Genetic analysis of bacteriophages from clinical and environmental samples

University College Cork
Coláiste na hOllscoile Corcaigh

A Thesis presented to the National University of Ireland for the Degree of Doctor of Philosophy
by

Kamila Knapik

Department of Microbiology
National University of Ireland
Cork

Head of Department: Prof. Gerald F. Fitzgerald

Supervisor: Prof. Michael Prentice

July 2013

Table of Contents

Declaration

This thesis is my own work and has not been submitted for another degree, either at University College Cork or elsewhere.

Kamila Knapik
5 July 2013

Acknowledgements

I would like to acknowledge my supervisor Professor Michael B. Prentice for giving me the opportunity and for his support and guidance throughout this study. I also would like to thank to my internal and external examiners Dr. Marcus Claesson (University College Cork) and Dr. Leonid Kulakov (Queen's University Belfast) for reviewing my thesis and for helpful discussion.

This study was funded by the Department of Microbiology, University College Cork. I would like to thank University College Cork for financial support and facilities.

I am also very grateful to many people who have helped me through the completion of this study. I thank Dr. Barry Plant (Cork University Hospital) and Kerry Ingredients & Flavours for gift of samples, and to Dr. Nicholas Thomson (Wellcome Trust Sanger Institute) for performing the Illumina sequencing. I also would like to thank Paweł Wasyłyszyn (Sr. Technical Support Engineer at EMC) and Dr. Florent Angly (Australian Centre for Ecogenomics) for their technical support in bioinformatics and Dr. Heinrich Lünsdorf (Helmholtz Centre for Infection Research) for electron microscopy.

Finally, I would like to thank to my family and friends. Thank you for your continuous support and putting your faith in me.

List of Tables

List of Figures

List of Abbreviations

aa	amino acid
BLAST	Basic Local Alignment Search Tool
BLASTn	BLAST (search a nucleotide database using a nucleotide query)
BLASTx	BLAST (search protein database using a translated nucleotide query)
bp	base pair
CsCl	Caesium chloride
CTAB	cetyltrimethylammonium bromide
Cm	chloramphenicol
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CDS	coding DNA sequence
CF	cystic fibrosis
DNase	Deoxyribonuclease
dNTP	deoxyribonucleotide 5‘‘-triphosphate
DTT	Dithiothreitol
ds	double-stranded
GOS	Global Ocean Sampling
gfp / GFP	green fluorescent protein (gene / protein)
HMM	Hidden Markov Model
h	hour(s)
Kn	kanamycin
kb	kilobase
kDa	kilodalton
l	litre
LB	Luria-Bertani medium
g23 / gp23	major capsid protein (gene / protein)
Mb	Megabases
µ	micro
µg	microgram
µl	microlitre(s)
µm	micrometre
ml	millilitre(s)

mM	millimolar
min	minute(s)
MDA	Multiple displacement amplification
TPEN	N,N,N′,N′-Tetrakis(2-pyridylmethyl)ethylenediamine
ng	nanogram
NCBI	National Centre of Biotechnology Information
NGS	Next-Generation Sequencing
nr	non-redundant
nt	nucleotides
ORF	Open Reading Frame
ppm	parts per million
PCR	polymerase chain reaction
Rep	replication-associated protein
RNase	Ribonuclease
rRNA	ribosomal RNA
rpm	rotations per minute
ss	single-stranded
SDS	sodium dodecyl sulphate
UV	ultra violet
VLP(s)	Virus-like particle(s)
v/v	volume/volume
w/v	weight/volume
°C	degree(s) Celsius

Abstract

Bacteriophages, viruses infecting bacteria, are uniformly present in any location where there are high numbers of bacteria, both in the external environment and the human body. Knowledge of their diversity is limited by the difficulty to culture the host species and by the lack of the universal marker gene present in all viruses. Metagenomics is a powerful tool that can be used to analyse viral communities in their natural environments. The aim of this study was to investigate diverse populations of uncultured viruses from clinical (a sputum of patient with cystic fibrosis, CF) and environmental samples (a sludge from a dairy food wastewater treatment plant) containing rich bacterial populations using genetic and metagenomic analyses. Metagenomic sequencing of viruses obtained from these samples revealed that the majority of the metagenomic reads (97-99%) were novel when compared to the NCBI protein database using BLAST. A large proportion of assembled contigs were assignable as novel phages or uncharacterised prophages, the next largest assignable group being single-stranded eukaryotic virus genomes. Sputum from a cystic fibrosis patient contained DNA typical of phages of bacteria that are traditionally involved in CF lung infections and other bacteria that are part of the normal oral flora. The only eukaryotic virus detected in the CF sputum was Torque Teno virus (TTV). A substantial number of assigned sequences from dairy wastewater could be affiliated with phages of bacteria that are typically found in the soil and aquatic environments, including wastewater. Eukaryotic viral sequences were dominated by plant pathogens from the *Geminiviridae* and *Nanoviridae* families, and animal pathogens from the *Circoviridae* family. Antibiotic resistance genes were detected in both metagenomes suggesting phages could be a source for transmissible antimicrobial resistance. Overall, diversity of viruses in the CF sputum was low, with 89 distinct viral genotypes predicted, and higher (409 genotypes) in the wastewater.

Function-based screening of a metagenomic library constructed from DNA extracted from dairy food wastewater viruses revealed candidate promoter sequences that have ability to drive expression of GFP in a promoter-trap vector in *Escherichia coli*. The majority of the cloned DNA sequences selected by the assay were related to ssDNA circular eukaryotic viruses and phages which formed a minority of the metagenome assembly, and many lacked any significant homology to known database sequences.

Natural diversity of bacteriophages in wastewater samples was also examined by PCR amplification of the major capsid protein sequences, conserved within T4-type bacteriophages from *Myoviridae* family. Phylogenetic analysis of capsid sequences revealed that dairy wastewater contained mainly diverse and uncharacterized phages, while some showed a high level of similarity with phages from geographically distant environments.

