) Distribution and abiotic influences on hypolithic microbial communities in an Antarctic Dry Valley. *Polar Biol.* **34:** 307–311

Cowan DA, Sohm JA, Makhalanyane TP, Capone DG, Green TG a., Cary SC & Tuffin IM (2011b) Hypolithic communities: important nitrogen sources in Antarctic desert soils. *Environ. Microbiol. Rep.* **3:** 581–586 Available at: http://doi.wiley.com/10.1111/j.1758-2229.2011.00266.x

Dunigan DD, Fitzgerald LA & Van Etten JL (2006) Phycodnaviruses: A peek at genetic diversity. *Virus Res.* **117:** 119–132

Van Etten JL, Graves M V., Müller DG, Boland W & Delaroque N (2002) Phycodnaviridae - Large DNA algal viruses. *Arch. Virol.* **147:** 1479–1516

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards R a., Felts B, Rayhawk S, Knight R, Rohwer F & Jackson RB (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73:** 7059–7066

Fortuin JPF & Oerlemans J (1990) Parameterization of the annual surface temperature and mass balance of Antarctica. *Ann. Glaciol.* **14:** 78–84

Friedmann EI & Thistle AB (1993) Antarctic microbiology.

Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D & La Scola B (2014) Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS One* **9:** 1–8

Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41:** 95–98 Available at: http://jwbrown.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf

Hansen MC, Tolker-Nielsen T, Givskov M & Molin S (1998) Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiol. Ecol.* **26:** 141–149

Hurst CJ (2011) Studies in Viral Ecology: Microbial and botanical host systems John Wiley & Sons

Iyer LM, Aravind L & Koonin E V (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **75:** 11720–11734

Kennedy S, Kuiken T, Jepson PD, Deaville R, Forsyth M, Barrett T, Van De Bildt MWG, Osterhaus ADME, Eybatov T, Duck C, Kydyrmanov A, Mitrofanov I & Wilson S (2000) Mass die-off of Caspian seals caused by canine distemper virus. *Emerg. Infect. Dis.* **6:** 637–639

Khan N, Tuffin M, Stafford W, Cary C, Lacap DC, Pointing SB & Cowan D (2011) Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. *Polar Biol.* **34:** 1657–1668

Laybourn-Parry J (2009) Microbiology. No place too cold. *Science* **324:** 1521–1522

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A & Alcamí A (2009) High diversity of the viral community from an Antarctic lake. *Science* **326:** 858–861

Makhalanyane T, Valverde A, Birkeland N-K, Cary SC, Tuffin IM & Cowan DA (2013) Evidence for successional development in Antarctic hypolithic bacterial communities. *ISME J.* **7:** 2080–90 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23765099

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

Meiring TL, Marla Tuffin I, Cary C & Cowan D a. (2012) Genome sequence of temperate bacteriophage Psymv2 from Antarctic Dry Valley soil isolate Psychrobacter sp. MV2. *Extremophiles* **16:** 715–726

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R & Wilke A (2008) The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386

Pearce DA, Newsham KK, Thorne MAS, Calvo-Bado L, Krsek M, Laskaris P, Hodson A & Wellington EM (2012) Metagenomic analysis of a southern maritime Antarctic soil. *Front. Microbiol.* **3:** 1–13

Pointing SB, Chan Y, Lacap DC, Lau MCY, Jurgens JA & Farrell RL (2009) Highly specialized microbial diversity in hyper-arid polar desert. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 19964–19969

Reysenbach A, Pace N, Robb F & Place A (1995) Archaea: a laboratory manual—thermophiles Cold Spring Harbour Laboratory Press

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D & Enault F (2011) Metavir: A web server dedicated to virome analysis. *Bioinformatics* **27:** 3074–3075

Säwström C, Lisle J, Anesio AM, Priscu JC & Laybourn-Parry J (2008) Bacteriophage in polar inland waters. *Extremophiles* **12:** 167–175

Schattner P, Brooks AN & Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33:**

La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M,

Forterre P, Koonin E & Raoult D (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* **455:** 100–104

Smith JJ, Tow LA, Stafford W, Cary C & Cowan DA (2006) Bacterial diversity in three different antarctic cold desert mineral soils. *Microb. Ecol.* **51:** 413–421

Smith MC, Bowman JP, Scott FJ & Line MA (2000) Sublithic bacteria associated with Antarctic quartz stones. *Antarct. Sci.* **12:**

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T & Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* **159:** 349–357

Swanson MM, Reavy B, Makarova KS, Cock PJ, Hopkins DW, Torrance L, Koonin E V. & Taliansky M (2012) Novel bacteriophages containing a genome of another bacteriophage within their genomes. *PLoS One* **7:**

Tracy CR, Streten-Joyce C, Dalton R, Nussear KE, Gibb KS & Christian KA (2010) Microclimate and limits to photosynthesis in a diverse community of hypolithic cyanobacteria in northern Australia. *Environ. Microbiol.* **12:** 592–607

Treonis AM, Wall DH & Virginia RA (1999) Invertebrate biodiversity in Antarctic Dry Valley soils and sediments. *Ecosystems* **2:** 482–492

Wallensten A, Munster VJ, Osterhaus ADME, Waldenström J, Bonnedahl J, Broman T, Fouchier RAM & Olsen B (2006) Mounting evidence for the presence of influenza A virus in the avifauna of the Antarctic region. *Antarct. Sci.* **18:**

Wang G, Asakawa S & Kimura M (2011) Spatial and temporal changes of cyanophage communities in paddy field soils as revealed by the capsid assembly protein gene g20. *FEMS Microbiol. Ecol.* **76:** 352–359

Wierzchos J, de los Ríos A & Ascaso C (2012) Microorganisms in desert rocks: The edge of life on Earth. *Int. Microbiol.* **15:** 173–183

Williamson KE, Radosevich M, Smith DW & Wommack KE (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9:** 2563–2574

Wood SA, Rueckert A, Cowan DA & Cary SC (2008) Sources of edaphic cyanobacterial diversity in the Dry Valleys of Eastern Antarctica. *ISME J.* **2:** 308–320

Yau S, Lauro FM, DeMaere MZ, Brown M V, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson J a & Cavicchioli R (2011) Virophage control of antarctic algal host-virus dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **108:** 6163–6168

Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, Pan Y & Wang Y (2013) Diversity of virophages in metagenomic data sets. *J. Virol.* **87:** 4225–36 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3624350&tool=pmcentrez&rendertype=abstract

Chapter 4:
Niche-dependent genetic diversity in Antarctic metaviromes

4.1 Chapter summary

The metaviromes from two different Antarctic terrestrial soil niches have been analysed. Both hypoliths (microbial assemblages beneath transluscent rocks) and surrounding open soils showed a high level diversity of tailed phages, viruses of algae and amoeba, and virophage sequences. Comparisons of other global metaviromes with the Antarctic libraries showed a niche-dependent clustering pattern, unrelated to the geographical origin of a given metavirome. Within the Antarctic open soil metavirome, a putative circularly permuted, ~42kb dsDNA virus genome was annotated, showing features of a temperate phage possessing a variety of conserved protein domains with no significant taxonomic affiliations in current databases.

4.2 Introduction

The hyperarid soils of the Antarctic Dry valleys were for long thought to harbour low numbers of microorganisms (Wynn-Williams, 1996). However, molecular tools such as 16S rRNA analysis have provided a more realistic view of the true microbial diversity within this polar desert ecosystem (Pointing *et al*, 2009; Makhalanyane *et al*, 2013). Cyanobacterial communities, in particular, have been attributed key roles within this ecosystem (Cowan *et al*, 2011), and are commonly found on the ventral surface of translucent rocks, termed hypoliths (Wood *et al*, 2008; Smith *et al*, 2000). While these have received much attention and have now been well characterized in terms of taxonomic diversity(Khan *et al*, 2011), they do not necessarily form the basal tier of the food chain, as associated bacteriophages may be involved in the regulation, survival and evolution of these communities. In Chapter 3, the viral component of cyanobacterial-dominated bacterial communities associated with quartz rocks (i.e. Type I hypoliths) and the surrounding open soils were characterized using a shotgun metagenomic approach. A high diversity of viruses was found in both habitats, while cyanophage marker genes were poorly represented. In this chapter, a global comparison of the hypolith and open soil metaviromes is presented, using additional publicly available data from a range of habitats, including a hot hyperarid biome, the Namib Desert. We also present a complete, circular, dsDNA temperate phage genome.

4.3 Globally related, niche-specific microbial communities

The hypolith metavirome showed greater viral diversity than the open soil metavirome, whereas the latter contained greater sequence abundance and lower sequence diversity. However, rarefaction curves suggested that sequence diversity may be under-estimated. Reads from both habitat libraries shared a 33% sequence identity overlap, indicating very genetically distinct communities despite their close habitat proximity.

A BLASTp-based comparison (10^{-5} threshold for the E-value) between hypolith and open soil metaviromes has already been discussed in detail in Chapter 3. However, our dataset contained a large

number of unknown/unaffiliated sequences (58.5% and 81.3 % for open soil and hypolith samples, respectively). For such datasets, whole metavirome nucleotide frequency (di-, tri- and tetranucleotide) comparisons can be a valuable tool for assigning putative ecological classifications, without the requirement of homology against reference databases (Willner *et al*, 2009). Contig datasets from several metaviromes from a range of habitats (freshwater (Roux *et al*, 2012), seawater (Hurwitz & Sullivan, 2013), plant-associated (Coetzee *et al*, 2010), Namib Desert open soil (unpublished data) and hypoliths (Adriaenssens *et al*, 2015), available from the MetaVir server (Roux *et al*, 2014)) were selected for dinucleotide frequency comparisons (Willner *et al*, 2009) with the Antarctic sequence datasets.

Figure 14 shows that the 9 metaviromes clustered in two separate groups. Group 1, composed of plant and water metaviromes, were further sub-divided into freshwater and seawater clades. Group 2 was composed of soil-associated habitats, sub-divided into open soil and soil-associated rock (i.e. hypolith) clades. Interestingly, despite their widely differing habitat-associated environmental characteristics and equally great spatial separation (i.e., Antarctic and Namib Desert), both hypolithic viromes clustered at a single node. The same was true for the open soil metaviromes, with the Antarctic dataset clustered with the Namib Desert dataset. The implications of this result are that the widely differing environmental temperatures do not define the genetic relatedness of the communities, but that aridity may be a strong driver of host and viral diversity.

In both habitats, the majority of predicted virus genes with assigned taxonomy belonged to the order *Caudovirales*. This was expected, as both habitats have been previously shown to be dominated by typical soil prokaryotic taxa. However, several virus families displayed variation in sequence abundance according to niche. In the open soil sample, invertebrate viruses (*Ascoviridae*, *Baculoviridae*, *Iridoviridae*) and large dsDNA viruses (*Mimiviridae*, *Phycodnaviridae*) showed higher abundance than in hypoliths. These results suggest that these viral hosts are less established in hypolithic habitats. Sequences closely related to virophages were found exclusively in the open soil sample. To date, all but one virophage have been isolated from aquatic habitats (Gaia *et al*, 2014), with this study representing only the second report of such a sub-viral entity in soil. Despite the deep sequencing approach employed in the study, the low number of virophage-related sequences identified would suggest that virophages were either present in low abundance or were poorly enriched by the protocol employed.

4.4 Metagenomic assembly of a circularly permuted, temperate phage genome

In the previous chapter, a deeper analysis of a subset of annotated contigs was presented, which showed that genes associated with phylogenetically distant virus families occurred within continuous nucleotide sequences (e.g. AntarOS_17). Here we describe another annotated contig (assembled with CLC Genomics/DNASTAR Lasergene NGen, method outlined in Chapter 3), from the open soil read dataset, in the form of a circularly permuted sequence.

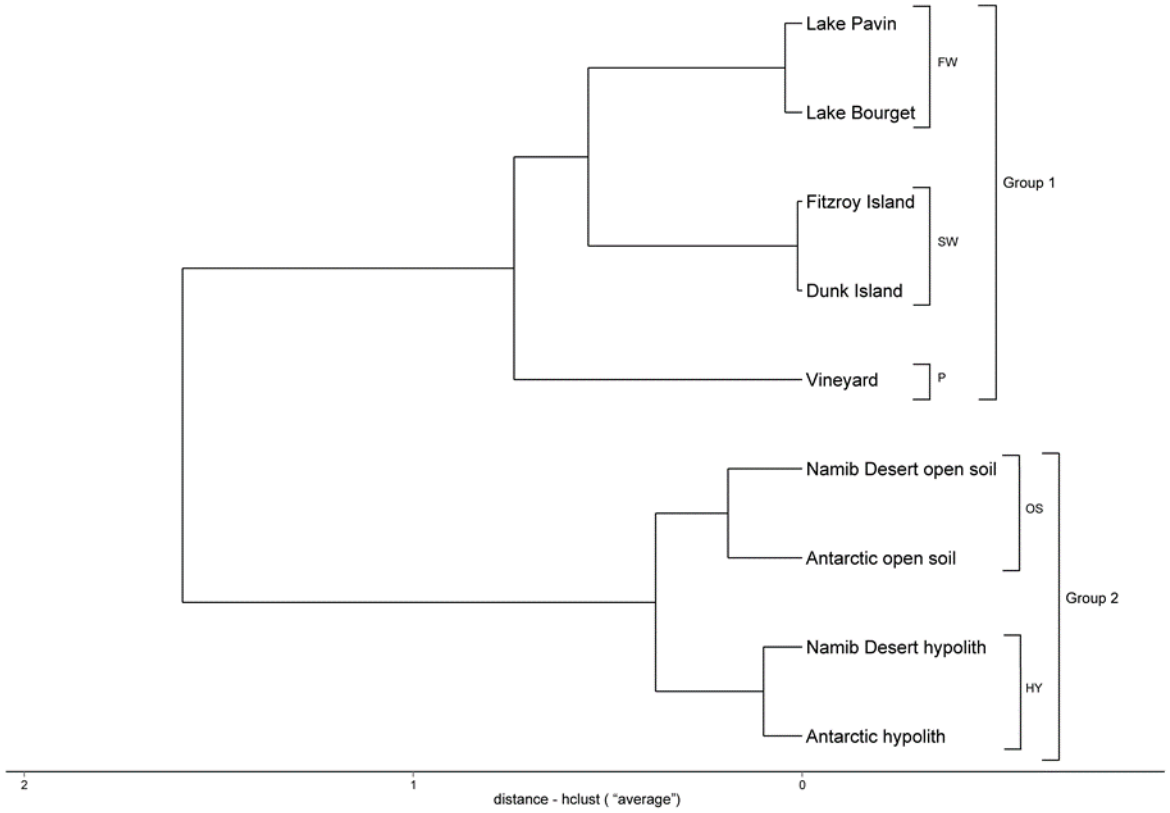

Figure 14. Hierarchical clustering of various metaviromes (assembled into contigs) based on dinucleotide frequencies. The x-axis denotes eigenvalues distances. The tree was constructed using the MetaVir server pipeline according to the method of Willner et al. (2008). FW: freshwater; SW: seawater; P: plant; OS: open soil; HY: hypolith.

Gene prediction with MetaGeneAnnotator on the MetaVir pipeline yielded 62 open reading frames (ORFs) along the 42,417 base-pair (bp) contig (named AntarOS_8, Figure 15). Fifty-seven percent of the ORFs identified had a database homolog using BLASTp against the non-redundant (nr) GenBank database or HHPred (http://toolkit.tuebingen.mpg.de/hhpred) with HMM-HMM scans against several protein databases (PDB, SCOP, pfamA, smart, PANTHER, TIGRfam, PIRSF and CD). Gene functions and/or taxonomic affiliations for these putative ORFs are summarized in Table 1.

The mean G+C content of the putative genome was 65.4%. Both ARAGORN (Laslett & Canback, 2004) and tRNAscan-SE (Lowe & Eddy, 1997) were used to search for tRNAs and tmRNAs. One tRNA gene was identified (tRNA-Gly, anticodon: TCC, 73bp). The gene was positioned at nucleotide position 18,059-18,131, corresponding to a region upstream of gene 27 (position 15,833-17,173), an integrase domain, which together may be involved in the integration of the phage genome by site-specific host genome integration. A XerD tyrosine recombinase domain within gene 45 (similar to phage P1 Cre recombinase[18]) and a putative transcription regulator (gene 34) with a helix-turn-helix (HTH) xenobiotic response element domain (containing 3 non-specific and 6 specific DNA-binding sites), was most similar to the generic structure of the prophage repressor family of

transcriptional regulation proteins (Wood *et al*, 1990). This composition of genes strongly suggests that this putative virus genome is that of a temperate bacteriophage.

The presence of the small and large terminase subunit protein domains (gene 48 and 49, respectively), a tail tape measure domain (gene 37), a portal protein (gene 51) and a tail collar protein domain (gene 7) together support a tailed phage virion structure. A BLASTp search using the *terL* gene showed closest homology (41% identity with 90% sequence coverage, E-value $1e^{-94}$) to *Xanthomonas* phage Xp15 (accession YP_239275.1). This phage isolate belongs to the order *Caudovirales*, but has not been classified at family or genus level. BLASTp analysis of *Xanthomonas* phage XP15 AAX84861.1 (putative tail tape measure protein) revealed its closest homologs belongs to the family *Siphoviridae*, including *Burkholderia* phage BcepGomr. Similarly, the *terL* gene phylogeny generated by a BLASTp search indicated that the closest homologs belonged to siphoviruses (data not shown). Several genes on contig AntarOS_8 could not be assigned to any known viral genome, but nevertheless contained conserved sequences translating to proteins with homologs in current databases. These included a site-specific RNA endonuclease (RecA-like, gene 16) and a J-domain (DnaJ) at the C-terminus of a larger peptide of unknown function (possible role in protein folding and degradation).

The remainder of predicted proteins were genes coding for an unknown structural protein, an internal virion protein, the major capsid protein (*mcp*), acyl-CoA-N-acyltransferase, AAT superfamily (PLP-dependent) and hypothetical protein related to diverse phage isolates infecting *Brucella*, *Burkholderia*, *Sulfitobacter*, *Pseudomonas*, *Bacillus*, *Mycobacterium*, *Myxococcus* and *Rhizobium*. Several of the conserved genes of contig AntarOS_8 show a distant relation to the siphoviruses lambda (integrase (gene 27) and HK97 (MCP) and the podovirus P22 (*terL*, c2). This indicates that this contig represents a novel mosaic phage genome which is a potential new member of the hybrid supercluster of Lambda-like phages recently described by Grose and Casjens and it is speculated that the host belongs to the family *Enterobacteriaceae* (Grose & Casjens, 2014) .

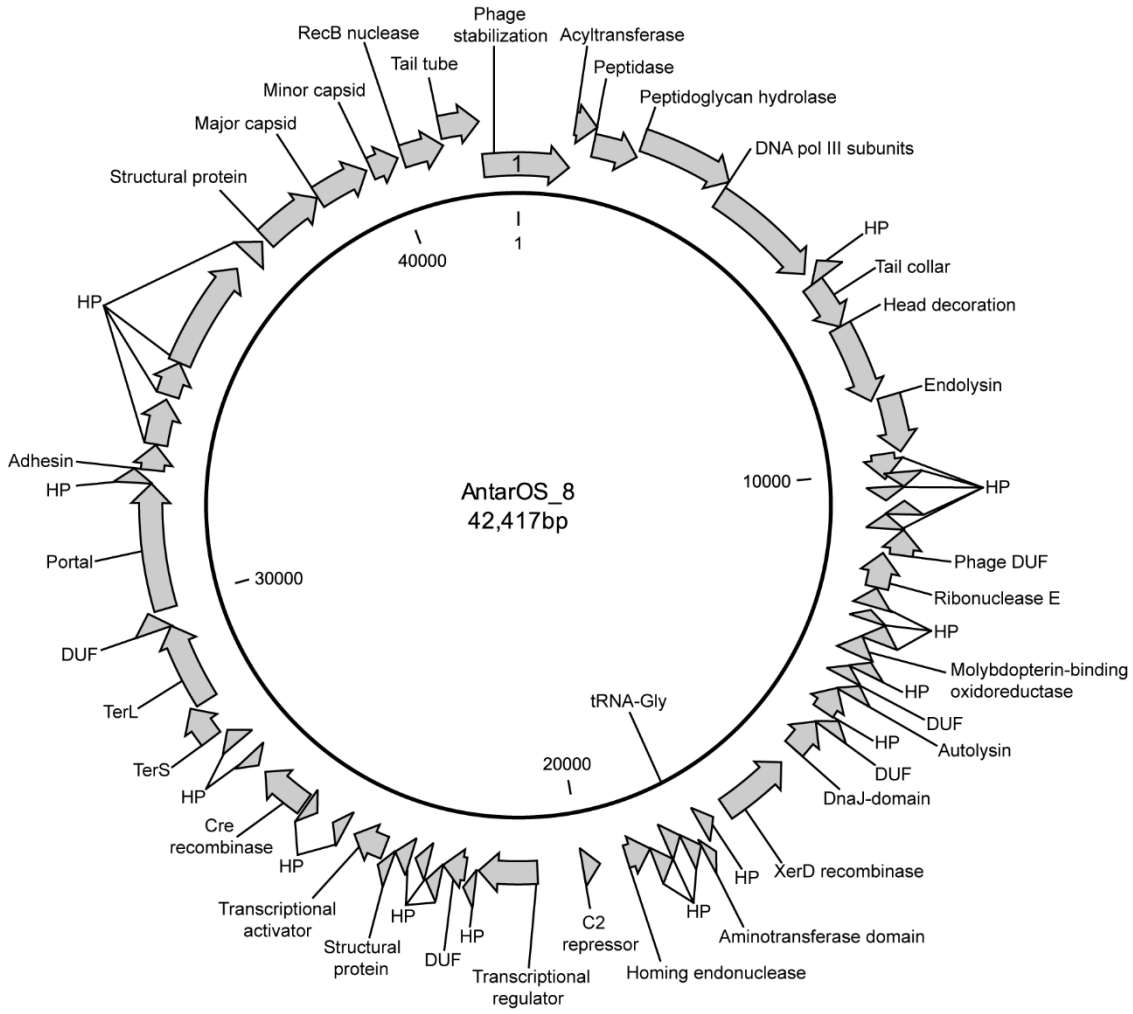

Figure 15. Gene annotation of contig AntarOS_8. Arrowed blocks are open reading frames (ORFs) and their orientation. HP denotes hypothetical proteins and DUF denotes conserved protein domains of unknown function. Numbers within the circle are nucleotide positions, starting within gene number 1 (indicated within the top arrow as "1") and onwards in a clockwise orientation.

4.5 Conclusions

Metaviromic analyses of Antarctic hyperarid niche habitats has revealed that these are highly novel, but not necessarily geographically distinct. Dinucleotide clustering of the metavirome contigs grouped hypolith niche habitats together, as well as open soil, rather than the cold versus warm hyperarid environments. Detailed analysis of a specific contig revealed a distant lambda-like genome, linking it in with a highly diverse group of temperate phages isolated from all over the world.

Table 10. Detailed ORF list including peptide length, associated taxonomy and predicted function. "TM" denotes the detection of a transmembrane domain within the predicted protein; a.a: amino acid.

ORF	Size (a.a)	Best BLASTp taxonomic hit	E- value	Accession	% identity; % query coverage	Conserved protein domain (accession)	Predicted function
1	566	*Sulfitobacter* phage NYA-2014a	5e-118	AIM40653.1	40; 99	Phage_stabilize (pfam11134)	Phage stabilisation protein
2	127	*Burkholderia* phage Bcep22	3e-06	NP_944298.1	31; 100	Ribosomal-protein-alanine acetyltransferase (2z10_A)	Acyl-CoA N-acyltransferase
3	274	*Sulfitobacter* phage pCB2047-C	5e-54	YP_007675267.1	44; 91	-	Peptidase domain-containing hypothetical protein
4	557	*Brucella* phage Tb	2e-04	YP_007002033.1	33; 23	-	Peptidoglycan hydrolase
5	705	-	-	-	-	DNA polymerase III subunits gamma and tau (PRK14954)	Gene expression
6	96	-	-	-	-	-	Hypothetical protein
7	307	*Brucella* phage S708	1e-14	AHB81257.1	29; 70	Phage Tail Collar Domain (pfam07484)	Collar protein for virion assembly
8	493	Prophage MuMc02	0.005	ADD94511.1	34; 25	-	Head decoration protein
9	332	Stenotrophomonas phage Smp131	40;39	YP_009008370.1		Lysozyme_like domain (cl00222)	Endolysin (TM)
10	153	-	-	-	-	Phage_HK97_TLTM (PF06120)	Tail length tape measure protein
11	84	-	-	-	-		Hypothetical protein
12	81	-	-	-	-	Orn_Arg_deC_N: Pyridoxal-dependent decarboxylase (PF02784)	Hypothetical protein (TM)
13	79	-	--	-	-	-	Hypothetical protein
14	90	-	-	-	-	-	Hypothetical protein
15	149	*Vibrio* phage pYD38-A	2e-19	YP_008126176.1	42; 65	Phage gp49 66 superfamily (cl10351)	Unknown function

16	214	Mycobacteriophage MacnCheese	0.010	AFN37792.1	38; 35	ribonuclease E (PRK10811)	Site-specific RNA endonulcease
17	115					-	Hypothetical protein
18	64	-	-	-	-	-	Hypothetical protein
19	105	-	-	-	-	-	Hypothetical protein
20	120	-	-	-	-	MopB_4 CD (cd02765)	Molybdopterin-binding oxidoreductase-like domains
21	71	-	-	-	-		Hypothetical protein
22	77	-	-	-	-	cas_TM1812 CRISPR-associated protein (TIGR02221)	Unknown function
23	76	-	-	-	-	Autolysin_YrvJ (PIRSF037846)	N-acetylmuramoyl-L-alanine amidase
24	166	-	-	-	-	-	Hypothetical protein
25	80					AAA (cl18944)	Unknown function
26	213	Archaeal BJ1 virus	8e-08	YP_919032.1	28; 92	DnaJ domain (cd06257)	DnaJ-containing protein
27	446	-	-	-	-	XerD Site-specific recombinase (COG4974)	Integrase/recombinase
28	90	-	-	-	-	-	Hypothetical protein
29	57	-	-	-	-	4-aminobutyrate aminotransferase (PRK06058)	Unknown function
30	88	-	-	-	-	-	Hypothetical protein
31	98	-	-	-	-	-	Hypothetical protein
32	103	*Pseudomonas* phage DMS3	7e-15	YP_950436.1	40; 87	-	Hypothetical protein
33	153	-	-	-	-	HNH endonuclease (PF01844)	HNH homing endonuclease
34	106	-	-	-	-	P22 C2 repressor (d2r1jl_)	C2 repressor
35	360	*Lactobacillus* phage c5	1e-05	YP_007002377.1	32; 27	DnaA N-terminal domain (cl13142); Arsenical	Transcriptional regulator

						Resistance Operon Repressor (smart00418)	
36	67	-	-	-	-	-	Hypothetical protein
37	140	*Azospirillum* phage Cd	9e-24	YP_001686851.1	62; 52	Domain of unknown function (pfam10073)	Conserved hypothetical protein
38	80	-	-	-	-	-	Hypothetical protein
39	83	-	-	-	-	-	Hypothetical protein
40	100	-	-	-	-	-	Hypothetical protein
41	70	-	-	-	-	RHH_1: Ribbon-helix-helix protein, copG family (PF01402)	Structural protein
42	203	-	-	-	-	rfaH transcriptional activator (PRK09014)	transcriptional activator protein
43	82	-	-	-	-	-	Hypothetical protein
44	79	-	-	-	-	-	Hypothetical protein
45	308	Enterobacteria phage D6	1e-12	AAV84949.1	25; 88	Cre recombinase, C-terminal catalytic domain (cd00799)	Cre recombinase
46	82	-	-	-	-	-	Hypothetical protein
47	98	-	-	-	-	-	Hypothetical protein
48	211	*Sinorhizobium* phage phiLM21	2e-12	AII27789.1	31; 82	Terminase_2 Terminase small subunit (pfam03592)	TerS
49	506	*Xanthomonas* phage Xp15	1e-94	YP_239275.1	41; 90	Terminase_3 Phage terminase large subunit (pfam04466)	TerL (TM)
50	110	*Sulfitobacter* phage pCB2047-C	2e-23	YP_007675320.1	49; 85	Protein of unknown function (DUF3307)	Conserved hypothetical protein
51	774	*Brucella* phage Tb	2e-163	YP_007002020.1	40; 88	Portal protein (3lj5_A)	Portal protein
52	82					-	Hypothetical protein
53	155	*Rhizobium* phage vB_RleM_P10VF	3e-09	AIK68325.1	34; 60	Tfp pilus assembly protein (COG3419)	tip-associated adhesin PilY1
54	281	-	-	-	-	-	Hypothetical protein

55	208	-	-	-	-	-	Hypothetical protein
56	678	-	-	-	-	-	Hypothetical protein
57	115	-	-	-	-	-	Hypothetical protein
58	400	*Brucella* phage Tb	3e-12	YP_007002024.1	25; 90	-	Structural protein
59	339	*Brucella* phage Tb	7e-60	YP_007002025.1	37; 94	Major capsid protein gp5 (d2ft1a1)	Major capsid protein
60	178	*Myxococcus* phage Mx8	9e-24	NP_203463.1	46; 80		Major virion structural protein
61	258	*Brucella* phage Pr	1e-17	YP_007002027.1	32; 92	RecB family nuclease (TIGR03491)	DNA-binding protein
62	233	*Brucella* phage Tb	5e-33	YP_007002028.1	35; 93	Tail tubular protein A (PHA00428)	Tail tube protein

4.6 References

Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M & Cowan DA (2015) Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* **17:** 480–495 Available at: http://europepmc.org/abstract/MED/24912085

Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJG & Burger JT (2010) Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology* **400:** 157–163

Cowan DA, Sohm JA, Makhalanyane TP, Capone DG, Green TG a., Cary SC & Tuffin IM (2011) Hypolithic communities: important nitrogen sources in Antarctic desert soils. *Environ. Microbiol. Rep.* **3:** 581–586 Available at: http://doi.wiley.com/10.1111/j.1758-2229.2011.00266.x

Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D & La Scola B (2014) Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS One* **9:** 1–8

Grose JH & Casjens SR (2014) Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* **468:** 421–443

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

Khan N, Tuffin M, Stafford W, Cary C, Lacap DC, Pointing SB & Cowan D (2011) Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. *Polar Biol.* **34:** 1657–1668

Laslett D & Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32:** 11–16

Lowe TM & Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964

Makhalanyane T, Valverde A, Birkeland N-K, Cary SC, Tuffin IM & Cowan DA (2013) Evidence for successional development in Antarctic hypolithic bacterial communities. *ISME J.* **7:** 2080–90 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23765099

Pointing SB, Chan Y, Lacap DC, Lau MCY, Jurgens JA & Farrell RL (2009) Highly specialized microbial diversity in hyper-arid polar desert. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 19964–19969

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T & Debroas D (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7:**

Roux S, Tournayre J, Mahul A, Debroas D & Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15:** 76 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4002922&tool=pmcentrez&rendertype=abstract

Smith MC, Bowman JP, Scott FJ & Line MA (2000) Sublithic bacteria associated with Antarctic quartz stones. *Antarct. Sci.* **12:**

Willner D, Thurber RV & Rohwer F (2009) Metagenomic signatures of 86 microbial and viral

metagenomes. *Environ. Microbiol.* **11:** 1752–1766

Wood HE, Devine KM & McConnell DJ (1990) Characterisation of a repressor gene (xre) and a temperature-sensitive allele from the Bacillus subtilis prophage, PBSX. *Gene* **96:** 83–88

Wood SA, Rueckert A, Cowan DA & Cary SC (2008) Sources of edaphic cyanobacterial diversity in the Dry Valleys of Eastern Antarctica. *ISME J.* **2:** 308–320

Wynn-Williams DD (1996) Antarctic microbial diversity: the basis of polar ecosystem processes. *Biodivers. Conserv.* **5:** 1271–1293

Note: this chapter was adapted from an addendum article to Chapter 3. Therefore, some of the content is repeated, but additional analyses have been performed based on the same dataset used in Chapter 3. The content of this chapter was not merged with Chapter 3, as it was considered out of place.

SECTION III: VIRUSES IN NAMIB DESERT SOILS

Chapter 5:

Extracellular virus diversity and biogeography in Namib Desert soils

5.1 Chapter summary

The taxonomic composition of soil viruses along the Namib Desert aridity gradient was assessed using deep sequencing of metavirome libraries extracted from surface soils. Soil physicochemical data were also used in bivariate correlations to determine influences, if any, on the observed soil-specific taxonomic compositions of virus communities. Within the low proportion of identified viruses (~20% of the dataset), less than 1% of the total virus diversity was shared across the soil samples. The spatial distributions of Namib Desert soil virus communities were significantly correlated to soil texture (sand particle size) and soil chemistry. Relative virus abundance showed a positive correlation with microbial activity and increased as a function of distance from the coast. However, we could not identify any significant correlation between viral abundance and mean annual soil relative humidity. The ssDNA virus fraction was spatially restricted within inland soils, and sequences identified as *Microviridae* spanned three sub-families. From the low virus species overlap between samples, we suggest that limited virus dispersal occurs across the Namib Desert. We propose that the biogeographical patterns of the extracellular virus fraction are primarily determined by host distribution. Increased viral abundance with microbial activity is the first, albeit indirect, indication of active viral predation, and the possibility that soil viruses may contribute to nutrient cycling (via viral lysis) in a hot hyperarid desert soil environment.

5.2 Introduction

In contrast to the recognized key system services provided by viruses in marine habitats (Fuhrman, 1999; Wommack & Colwell, 2000; Bettarel *et al*, 2004; Suttle, 2007; Rodriguez-Valera *et al*, 2009), the primary ecological roles of viruses within terrestrial ecosystems remain uncertain. Despite high virus-like particle (VLP) counts (8.7×10^8 - 4.17×10^9 virions per gram dry soil) reported in several pilot studies (Ashelford *et al*, 2003; Williamson *et al*, 2005, 2007), soil viruses remain substantially understudied (Williamson *et al*, 2013). Nevertheless, preliminary virus community analyses within temperate and 'extreme' soils have indicated that ecological dynamics (e.g. virus-host interactions) and viral taxonomic compositions substantially differ from those reported in marine and fresh water habitats (Srinivasiah *et al*, 2008).

Desert environments are one of the largest terrestrial ecosystems on Earth, and account for ~33% of terrestrial land masses (Goudie, 2002). However, the diversity and ecological functions of viruses in these arid ecosystems are mostly unknown (Makhalanyane *et al*, 2015). Thus far, research associated with viruses in arid soils has primarily focused on the taxonomic composition associated

with surface soils and soil-lithic interfaces (Prigent et al. 2005; Fierer et al. 2007; Adriaenssens et al. 2015; Williamson et al. 2007; Prestel et al. 2008; Pearce et al. 2012; Prestel et al. 2013; Zablocki et al. 2014; Wei et al. 2015) and has revealed a wide diversity of virus taxa, mostly composed of tailed virus families (Order: *Caudovirales*). Virus discovery in hot desert soils has thus far necessitated mitomycin C treatment of soil slurries for the release of the SOS-inducible phage fraction (Prigent *et al*, 2005; Prestel *et al*, 2013, 2008). The authors suggested that low numbers of extracellular VLPs and/or difficulties associated with soil extraction protocols were responsible for the lack of free viruses in the soil samples.

One report has described the taxonomic diversity of SOS-induced proviruses extracted from Namib Desert surface soils (Prestel *et al*, 2008). In that study, sample soil collection was spatially restricted to the coastal area of the desert. Thus, as this desert further extends ~140 km inland (Lancaster, 1984), a large area of this desert has not been investigated for the presence of extracellular virus particles. Moreover, Namib Desert soils are subjected to the effects of a longitudinal aridity (and water availability) gradient, derived from sporadic fog and rainfall events (Eckardt *et al*, 2013). Fog originates from the coast, and may reach as far as 60 km inland. Precipitation measurements have shown a gradual rainfall increase following a West-East direction, from the coast (10 mm) to the inland escarpment (60 mm). The most arid section along this transect is situated between 60 and 100 km from the coast, at the boundary between the fog and rainfall zones (Lancaster & Seely, 1984). This inverse fog-rainfall gradient has been shown to significantly contribute to the spatial composition and community structure of open surface soil and rock-associated (hypolith) microbial communities (Stomeo *et al*, 2013). However, the extent of viral diversity, and how an environmental gradient may impact desert soil viral communities, are unknown.

In this study, Ion Proton deep sequencing and soil physicochemical data were used to assess soil viral diversity and spatial distribution in soil surfaces samples, collected along a West-East transect in the Namib Desert. We hypothesized that an increase in soil water availability across the transect (Lancaster & Seely, 1984) would constitute a major factor in determining the abundance and taxonomic composition of soil virus communities.

5.3 Materials and methods

Sampling strategy

Surface soil samples were collected during April 2013 at 30 km intervals along the C14 road, extending from Walvis Bay inland towards Windhoek, across the Namib Desert (Figure 16). Five hundred grams of surface soil (0-2 cm depth) were collected at each site (GPS coordinates: "NAM10": 22 58.773 S, 014 34.994 E; "NAM40": 23 01.080 S, 014 51.587 E; "NAM70": 23 06.004 S, 015 07.676 E; "NAM100": 23 14.776 S, 015 16.881 E; "NAM130": 23 19.243 S, 015 37.370 E) and transferred to sterile Whirl-Pak bags (product no.B01445WA; Nasco). Samples were kept at 4°C in the dark in the laboratory.

Figure 16. Sampling locations. Numbers within blue inverted drops represent sampling locations, named throughout this study with a 'NAM'prefix (e.g. NAM10).

Soil physicochemical analysis

Eighteen soil variables were measured for each sample, and values were averaged from four replicate measurements (Table 11). Standard protocols were used for the determination of soil texture (ASTM, 2007; Ashworth *et al*, 2001) and pH (Pansu, 2006). Organic matter content (OM%) was determined by weight loss-on-ignition (360 °C for 2 h), according to Schulte & Hopkins (1996), supplemented with a temperature pre-treatment of 150 °C. Soil calcium (Ca+), potassium (K+) magnesium (Mg+), phosphorus (P), sodium (Na+), and sulfur (S) were extracted by ammonium acetate, and the concentrations measured by inductively coupled plasma atomic emission spectroscopy (ICP OES; Spectro Genesis, Spectro Analytical Instruments GmbH, Germany) (Rhoades, 1982). A 15 g subsample was taken from each site for enzymatic assays and stored at -20 °C within 12 hours of collection. Soil surface relative humidity (%RH) was recorded using iButton® data loggers (model DS1923, Maxim Integrated Products, San Jose, CA) placed at a depth of 1 cm. Readings were recorded at 4 h intervals for 12 months.

Fluorescein diacetate assay

Fluorescein diacetate (FDA) hydrolysis assays (used as a proxy of overall microbial activity) were performed on all samples by the method reported by Green et al. (2006), with minor modifications. Briefly, 0.5 g of soil was combined with 12.5 mL of 1 X PBS buffer (pH 7.4) and 0.25

mL 4.9 mM FDA dissolved in acetone, and incubated at 43 °C for 2 h under constant agitation. FDA hydrolysis was denatured by adding 40 µl of acetone to 1 ml of soil slurry, and the samples were centrifuged at 8, 800 ×g for 5 min. Fluorescence was measured with a portable fluorometer (Quantifluor™, Promega, Madison, USA).