Chapter 1:

Literature review

1.1. Introduction to bacteriophages

1.1.1. Phage discovery and research

Bacteriophages were independently discovered in 1915 by Frederick Twort, who first described “glassy transformation” of micrococci colonies, and in 1917 by Felix d’Herelle, who noticed that a culture filtrate had the potential to kill bacteria (Duckworth, 1976). They were first imaged by electron microscopy in 1940 (Ackermann, 2011). Following their discovery, the potential of phages to kill bacteria was successfully exploited from 1919 onwards and was the basis of the earliest successful specific treatments of bacterial infections in humans (Sulakvelidze *et al.*, 2001). With the invention of antibiotics after the Second World War, phage therapy declined in the West, but continues in countries such as Russia, Georgia and Poland (Pirnay *et al.*, 2010; Sulakvelidze *et al.*, 2001). Recently, there has been worldwide renewed interest in phage therapy, due to concerns about the rise of antibiotic resistance among many strains of bacteria (Alisky *et al.*, 1998; Harper & Morales, 2012).

Bacteriophage functions in bacterial hosts were intensively studied in the decades after their discovery, and the knowledge gained formed much of the foundations of the modern molecular biology. For example, experiments conducted on bacteriophage T4 infecting *Escherichia coli* proved that DNA is a genetic material (Hershey & Chase, 1952), whereas other phage experiments revealed the triplet nature of the genetic code (Crick *et al.*, 1961). In 1977, the DNA of bacteriophage φX174 was the first fully sequenced genome to be reported (Sanger *et al.*, 1977a). The application of electron and fluorescence microscopy for enumeration of phage particles in the environment showed the abundance of phages in nature, including all aquatic environments (from the Antarctic to thermal springs at 80 °C), soil, and subsurface samples (Ackermann & Prangishvili, 2012; Bergh *et al.*, 1989; Noble & Fuhrman, 1998; Wen *et al.*, 2004). Recent advances in viral particle purification (Casas & Rohwer, 2007; Thurber *et al.*, 2009) and progress in sequencing technology have revealed that environmental bacteriophages are not only numerous in the environment but also extremely diversified with the vast majority of coding sequences identified being of unknown function (Clokie *et al.*, 2011).

1.1.2. Phage structure and classification

Bacteriophages are composed of nucleic acid protected by a protein or lipoprotein shell. The virion shape can be tailed, polyhedral, filamentous, or pleomorphic (Figure 1.1). The genome can be ssDNA, dsDNA, ssRNA or dsRNA, either linear or circular. The known genome sizes range from the smallest *Leuconostoc oenos* phage L5 (2,435 bp) (Hatfull, 2008) to the gigantic *Bacillus megaterium* phage G (497,513 bp) (Hatfull & Hendrix, 2011).

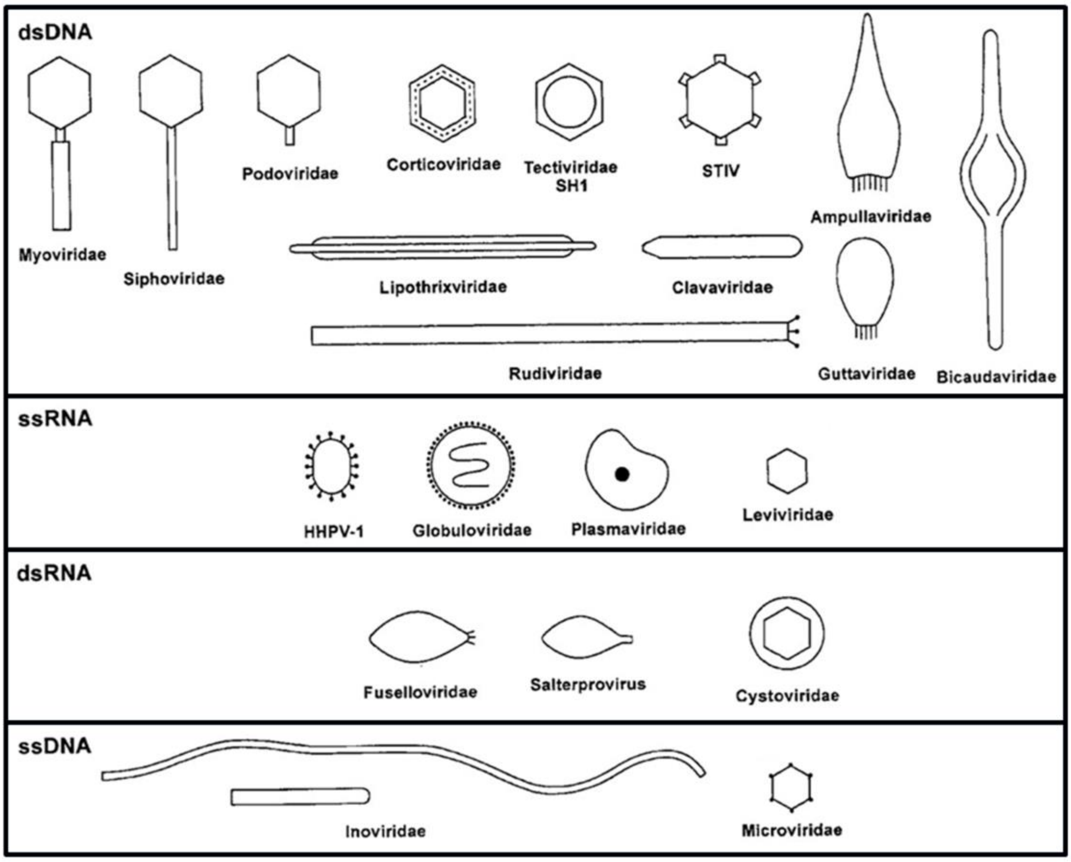


Figure 1.1. Morphology of viruses infecting Bacteria and Archaea.Modified from (Ackermann & Prangishvili, 2012). See Table 1.1 for more details on phage classification.

Most bacteriophages analysed to date belong to the *Caudovirales* order (from Latin *cauda* meaning “tail”), which comprises polyhedral phages with dsDNA genomes and tail. The members of this order are further divided based on tail morphology into three families: *Myoviridae* (from Greek *myos* meaning “muscle”, referring to the contractile tail), *Siphoviridae* (from Greek *siphon* meaning “tube”, referring to the long tail) and *Podoviridae* (from Greek *podos* meaning “foot”, referring to the short tail) (Figure 1.2). The tail can be contractile or non-contractile and vary from 3 to 825 nm in length (King *et al.*, 2011). The tail is involved in recognition, attachment and injection of the nucleic acid into host bacteria (Leiman *et al.*, 2003). The virion has no envelope and consists of the icosahedral or elongated head connected to the tail. Icosahedral heads varies in diameter between 45 and 170 nm, while elongated heads can be up to 230 nm long.

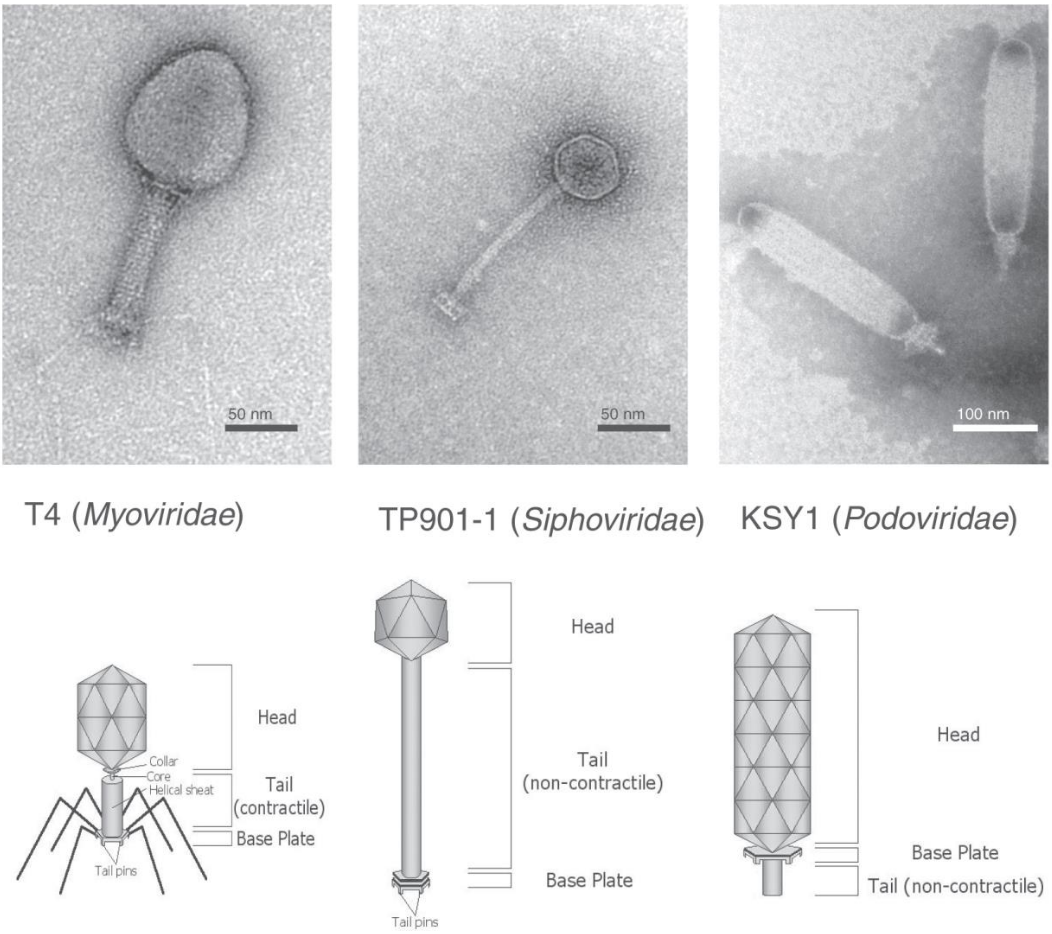


Figure 1.2 Transmission electron micrographs (upper panel) and schematic representations (lower panel) showing different morphotypes of the tailed phages from the order Caudovirales. The phage T4 of *E. coli* has an icosahedral head and a long contractile tail (*Myoviridae*); the phage TP901-1 of *L. lactis* has an icosahedral head and a long non-contractile tail (*Siphoviridae*); and the phage KSY1 of *L. lactis* has an elongated head and a short non-contractile tail (*Podoviridae*). Taken from (Emond & Moineau, 2007).

The genome of tailed bacteriophages range from 18 to 500 kb and encode from 27 to over 600 genes. Virions contains linear dsDNA, which may be circularly permuted (King *et al.*, 2011). A single DNA molecule condensed into a capsid is subject to very high osmotic pressure, which provides some of the force needed to eject it from the capsid into the cytoplasm of the host bacterium (Molineux & Panja, 2013). DNA often contains modified nucleotides e.g. 5-hydroxymethyl cytosine instead of cytosine (Warren, 1980). The genome encodes genes involved in host cell modification and viral DNA replication (“early genes”), and virion structural proteins and lysis proteins (“late genes”) (King *et al.*, 2011). The genome is organized into modules (Figure 1.3). Each module consists of sets of genes involved in similar function, e.g. replication, head assembly or tail formation. Phages exchange the modules with other phages and their hosts via homologous recombination, which results in genomic mosaicism (Hendrix *et al.*, 1999; Veesler & Cambillau, 2011). Tailed bacteriophages differ in host range and infect members of the *Enterobacteriaceae* family and genera such as *Pseudomonas*, *Haemophilus*, *Vibrio*, *Aeromonas*, *Bacillus*, *Burkholderia* and *Mycobacterium*. The best known example of tailed phage is bacteriophage T4 that infects *Escherichia coli* (Figure 1.2).