Viral DNA extraction, amplification and sequencing

Direct extraction of viruses from soil samples was conducted according to Williamson et al. (2005), with slight modifications. Five grams of soil sample were added to 15 ml 1% potassium citrate buffer and vortexed at full speed for 15 seconds. The mixture was incubated on ice for 25 min, followed by three cycles (30% amplitude for 59 sec) of sonication with an ultrasonic processor (Qsonica sonicator) using a 1/16" probe tip. Samples were centrifuged at 3000 x g at 4°C for 30 minutes. The supernatant was decanted, transferred to a new tube and passed through a 0.20 µm cellulose acetate sterile syringe filter (GVS). Viruses were concentrated by adding 25% PEG8000 (in 1M NaCl) to the filtrate to a final concentration of 10% (w/v) and incubated overnight at 4°C. Concentrates were centrifuged for 30 minutes at 3,200 × g at 4°C. The supernatant was decanted and the viral pellet resuspended in 300 µl phage buffer (10mM Tris-HCl, 10 mM MgSO4, 150 mM NaCl, pH 7.5). Viral concentrates were treated with DNase I (Thermo Scientific, cat#EN0523) and RNase A (Thermo Scientific, #EN0531) according to the manufacturer's instructions. Viral DNA was purified using the Quick-gDNA MiniPrep kit (Zymo Research, cat# D3025) according to the manufacturer's instructions. Viral DNA was randomly amplified using the REPLI-g Midi kit (Qiagen, cat# 150043) according to the manufacturer's instructions. Amplified DNA was precipitated with isopropanol, washed with 70% ethanol and resuspended in 25µl milli-Q water.

The amplified metaviromes were checked for bacterial contamination by assessing the presence of the 16S rRNA gene, using primer pair E9F and U1510R (Hansen et al. 1998; Reysenbach, 1995) as follows: 1 µl of genomic DNA was mixed with 2.5 µl of each primer (10mM), 2.5 µl of 2 mM deoxynucleoside triphosphates (dNTPs), 2.5 µl of 10X DreamTaq buffer (ThermoFisher Scientific, MA, USA), 1 µl of 10- mg/ml bovine serum albumin (BSA), 0.125 µl DreamTaq polymerase (ThermoFisher Scientific, MA, USA), and Milli-Q water to a total volume of 25 µl. PCR was conducted under the following thermal regime: (i) 5 min at 95°C; (ii) 30 cycles, with 1 cycle consisting of 30 s at 95°C, 30 s at 52°C, and 85 s at 72°C; and (iii) 10 min at 72°C. Library building for sequencing was done using the Ion Xpress™ Plus and Ion Plus Library Preparation for the AB Library Builder™ System (Publication Number MAN0006946). Template amplification was done using the Ion OneTouch™ 2 System (OT2) Ion PI™ Hi-Q™ OT2 200 Kit (Number MAN0010857). The metavirome libraries were multiplexed and sequenced using the Ion PI™ Hi-Q™ Sequencing 200 Kit (Number MAN0010947) using the Ion PI™ Chip Kit v3. Sequencing was performed on the Ion Proton platform, located at the Central Analytical Facilities, Stellenbosch University, South Africa.

Table 11. Physicochemical and environmental properties of the Namib Desert soil samples. All values were averaged from four replicate measurements, except for elevation and DNA concentration. Standard deviation values are shown were appropriate. Soil texture, organic carbon and organic matter content were measured as a percentage per soil sample. VCS: very coarse sand; CS: coarse sand; MS: medium sand; FS: fine sand; VFS: very fine sand. The proxy measure for total microbial activity (FDA) is expressed as µmol fluorescin/ g dry soil. Elemental quantification (Ca, S, Mg, K, Na, P) is expressed as mg per kg of dry soil. DNA concentration values are ng/µl.

	West samples			East samples	
	NAM10	**NAM40**	**NAM70**	**NAM100**	**NAM130**
Soil texture					
VCS	9.8±3.2	4.43±0.59	2.38±0.43	6.33±1.94	5.30±2.39
CS	6.66±2.13	7.23±0.41	5.82±1.23	6.22±0.32	6.93±2.70
MS	8.98±1.81	14.16±0.58	12.93±1.73	11.74±1.73	6.91±1.28
FS	29.05±3.76	37.47±3.09	45.99±2.77	26.30±4.48	35.15±4.78
VFS	30.69±5.96	15.03±3.34	22.04±0.82	23.49±0.76	21.39±3.02
Silt	8.3±2.4	14.64±1.79	5.31±1.83	18.88±3.59	17.28±2.15
Clay	6.5±1	7±2.5	5.5±1.91	7±1.15	7±2
Soil Physicochemistry					
pH	6.7±0.03	7.24±0.22	7.56±1.34	8.04±0.97	7.44±0.85
Organic carbon	0.25±0.07	0.06±0.02	0.03±0.01	0.28±0.03	0.24±0.05
Calcium	18525±6174	25573±1803	4734±3084	4427±73	1308±173
Sulphur	12278±5300	15270±111	2375±2005	10±3	45±10
Magnesium	459±39	168±92	23±3	76±8	101±7
Potassium	522±61	339±54	80±13	232±22	284±83
Sodium	11357±3546	621±172	23±3	17±4	35±9
Phosphorous	12.34±4.57	6.69±3.37	12.92±2.01	5.69±1.86	10.24±2.29
Miscellaneous					
FDA	0.07±0.05	0.70±0.24	3.29±2.02	14.45±2.03	19.72±1.60
Organic matter	0.42±0.08	0.37±0.07	0.19±0.01	0.61±0.08	0.74±0.12
Virus DNA concentration	<0.5	3.8	14.2	9.46	2.14
Elevation (m)	23	311	568	749	931
Distance from coastline (DCL)(km)	10	40	70	100	130
Mean annual relative humidity (%RH)	63.05±8.41	45.21±15.64	46.27±21.45	37.66±20.33	49.91±21.57

Metavirome analyses

Sequence reads were curated for quality and adapter trimmed using CLC Genomics version 6.0.1 (CLC, Denmark), using the default parameters. *De novo* assembly for each read dataset was performed with CLC Genomics assembler suite using the default parameters. Contigs were uploaded to the MetaVir version 2 server (Roux et al., 2014; http://metavir-meb.univ-bpclermont.fr/), VIROME (Wommack et al., 2012; http://virome.dbi.udel.edu/) and MG-RAST (Glass et al. 2010; https://metagenomics.anl.gov/) for analysis. Taxonomic composition by MetaVir was computed from a BLASTp comparison with the Refseq complete viral genomes protein sequences database from NCBI (release of 2015-01-05) with a threshold of 10^{-5} for the E-value. Assembled reads were searched for

open reading frames and compared to the RefSeq complete viral database (through the MetaVir pipeline) and MG-RAST, which include annotations using, for functional and organism assignment: GenBank, IMG, KEGG, PATRIC, RefSeq, SEED, SwissProt, trEMBL and eggnog. Unique and shared virus hits were determined by recording the occurrence of all virus isolate hits (contig best blast hit number, 10-5 threshold for the E-value, MetaVir) in each soil sample dataset, and visualized using the Venn diagram online tool, available from the Bioinformatics and Evolutionary Genomics group website (http://bioinformatics.psb.ugent.be/webtools/Venn/).

VP1 phylogeny reconstruction

A reference dataset for protein F (*Microviridae* major capsid protein, VP1) was obtained from doi:10.5061/dryad.8ht80 (Roux *et al*, 2012). Additional VP1 sequences were retrieved from the (nr) NCBI database, consisting of sequences from *Microviridae* isolates which best matched Namib sequences with BLASTp. In total, 111 sequences (including Namib VP1 sequences) were aligned with MAFFT (Katoh & Standley, 2013). Neighbour-joining trees were drawn using MEGA6 (Tamura *et al*, 2013), with a 100 bootstrap cycles.

Statistical analyses

Putative relationships between soil properties and virus family distribution ratios were tested by using two-tailed, bivariate Pearson correlations computed with SPPS 20.0 (IBM Corp., Armonk, NY).

5.4 Results

Overview of the environmental gradient and virus DNA extraction

Soil parameter measurements showed that pH and % clay remained relatively consistent across all soil samples (Table 11). Significant correlations were found between the distance from the coastline (DCL) and ground elevation ($r = 0.996$, $p = 0.004$) and total microbial activity ($r = 0.975$, $p = 0.025$). The incremental increase in microbial activity was most pronounced between NAM10 and NAM40, which showed a 10-fold increase. Ground elevation was also positively correlated with microbial activity ($r = 0.956$, $p = 0.044$). Magnesium, potassium and sodium content generally tended to decrease as a function of distance of the coast.

Despite a random amplification step, only low DNA concentrations were obtained from all samples. This indicated that either very low viral biomass was present in the samples or that soil inhibitors may have resulted in sub-optimal DNA amplification. A significant correlation ($r = -0.897$; $p = 0.039$) was found between DNA concentration and potassium content in the soil samples. Sample NAM10 did not yield sufficient DNA (<1ng/µl) for sequencing library construction, and was therefore omitted from further analyses (Table 11).

Overview of the metaviromes

No detectable bacterial contamination was identified in the metaviromes, validated by PCR (no discernible DNA bands) and the MG-RAST pipeline (no BLAST results to ribosomal DNA sequences). Multiplexed sequencing of the four metaviromes produced 93,519,306 reads (~13.4 Gb), which resulted in each metavirome dataset totalling ~ 22 million reads, with a mean read length of 142.5 bp and a GC content ranging from 54 to 62%. A summary of the sequencing output and read assemblies are provided in Table 12. Between the three pipelines used (MetaVir, MG-RAST and VIROME), the greatest number of virus hits to sequence databases was achieved by MetaVir (Table 13). Therefore, the output from MetaVir was used exclusively throughout this study for assessing viral taxonomic composition. The term viral operational taxonomic unit (vOTU) will be used here as a proxy for defining a single virus species, based on individual contig best BLASTp hit (10^{-5} threshold for the E-value) against the RefSeq database. Across all soil samples, the ratio of sequences with taxonomic assignment (i.e. having a database homologue) ranged from 9.18 to 18.91 %, indicative of a highly uncharacterized pool of viral diversity. Rarefaction curves (Figure 17) for all metaviromes did not reach a plateau, indicating that each sequence dataset did not represent the full viral diversity within each sample.

Table 12. Sequencing and read assemblies metadata.

Statistic	**Transect samples**			
	NAM40	**NAM70**	**NAM100**	**NAM130**
Raw reads				
# sequences	19,572,328	22,794,083	23,874,886	23,397,857
Mean sequence length (bp)	137.11 ± 29.93	114.43 ± 45.09	157.69 ± 36.80	160.55 ± 28.03
Mean GC content (%)	61.95 ± 3.22	58.42 ± 6.75	57.96 ± 5.45	54.45 ± 6.38
Read assemblies				
# of contigs	2767	4700	10795	9951
Min length (bp)	200	200	200	200
Max length (bp)	4966	10765	37481	10714
N25	334	561	1337	1183
N50	256	268	298	330
N75	221	224	236	241

Table 13. Number of virus hits compared between three online pipelines.

# of virus-only hits (% from total number of uploaded sequences)	NAM40	NAM70	NAM100	NAM130
MG-RAST	2 (0.5%)	18 (7.7%)	41 (7.3%)	55 (10.7%)
VIROME	5 (1%)	18 (9.5%)	193 (12.5%)	177 (11.7%)
MetaVir	43 (9.18%)	75 (13.76%)	548 (18.91%)	366 (11.81%)

Viral richness and taxonomic distribution across metaviromes

Individual soil sample viral richness (relative sequence abundance, which include vOTUs occurring more than once) was lowest in the West transect samples (NAM40: 44 vOTUs and NAM70: 75 vOTUs) and markedly increased furthest inland (NAM100: 549 vOTUs and NAM130: 368 vOTUs). A total of 444 unique vOTUs (after the removal of duplicate vOTUs) inclusive of all four soil sample metaviromes was identified. The numbers of shared and sample-specific viral OTUs distributed amongst the four soils were plotted as a Venn diagram, shown in Figure 18. Only 3 OTUs (0.67%) were shared between all four metaviromes, and 295 vOTUs (66.44%) were exclusive to their respective soil sample. The three globally shared vOTUs were *Streptomyces* phage mu1/6, *Yersinia* phage phiR1-37 and *Cellulophaga* phage phi19:1. The highest number of shared vOTUs was between NAM100 and NAM130 (85). Only one vOTU was shared between NAM40 and NAM70 (Pithovirus sibericum, a giant dsDNA amoeba virus), while 12 vOTUs were shared between the most spatially distant samples (NAM40 and NAM130). *Mycobacterium* phages represented the top sequence hits across all soil samples (11.4-19.3 %), except for NAM40, where *Enterobacteria* phages were most numerous (18.6%). The next most represented phage isolates across the transect samples were to *Rhodococcus* viruses in NAM130 (9.83 %), *Bacillus* viruses in NAM100 (8%) and *Geobacillus* viruses in NAM70 (5.3%).

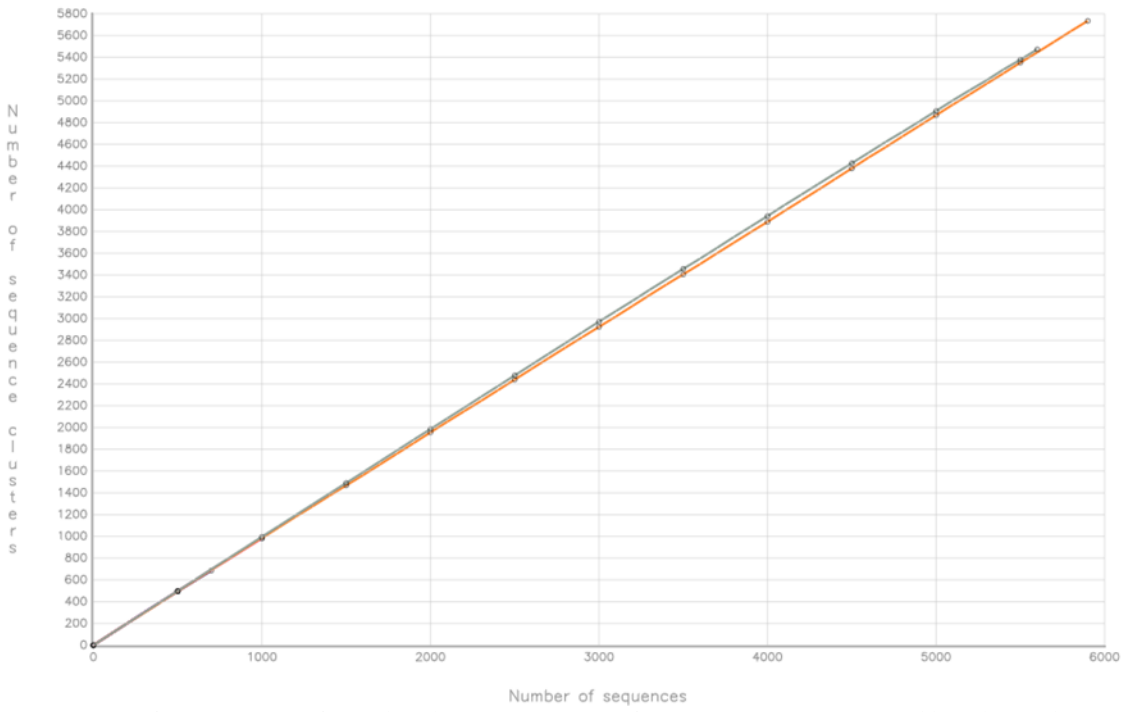

Figure 17. Rarefactions curves of Namib soil metaviromes. Rarefaction curves were generated based on a clustering of the predicted protein genes. Clustering (i.e. grouping) of predicted protein sequences was done through the detection of conserved domain (using the PFAM database) with a similarity threshold of 75%). The curve represents the number of different clusters created (y-axis) from a given number of sequences (x-axis).

A relatively equal distribution (mean: 10.35 ± 2.01%) of unclassified viruses was identified in all samples (Figure 19). In contrast, the frequency values of virus families were unevenly distributed along the transect samples. The *Myoviridae* and *Podoviridae* families (Order: *Caudovirales*) were most prevalent in NAM40, while siphoviruses were most abundant in sample NAM70. Family members of the nucleocytoplasmic large DNA viruses (NCLDVs), *Mimiviridae* and *Phycodnaviridae*, were also primarily identified in sample NAM40. Single-stranded DNA (ssDNA) viruses were mainly identified in samples NAM 100 and NAM130 (desert interior region) and represented the lower fraction compared to dsDNA viruses; ~3.3 % and ~92.5 % respectively. Homologous sequences to viruses of the *Inoviridae*, *Microviridae* and *Circoviridae* were detected. Dominant sequences belonged to *Propionibacterium* phage B5 (*Inoviridae*: genus *Inovirus*) and to chlamydiamicroviruses (*Microviridae:* subfamily *Gokushovirinae*).

Microviridae-related sequence hits were unique to sample NAM130 and identified by 11 conserved structural domains matching the major capsid protein (VP1) gene (phage_F; PF02305.12), a hallmark gene for the classification of the family (Hopkins *et al*, 2014). A dendrogram (Figure 20) was constructed based on aligned VP1 amino acid sequences retrieved from the NAM130 metavirome and the NCBI (nr) database, encompassing major family representatives from characterized and uncultured *Microviridae* isolates. The tree displayed a topology similar to the currently proposed *Microviridae* classification scheme, composed of 5 sub-groups (Quaiser *et al*, 2015; Roux *et al*, 2012). This includes the *Microvirus* genus and four (proposed) sub-families: Pichovirinae, *Gokushovirinae*, Aravirinae and Alpavirinae. Most Namib sequences (contig # 28, 459, 460, 650) formed a distinct clade within the *Pichovirinae* group. Contig 351 belonged to the *Gokushovirinae*, closely related to an uncultured marine microvirus (AGT39863.1), isolated from surface Atlantic Ocean waters (Labonté & Suttle, 2013). The last VP1 sequence (contig #555) identified in the Namib metavirome formed a very distant clade with a pequenovirus isolate from cold methane seep sediments (AJK28316.1, Bryson et al. 2015), located in the tree at the root of the *Gokushoviridae/Pichovirinae/Aravirinae* clades.

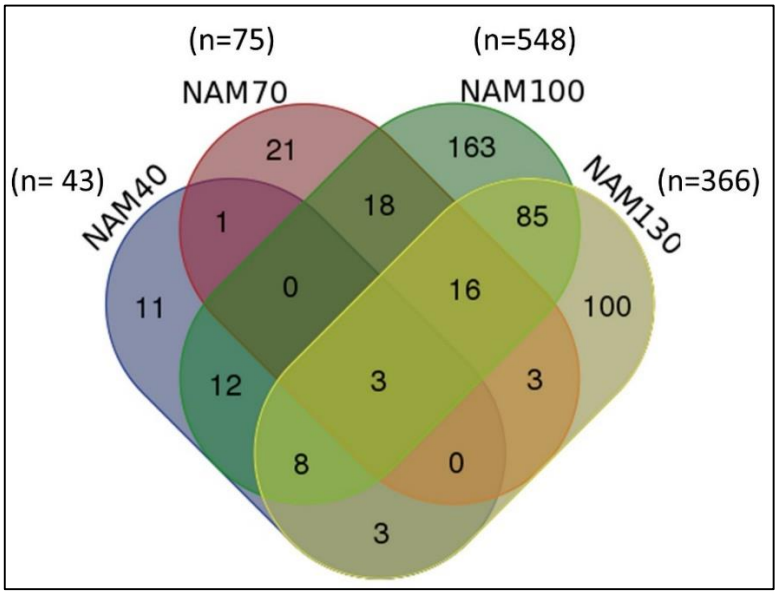

Figure 18. Venn diagrams showing the distribution of unique and shared viral OTUs across the four transect soil samples. Numbers in brackets represent the number (de-replicated) of identified vOTUs per sample.

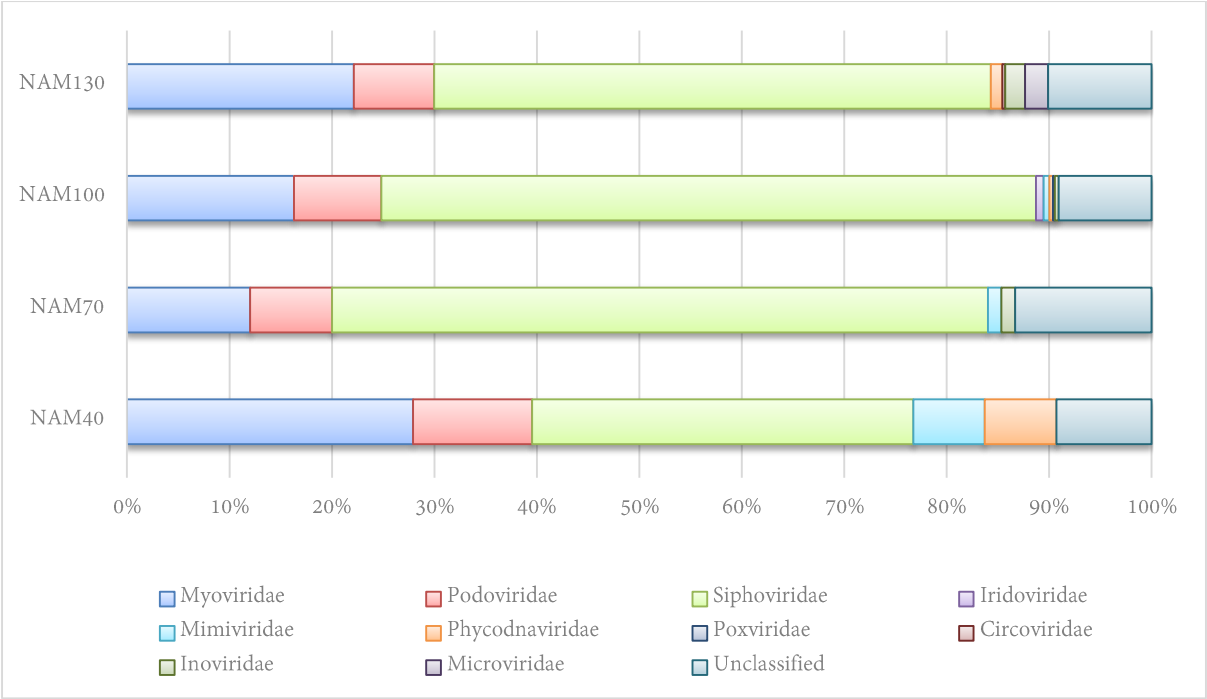

Figure 19. Family level taxonomic compositions across all transect samples computed from a BLAST comparison with NCBI RefSeq complete viral genomes proteins using BLASTp (threshold of 10^{-5} for the e-value). Virus hit numbers were normalized and converted into ratios. The unclassified category includes all dsDNA and ssDNA viruses.

Links between soil physicochemical properties, the environment, and viruses.

Using bivariate correlations, we examined the factors that may shape the viral family distributions in relation to soil abiotic and biotic properties (Table 14, found ON supplementary CD, due to size). The total number of identified vOTUs per sample was correlated to FDA ($r = 0.884$, $p = 0.047$). Recorded frequencies of virus sequences counts were organized at the family level (e.g. *Myoviridae*) and compared with environmental variables (Table 14). Correlation data showed that the virus family spatial distributions were mostly related to sand particle size and the presence of several chemical elements. For example, the occurrence of *Siphoviridae* was strongly associated with very fine sand (VFS; ≤100 μm) and *Myoviridae* with coarse sand (CS; ≤ 1mm) and soil magnesium content. The presence of ssDNA viruses (*Circoviridae* and *Microviridae*) was negatively correlated with the medium size range of sand particles (MS; ≤ 500 μm). Podoviruses and NCLDVs (mimiviruses and phycodnaviruses) were positively correlated to soil sodium, calcium and sulphur content. Soil relative humidity (%RH) showed a negative correlation with soil pH.

5.5 Discussion

A previous study has reported that no extracellular VLPs could be detected in Namib Desert soils (Prestel *et al*, 2008). However, the present study demonstrated that, with direct soil extraction, viruses were present in increasing abundance across the West-East Namib Desert transect (based on DNA yield per sample). Comparisons of the phylogenetic diversity of viruses between the four soil metaviromes showed strong sample-dependent specificity in terms of vOTU richness. Furthermore, the spatial composition of several dsDNA (e.g. *Siphoviridae*), NCLDVs and ssDNA (e.g. *Microviridae*) virus families were significantly correlated to soil texture (particularly sand particle size) and soil chemical composition. However, contrary to our initial hypothesis, we found no statistically significant correlation between soil relative humidity (as a proxy for water bioavailability) and with the abundance or spatial distribution of virus families.

Effects of soil chemistry on metavirome library construction

Very low DNA concentrations were obtained from all soil samples, requiring a random amplification step for all metavirome DNA extracts. Despite a random amplification step of sample NAM10, DNA yield (<1ng/μl) was insufficient for library construction and high-throughput sequencing. Incidentally, sample NAM10 originated from the same Namib Desert latitudinal segment (i.e., soil in close proximity to the coastline) that has previously been shown to be free of detectable extracellular viruses (Prestel *et al*, 2008). Sample NAM10 contained the highest quantities of magnesium, potassium and sodium salts, probably due to fog-derived deposits originating from the Atlantic Ocean (Cunningham & Jankowitz, 2011). This soil chemical composition may have contributed either to low soil extraction efficiency and/or acted as PCR inhibitors, in particular for DNA polymerases (Hebda & Foran, 2015).

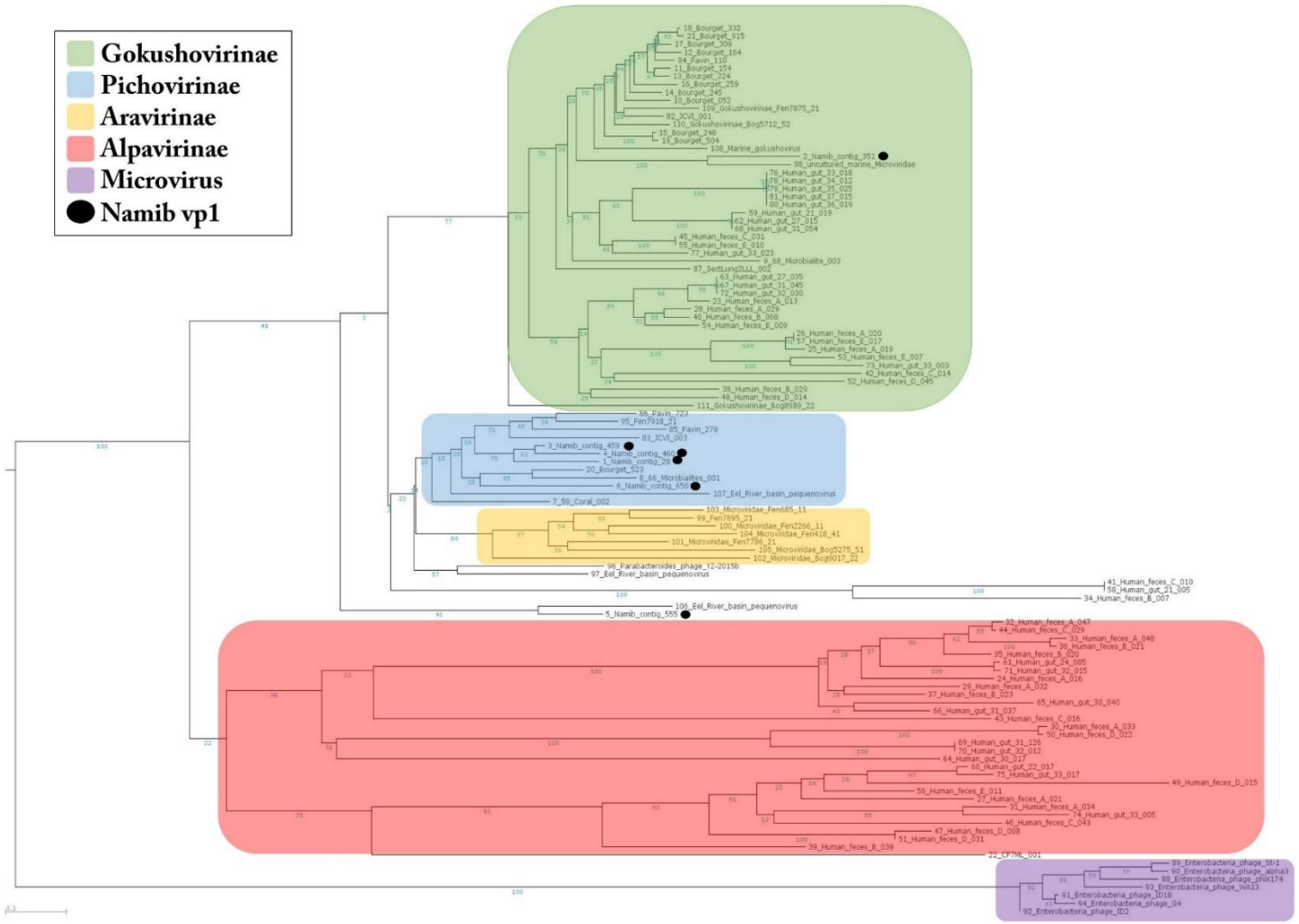

Figure 20. VP1 phylogeny in NAM130. The dendrogram was constructed using a neighbour-joining method with100 bootstraps. Coloured boxes represent Microviridae sub-families and genus. Black dots next to the external nodes show Namib Desert VP1 sequences from this study.

Impact of soil composition on virus community structure

Comparisons of virus community structure between soil samples suggested that the virus communities are phylogenetically distinct on a large spatial scale. While the effects of soil composition on large scale spatial partitioning of virus communities across the soil samples are unclear, our data suggested that soil chemistry and sand particle size may play a role. Surface soil physicochemical properties have been shown to shape bacterial community structure in the Namib Desert (Gombeer *et al*, 2015). Furthermore, microcosm experiments using Namib Desert soils have demonstrated that soil origin (in terms of water regime history) is a major driver for microbial activity (estimated by FDA) in simulated wetting events (Frossard *et al*, 2015). In addition, soil sodium content has been shown to be an important variable in shaping the composition of soil microbial communities along the Namib Desert aridity transect (Stomeo *et al*, 2013).Thus, the observed soil-specific virus spatial composition may reflect the host community structure, and supports the notion that, in extreme hot deserts, abiotic selection of microbial host species composition may be a major driver in shaping phage biogeography (Thurber, 2009). Similarly, an inverse relationship between ssDNA viruses (family: *Microviridae*) and medium-sized soil particles (≤ 500 μm) suggested that phylogenetic clustering of ssDNA viruses may be defined by soil properties and microbial diversity (Reavy *et al*, 2015).

As the soil samples were collected along an aridity gradient (Eckardt *et al*, 2013), water availability across the transect might be expected to significantly impact the diversity and spatial distribution of virus communities. Soil relative humidity showed a general increase as a function of increased distance from the coast, consistent with previous climatic surveys of the region (Eckardt et al., 2013). However, annual mean soil relative humidity (%RH) measurements from each of the sampling sites showed no statistically significant correlation with either total microbial activity (FDA) or virus abundance. The absence of this predicted correlation is surprising, given that a positive correlation between microbial activity and water availability changes has been demonstrated for Namib soil bacterial communities (Frossard *et al*, 2015). We believe that there may be two possible causes of this apparent lack of correlation. Firstly, the use of mean annual soil %RH values provides valid long-term average estimates of the soil atmosphere, but may not truly reflect water bioavailability (particularly at levels which trigger changes in viral particle numbers). Major precipitation events in the Namib Desert are highly stochastic and infrequent, and major changes (both increases and decreases) in extracellular virus numbers may only be responsive on short timescales, which are not compatible with mean annual soil %RH data. Secondly, we suggest that potentially observable changes in some virus taxa might be missed in our analyses as the current virus genome database provides only a fragmented and incomplete representation of total virus diversity, with a focus on pathogenic and industrially-relevant hosts rather than environmentally abundant hosts. This is especially true for viruses of terrestrial cyanobacteria, for which very limited genomic sequence data are available (Adriaenssens *et al*, 2015). Cyanobacteria in hyperarid desert soils are considered as keystone taxa,

with important roles in C and N cycling (Warren-Rhodes *et al*, 2013; Stomeo *et al*, 2013). The species richness and community structure of desert-soil cyanobacterial species (e.g., *Microcoleus, Chroococcidiopsis*) are known to respond significantly to shifts in water availability changes (Warren-Rhodes *et al*, 2007; Pointing *et al*, 2007). Thus, it is possible that the cognate cyanophages may display significant population changes (in terms of abundance and/or diversity) in response to host community shifts due to water bioavailability changes, but are not detectable using current sequence database limitations.

Viral propagation and links to soil-lithic niches

From the positive correlation between virus abundance and total microbial activity (FDA), it is speculated that this may represent active viral propagation. Previous studies on hot hyperarid desert soils communities have not reported the presence of extracellular virus particles, nor have suggested the effects of viral populations on autochthonous soil microbial communities (Prigent *et al*, 2005; Prestel *et al*, 2008, 2013). To our knowledge, this is the first, albeit indirect, indication that soil viruses may contribute to nutrient cycling (via viral lysis) in a hot hyperarid desert soil environment. Our observations therefore to support the "killing the winner" model proposed for phage-host interactions (Weinbauer & Rassoulzadegan 2004), whereby phage propagation (i.e., increasing extracellular abundance) increases as a function of host density (in this case, represented by FDA measurements).

The three universally shared vOTUs between the soil samples were assigned to viruses which infect three different bacterial phyla: Actinobacteria, Proteobacteria and Bacteroidetes. These phyla are known to dominate hot desert soils around the world (Makhalanyane *et al*, 2015), and this result strongly supports the conclusion that virus communities reflect the microbial host diversity.

Phages infecting *Rhodococcus*, *Bacillus* and *Geobacillus* were also amongst the most abundant sequences found within all metaviromes**.** This soil virus distribution partially supports the hypothesis of viral community recruitment by hypoliths from surrounding open soil (Makhalanyane *et al*, 2013; Adriaenssens *et al*, 2015). Hypoliths are cyanobacteria-dominated translucent rocks, and are a common feature of hyperarid ecosystems (Thomas, 2011). The viral component of the desert hypolith niche has been recently investigated (Adriaenssens *et al*, 2015), and showed that a large fraction of known viral contigs showed high similarity to *Bacillus*, *Geobacillus*, *Mycobacterium*, *Rhodococcus* and *Propionibacterium* phages. Our results showed a similar composition across the surface soil samples. Thus, a similar trend of hypolithic recruitment may also apply to virus communities due to overlapping vOTU distributions between the two soil niches. Clarifying this hypothesis would require the analysis of single hypolith and surrounding open soil metavirome datasets, in close spatial proximity, as well as comparative analyses of the shared vOTU fraction.

Soil-specific ssDNA sequence signals

The diversity and ecological functions of ssDNA viruses in soil ecosystems are still poorly understood (Quaiser *et al*, 2015). Despite the known amplification bias towards single-stranded circular DNA genomes associated with MDA (Kim & Bae, 2011), low counts of VP1 sequences were detected in only 2 of the 4 soil samples (NAM100 and NAM130). This reflects the very low abundance or even absence of ssDNA viruses in Namib Desert soils, and suggested that ssDNA virus hosts represent a minor fraction of the autochthonous soil microbial communities. Most VP1 sequences belonged to the Pichovirinae sub-family (Roux *et al*, 2012), and formed a distinct Namib sub-group within the clade. The addition of these isolates further expands the known diversity of pichoviruses and demonstrates their presence in a hot desert edaphic environment. Abundance estimates of ssDNA viruses across a diversity of viromes from multiple ecosystems have shown that gokushoviruses are the most abundant (Quaiser *et al*, 2015). In this study, gokushoviruses were only represented by a single sequence, most closely related to an uncultured marine isolate from marine *Microviridae* (Figure 20), but not closely related to other soil-isolated *Gokushoviridae* members. It is unclear whether the Namib VP1 groupings within Pichovirinae and *Gokushovirinae* reflect selection by environmental conditions or on host-presence/absence, or a combination of both, since the majority of these virus sequences are environmental in origin.

Conclusions

This study has highlighted the extracellular virus diversity and biogeography of previously uncharacterized virus communities in a hyperarid desert biome. The number of metavirome sequences with no database homologs (~80%) indicated that these soils contain a novel pool of virus diversity. While this limits ecological interpretation (since only 20 % of the virus fraction can be used in developing ecological inferences), the uncharacterized gene fraction does represent a pool of potentially valuable, novel and interesting, genes, gene functions and virus taxonomic groups.

Given the very small overlap in vOTU numbers between the soil samples (0.67%), we conclude that very limited dispersal occurs along the 120 km Namib Desert transect. Contrary to our initial hypothesis (i.e., that an increase in water availability would impact on virus abundance), we suggest that the observed patterns of virus biogeography are primarily determined by host diversity and spatial structure, but that the known linkage between host community structure and water bioavailability (Stomeo *et al*, 2013) is not clearly delineated in the viral population. We attribute this to two possible causes: firstly that averaged long term (annual) soil water data provide a poor guide to changes in viral population structure, which are likely to be only responsive to short-term stochastic precipitation events, and that the paucity of soil phage sequences in public databases may further obscure observable changes in certain highly water-responsive group of host organisms (e.g., cyanobacteria) and their cognate viruses.

5.6 References

Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M & Cowan DA (2015) Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* **17:** 480–495 Available at: http://europepmc.org/abstract/MED/24912085

Ashelford KE, Day MJ & Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. *Appl. Environ. Microbiol.* **69:** 285–289

Ashworth J, Keyes D, Kirk R & Lessard R (2001) Standard procedure in the hydrometer method for particle size analysis. *Commun. Soil Sci. Plant Anal.* **32:** 633–642

ASTM (2007) Standard test method for particle-size analysis of soils.

Bettarel Y, Sime-Ngando T, Amblard C & Dolan J (2004) Viral activity in two contrasting lake ecosystems. *Appl. Environ. Microbiol.* **70:** 2941–2951

Bryson SJ, Thurber AR, Correa AMS, Orphan VJ & Vega Thurber R (2015) A novel sister clade to the enterobacteria microviruses (family Microviridae) identified in methane seep sediments. *Environ. Microbiol.***:** n/a–n/a Available at: http://dx.doi.org/10.1111/1462-2920.12758

Cunningham PL & Jankowitz W (2011) Sabkha Ecosystems. **46:** 9–18 Available at: http://link.springer.com/10.1007/978-90-481-9673-9

Eckardt FD, Soderberg K, Coop LJ, Muller a. a., Vickery KJ, Grandin RD, Jack C, Kapalanga TS & Henschel J (2013) The nature of moisture at Gobabeb, in the central Namib Desert. *J. Arid Environ.* **93:** 7–19 Available at: http://dx.doi.org/10.1016/j.jaridenv.2012.01.011

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards R a., Felts B, Rayhawk S, Knight R, Rohwer F & Jackson RB (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73:** 7059–7066

Frossard A, Ramond J-B, Seely M & Cowan D a. (2015) Water regime history drives responses of soil Namib Desert microbial communities to wetting events. *Sci. Rep.* **5:** 12263 Available at: http://www.nature.com/doifinder/10.1038/srep12263

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

Glass EM, Wilkening J, Wilke A, Antonopoulos D & Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.* **5:**

Gombeer S, Ramond J-B, Eckardt FD, Seely M & Cowan DA (2015) The influence of surface soil physicochemistry on the edaphic bacterial communities in contrasting terrain types of the Central Namib Desert. *Geobiology*

Goudie AS (2002) Great Warm Deserts of the World: Landscapes and Evolution Oxford University Press Available at: http://books.google.com/books?id=v911L1sq1FkC&pgis=1

Green VS, Stott DE & Diack M (2006) Assay for fluorescein diacetate hydrolytic activity: Optimization for soil samples. *Soil Biol. Biochem.* **38:** 693–701

Hansen MC, Tolker-Nielsen T, Givskov M & Molin S (1998) Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiol. Ecol.* **26:** 141–149

Hebda LM & Foran DR (2015) Assessing the Utility of Soil DNA Extraction Kits for Increasing DNA Yields and Eliminating PCR Inhibitors from Buried Skeletal Remains. *J. Forensic Sci.*

Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M & Breitbart M (2014) Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *ISME J.***:** 1–11 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24694711

Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30:** 772–780

Kim K-H & Bae J-W (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77:** 7663–8

Labonté JM & Suttle C a. (2013) Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Front. Microbiol.* **4:** 1–11

Lancaster N & Seely MK (1984) central Namib Desert. **14:**

Lancaster NL (1984) CUm ate of the central Namib Desert.