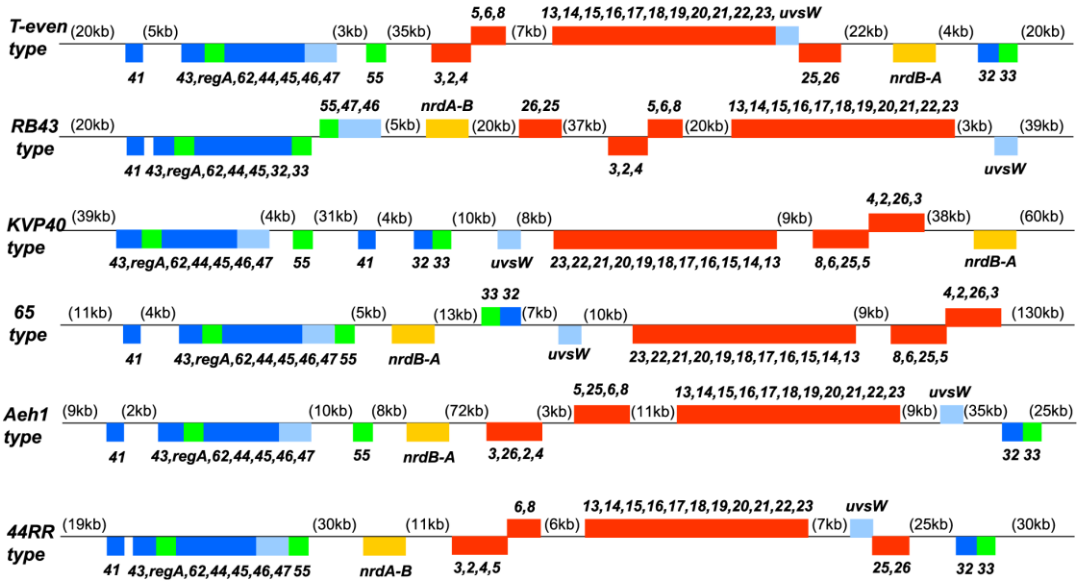


Figure 1.3. Genomic organization of T4-type phages showing mosaicism between related phages. The numbers refer to phage genes annotated in reference (Petrov *et al.*, 2010a). Gene organised in modules are shown in colours: in dark blue are DNA replication genes, in light blue are the recombination/repair genes, in green are the transcription and translation genes, in red are the morphogenetic genes and the genes for aerobic nucleotide reductase (nrdAB) are shown in orange. From (Petrov *et al.*, 2010a).

Table 1.1. Prokaryotic (bacterial and archaeal) virus families based on ICTV classification (King *et al.*, 2011).

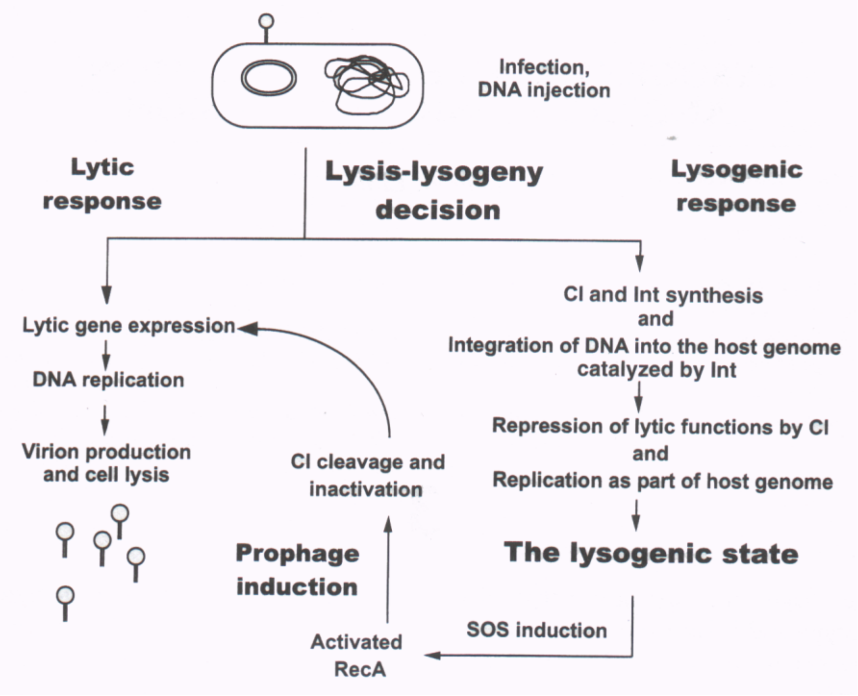
Shape	**Family or unassigned genus**	**Genome type**	**Morphology**	**Example**
Tailed	*Myoviridae*	Linear dsDNA	Non-enveloped, icosahedral head with long contractile tail	T4
	Siphoviridae	Linear dsDNA	Non-enveloped, icosahedral head with long non-contractile tail	λ
	Podoviridae	Linear dsDNA	Non-enveloped, icosahedral head with short non-contractile tail	T7
Polyhedral	*Tectiviridae*	Linear dsDNA	Non-enveloped, icosahedral	PRD1
	SH1, group*	Linear dsDNA	Icosahedral, lipid-containing	SH1
	Corticoviridae	Circular dsDNA	Non-enveloped, icosahedral	PM2
	Microviridae	Circular ssDNA	Non-enveloped, icosahedral	φX174
	STIV group*	Circular dsDNA	Icosahedral, turret-shaped	STIV
	Leviviridae	Linear ssRNA	Non-enveloped, icosahedral	MS2
	Cystoviridae	Segmented dsRNA	Enveloped, spherical	φ6
Pleomorphic	*Ampullaviridae*	Linear dsDNA	Enveloped, bottle-shaped	ABV
	Globuloviridae	Linear dsDNA	Enveloped, spherical	PSV
	*Salterprovirus**	Linear dsDNA	Enveloped, lemon-shaped	His1
	Bicaudaviridae	Circular dsDNA	Enveloped, lemon-shaped, two-tailed	ATV
	Fuselloviridae	Circular dsDNA	Enveloped, lemon-shaped	SSV1
	Plasmaviridae	Circular dsDNA	Enveloped, pleomorphic	L2
	Guttaviridae	Circular dsDNA	Enveloped, droplet-shaped	SNDV
	HHPV-1 group*	Circular dsDNA	Pleomorphic, contain lipids	HHPV-1
Filamentous	*Clavaviridae*	Circular dsDNA	Non-enveloped, rod-shaped	APBV1
	Inoviridae	Circular ssDNA	Non-enveloped, filamentous or rod-shaped	M13
	Lipothrixviridae	Linear dsDNA	Enveloped, filamentous	TTV1
	Rudiviridae	Linear dsDNA	Non-enveloped, rod-shaped	SIRV-1

*Awaiting classification (Ackermann & Prangishvili, 2012)

Phage classification is complicated because phage genomes are mosaic (Hendrix *et al.*, 1999). The International Committee on Taxonomy of Viruses (ICTV) classifies phages based on their morphology (such as capsid size and shape, presence of tail) and type of nucleic acid. The current ICTV classification of prokaryotic viruses (http://ictvonline.org/virusTaxonomy.asp?version=2012) divides them onto 18 families, 16 families contain DNA, and 2 RNA as a nucleic acid (Table 1.1). Four groups have not been yet classified into any family (Table 1.1). Over 96% of all prokaryotic viruses described in the literature are tailed dsDNA phages. Polyhedral, filamentous or pleomorphic viruses compromise only 4% of the prokaryotic viruses of known morphology (Ackermann & Prangishvili, 2012). Alternative method of phage classification called the phage proteomic tree (PPT) was proposed by (Rohwer & Edwards, 2002). PPT classifies phages based on the overall sequence similarity of complete phage genomes.

1.1.3. Phage life cycles

Phages, like all viruses, lack independent metabolism and require the host biosynthetic machinery to propagate (Christie & Dokland, 2012). Once the nucleic acid has entered the host, there are four possible life cycles: lytic cycle, lysogenic cycle, pseudolysogenic cycle or chronic infection. In the lytic cycle (carried out by a virulent phage, for example phage T4), phage DNA is replicated and transcribed by the host, programming the synthesis of new phage. Following lysis of infected bacteria by phage-specified enzymes and cell death, the newly assembled phage particles are released into the environment (Calendar & Inman, 2005). A lysogenic cycle (carried out by a temperate phage, for example phage λ) results in integration of the phage DNA into host genome. Integrated phage (called a prophage) is replicated along with the host genome and is passed along to daughter cells (vertical transmission). Under some circumstances such as UV-induced DNA damage, heat shock or starvation, prophages are able to switch to the lytic cycle (Little, 2005) (Figure 1.4). Pseudolysogeny is where phage doesn’t multiply or integrate into host chromosome after cell entry, but exists in an inactive state, in which it cannot enter the lytic or lysogenic cycle due to unfavourable growth conditions for the host cell (such as low level of nutrient availability). Pseudolysogenic phage enter the lytic or lysogenic cycle when nutrients become available to the host (Los & Wegrzyn, 2012; Ripp & Miller, 1997). A chronic infection occurs when newly assembled phage particles are released from bacterium by budding or extrusion (secretion) without causing cell death (Weinbauer, 2004).



Infection,
DNA injection
Lytic response
Lysis-lysogeny decision
Lysogenic response
Lytic gene expression
DNA replication
Virion production and cell lysis
CI and Int synthesis and Integration of DNA into the host genome catalyzed by Int
Repression of lytic functions by CI and Replication as part of host genome
The lysogenic state
CI cleavage and inactivation
Prophage induction
Activated RecA
SOS induction

Figure 1.4. Life cycle of the temperate λ phage. The phage particle attaches to the bacterial cell surface and injects the DNA, which ends join to form a circle. Up to 15 min after infection, the decision is made between two alternative pathways. During the lysogenic response, phage development is repressed by CI repressor and phage DNA integrates into the host chromosome with the aid of the viral integrase (Int). The resulting lysogenic cell can replicate indefinitely until the lytic cycle is induced and phage DNA is excised from the chromosome. From (Little, 2005).

1.1.4. Phage applications and significance

Phages and their components have many practical applications in biotechnology and medicine. They have been used in diagnostics for identification of pathogenic bacteria (phage typing); in proteomics for identification of peptides, which have affinity for proteins displayed on the surface of the phage (phage display); in gene therapy as gene delivery vehicles; or in medicine for treatment of bacterial infections (phage therapy) (Haq *et al.*, 2012). Although phage therapy in humans has not been yet approved in most countries and only a few clinical trials of phage therapy are currently ongoing (Abedon *et al.*, 2011; Harper & Enright, 2011; Parracho *et al.*, 2012), recent studies show their potential application in the treatment of acute lung infections in an *in vivo* murine lung model with *Pseudomonas aeruginosa* (Alemayehu *et al.*, 2012b; Debarbieux *et al.*, 2010) and *Burkholderia cenocepacia* (Carmody *et al.*, 2010). Because they have the capacity to exert selection on bacteria phages play an important role in bacterial evolution e.g. pathogenesis (Breitbart *et* *al.*, 2005) and in ecological processes e.g. in cycling of nutrients and energy in ecosystems (Fuhrman, 1999).

1.1.4.1. Role in the environment

Phages are responsible for significant bacterial mortality in the environment (Fuhrman, 1999). The rate of bacterial predation by phage influences bacterial community composition diversity (Fuhrman & Schwalbach, 2003). The impact of phages on microbial diversity is usually explained by the “kill the winner” model (Thingstad & Lignell, 1997). This model assumes that predominant phages present are those which preferentially infect the dominant bacterial species present at any time point in the community (“the winner”). This phage multiplication results in a collapse in the abundance of the bacterial host giving rise to successive peaks of individual host strains and their specific phage parasite, followed by the emergence of a different bacterial host strain as numerically predominant before it in its turn is reduced by phage attack. This constant turnover of dominant species helps to maintain the host diversity (Rohwer *et al.*, 2009; Thingstad & Lignell, 1997; Thingstad *et al.*, 2008).

Phages, by lysing bacteria, play a significant role in global biogeochemical and ecological cycles (Fuhrman, 1999). Viral infection releases organic material from the lysed bacteria as dissolved organic matter (DOM) (Thingstad & Lignell, 1997). Released DOM is consumed by other heterotrophic bacteria. These bacteria are then lysed and DOM returns back to the dissolved nutrients pool, therefore it is cycled in a closed loop (Thingstad *et al.*, 2008). For example, in marine ecosystems it is estimated that “6–26% of photosynthetically fixed organic carbon is recycled back to dissolved organic material by viral lysis” (Wilhelm & Suttle, 1999).

1.1.4.2. Role in bacterial evolution

Phages contribute significantly to bacterial evolution and pathogenesis by moving genes between bacteria via horizontal gene transfer (HGT). This gene movement can occur through transduction or as a result of prophage integration (Brussow *et al.*, 2004). Two types of phage transduction can take place: generalized or specialized. Generalized transduction is when bacterial DNA is accidentally packaged into the phage head instead of phage DNA and transferred to another bacterium. Transferred foreign bacterial DNA can subsequently become incorporated into bacterial genome through homologous recombination event (Brussow *et al.*, 2004). The best known example of generalized transducing phage is *Escherichia coli*-phage P1 and *Salmonella enterica*-phage P22. Specialized transduction is when bacterial genes are acquired due to imprecise excision of prophage from a specific integration site, packaging both phage DNA and adjacent DNA from the bacterial genome into a single phage particle. Such phage can subsequently infect another host and integrate into its genome (Brussow *et al.*, 2004). Transduction occurs in the natural environment, e.g. in the marine environment transduction occurs at a rate of 10^{14} transduction events per year (Jiang & Paul, 1998) and (Paul *et al.*, 2002) estimated that marine phages transduce 10^{28} bp of DNA per year in the world's oceans.