Makhalanyane T, Valverde A, Lacap D, Pointing S, Tuffin M & Cowan D (2013) Evidence of species recruitment and development of hot desert hypolithic communities. *Environ. Microbiol. Rep.* **5:** 219–224

Makhalanyane TP, Valverde A, Gunnigle E, Frossard A, Ramond J-BJ-B & Cowan DA (2015) Microbial ecology of hot desert edaphic systems. *FEMS Microbiol. Rev.* **39:** 203–221 Available at: http://femsre.oxfordjournals.org/cgi/doi/10.1093/femsre/fuu011

Pansu M (2006) Handbook of Soil Analysis

Pearce DA, Newsham KK, Thorne MAS, Calvo-Bado L, Krsek M, Laskaris P, Hodson A & Wellington EM (2012) Metagenomic analysis of a southern maritime Antarctic soil. *Front. Microbiol.* **3:** 1–13

Pointing SB, Warren-Rhodes K a, Lacap DC, Rhodes KL & McKay CP (2007) Hypolithic community shifts occur as a result of liquid water availability along environmental gradients in China's hot and cold hyperarid deserts. *Environ. Microbiol.* **9:** 414–24 Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17222139\nhttp://www.ncbi.nlm.nih.gov/pubmed/17222139

Prestel E, Regeard C, Salamitou S, Neveu J & Dubow MS (2013) The bacteria and bacteriophages from a Mesquite Flats site of the Death Valley desert. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **103:** 1329–1341

Prestel E, Salamitou S & Dubow MS (2008) An examination of the bacteriophages and bacteria of the Namib desert. *J. Microbiol.* **46:** 364–372

Prigent M, Leroy M, Confalonieri F, Dutertre M & DuBow MS (2005) A diversity of bacteriophage

forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles* **9:** 289–296

Quaiser A, Dufresne A, Ballaud F, Roux S, Zivanovic Y, Colombet J, Sime-Ngando T & Francez A-J (2015) Diversity and comparative genomics of Microviridae in Sphagnum- dominated peatlands. *Front. Microbiol.* **6:** 1–10 Available at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00375

Reavy B, Swanson MM, Cock P, Dawson L, Freitag TE, Singh BK, Torrance L, Mushegian AR & Taliansky M (2015) Distinct circular ssDNA Viruses Exist in Different Soil Types. *Appl. Environ. Microbiol.***:** AEM.03878–14 Available at: http://aem.asm.org/lookup/doi/10.1128/AEM.03878-14

Reysenbach A, Pace N, Robb F & Place A (1995) Archaea: a laboratory manual—thermophiles Cold Spring Harbour Laboratory Press

Rhoades JD (1982) Methods of Soil Analysis Part 2 Soil Science Society of America Madison. *Wisconsin, USA* **238:**

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F & Mira A (2009) Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7:** 828–836

Roux S, Krupovic M, Poulet A, Debroas D & Enault F (2012) Evolution and diversity of the microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* **7:** 1–12

Roux S, Tournayre J, Mahul A, Debroas D & Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15:** 76 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4002922&tool=pmcentrez&rendertype=abstract

Schulte EE & Hopkins BG (1996) Estimation of soil organic matter by weight loss-on- ignition. *Soil Org. Matter Anal. Interpret.* **049:** 21–31

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T & Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* **159:** 349–357

Stomeo F, Valverde A, Pointing SB, McKay CP, Warren-Rhodes K a., Tuffin MI, Seely M & Cowan D a. (2013) Hypolithic and soil microbial community assembly along an aridity gradient in the Namib Desert. *Extremophiles* **17:** 329–337

Suttle CA (2007) Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5:** 801–812

Tamura K, Stecher G, Peterson D, Filipski A & Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30:** 2725–2729

Thomas DSG (2011) Arid environments: their nature and extent. *Arid Zo. Geomorphol. Process. Form Chang. Drylands, Third Ed.***:** 1–16

Thurber RV (2009) Current insights into phage biodiversity and biogeography. *Curr. Opin. Microbiol.* **12:** 582–587

Warren-Rhodes K a., McKay CP, Boyle LN, Wing MR, Kiekebusch EM, Cowan D a., Stomeo F, Pointing SB, Kaseke KF, Eckardt F, Henschel JR, Anisfeld A, Seely M & Rhodes KL (2013) Physical ecology of hypolithic communities in the central Namib Desert: The role of fog, rain, rock habitat, and light. *J. Geophys. Res. Biogeosciences* **118:** 1451–1460

Warren-Rhodes KA, Rhodes KL, Boyle LN, Pointing SB, Chen Y, Liu S, Zhuo P & McKay CP (2007) Cyanobacterial ecology across environmental gradients and spatial scales in China's hot and cold deserts. *FEMS Microbiol. Ecol.* **61:** 470–482

Wei STS, Higgins CM, Adriaenssens EM, Cowan D a. & Pointing SB (2015) Genetic signatures indicate widespread antibiotic resistance and phage infection in microbial communities of the McMurdo Dry Valleys, East Antarctica. *Polar Biol.*

Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28:** 127–181

Williamson K, Corzo K, Drissi C, Buckingham J, Thompson C & Helton R (2013) Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. *Biol. Fertil. Soils* **49:** 857–869 Available at: http://dx.doi.org/10.1007/s00374-013-0780-z

Williamson KE, Radosevich M, Smith DW & Wommack KE (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9:** 2563–2574

Williamson KE, Radosevich M & Wommack KE (2005) Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol.* **71:** 3119–3125

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S & Nasko DJ (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.* **6:** 427–439

Wommack KE & Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64:** 69–114

Zablocki O, van Zyl L, Adriaenssens E, Rubagotti E, Tuffin M, Cary S & Cowan D (2014) High-level diversity of tailed Phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. *Appl. Environ. Microbiol.* **80:** 6888–6897

Chapter 6:
First report of a potyvirus infecting *Albuca rautanenii* in the Namib Desert

6.1 Chapter summary

To date, the occurrence of plant viruses in the Namib Desert has not been reported. A combination of electron microscopy and PCR were used to assess whether leaf mosaic symptoms observed on *Albuca rautanenii*, collected in the Namib Desert, were caused by a virus infection. Morphological observation from plant leaf extracts revealed flexuous, rod-shaped virus particles. PCR amplification of the nuclear inclusion b gene (a potyvirus marker gene) and phylogenetic analysis have revealed the isolated virus genotype represented a novel member within the genus *Potyvirus*, family *Potyviridae*.

6.2 Introduction and methods

Chlorotic, streak-like symptoms (Figure 21) were observed in April 2013 on a single specimen of *Albuca rautanenii* (Schinz) J.C.Manning & Goldblatt (Family: *Hyacinthaceae*) found among other plants near Homeb in the Namib Desert, Namibia. No potential insect vectors (e.g., aphids) were observed on or around the infected plant. An extract from symptomatic leaves was assessed by transmission electron microscopy (leaf dip method) to ascertain if the symptoms were viral in origin. Total RNA was extracted from symptomatic leaf tissue (Qiagen RNeasy Plant Mini Kit), and RT-PCR was conducted to amplify part of the nuclear inclusion b gene (Nib), a universal potyvirus signature gene (Zheng *et al*, 2010). The triplicated reaction yielded amplicons of the expected size (~350 bp), which were cloned into the pJET 1.2 vector (Thermo Scientific, Waltham, MA) according to manufacturer's instructions. The sequences of 10 clones were trimmed to remove vector and primer ends and were deposited in the EBI database under the accession numbers LK995422 to LK995431. Curated sequences were used to search the GenBank database using BLASTn and tBLASTx, as well as for phylogenetic analysis. Alignment and tree building parameters were performed according to Zheng et al. (2010).

6.3 Results and Discussion

TEM observations showed long, flexuous thread-like particles 687 to 825 nm in length and 12.5 nm in diameter were observed (Figure 21). The morphology and size of the particles were indicative of a putative member of the taxonomic family *Potyviridae*. For this reason, universal potyvirus primers were used to confirm morphological data.

Figure 21. Mosaic symptoms on the leaves of *Albuca rautanenii.*

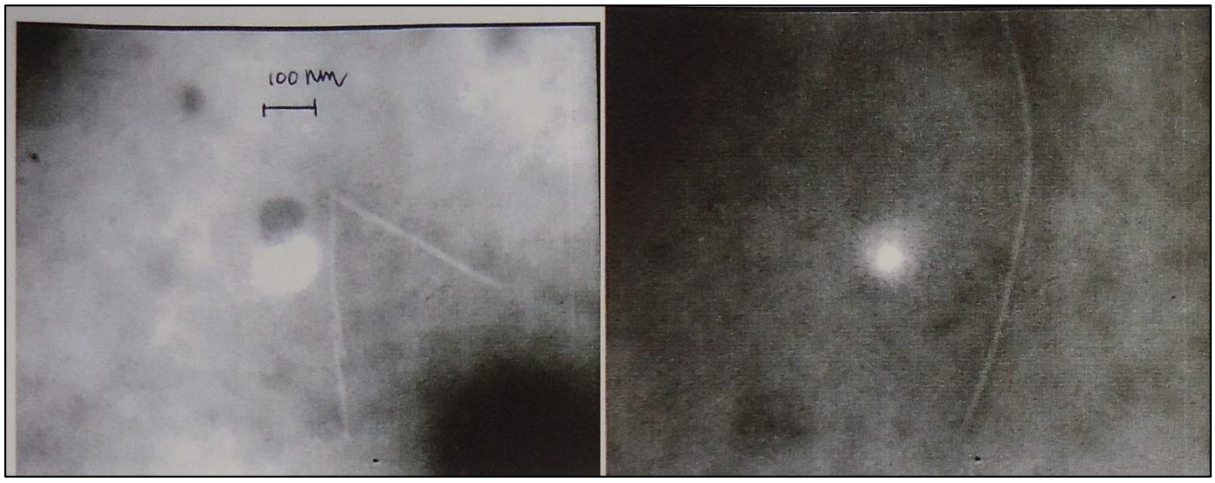

Figure 22. Virion photomicrographs (80,000 × magnification). The scale bar on the left photograph represents 100 nanometres. This scale also applies to the photograph on the left.

Intraclonal nucleotide sequence similarity ranged from 97.99 to 99.72%. BLASTn searches showed all clones were 72% identical to Papaya ringspot virus isolate 1 accession JQ314105.1 (87% coverage),

followed by Bean yellow mosaic virus clone Brn167 accession JF707769.1 (72% identity with 86% sequence coverage). The translated peptide fragment was most similar to Sugarcane mosaic virus isolate Beijing (AY042184.1), with a query cover of 98% and a similarity of 81%. Phylogenetic analysis was performed with a set of 57 reference potyvirus genomes, with their NIb regions aligned with the cloned nucleotide sequences. The clones formed a distinct cluster, at a node with Cocksfoot streak virus (CSV) (NC_003742.1). An identity matrix of the aligned NIb clones and CSV showed a nucleotide identity range of 68.79 to 70.23%. These results suggest that the virus isolate belongs to the family *Potyviridae*, genus *Potyvirus*, supported by the characteristic morphological features of the virion and its relatedness to CSV. Moreover, the clustering of all sequences at a single node suggests a homogeneous viral population, without significant strain variation. Genetic distance inferred by phylogenetic analysis further suggests that the isolate is a novel species within the genus, which we tentatively name *Albuca* mosaic virus, AlbMV. To our knowledge, this is the first report of any plant virus infection in the native Namib Desert ecosystem. This is particularly relevant due to the scarcity and uniqueness of plant life in this hyperarid desert environment, and additional monitoring of this virus infection and other desert plant species is encouraged.

6.4 References

Zheng L, Rodoni BC, Gibbs MJ & Gibbs a. J (2010) A novel pair of universal primers for the detection of potyviruses. *Plant Pathol.* **59:** 211–220

Chapter 7:

Concluding remarks

7.1 Introduction

The most numerous and genetically diverse members of the Earth's biosphere are most certainly viruses. Yet, despite the demonstrated ubiquity of viruses in a wide range of environments, there remain substantial gaps in our understanding of their ecological roles, taxonomic diversity and distribution. From the wealth of information on the ecology of viruses in the oceans (Suttle, 2007, 2005), it has been demonstrated that viruses have key roles within the global hydrosphere (see Chapter 2). However, it is still a matter of debate whether the same ecological importance holds true for viruses within terrestrial ecosystems. To provide a starting point in answering this question, this study aimed to explore the taxonomic diversity and function of viruses in a major terrestrial ecosystem: the desert ecosystem. Two hyperarid, thermally- contrasted desert biomes were studied: the Dry Valleys of Antarctica and the Namib Desert in Namibia. Prior to this research project, only very limited data were available on the presence of viruses in these ecosystems. Therefore, this study sought to answer two main research questions:

1) What is the viral diversity in hypolith communities found underneath rocks and surrounding surface soils in the Dry Valleys of Antarctica?
2) What is the viral diversity associated with surface soils of the Namib Desert, and how is community structure influenced by environmental gradients across the desert ecosystem?

The analysis of desert soils using a NGS-based metaviromics approach has provided an answer to both research questions. The implications of these answers is discussed in Section 7.2. This study has highlighted the effects of insufficiently populated virus sequence databases. To expand virus databases with annotated virus genomes, a novel laboratory method is proposed. The protocol permits the screening and isolation of environmental viruses, filtered by host range (see Section 7.3). It is also argued that in order to gain an ecosystem-wide understanding on the ecological roles of viruses, analysis of prokaryotic viruses alone is insufficient, and should include eukaryotic viruses (Section 7.4). Lastly, the key findings and implications gathered throughout this study are summarized in Section 7.5.

7.2 Implications of results

The research hypotheses postulated in Chapter 2 (section 2.6) were all demonstrated to be true. The first hypothesis predicted that the majority of metavirome sequences would belong to bacterial viruses. Because of limited animal and plant life in hyperarid deserts, desert environments are considered microbially-dominated (Geyer *et al*, 2013; Makhalanyane *et al*, 2015). Indeed, whether

in the soil metaviromes of the Namib Desert or the Dry Valleys, 77-96% of the identified virus fraction were to viruses of dominant desert bacterial groups (and surprisingly, close to none to archaea). In both deserts, the second most dominant virus group was the large dsDNA viruses which infect micro-eukaryotes (e.g. mimiviruses which infect amoeba). This demonstrated that a poorly characterized pool of micro-eukaryotes (including fungi, insects and unicellular eukaryotes) were present in these soils, and have not been extensively described (Fell *et al*, 2006; Vyverman *et al*, 2010; De Wever *et al*, 2009; Arenz & Blanchette, 2011). Thus, metaviromics not only serves as a virus discovery tool, but can also be used as a proxy for the assessment of host species diversity and their possible impact of soil-associated foodwebs. This approach may be useful in cases where the recovery of specific host species is not easily achievable.

The second hypothesis predicted that it would be possible to assemble novel virus genomes. Virus discovery through whole genome *de novo* assembly was most successful in the metaviromes of Antarctic soils, particularly within the hypolith dataset. The study showed that a contig contained chimeric genomic regions, belonging to viruses which infect bacteria and algae. This could indicate that accidental gene acquisition between distantly-related viruses is possible, albeit at a probable very low frequency of occurrence. Furthermore, full-genome assembly allowed the estimation of the proportion of temperate and virulent bacteriophages within the virus fraction. This has important ecological implications, because, as our results from the Antarctic metaviromes indicated, most assembled virus genomes (61.5%) were temperate (i.e., able to follow the lysogenic cycle). Yet, as the sampling strategy was designed to recover extracellular virus particles, it could indicate that a lytic lifestyle in cold hyperarid soils predominates, despite a majority of temperate bacteriophages. Indeed, a previous study (Williamson *et al*, 2007) has indicated that Antarctic soils contained a much lower prevalence of bacterial lysogens in comparison to hot desert soil microbial communities (~20% vs. ~84%). Collectively, these observations suggest that viruses in Antarctic soils are actively infecting their microbial hosts. This has important implications in the cycling of nutrients in these soils, mediated by bacterial lysis. This is the first indication that a viral shunt process (viral lysis- mediated release of dissolved organic matter) may play a role in nutrient cycling in cold hyperarid soils.

The third hypothesis predicted that a large number of metavirome sequence would have no similarity to sequences in public databases, as a result of novel virus gene pools. In all metaviromes analysed for this project, 60-80% of sequences were completely unknown. This is a common occurrence in many metavirome studies (Hatfull, 2015), and represents one of the biggest limitations in ecological metaviromics. Further accumulation of unknown environmental metagenomic sequence data will likely not help assigning taxonomic information to new sequences. Thus, future NGS-based studies, on the short term, will likely encounter the same problem (i.e., large number of unknown sequences). A methodological change is required in order to make high-throughput sequencing taxonomically informative. In the section below, a laboratory method termed "environmental lysis assay" is proposed, to increase the number of annotated, full-length phage genomes in public databases.

7.3 Achieving the full potential of NGS-based metaviromics

Given the estimated very large number (~10^{31}) of viruses in the biosphere (Suttle, 2005; Ashelford *et al*, 2003), the ability to sequence thousands of virus genomes from one sample in a single experiment seemed the ideal technique for gaining in-depth insights into the vast viral genomic pool. While NGS-based metaviromics has proven this potential, its application has also revealed how little is known about environmental virus communities. This project has demonstrated that the sequence data generated from high-throughput metaviromics is only as informative as the sequence databases it uses for taxonomic assignment. To reduce the number of unknown sequences, an obvious solution would be to first populate sequence databases with annotated sequences. Achieving this goal by generating more metavirome data from random amplification will not solve this problem. Instead, it is suggested that, before a high-throughput sequencing approach is considered, the most important factor to be taken into account is the *history* of the sample. In the context of metaviromics, history is referred to as in terms of having prior knowledge on the virus composition to be sequenced (e.g. if a sample is dominated by Proteobacteria species, then a range of phages infecting members of this group can be predicted to be in the metavirome data). This does not mean that environmental samples cannot be used. Furthermore, in Chapter 2, it was emphasized that amplification of metaviromes leads to several levels of bias, and that methods should be developed to avoid an amplification step prior to NGS.

A method called "environmental lysis assay" is proposed. This method would enable: 1) high-throughput screening of environmental viruses, 2) the filtering of a virus extract according to host range, and 3) virus genome amplification without PCR. The principle of operation is depicted in Figure 23. A virus extract (purified from an environmental sample) is deposited into micro-wells (similar in size to a 96-well plate) containing a liquid pure culture of a bacterial species, representative of a user-selected genus (e.g., *Bacillus*). Each well contains a single representative bacterial strain. Prior to virus inoculation into the wells, the optical density (OD) of each well is measured, so as to obtain a baseline measurement with regards to the number of bacterial cells present in each well. Once baseline OD measurement are recorded, the virus concentrate is inoculated into each well (each well is inoculated with the same virus extract). Following incubation, if virus lysis occurs in a given well, the number of bacterial cells will decrease, which can be empirically determined by subsequent OD measurements. If a well contains a successfully infected bacterial population, this effectively achieves the *in vivo* amplification of a given virus that may have been originally present in low titre in the original virus extract. This step therefore removes the need for PCR amplification. Once a well is successfully "infected", further virus concentration and DNA extraction can be performed. This method allows to reasonably predict that, if for example, the "*Bacillus* well" is infected, then the extracted viruses from this well will most likely be in majority composed of *Bacillus* phages. Knowing the history of the extract now permits enhanced sequence annotation if the extract is subjected to NGS. The DNA from every well may be combined into one sequencing tube, if individual "well DNAs" are bar-coded.

However, there are several major limitations to this method. It is designed for the recovery

and preliminary identification of extracellular viruses only. Therefore, environmental viruses in a prophage state will not be detected. There is also bias in terms of bacterial species selected in the assay, as well as their ability to grow *in vitro*. However, even if a limited number of bacterial strain are tested, the technique still has the potential for the characterization of many phages. As an example, a similar methodology (using conventional plaque assays, in petri dishes) has proven successful at isolating a large number of *Mycobacterium* phages by using a single representative *Mycobacterium* species (Pope & Hatfull, 2015). Despite the use of a single bacterial type strain, this method has allowed considerable accumulation of well-annotated *Mycobacterium* phage genomes in the NCBI RefSeq database (a total of 246 genomes, accounting for 15% from the total number of bacteriophage genomes). Furthermore, the lysis assay has the potential to provide virus morphological and genomic information across a wide range of culturable bacterial host species. It would surely represent a huge leap forward if a large number of viruses which infect the known culturable bacterial fraction was to be characterized. Furthermore, the lysis assay could be extended to other organisms such as micro-eukaryotes (e.g., phytoplankton and protists).

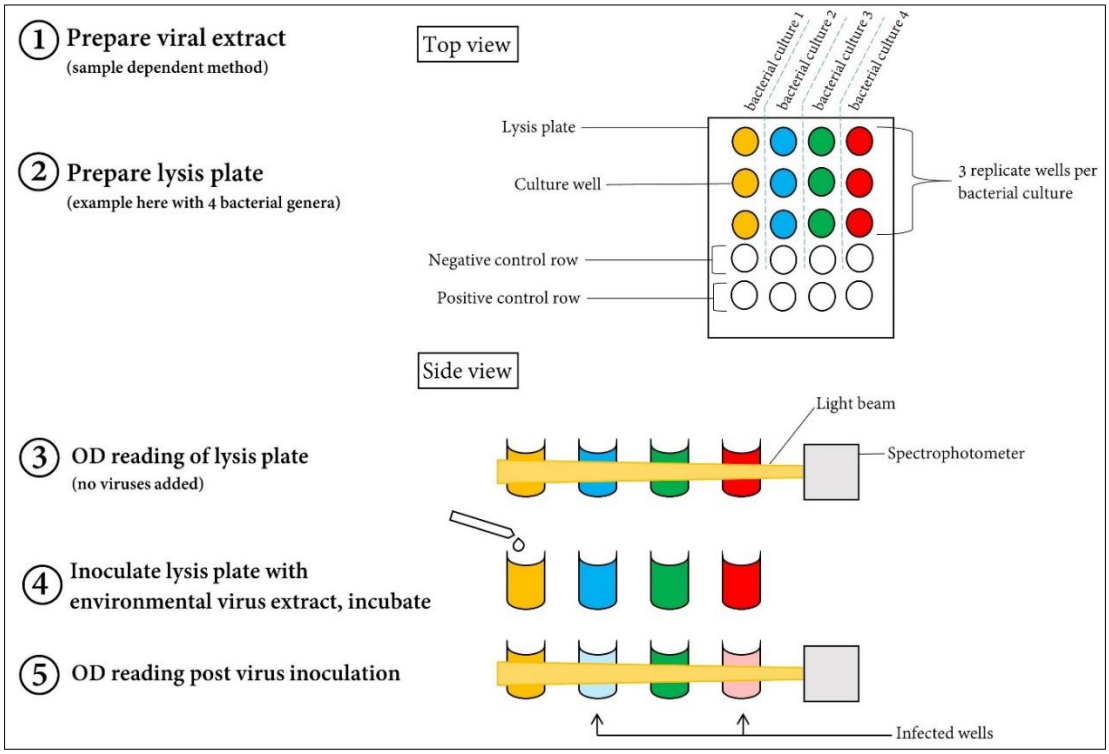

Figure 23. Overview of the lysis assay method. First, a virus environmental extract is prepared. Second, the lysis plate is prepared by inserting into each well a pure bacterial culture (in this example, four separate cultures, each represented by a different colour). Each pure culture has five wells associated with it. Three replicate wells are used for the inoculation of the virus extract. The fourth well is the negative control, which only contains distilled water and the virus extract. This ensures the detection of external or cross-sample contamination, and this well should remain clear throughout the experiment. The fifth well contains only the bacterial culture, free of virus inoculation. Third, all the wells are measured in terms of optical density (OD). Fourth, the environmental virus extract is inoculated in the replicate wells of each bacterial pure culture. Following a defined incubation period, OD is re-measured in order to determine if loss of bacterial cell density has occurred, indicative of a successful virus infection. Further virus characterization can then be conducted (electron microscopy, virus DNA extraction, and sequencing).

7.4 Beyond bacterial viruses

This study has demonstrated that prokaryotic viruses constituted the majority of known sequences in both the Namib and Antarctic desert soil samples. However, I argue that, in order to gain an ecosystem-wide understanding on the roles of viruses, additional representative species such a micro-eukaryotes should be investigated in future research. In this study, eukaryotic virus sequences were found in very low numbers, probably caused by an inappropriate sampling method, not originally intended to recover eukaryotic viruses. In the hyperarid Antarctic Dry Valley soils for example, a wide diversity of eukaryotic taxa including mosses, lichens, microarthropods (e.g. mites), nematodes, fungi, protozoa, tardigrades, rotifers and unicellular algae has been described (Mcknight *et al*, 1998; Powers *et al*, 1998; Treonis *et al*, 1999; Stevens & Hogg, 2003; Robinson *et al*, 2003). Collectively, these organisms contribute to biogeochemical and ecological processes in temperate soil systems (Paul, 2014). However the contribution of these species in cold desert soils is largely unknown (Wall & Virginia, 1999; Wall, 2005). The putative impacts viruses may have on the lifecycles of these organisms are yet to be investigated.

Unlike the Antarctic Dry Valleys, the Namib Desert is capable of sustaining the growth of higher vascular plants (Green *et al*, 2012; Juergens *et al*, 2013). In Chapter 6, a potyvirus infecting *Albuca rautanenii* was described. It is most likely that the potyvirus was carried by an aphid vector, as most plant viruses require an intermediate insect vector for virus-host transmission (Fereres & Raccah, 2015; Whitfield *et al*, 2015). Therefore, the presence of viruses carried by soil-borne eukaryotes has the potential to influence the survival and distribution of plants in the Namib Desert. This is especially relevant in a low biodiversity ecosystem. Therefore, additional monitoring of plant and insect viruses within plant-colonized areas of the Namib Desert is strongly encouraged.

The analysis of viruses of additional soil-associated organisms (e.g. micro-eukaryotes, insects and plants) in low biodiversity environments such as hyperarid deserts may provide insights into local host distribution and/or biotic interactions, to analyses of ecosystem functions, and will certainly contribute to our understanding of the global phylogenetic diversity of viruses.

7.5 Conclusion

This thesis has provided an initial assessment of the diversity and function of soil viruses in two hyperarid deserts. The analysis of viruses at the community level using metagenomics heavily depends on the quality and quantity of virus sequences in sequence databases. It is expected that the gradual addition of annotated virus genomes to databases will eventually allow the metaviromics approach to estimate the true virus diversity within an environmental sample. It is therefore a real possibility that within a short time span (~10-15 years), as a result of expanded virus genome reference databases, the diversity estimations and associated ecological inferences made in this study will be totally different. Finally, the key research findings of this project are summarized below:

- Both deserts contained a majority of tailed bacteriophages, spanning members of the three *Caudovirales* families. Also, to a lesser extent, eukaryotic viruses of algae, amoeba and insects were detected.
- Soil viruses in all desert soil samples analysed (including hypoliths) displayed strong sample specificity, reflective of host distribution and probable limited dispersal. In Namib Desert soils, edaphic properties do not appear to significantly influence the spatial distribution of virus groups.
- Viral metagenomics can be used as proxy for the estimation of host diversity in soils (e.g. bacteria, archaea and micro-eukaryotes).
- An ecosystem-wide understanding on the ecological roles of viruses in the soil environment requires more than the analysis of prokaryotic viruses alone.
- Enhanced soil virus isolation techniques are required to populate sequence databases.

7.6 References

Arenz BE & Blanchette R a. (2011) Distribution and abundance of soil fungi in Antarctica at sites on the Peninsula, Ross Sea Region and McMurdo Dry Valleys. *Soil Biol. Biochem.* **43:** 308–315 Available at: http://dx.doi.org/10.1016/j.soilbio.2010.10.016

Ashelford KE, Day MJ & Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. *Appl. Environ. Microbiol.* **69:** 285–289

Fell JW, Scorzetti G, Connell L & Craig S (2006) Biodiversity of micro-eukaryotes in Antarctic Dry Valley soils with< 5% soil moisture. *Soil Biol. Biochem.* **38:** 3107–3119

Fereres A & Raccah B (2015) Plant Virus Transmission by Insects. *eLS*: 1–12 Available at: http://doi.wiley.com/10.1002/9780470015902.a0000760.pub3

Geyer KM, Altrichter AE, Van Horn DJ, Takacs-Vesbach CD, Gooseff MN & Barrett JE (2013) Environmental controls over bacterial communities in polar desert soils. *Ecosphere* **4:** art127 Available at: http://www.esajournals.org/doi/abs/10.1890/ES13-00048.1

Green AG, Brabyn L, Beard C & Sancho LG (2012) Extremely low lichen growth rates in Taylor Valley, Dry Valleys, continental Antarctica. *Polar Biol.* **35:** 535–541 Available at: http://link.springer.com/10.1007/s00300-011-1098-7

Hatfull GF (2015) Dark matter of the biosphere: The amazing world of bacteriophage diversity. *J. Virol.*: JVI–01340

Juergens N, Oldeland J, Hachfeld B, Erb E & Schultz C (2013) Ecology and spatial patterns of large-scale vegetation units within the central Namib Desert. *J. Arid Environ.* **93:** 59–79 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0140196312002522

Makhalanyane TP, Valverde A, Gunnigle E, Frossard A, Ramond J-BJ-B & Cowan DA (2015) Microbial ecology of hot desert edaphic systems. *FEMS Microbiol. Rev.* **39:** 203–221 Available at: http://femsre.oxfordjournals.org/cgi/doi/10.1093/femsre/fuu011

Mcknight DM, Alger A, Tate C, Shupe G & Spaulding S (1998) Longitudinal patterns in algal abundance and species distribution in meltwater streams in Taylor Valley, Southern Victoria Land, Antarctica Wiley Online Library

Paul EA (2014) Soil microbiology, ecology and biochemistry Academic press

Pope WH & Hatfull GF (2015) Adding pieces to the puzzle: New insights into bacteriophage diversity from integrated research-education programs. *Bacteriophage***:** 0

Powers LLE, Ho M, Freckman DWD & Virginia RRA (1998) Distribution, Community Structure, and microhabitats of soil invertebrates alaong an elevational gradient in Taylor valley, Antarctica. *Arct. Alp. Res.* **30:** 133–141 Available at: http://www.jstor.org/stable/1552128

Robinson SA, Wasley J & Tobin AK (2003) Living on the edge - Plants and global change in continental and maritime Antarctica. *Glob. Chang. Biol.* **9:** 1681–1717

Stevens MI & Hogg ID (2003) Long-term isolation and recent range expansion from glacial refugia revealed for the endemic springtail Gomphiocephalus hodgsoni from Victoria Land, Antarctica. *Mol. Ecol.* **12:** 2357–2369

Suttle CA (2005) Viruses in the sea. *Nature* **437:** 356–361

Suttle CA (2007) Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5:** 801–812

Treonis AM, Wall DH & Virginia RA (1999) Invertebrate biodiversity in Antarctic Dry Valley soils and sediments. *Ecosystems* **2:** 482–492

Vyverman W, Verleyen E, Wilmotte A, Hodgson D a., Willems A, Peeters K, Van de Vijver B, De Wever A, Leliaert F & Sabbe K (2010) Evidence for widespread endemism among Antarctic micro-organisms. *Polar Sci.* **4:** 103–113

Wall DH (2005) Biodiversity and ecosystem functioning in terrestrial habitats of Antarctica. *Antarct. Sci.* **17:** 523

Wall DH & Virginia RA (1999) Controls on soil biodiversity: Insights from extreme environments. In *Applied Soil Ecology* pp 137–150.

De Wever A, Leliaert F, Verleyen E, Vanormelingen P, Van der Gucht K, Hodgson DA, Sabbe K & Vyverman W (2009) Hidden levels of phylodiversity in Antarctic green algae: further evidence for the existence of glacial refugia. *Proc. Biol. Sci.* **276:** 3591–9 Available at: http://rspb.royalsocietypublishing.org/content/early/2009/07/20/rspb.2009.0994.abstract

Whitfield AE, Falk BW & Rotenberg D (2015) Insect vector-mediated transmission of plant viruses. *Virology* **479-480:** 278–289 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0042682215001622

Williamson KE, Radosevich M, Smith DW & Wommack KE (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9:** 2563–2574

APPENDIX I: Thesis outputs

1. Publications

- **Zablocki, O.**, Rybicki, E., and Cowan, D.A., 2014. First report of a potyvirus infecting *Albuca rautanenii* in the Namib Desert. *Plant Disease* 98(12):1749.

- **Zablocki, O.**, van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, C., and Cowan, D.A., 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(22):6888-6897.

- **Zablocki, O.,** van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, C., and Cowan, D.A., 2014. Niche-dependent genetic diversity in Antarctic metaviromes. *Bacteriophage* 4(4):e980125.

- **Zablocki, O.**, Adriaenssens, E.M., Cowan, D.A., 2015. Diversity and ecology of viruses in hyperarid desert soils. *Applied and Environmental Microbiology* (mini-reviews). Manuscript in production, DOI: 10.1128/AEM.02651-15

- **Zablocki, O.**, Adriaenssens, E.M., Ramond, J-B., Frossard, A.,Vikram, S., Scola, V., Seely, M. and Cowan, D.A., 2015. Extracellular virus diversity and biogeography in Namib Desert soils. Manuscript submitted.

2. Conference outputs

- **Olivier Zablocki**, Evelien M. Adriaenssens, Pieter De Maayer, Lonnie van Zyl, Marla I. Tuffin, Mary Seely and Don Cowan. 2013. Metaviromics of extreme terrestrial habitats. *2nd International Conference on Functional Metagenomics*. Poster presentation.

- **Olivier Zablocki**, Lonnie van Zyl, Evelien Adriaenssens, Enrico Rubagotti, Marla Tuffin, Craig Cary and Don Cowan. 2014. Viruses from a cold hyperarid desert: a comparative metagenomic analysis of hypolithic and open soil biotopes. *15th International Symposium on Microbial Ecology*. Oral presentation.

- **Olivier Zablocki**, Lonnie van Zyl, Evelien Adriaenssens, Enrico Rubagotti, Marla Tuffin, Craig Cary and Don Cowan. 2014. High-level virus diversity in Antarctic soils. *Genome Research Institute (University of Pretoria) seminar day*. Poster presentation.

- **Olivier Zablocki**, Evelien M. Adriaenssens, Jean-Baptiste Ramond, Aline Frossard, Surendra Vikram, Vincent Scola, Mary Seely, and Don Cowan. 2015. Viral diversity and biogeography in the hyperarid Namib Desert. *Virology Africa 2015*. Oral presentation.

- **Olivier Zablocki**, Evelien M. Adriaenssens, Jean-Baptiste Ramond, Aline Frossard, Surendra Vikram, Vincent Scola, Mary Seely, and Don Cowan. 2015. Viral diversity and biogeography in Namib Desert soils. *Genome Research Institute (University of Pretoria) seminar day*. Oral presentation.

Future conference attendance/presentations related to this research project:

- **Olivier Zablocki**, Evelien M. Adriaenssens, Jean-Baptiste Ramond, Aline Frossard, Surendra Vikram, Vincent Scola, Mary Seely, and Don Cowan. 2016. Viral diversity and biogeography in Namib Desert soils. *16th International Symposium on Microbial Ecology*. Poster/Oral presentation to be determined.

Diversity and ecology of viruses in hyperarid desert soils

Olivier Zablocki, Evelien Adriaenssens and Don Cowan

Centre for Microbial Ecology and Genomics, University of Pretoria, South Africa[a]

Abstract

In recent years, remarkable progress has been made in the field of virus environmental ecology. In marine ecosystems for example, viruses are now thought to play pivotal roles in the biogeochemical cycling of nutrients and to be mediators of microbial evolution through horizontal gene transfer. In soils, the diversity and ecology of viruses is poorly understood, but evidence supports the view that these differ substantially from aquatic systems. Desert biomes cover ~33% of global land masses, yet the diversity and roles of viruses in these dominant ecosystems remain poorly understood. There is evidence that hot hyperarid desert soils are characterised by high levels of bacterial lysogens and low extracellular virus counts. In contrast, cold desert soils contain high extracellular virus titres. We suggest that the prevalence of microbial biofilms in hyperarid soils, combined with extreme thermal regimes, constitute strong selection pressures on both temperate and virulent viruses. Many desert soil virus sequences show low identity values to virus genomes in public databases, suggesting the existence of distinct and as yet uncharacterised soil phylogenetic lineages (e.g. cyanophages). We strongly advocate for amplification-free metavirome analyses while encouraging the classical isolation of phages from dominant and culturable microbial isolates in order to populate sequence databases. This review provides an overview of recent advances in the study of viruses in hyperarid soils, the factors that contribute to viral

abundance and diversity in hot and cold deserts and suggests technical recommendations for future studies.

Introduction

Over recent decades, the critical roles that viruses play in the environment have become increasingly recognized by the research community (1). It has been estimated by direct counts of extracellular ('free floating') virus-like particles (VLPs) that the global "virosphere" may contain up to 10^{31} viral particles (2), suggesting that viruses may be the most abundant biological entities on the planet and, potentially, the greatest reservoir of genetic diversity (3–5). The ecological importance of viruses on a global scale has predominantly emerged from studies of marine and fresh water microbial communities (6–12), where viruses have been linked to core processes such as biogeochemical nutrient cycling (6, 7, 10), microbial population control through viral lysis (7, 8) and microbial evolution via horizontal gene transfer (11).

Research on the virus ecology of soil environments has progressed more slowly and has received proportionally less attention (12–14). However, enumeration of virus particles by electron microscopy (EM) on several soil types (15–17) has shown high viral abundance values ranging from 1.5 $x10^8$ to 6.4 $x10^8$ per gram dry weight soils. Soil ecosystems are subject to unique abiotic ecological pressures, in part due to their wide compositional spectrum and spatial heterogeneity in terms of physicochemical properties (18, 19). Environmental stresses are even greater in extremely arid soil systems, where soil organisms and communities may be simultaneously exposed to pulsed water events, and to the effects of desiccation-, solute- and UV-B radiation induced oxidative-stresses (20, 21). Deserts represent the single largest terrestrial ecosystem type on Earth, covering ~33.6% of the global land mass, excluding Antarctica (22), and are classified in terms of their aridity index, a ratio between precipitation (P) and potential of evapotranspiration (PET) (23). This results in four

desert categories, as dry-semiarid (0.5<P/PET<0.65), semiarid (0.2<P/PET<0.5), arid (0.05<P/PET<0.2) and hyperarid (P/PET<0.05). Hyperarid deserts generally receive annual precipitation of ≤ 70 mm and are often associated with intrinsic characteristics such as high pH (~7-9), high salinity levels, high surface radiation fluxes, long periods of desiccation and low water activity (24). Desert soil microbial ecology research has primarily focused on bacterial communities, which have been shown to be largely responsible for primary production and the provision of key ecosystem services (25–28). Soil virus populations and functions are seldom taken into consideration, thereby omitting a crucial variable within ecological models designed to predict microbial population dynamics. As a result, the ecological roles, survival mechanisms (against biotic and abiotic factors), the spatial and temporal changes in viral community structures (virus biogeography) and viral phylogenetic diversity, are still poorly understood in desert soils.