Phages often transfer genes that change the phenotype or fitness of their host (Brussow *et al.*, 2004; Canchaya *et al.*, 2003). These genes encode virulence factors, which play roles in bacterial attachment, colonization, invasion, and immunosuppression, are responsible for resistance to antibiotics, biofilm formation, and production of toxins (Wu *et al.*, 2008). Toxin genes are responsible for pathogenesis of cholera, diphtheria, and shigellosis (Mekalanos *et al.*, 1997; Wagner & Waldor, 2002). For example, infection of nontoxigenic *Vibrio* spp. with cholera toxin-encoding phage CTXφ from *V. cholerae* converts them to toxigenic strains (Boyd *et al.*, 2000). Phages have the ability to transduce resistance to antibiotics in many different bacterial species with clinical importance e.g. *Actinobacillus actinomycetemcomitans*, *Staphylococcus aureus* and *Pseudomonas aeruginosa* (Blahova *et al.*, 1993; Pereira *et al.*, 1997; Willi *et al.*, 1997). Transfer of antibiotic resistance through phage infection may constitute a serious problem if pathogenic bacteria are rendered resistant to antimicrobials. Studies show that bacteriophages carrying antibiotic resistance genes (Colomer-Lluch *et al.*, 2011; Muniesa *et al.*, 2004) and toxin genes (Casas *et al.*, 2006) are widely distributed in the environment.

1.1.5. Phage abundance and diversity

Phages can be found everywhere where bacteria exist and have been isolated from many types of environments. They naturally occur in soil, water, food, plants and animals, as well as parts of human body colonised by a normal microbiota, e.g. mouth, skin or the gastrointestinal tract. It has been estimated that there are approximately 10^{30} prokaryotic cells in the world (Whitman *et al.*, 1998), and direct counts of viruses using epifluorescence microscopy indicate that they typically outnumber bacteria about 10-fold (Brussow & Hendrix, 2002; Fuhrman, 1999). Phage abundance depends on abundance of their hosts, therefore environments rich in bacteria will contain more bacteriophages. There are typically 10^4 – 10^8 virus-like particles (VLPs) per ml in aquatic environments (Bergh *et al.*, 1989; Wommack & Colwell, 2000), 10^8 – 10^9 VLPs per gram of soil (Williamson *et al.*, 2005; Williamson *et al.*, 2007), and up to 10^{11} VLPs per g of marine sediment (Danovaro *et al.*, 2001; Danovaro *et al.*, 2005; Helton *et al.*, 2006). Sites in the human body such as gut including the oral cavity contain approximately 10^8 VLPs per 1 ml of fluid (Haynes & Rohwer, 2011). Since 1959, over 6300 bacteriophages have been examined using electron microscopy, of which 96% are tailed (Ackermann & Prangishvili, 2012). More than 900 phage genomes have been sequenced and deposited in the NCBI phage genome database http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi). These phages infect a broad range of hosts (Figure 1.5) and presumably represent only a small fraction of the total population that has been estimated as above to be approximately 10^{31} particles. Comparative phage genomics revealed that phage genetic diversity is extraordinary high. Phages isolated from different bacterial hosts typically have little or no recognizable sequence similarity, and even phages infecting the same host may exhibit great diversity (Comeau *et al.*, 2008; Hatfull *et al.*, 2006; Kwan *et al.*, 2005; Kwan *et al.*, 2006).

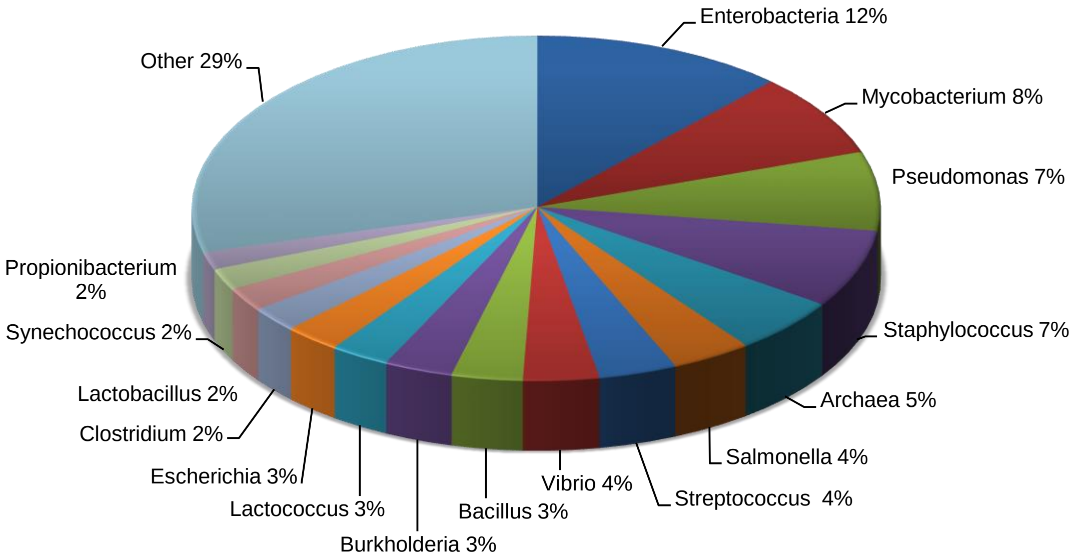


Figure 1.5. Distribution of 904 complete phage genomes deposited in the GenBank database on 5th February, 2013, grouped according to host they infect.

1.1.5.1. Methods to measure abundance and phage diversity

Traditional approaches used to characterise phage diversity involve phage culture. The phage is isolated from environmental samples using a plaque assay method, in which dilutions of phage suspensions are mixed with a susceptible host which is then grown as a lawn on an agar plate, resulting in focal lysis of that host and formation of clear areas within the lawn of bacterial growth, called a plaque (Kropinski *et al.*, 2009; Millard, 2009). Once phage are cultured, phenotypic characterisation can be applied, including host-range determination (Abedon, 2008). However, since only a small fraction of bacteria from natural environments can be cultured by current standard laboratory methods (Stewart, 2012), culture-based techniques largely underestimate environmental phage diversity.