Within the field of soil virus ecology (13), several desert soil ecosystems have been recently investigated (Table 1). With the advances in next generation sequencing (NGS) technologies, culture-independent methods have become the standard for determination of viral diversity (29). However, the rapidly growing volume of viral environmental sequence data has revealed that most sequences (~70%) have no homologs in public databases, and are typically labelled “viral dark matter” (30, 31). Here, we discuss the current understanding of hot and cold desert soil virus diversity and function, propose alternative technical approaches to virus concentration methods and identify key areas of future research.

Diversity and abundance of viruses in desert soils

Hot deserts. Viral community analyses have been conducted on surface soil samples from three hot hyperarid deserts: the Sahara (32), Namib (33, 34) and Mojave (35, 36). In each of these studies, difficulties in detecting extracellular VLPs by electron microscopy (EM) or pulse field gel electrophoresis (PFGE) profiling were reported, suggesting a very low

viral abundance within these soils. However, the inclusion of a lytic induction step (prophage excision stimulated by the addition of Mitomycin C (37)) in the soil extraction protocol substantially increased the recovery of virus particles (32, 33). For Sahara Desert surface sand samples, induced phage genomes were estimated to range in size from 45 to 270 kb. Electron microscopy (EM) of the induced phage fraction showed a majority of tailed virus morphotypes belonging to the *Myoviridae* family, some of which showed peculiar ribbon-like structures located at the tail tip of the virions (38). In the Namib Desert soil samples, twenty distinct morphotypes were identified, all members of the *Myoviridae* and *Siphoviridae* families with no apparent *Podoviridae*-like virions (33). PFGE profiles from Namib soils indicated an average genome size of 55 to 65 kb, with several genomes of up to 350 kb in size (33). EM visualisation of Mojave Desert sand samples showed eleven distinct tailed morphotypes, belonging to all three families of the *Caudovirales* (36).

Sanger sequencing of randomly selected cloned phage fragments from the Mojave Desert soil virus communities showed that 36% of sequenced clones had no homologs in public sequence databases (36). Within the identified virus sequences, the majority were homologous to bacteriophages infecting common soil bacteria such as members of the Proteobacteria, including *Bacillus* and *Rhizobium*. From the same samples, 38 bacterial isolates were grown in pure culture and 84% were shown to harbour at least one SOS-inducible phage. A similar study on loamy sand from a different area of the Mojave Desert showed that a large majority of randomly selected metaviral clone sequences had no database homologs (35). Of those clones with significant sequence identity (tBLASTx search using an E-value cut-off of 10^{-3}), phages associated with *Actinoplanes*, *Mycobacterium*, *Myxococcus* and *Streptomyces* were the most common. Other virus signals detected included archaeal (*Haloarcula* phage) and herpes-like viruses. Using a similar methodology, 50% of the viral sequences from three Namib Desert surface sand samples had no homologs in public

sequence databases, with most positive hits showing homology to *Siphoviridae* phages linked to Gram-positive bacteria (33). Most recently, a shotgun NGS approach was used to investigate the metavirome of Namib Desert hypoliths (34), cyanobacteria-dominated microbial niche communities on the ventral surfaces of translucent rocks (39). The most abundant sequences belonged to *Geobacillus*- and *Bacillus*-infecting phages, while cyanophage markers were unexpectedly found only in low numbers. The distinct phylogenetic clustering of assembled *phoH* genes (a cyanophage marker (40)) suggested that desert soil cyanophages were only distantly related to their well-studied marine counterparts (34), and that the dominance of marine cyanophage sequences in sequence databases might account for the low cyanophage hit rate of homologous sequences in the Namib Desert hypolithon metavirome. This observation has wider implications for studies of soil metaviromics, where an underestimation of cyanophage abundance and diversity may skew estimates of the functional importance (and population dynamics) of soil cyanobacteria, arguably the most important taxonomic group in desert soil microbial communities (27, 28).

Cold deserts. Studies of viral communities in cold hyperarid desert soils have been almost exclusively conducted in the major ice-free regions of Antarctica (e.g., the East Antarctic McMurdo Dry Valleys). Direct viral counts by epifluorescence microscopy (17) showed high VLP densities, in the range of 2.3 - 6.4 × 10^8 extracellular VLPs per gram of dry soil. The prevalence of bacterial lysogens within these soils was between 4.6 and 21.1%, a much lower occurrence level than estimated for bacteria in hot desert soils (84% (35)). Using epifluorescence direct counts of extractable bacteria and extracellular virus particles, virus-to-bacteria ratios (VBR) ranging from 170 to 8200 were calculated, the highest recorded for any soil ecosystem (17).

Antarctic soil bacterial isolates have yielded several unique virus genomic structures.

The distinct temperate siphoviruses (SpaA1 and BceA1) isolated from *Staphylococcus pasteuri* and *Bacillus cereus* both contained almost complete additional phage genomes (MZTP02) (41). This "Russian doll" gene arrangement had not been previously described for soil bacteriophages, and has led to speculation that it may represent a 'fast-track' route for virus evolution and horizontal gene transfer, with a possible role in host range expansion.

Pyrosequencing of Antarctic soil metagenomic DNA has identified a wide diversity of bacteria, archaea, microeukaryotes and viruses (42). From the total sequence dataset, 494 phage-related hits (0.18% of the total number of sequences) were identified. Top BLAST hits against public databases were related to phages known to infect to *Mycobacteria*, *Burkholderia*, *Bordetella*, *Pseudomonas*, *Enterobacteria*, *Flavobacterium*, *Myxococcus*, *Synechococcus*, *Prochlorococcus* and *Sinorhizobium*. However, viral DNA was not specifically enriched in this study, and this may have resulted in an underestimation of viral diversity.

The spatial composition and dynamics of viral communities along an Antarctic soil transect have been recently reported (43). Using random PCR amplification of polymorphic DNA (RAPD-PCR) assays, viral community fingerprints were used to assess short-term changes in the composition of viral communities. To maximize the number of viruses sequences amplified, RAPD-PCR primer design was based on the identification of recurring dodecamer sequences (G+C content ≥70%) within 22 selected viral metagenomes. Qualitative comparisons of the Antarctic fingerprint patterns demonstrated that heterogeneous soil conditions and associated environmental factors (e.g., carbon levels, moisture content, pH and light exposure frequency) impacted the composition of viral assemblages across geographic distances as short as 20 metres. The RAPD-PCR fingerprint data also suggested that virus assemblages were not present as inactive, inert particles, but were dynamically involved in infection of co-existing microbial hosts. Furthermore, the

authors suggested that environmental pressures (e.g. low moisture) known to influence bacterial community structures in the Antarctic desert (17) were shown to have a similarly influential role on virus community dynamics.

Abundance estimates (17) suggest that Antarctic desert soils contain a substantially higher proportion of free extracellular VLPs than hot hyperarid desert soils, where a lysogenic lifestyle appears to be prevalent (32, 33, 36). A sequence-based metagenomic comparison of viral assemblages (single- and double-stranded DNA viruses only) in surface soils and hypolithic communities in the Antarctic McMurdo Dry Valleys (44) demonstrated that bacteriophages constituted the majority of the identified viruses, representing all *Caudovirales* families. *Mycobacterium* phage sequences were the most highly represented in the viral fraction (42). No archaeal virus sequences were recorded, in line with previous observations that archaea are either absent or present in very low numbers in this environment (26, 45). Within the hypolith metavirome dataset, the fraction of cyanophage sequences was under-represented, with low sequence similarities to known cyanophages. Dry Valley surface soils also contained a number of other virus signatures, including phycodnaviruses, mimiviruses and virophage capsid protein genes (44), many of which are most commonly identified in aquatic systems.

Factors shaping viral community structures in desert soils

Soil virus populations display different dynamics from marine and freshwater systems (43) (Figure 1). In marine systems, two major factors influence viral abundance: the biological productivity of the system and microbial diversity and abundance (3, 5). Viral abundance has been shown to increase as bacterial productivity in a system increases (46). Co-occurring virus host communities also influence viral abundance, as in microbial bloom events, which increase the number of lytic infections thereby releasing additional phage particles (47, 48). Marine-associated abiotic parameters such as temperature, salinity and pH

are stable on relatively large spatial scales (13), and do not appear to significantly affect viral abundance.

Soil systems, particularly desert soils, are inhomogeneous, in that soil particles are semi-discrete. Extended periods of desiccation and oligotrophy are typical characteristics of hyperarid desert soils. Under these environmental constraints, microbial populations often form discrete biofilms, where cells embedded in EPS matrices are adsorbed to particle surfaces., (49–51). The EPS matrix serves a protective role, sequesters nutrients and provides a defence barrier against virulent phages (52). Temperate phages in their prophage state have been shown to stabilize biofilms, whereas a switch to the lytic cylce aids in biofilm dispersalWithin the biofilm, temperate phages have been shown to contribute to the lifecycle of biofilm by aiding in biofilm dispersal (52). While this has been described for Pseudomonas aeruginosa biofilms (53), we argue that a similar mechanism may be present in hyperarid desert soil biofilms. Such a mechanism would drive the positive selection of temperate phages in this ecosystem and negatively influence the presence of virulent phages. This suggestion is consistent with the observation that extracellular VLPs are only readily extracted after induction of prophages in hot desert soils (32, 33, 36, 54).

In Antarctic hyperarid desert soils, where biofilm communities also frequently occur (55), high VLP counts have been recorded. We suggest that the effects of temperature may explain the apparent difference between hot and cold desert soil systems. Temperature has been shown to be one of the major factors controlling viral survival rates in soils (56, 57), with lower temperatures enabling longer survival rates, extended latent periods and reduced burst sizes. Warmer temperatures have been associated with reduced virus proliferation and greater inactivation rates (58, 59). Thus, in Antarctic soils, colder temperatures may allow for the preservation of extracellular VLPs, making them more abundant and detectable (60). In

198 contrast, the high temperature regimes (e.g., maxima of ≥ 50°C in the Namib Desert (45)) of
199 hot deserts may increase the rate of degradation of extracellular virus particles.

200 Viral operational taxonomic unit (OTU) abundance estimates from low-throughput
201 Sanger sequencing of metaviromes have provided some insights into the factors that shape
202 the diversity of viral communities in desert soils (35). Comparisons of viral community
203 compositions across three contrasting soil ecosystems (prairie, desert and rainforest) have
204 demonstrated that microbial communities were both locally and globally diverse.
205 Comparative phylogenetic analyses showed little taxonomic overlap between soils sampled
206 from the three different habitats, as well as low identity values to annotated sequences in
207 public databases. However, the factors that may be responsible for the observed niche
208 specialization are, as yet, unknown.

209 Similar habitat-specific viral community compositions have been reported through the
210 use of hierarchical clustering of metaviromes, based on dinucleotide frequencies (61). This
211 method is especially useful for gaining ecological insights from metagenomic datasets
212 containing a majority of unaffiliated reads to public databases. Dinucleotide frequencies
213 within metaviromes have showed distinct virus community clustering within single habitat
214 types such as desert soils. Although reported from a single study which analysed two sets of
215 pooled samples, cluster analysis of hypolith and open soil metaviromes from Antarctic and
216 Namib Desert soil has shown that both hypolith metaviromes clustered at a single node while,
217 conversely, both open soil metaviromes displayed an identical pattern (62). Despite the great
218 geographic distances or differing environmental conditions, similar habitat types harboured
219 more closely related viral communities. The most obvious common factor between the two
220 contrasting deserts is very limited water availability, which may be a key driver of
221 community speciation and recruitment in these soils.

222 **Technical recommendations and future research**

223 Research on desert soil viruses is technically challenging, partly due to the physical
224 properties of soil. Desert soils frequently produce sub-optimal viral DNA yields (≤10 ng/µl)
225 (63), forcing the inclusion of a random PCR amplification step for NGS library construction.
226 The use of whole genome amplification (WGA) by multiple displacement amplification
227 (MDA) or random-priming, sequence independent, single primer amplification (RP-SISPA)
228 (64) almost certainly results in biased amplification of certain virus groups (65–68) and
229 prevents the accurate determination of viral abundances and diversity. While viral
230 amplification is widely accepted as a necessity in metaviromic studies (69), we argue that
231 amplification of virus metagenomic DNA should be avoided where possible. It would be
232 preferable to focus efforts on improving virus concentration methods in order to reach the
233 minimum concentrations required for sequencing. Sequential rounds of centrifugation and
234 the pooling of samples should increase the number of viruses recovered. Methodological
235 improvements in virus concentration would also allow for more precise virus counts using
236 microscopy (70). Thus, the development of more efficient and effective metaviromic DNA
237 extraction technologies, so as to obviate the need for WGA, would represent a substantial
238 advance in the field. This goal is further facilitated by recent technical improvements in
239 sequencing chemistries where, for example, Illumina paired-end sequencing library
240 construction kits have reduced the minimum genomic DNA requirement to around 50
241 picograms (ThruPLEX, Rubicon Genomics).

242

243 Sequence-based identification of viral communities, using either multiple gene
244 markers (71) or full virome sequencing (72), is becoming more routine. In marine virus
245 ecology, the use of conserved viral marker genes such as DNA polymerases (73),

ribonucleotide reductases (74) and T4-related structural proteins (75, 76) has provided detailed data on viral biodiversity, on intra- and inter-viral evolutionary relationships and on oceanic viral turnover rates. The use of these methods to study virus diversity and biogeography in desert soils is relatively new and most commonly involves the sequencing of whole metaviromes (34, 43, 44). However, metaviromic approaches generally result in a large number of unknown sequences (31). In addition, we warn that the taxonomic affiliation of single genes and/or virus genome fragments (using BLAST against public databases) in metavirome datasets may not be evidence for the presence of these viruses in the sample (77), and should be carefully inspected by additional read mapping to a reference genome. While a metaviromics approach provides the opportunity for virus discovery, it may be also valuable to use, in parallel, a high-throughput sequencing approach focusing on conserved signature genes. Such a combinatorial strategy could provide both informative data on viral richness and insights into the functional roles of viruses in soil ecosystems.

A common feature of many desert ecosystems is the occurrence of hypolithic niches (78). These rock-associated cryptic microbial communities are usually dominated by photosynthetic cyanobacteria, but contain a wide diversity of members of the phyla Actinobacteria, Acidobacteria, Bacteroidetes and Proteobacteria (26, 79–81). Cyanobacteria are of particular importance, due to their key roles in primary productivity and nitrogen input in depauperate ecosystems (27, 28). To date, no fully characterized desert soil-associated cyanophage isolates have been reported. Preliminary metagenomic data on Antarctic and Namib hypoliths (34, 44) have shown evidence of novel soil cyanophage lineages, the sequences of which have low identities to characterized marine cyanophage genomes. As cyanobacteria are readily amenable to culturing (82), this provides opportunities for the isolation of their phages and access to full-length soil cyanophage genomes. Such data would

support downstream applications such as primer design for targeted amplification of related taxa, and monitoring of these assemblages within desert soil ecosystems.

Conclusion

Research on phage ecosystem ecology in hyperarid desert soils has demonstrated that desert soil viruses are numerous, diverse, and encode novel genes whose function are yet to be determined. In order to understand how viruses contribute to desert soil ecosystem functioning, critical research questions, addressing both micro- and macro-scale issues, must be addressed. The microscale complexity of the soil matrix drives the distribution, maintenance, metabolic state and biodiversity of microbial communities (83). Consequently, investigating the dynamics of virus-host interactions at the microscale level will contribute significantly to our understanding on the factors which determine virus distribution and diversity of bacteriophages in soil systems. In addition, the effects of extreme physicochemical conditions (e.g., intense UV radiation and temperature) on the preservation of virion particles and the kinetics of virus decay remain unexplored in hyperarid deserts. At the macroscale level (i.e., ecosystem-scale), the contributions of viruses to ecosystem services such as nutrient cycling, energy flow and the sequestration of nutrients remain open questions. It would also be highly informative to understand the kinetics and scales of virus transport processes within and between hyperarid ecosystems, potentially important factors in understanding phage phylogeography.

Acknowledgements

This work was funded by the South African National Research Foundation and the University of Pretoria Genomics Research Institute. EMA is funded by the Claude Leon Foundation Postdoctoral Fellowship program.

References

1. **Rohwer F**, **Prangishvili D**, **Lindell D**. 2009. Roles of viruses in the environment. Environ Microbiol **11**:2771–2774.

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

3. **Weinbauer MG**, **Rassoulzadegan F**. 2004. Are viruses driving microbial diversification and diversity? Environ Microbiol **6**:1–11.

4. **Frost LS**, **Leplae R**, **Summers AO**, **Toussaint A**. 2005. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol **3**:722–732.

5. **Dennehy JJ**. 2013. What Ecologists Can Tell Virologists. Annu Rev Microbiol 117–135.

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

7. **Wommack KE**, **Colwell RR**. 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev **64**:69–114.

8. **Bettarel Y**, **Sime-Ngando T**, **Amblard C**, **Dolan J**. 2004. Viral activity in two contrasting lake ecosystems. Appl Environ Microbiol **70**:2941–2951.

9. **Winget DM**, **Williamson KE**, **Helton RR**, **Wommack KE**. 2005. Tangential flow diafiltration: An improved technique for estimation of virioplankton production. Aquat Microb Ecol **41**:221–232.

10. **Suttle CA**. 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol **5**:801–812.

11. **Rodriguez-Valera F**, **Martin-Cuadrado A-B**, **Rodriguez-Brito B**, **Pasić L**, **Thingstad TF**, **Rohwer F**, **Mira A**. 2009. Explaining microbial population genomics through phage predation. Nat Rev Microbiol **7**:828–836.

12. **Breitbart M**. 2012. Marine Viruses: Truth or Dare. Ann Rev Mar Sci.

13. **Kimura M**, **Jia ZJ**, **Nakayama N**, **Asakawa S**. 2008. Ecology of viruses in soils:

Past, present and future perspectives. Soil Sci Plant Nutr **54**:1–32.

14. **Srinivasiah S**, **Bhavsar J**, **Thapar K**, **Liles M**, **Schoenfeld T**, **Wommack KE**. 2008. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Res Microbiol **159**:349–357.

15. **Ashelford KE**, **Day MJ**, **Fry JC**. 2003. Elevated abundance of bacteriophage infecting bacteria in soil. Appl Environ Microbiol **69**:285–289.

16. **Williamson KE**, **Radosevich M**, **Wommack KE**. 2005. Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol **71**:3119–3125.

17. **Williamson KE**, **Radosevich M**, **Smith DW**, **Wommack KE**. 2007. Incidence of lysogeny within temperate and extreme soil environments. Environ Microbiol **9**:2563–2574.

18. **Schlesinger WH**, **Reynolds JF**, **Cunningham GL**, **Huenneke LF**, **Jarrell WM**, **Virginia RA**, **Whitford WG**. 1990. Biological feedbacks in global desertification. Science **247**:1043–1048.

19. **Palmer TM**. 2003. Spatial habitat heterogeneity influences competition and coexistence in an African acacia ant guild. Ecology **84**:2843–2855.

20. **Austin AT**, **Yahdjian L**, **Stark JM**, **Belnap J**, **Porporato A**, **Norton U**, **Ravetta DA**, **Schaeffer SM**. 2004. Water pulses and biogeochemical cycles in arid and semiarid ecosystems. Oecologia **141**:221–235.

21. **Chen LZ**, **Wang GH**, **Hong S**, **Liu A**, **Li C**, **Liu YD**. 2009. UV-B-induced oxidative damage and protective role of exopolysaccharides in desert cyanobacterium Microcoleus vaginatus. J Integr Plant Biol **51**:194–200.