Several methods have been used to estimate phage abundance and diversity in the environment. These include epifluorescence microscopy (Hara *et al.*, 1991; Noble & Fuhrman, 1998) and flow cytometry (Brussaard *et al.*, 2000; Marie *et al.*, 1999) for phage enumeration, and transmission electron microscopy for examining morphological diversity (Ackermann & Prangishvili, 2012; Bergh *et al.*, 1989). Pulsed field gel electrophoresis (PFGE) of extracted DNA can be used to determine genome size (Fuhrman *et al.*, 2002; Steward *et al.*, 2000). Recently, culture-independent studies such as PCR assays and metagenomic sequencing have greatly expanded knowledge about the genetic diversity of viruses in their natural environment.

1.1.5.2. Diversity based on conserved gene studies

Currently, there is no single gene universally present within all phages suitable for use as a phylogenetic marker, in an analogous way to the use of 16S ribosomal RNA gene sequences in bacteria (Rohwer & Edwards, 2002). However, several widely distributed genes such as those encoding capsid proteins and DNA polymerases have been used to study the genetic diversity of specific phage groups. The sequence conservation of these genes is sufficient to allow design of PCR primers that can be used to assess genetic diversity by cloning and sequencing PCR products amplified directly from environmental samples. Degenerate primers have been used to amplify the DNA of T4-type phages (Comeau & Krisch, 2008; Filee *et al.*, 2005; Short & Suttle, 2005; Zhong *et al.*, 2002) and the T7-type phages (Breitbart *et al.*, 2004b; Chen *et al.*, 2009) in samples from their natural environment. The diverse target sequences detected by these PCR assays in various marine and freshwater environments compared with the corresponding sequences in fully sequenced phages, suggest that much of the phage genetic diversity remains uncharacterised. Diversity estimated using PCR-based molecular markers is also necessarily an underestimate as PCR primer design is based on sequences from cultured phages.

1.1.5.3. Metagenomic approach to measure viral diversity

Many viruses present in their natural environment still remain uncharacterised or poorly studied, because they cannot be cultured in the absence of their host, and most bacteria cannot be readily cultured in the laboratory. Additionally, molecular approaches such as PCR assays require prior knowledge about the sequence of the target gene to design primers for amplification, and are restricted to particular viral groups as no gene is universally present in all viruses. Metagenomic approaches potentially overcome these limitations allowing quantification of genetic diversity by direct extraction and sequencing of the nucleic acid of the entire viral community, with no prior culture or information about the sequences present being required

(Delwart, 2007; Edwards & Rohwer, 2005). Such analyses have been carried out on a number of environmental (seawater, marine sediments, soil) and clinical (faeces, respiratory secretions) samples (Delwart, 2007). The metagenomic analyses revealed the enormous scale of genetic diversity. For example, metagenomics sequencing of viral communities from marine water (Breitbart *et al.*, 2002) and human faeces (Breitbart *et al.*, 2003) demonstrated that the majority of sequences showed no significant similarity to any other sequences deposited in databases. The degree of community diversity (which is expressed as species richness and abundance) can be estimated based on mathematical modelling of sequence data variations (Angly *et al.*, 2005b). Biodiversity estimates indicate that viral diversity is different for samples from different environments (Table 1.2), ranging from only 8 viral types in infant faeces (Breitbart *et al.*, 2008), to more than 1,000,000 viral types in the case of rainforest soil (Fierer *et al.*, 2007).

Table 1.2. Examples of richness* estimates for published viral metagenomes.

Reference	**Sample type**	**Number of genotypes**	**Shannon index**
(Fierer *et al.*, 2007)	Soil	1,000,000	**NA
(Angly *et al.*, 2006)	Marine water	129,000	10.8
(Marhaver *et al.*, 2008)	Coral tissue	28,600	8.96
(Desnues *et al.*, 2008)	Stromatolite	19,520	8.9
(Breitbart *et al.*, 2004a)	Marine sediment	10,000	9.2
(Lopez-Bueno *et al.*, 2009)	Antarctic lake	9,730	8.15
(Park *et al.*, 2011)	Fermented shrimp	7,310	5.90
(Breitbart *et al.*, 2003)	Human adult faeces	1,930	6.43
(Fancello *et al.*, 2013)	Freshwater pond	977	4.19
(Willner *et al.*, 2011)	Human oropharyngeal swabs	236	NA
(Willner *et al.*, 2009)	CF sputum	105	4.17
(Breitbart *et al.*, 2008)	Human infant faeces	8	1.69

* Richness is defined as the total number of distinct species in a community
** Data not available

1.2. Introduction to metagenomics

1.2.1. Viral metagenomics

Viral metagenomics refers to the culture-independent analysis of viral DNA or RNA directly extracted from environmental samples (Edwards & Rohwer, 2005; Handelsman *et al.*, 1998) containing a population of different viruses. Meta-genomic DNA is directly cloned and/or sequenced, and characterised using computational approaches, and RNA is similarly analysed after reverse transcription. Because metagenomics does not require prior virus cultivation, and does not rely on prior knowledge about the viral types present in the samples, this method provides insight into community diversity and can be used to address the challenge of studying unknown viruses (Delwart, 2007; Edwards & Rohwer, 2005). Since the publication of the first viral metagenomes in 2002 (Breitbart *et al.*, 2002), the number of viral metagenomic studies has increased exponentially in recent years with the availability of next-generation sequencing technologies (Figure 1.6). Viral metagenomics has been used to characterise viruses in a wide variety of marine and terrestrial and animal-associated environments (Table 1.3). The results obtained from metagenomic studies have provided a great amount of information on the diversity, abundance and metabolic potential of viruses in their natural environment and the constant novelty found suggests that the global viral metagenome is still largely uncharacterised.