22. **Meigs P**. 1952. Arid and semiarid climatic types of the world, p. 135–138. *In* Proceedings, VIII General Assembly and XVII International Congress, International Geographical Union, Washington DC.

23. **UNEP**. 2013. Global environment outlook 2000. Routledge.

24. **Shmida A**, **Wilson M**. 1985. Biological Determinants of Species Diversity. J Biogeogr **12**:1–20.

25. **Harel Y**, **Ohad I**, **Kaplan A**. 2004. Activation of photosynthesis and resistance to photoinhibition in cyanobacteria within biological desert crust. Plant Physiol

136:3070–3079.

26. **Pointing SB**, **Chan Y**, **Lacap DC**, **Lau MCY**, **Jurgens JA**, **Farrell RL**. 2009. Highly specialized microbial diversity in hyper-arid polar desert. Proc Natl Acad Sci U S A **106**:19964–19969.

27. **Tracy CR**, **Streten-Joyce C**, **Dalton R**, **Nussear KE**, **Gibb KS**, **Christian KA**. 2010. Microclimate and limits to photosynthesis in a diverse community of hypolithic cyanobacteria in northern Australia. Environ Microbiol **12**:592–607.

28. **Cowan D a.**, **Pointing SB**, **Stevens MI**, **Craig Cary S**, **Stomeo F**, **Tuffin IM**. 2011. Distribution and abiotic influences on hypolithic microbial communities in an Antarctic Dry Valley. Polar Biol **34**:307–311.

29. **Rosario K**, **Breitbart M**. 2011. Exploring the viral world through metagenomics. Curr Opin Virol **1**:289–297.

30. **Youle M**, **Haynes M**, **Rohwer F**. 2012. Scratching the surface of biology's dark matter, p. 61–81. *In* Viruses: Essential agents of life. Springer.

31. **Hatfull GF**. 2015. Dark matter of the biosphere: The amazing world of bacteriophage diversity. J Virol JVI.01340–15.

32. **Prigent M**, **Leroy M**, **Confalonieri F**, **Dutertre M**, **DuBow MS**. 2005. A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. Extremophiles **9**:289–296.

33. **Prestel E**, **Salamitou S**, **Dubow MS**. 2008. An examination of the bacteriophages and bacteria of the Namib desert. J Microbiol **46**:364–372.

34. **Adriaenssens EM**, **Van Zyl L**, **De Maayer P**, **Rubagotti E**, **Rybicki E**, **Tuffin M**, **Cowan DA**. 2015. Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ Microbiol **17**:480–495.

35. **Fierer N**, **Breitbart M**, **Nulton J**, **Salamon P**, **Lozupone C**, **Jones R**, **Robeson M**, **Edwards R a.**, **Felts B**, **Rayhawk S**, **Knight R**, **Rohwer F**, **Jackson RB**. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol **73**:7059–7066.

36. **Prestel E**, **Regeard C**, **Salamitou S**, **Neveu J**, **Dubow MS**. 2013. The bacteria and bacteriophages from a Mesquite Flats site of the Death Valley desert. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol **103**:1329–1341.

37. **Ackermann H-W**, **DuBow MS**. 1987. Viruses of prokaryotes. CRC press.

38. **Prestel E**, **Regeard C**, **Andrews J**, **Oger P**, **DuBow MS**. 2012. A Novel Bacteriophage Morphotype with a Ribbon-like Structure at the Tail Extremity. Res J Microbiol.

39. **Makhalanyane TP**, **Valverde A**, **Lacap DC**, **Pointing SB**, **Tuffin MI**, **Cowan D a.** 2013. Evidence of species recruitment and development of hot desert hypolithic communities. Environ Microbiol Rep **5**:219–224.

40. **Goldsmith DB**, **Crosti G**, **Dwivedi B**, **McDaniel LD**, **Varsani A**, **Suttle CA**, **Weinbauer MG**, **Sandaa RA**, **Breitbart M**. 2011. Development of phoH as a novel signature gene for assessing marine phage diversity. Appl Environ Microbiol **77**:7730–7739.

41. **Swanson MM**, **Reavy B**, **Makarova KS**, **Cock PJ**, **Hopkins DW**, **Torrance L**, **Koonin E V.**, **Taliansky M**. 2012. Novel bacteriophages containing a genome of another bacteriophage within their genomes. PLoS One **7**.

42. **Pearce DA**, **Newsham KK**, **Thorne MAS**, **Calvo-Bado L**, **Krsek M**, **Laskaris P**, **Hodson A**, **Wellington EM**. 2012. Metagenomic analysis of a southern maritime Antarctic soil. Front Microbiol **3**:1–13.

43. **Srinivasiah S**, **Lovett J**, **Polson S**, **Bhavsar J**, **Ghosh D**, **Roy K**, **Fuhrmann JJ**, **Radosevich M**, **Wommack KE**. 2013. Direct assessment of viral diversity in soils by random PCR amplification of polymorphic DNA. Appl Environ Microbiol **79**:5450–5457.

44. **Zablocki O**, **van Zyl L**, **Adriaenssens EM**, **Rubagotti E**, **Tuffin M**, **Cary SC**. 2014. High-level diversity of tailed phages, eukaryote-associated viruses and virophage-like elements in the metaviroms of Antarctic soils. Appl Environ Microbiololgy **80**:6888–6897.

45. **Makhalanyane TP**, **Valverde A**, **Birkeland N-K**, **Cary SC**, **Tuffin IM**, **Cowan DA**. 2013. Evidence for successional development in Antarctic hypolithic bacterial communities. ISME J **7**:2080–90.

46. **Maranger R**, **Bird DF**. 1995. Viral abundance in aquatic systems: A comparison between marine and fresh waters. Mar Ecol Prog Ser **121**:217–226.

47. **Ortmann a. C**, **Lawrence JE**, **Suttle C a.** 2002. Lysogeny and lyric viral production

during a bloom of the cyanobacterium Synechococcus spp. Microb Ecol **43**:225–231.

48. **Bratbak G**, **Heldal M**, **Norland S**, **Thingstad T**. 1990. Viruses as partners in spring bloom microbial trophodynamics. Appl Environ Microbiol **56**:1400–1405.

49. **Davey ME**, **O'toole GA**. 2000. Microbial biofilms: from ecology to molecular genetics. Microbiol Mol Biol Rev **64**:847–867.

50. **Donlan RM**. 2002. Biofilms: Microbial life on surfaces. Emerg Infect Dis.

51. **Gorbushina AA**. 2007. Life on the rocks. Environ Microbiol.

52. **McDougald D**, **Rice S a.**, **Barraud N**, **Steinberg PD**, **Kjelleberg S**. 2011. Should we stay or should we go: mechanisms and ecological consequences for biofilm dispersal. Nat Rev Microbiol **10**:39–50.

53. **Rice SA**, **Tan CH**, **Mikkelsen PJ**, **Kung V**, **Woo J**, **Tay M**, **Hauser A**, **McDougald D**, **Webb JS**, **Kjelleberg S**. 2009. The biofilm life cycle and virulence of Pseudomonas aeruginosa are dependent on a filamentous prophage. ISME J **3**:271–282.

54. **Thurber RV**. 2009. Current insights into phage biodiversity and biogeography. Curr Opin Microbiol **12**:582–587.

55. **Aguilera A**, **Souza-Egipsy V**, **Amils R**. 2012. Photosynthesis in extreme environments. Artif Photosynth InTech Eur Rijeka 271–288.

56. **Hurst CJ**, **Gerba CP**, **Cech I**. 1980. Effects of environmental variables and soil characteristics on virus survival in soil. Appl Environ Microbiol **40**:1067–1079.

57. **Nasser AM**, **Oman SD**. 1999. Quantitative assessment of the inactivation of pathogenic and indicator viruses in natural water sources. Water Res **33**:1748–1752.

58. **Straub TM**, **Pepper IL**, **Gerba CP**. 1992. Persistence of viruses in desert soils amended with anaerobically digested sewage sludge. Appl Environ Microbiol **58**:636–641.

59. **Leonardopoulos J**, **Papaconstantinou A**, **Georgakopoulou-Papandreou E**. 1996. The meaning of soil characteristics and temperature for the survival of bacteriophages in naturally contaminated soil samples. ACTA Microbiol Hell **41**:309–316.

60. **Williamson KE**. 2011. Soil phage ecology: abundance, distribution, and interactions with bacterial hosts, p. 113–136. *In* Biocommunication in Soil Microorganisms. Springer.

61. **Willner D**, **Thurber RV**, **Rohwer F**. 2009. Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol **11**:1752–1766.

62. **Zablocki O**, **van Zyl L**, **Adriaenssens EM**, **Rubagotti E**, **Tuffin M**, **Cary SC**, **Cowan D**. 2014. Niche-dependent genetic diversity in Antarctic metaviromes. Bacteriophage **4**:e980125.

63. **Kim KH**, **Bae JW**. 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol **77**:7663–7668.

64. **Weynberg KD**, **Wood-Charlson EM**, **Suttle C a.**, **van Oppen MJH**. 2014. Generating viral metagenomes from the coral holobiont. Front Microbiol **5**:1–11.

65. **Kim K-H**, **Chang H-W**, **Nam Y-D**, **Roh SW**, **Kim M-S**, **Sung Y**, **Jeon CO**, **Oh H-M**, **Bae J-W**. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol **74**:5975–5985.

66. **Polson SW**, **Wilhelm SW**, **Wommack KE**. 2011. Unraveling the viral tapestry (from inside the capsid out). ISME J **5**:165.

67. **Karlsson OE**, **Belák S**, **Granberg F**. 2013. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. Biosecur Bioterror **11 Suppl 1**:S227–34.

68. **Rosseel T**, **Van Borm S**, **Vandenbussche F**, **Hoffmann B**, **van den Berg T**, **Beer M**, **Höper D**. 2013. The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. PLoS One **8**.

69. **Delwart EL**. 2007. Viral metagenomics. Rev Med Virol.

70. **Williamson KE**, **Wommack KE**, **Radosevich M**. 2003. Sampling Natural Viral Communities from Soil for Culture-Independent Analyses. Appl Environ Microbiol **69**:6628–6633.

71. **Adriaenssens EM**, **Cowan DA**. 2014. Using signature genes as tools to assess environmental viral ecology and diversity. Appl Environ Microbiol **80**:4470–4480.

542

543 72. **Sullivan MB**. 2015. Viromes, Not Gene Markers, for Studying Double-Stranded DNA
544 Virus Communities. J Virol **89**:2459–2461.
545

546 73. **Chen F**, **Suttle CA**, **Short SM**. 1996. Genetic diversity in marine algal virus
547 communities as revealed by sequence analysis of DNA polymerase genes. Appl
548 Environ Microbiol **62**:2869–2874.
549

550 74. **Sakowski EG**, **Munsell E V**, **Hyatt M**, **Kress W**, **Williamson SJ**, **Nasko DJ**, **Polson**
551 **SW**, **Wommack KE**. 2014. Ribonucleotide reductases reveal novel viral diversity and
552 predict biological and ecological features of unknown marine viruses. Proc Natl Acad
553 Sci **111**:15786–15791.
554

555 75. **Hambly E**, **Tétart F**, **Desplats C**, **Wilson WH**, **Krisch HM**, **Mann NH**. 2001. A
556 conserved genetic module that encodes the major virion components in both the
557 coliphage T4 and the marine cyanophage S-PM2. Proc Natl Acad Sci U S A
558 **98**:11411–11416.
559

560 76. **Short CM**, **Suttle CA**. 2005. Nearly identical bacteriophage structural gene sequences
561 are widely distributed in both marine and freshwater environments. Appl Environ
562 Microbiol **71**:480–486.
563

564 77. **Emerson JB**, **Thomas BC**, **Andrade K**, **Allen EE**, **Heidelberg KB**, **Banfielda JF**.
565 2012. Dynamic viral populations in hypersaline systems as revealed by metagenomic
566 assembly. Appl Environ Microbiol **78**:6309–6320.
567

568 78. **Thomas DSG**. 2011. Arid environments: their nature and extent. Arid Zo Geomorphol
569 Process Form Chang Drylands, Third Ed 1–16.
570

571 79. **Schlesinger WH**, **Pippen JS**, **Wallenstein MD**, **Hofmockel KS**, **Klepeis DM**,
572 **Mahall BE**. 2003. Community composition and photosynthesis by photoautotrophs
573 under quartz pebbles, southern Mojave Desert. Ecology **84**:3222–3231.
574

575 80. **Warren-Rhodes KA**, **Rhodes KL**, **Pointing SB**, **Ewing SA**, **Lacap DC**, **Gómez-**
576 **Silva B**, **Amundson R**, **Friedmann EI**, **McKay CP**. 2006. Hypolithic cyanobacteria,
577 dry limit of photosynthesis, and microbial ecology in the hyperarid Atacama Desert.
578 Microb Ecol **52**:389–398.
579

580 81. **Warren-Rhodes KA**, **Rhodes KL**, **Boyle LN**, **Pointing SB**, **Chen Y**, **Liu S**, **Zhuo P**,
581 **McKay CP**. 2007. Cyanobacterial ecology across environmental gradients and spatial
582 scales in China's hot and cold deserts. FEMS Microbiol Ecol **61**:470–482.
583

82. **Garcia-Pichel F**, **López-Cortés A**, **Nübel U**. 2001. Phylogenetic and Morphological Diversity of Cyanobacteria in Soil Desert Crusts from the Colorado Plateau. Appl Environ Microbiol **67**:1902–1910.

83. **Vos M**, **Wolf AB**, **Jennings SJ**, **Kowalchuk GA**. 2013. Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev **37**:936–954.

84. **Swanson MM**, **Fraser G**, **Daniell TJ**, **Torrance L**, **Gregory PJ**, **Taliansky M**. 2009. Viruses in soils: Morphological diversity and abundance in the rhizosphere. Ann Appl Biol **155**:51–60.

85. **Meiring TL**, **Marla Tuffin I**, **Cary C**, **Cowan D a.** 2012. Genome sequence of temperate bacteriophage Psymv2 from Antarctic Dry Valley soil isolate Psychrobacter sp. MV2. Extremophiles **16**:715–726.

86. **Srinivasiah S**, **Lovett J**, **Ghosh D**, **Roy K**, **Fuhrmann JJ**, **Radosevich M**, **Wommack KE**. 2015. Dynamics of autochthonous soil viral communities parallels dynamics of host communities under nutrient stimulation. FEMS Microbiol Ecol **91**:fiv063.

Author biographies

Olivier Zablocki received his B.Sc. and M.Sc. in Microbiology from the University of Pretoria, South Africa. During his M.Sc., he trained as a plant virologist, and participated in the implementation of NGS for plant disease diagnosis for the South African citrus industry, with a focus on Citrus tristeza virus. In 2013, he started a Ph.D. degree under the supervision of Prof. Don Cowan, the current director of the Centre for Microbial Ecology and Genomics at the University of Pretoria. For his thesis, he used both metaviromics and soil physicochemical analyses to assess virus community structure and dynamics in hyperarid desert soil ecosystems. He has recently joined the Institute for Microbial Biotechnology and Metagenomics at the University of the Western Cape (South Africa) to continue research in virus ecology using 'omics' strategies.

Evelien Adriaenssens received her B.Sc. and M.Sc. from the faculty of Bioscience Engineering of the University of Leuven (KU Leuven), Belgium. During her Ph.D. research,

she worked on the isolation and characterization of bacteriophages of the potato pathogen Dickeya solani for applications in plant protection. This was a collaborative effort between the Laboratory of Gene Technology of Professor Rob Lavigne (KU Leuven), the Plant Production Laboratory of Professor Maurice De Proft (KU Leuven) and the Unit Plant – Crop Protection of the Institute for Agricultural and Fisheries Research with Dr. Martine Maes. After obtaining the degree in 2012 she joined the Centre of Microbial Ecology and Genomics at the University of Pretoria, South Africa in 2013 to work as a Postdoctoral Fellow heading the viral metagenomics project, investigating hot and cold desert viral communities.

Don Cowan was educated in New Zealand at the University of Waikato and completed a period of postdoctoral study there before moving to University College London (UK) as a Lecturer in 1985. After 16 years in London, he accepted the position as Professor of Microbiology in the Department of Biotechnology at the University of the Western Cape, Cape Town, where he was a Senior Professor and Director of the Institute for Microbial Biotechnology and Metagenomics. In May 2012 he moved to the University of Pretoria as Director of both the Genomics Research Institute and the Centre for Microbial Ecology and Genomics. His research activities encompass a wide range of projects in the field of Ecogenomics: the use of genomic and metagenomic methods to understand the diversity and function of microorganisms in different environments.

Table 1. Mid-to high throughput soil-based studies pertaining to viral ecology since 2005.

Authors, Year, (Reference number)	Soil type	Location/country	Methods used
Prigent *et al.*, 2005 (32)	Hot desert surface sand	Sahara Desert in Morocco and Tunisia	EM[1], PFGE[2], Lytic induction
Williamson *et al*., 2005 (16)	Agricultural, forest	Delaware,USA	Epifluorescence microscopy, EM[1]

Fierer *et al*., 2007 (35)	Hot arid desert, tallgrass prairie, tropical rainforest	USA, Peru	Sanger sequencing of random viral metagenomics clones
Williamson *et al*., 2007 (17)	Loamy and sandy soils, agricultural, forested wetlands	Antarctica (Tom and Obelisk pond); USA (Delaware)	Induction assays, Epifluorescence counting
Prestel *et al*., 2008 (33)	Surface sand	Namib Desert	EM[1], PFGE[2], Sanger sequencing of cloned DNA fragments (LASL[3])
Swanson *et al*., 2009 (84)	Dystric-Fluvic Cambisol soil	Dundee, Scotland	EM[1], epifluorescence counting
Meiring *et al*., 2012 (85)	Soil underneath hypoliths	Miers Valley, Antarctica	Lytic induction, EM[1], phage isolation from culture
Pearce *et al*., 2012 (42)	Surface soil	Alexander Island, Antarctica	Shotgun metagenome pyrosequencing
Swanson *et al*., 2012 (41)	Surface soil (Antarctica)	Antarctica	EM[1], lytic induction, phage isolation
Prestel *et al*., 2013 (36)	Dune surface sand	Mojave Desert, USA	EM[1], random amplification for viral DNA (Sanger)
Srinivasiah *et al*., 2013 (43)	Surface soil (Antarctica); Silt loamy soil (USA)	Antarctica (Tom and Obelisk pond); Delaware, USA	RAPD[4] viral community fingerprinting
Adriaenssens *et al*., 2015 (34)	Soil-associated rocks (hypoliths)	Namib Desert	Shotgun viral metagenome sequencing (Illumina)

Zablocki *et al*., 2014 (44)	Antarctic surface soil and hypoliths	Miers Valley, Antarctica	Shotgun viral metagenome sequencing (Illumina)
Srinivasiah *et al*., 2015 (86)	Silt loamy soil	Delaware, USA	Microcosms, RAPD[4] viral community fingerprinting, epifluorescence counting

643 [1]EM: electron microscopy; [2]PFGE: pulse field gel electrophoresis; [3]: LASL: linker amplified shotgun library;
644 [4]RAPD: random amplified polymorphic DNA.

645

646 **Figure 1.** Virus community dynamics in aquatic (A) and soil (B) ecosystems. Marine and freshwater
647 systems can be regarded as homogenous systems, where the distribution of virus particles (e.g.
648 phages) and host organisms (e.g. bacterioplankton) is relatively even. Such a continuous medium
649 allows for rapid phage/host dispersion, and increases the rates of phage-host collisions, leading to
650 high infection rates (A). In contrast, hyperarid soil microbial communities exist as discrete systems,
651 embedded in protective biofilms (B). The level of virus-host interactions (VHI) within and between
652 individual biofilm communities remains an open question, but diffusion rates are expected to be low
653 on both small and large spatial scales.

654