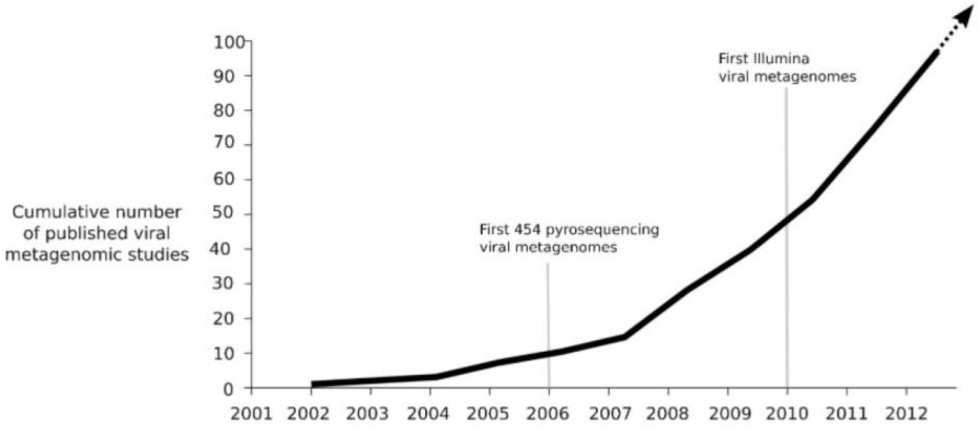


Figure 1.6. The number of published papers on viral metagenomics from 2002 to 2012. From (Willner & Hugenholtz, 2013).

Table 1.3. Examples of published viral metagenomic studies carried out in different environments.

Reference	**Sample type**	**Sampling location**	**Goal**	**Sequencing method**	**Classification method**	**Cut off (E-value)**	**Unknown hits**
(Breitbart *et al.*, 2002)	Marine water	San Diego Bay	Characterise the diversity of two near-shore marine DNA viral communities	Sanger	TBLASTX	1e-03	65-73%
(Breitbart *et al.*, 2003)	Human faeces	NA	Characterise the composition and population structure of a DNA viral community isolated from human faeces	Sanger	TBLASTX	1e-03	59%
(Breitbart *et al.*, 2004a)	Marine sediment	San Diego Bay	Characterise the diversity of a near-shore marine sediment DNA viral community and compare to marine communities	Sanger	TBLASTX	1e-03	75%
(Breitbart & Rohwer, 2005b)	Human blood	San Diego, USA	Develop a method for identification of DNA viruses in plasma samples	Sanger	TBLASTX	1e-03	10%
(Cann *et al.*, 2005)	Horse faeces	NA	Characterise the diversity of DNA viruses in the equine gut	Sanger	TBLASTX	1e-03	68%
(Angly *et al.*, 2006)	Marine water	Sargasso sea; Gulf of Mexico; British Columbia coast and Artic ocean	Characterise the diversity of DNA viruses isolated from four oceanic sites	454	TBLASTX	1e-05	>91%
(Zhang *et al.*, 2006)	Human faeces	San Diego, USA	Characterise RNA viruses isolated from human faeces	Sanger	TBLASTX	1e-03	8.5%
(Allander *et al.*, 2007)	Nasopharyngeal aspirate	Stockholm, Sweden	Identify viruses associated with respiratory tract infections	Sanger	TBLASTX	1e-04	2%
(Bench *et al.*, 2007)	Virioplankton	Chesapeake Bay	Characterise DNA viruses isolated from estuarine ecosystem	Sanger	T/BLASTX	1e-03	61%
(Fierer *et al.*, 2007)	Soil	Peru; California; Kansas	Characterise the diversity of prairie, desert, and rainforest soil viral communities and compare to soil bacterial, archaeal, and fungal communities	Sanger	TBLASTX	1e-03	NA
(Breitbart *et al.*, 2008)	Infant faeces	NA	Characterise the diversity of DNA viruses in the infant gut	Sanger	TBLASTX	1e-03	66%
(Desnues *et al.*, 2008)	Microbialites	Mexico and Bahamas	Characterise DNA viruses associated with marine stromatolite and two freshwater thrombolites and stromatolites	454	BLASTX	1e-02	>97%
(Finkbeiner *et al.*, 2008)	Human diarrhoea	Melbourne, Australia; Seattle, USA	Identification of RNA viruses in stool from paediatric patients suffering from diarrhoea	Sanger	TBLASTX	1e-05	28%

Reference	**Sample type**	**Sampling location**	**Goal**	**Sequencing method**	**Classification method**	**Cut off (E-value)**	**Unknown hits**
(Kim *et al.*, 2008)	Rice paddy soil	Daejeon, Korea	Characterise ssDNA viruses in soil	Sanger	T/BLASTX	1e-03	>64%
(Marhaver *et al.*, 2008)	Coral	Mount Irvine Bay, Tobago	Characterise the diversity of DNA viruses isolated from healthy and bleaching corals	Sanger	TBLASTX	1e-03	>18%
(McDaniel *et al.*, 2008)	Marine water	Tampa Bay	Identify viral genes involved in lysogeny in marine environment	454	TBLASTX	1e-03	93%
(Palacios *et al.*, 2008)	Human tissue	Australia	Identify cause of death of two organ transplant patients	454	BLASTX	NA	NA
(Schoenfeld *et al.*, 2008)	Hot springs	Yellowstone, USA	Characterise the diversity, composition, and adaptations of DNA viral communities in two hot springs	Sanger	BLASTX	1e-03	>33%
(Williamson *et al.*, 2008)	Marine water, freshwater and hypersaline	37 of aquatic sites from Halifax, Nova Scotia through the South Pacific Gyre	Characterise DNA viruses within aquatic microbial samples	Sanger	NA	NA	NA
(Vega Thurber *et al.*, 2008)	*Porites compressa* (finger coral)	Hawaii	Determine shifts in diversity of viral communities associated with coral in response to abiotic stressors	454	BLASTN	1e-04	>98%
(Djikeng *et al.*, 2009)	Freshwater lake	Lake Needwood, Maryland, USA	Characterise RNA viruses isolated from freshwater lake	Sanger and 454	BLASTX	1e-05	66%
(Lopez-Bueno *et al.*, 2009)	Freshwater lake	Limnopolar lake, Livingston Island, Antarctica	Characterise and compare DNA viral communities from Antarctic lake during spring and summer	454	BLASTX	1e-03	>81%
(Nakamura *et al.*, 2009)	Human faeces and nasopharyngeal aspirates	Osaka, Japan	Detect RNA viral pathogens in nasal and faecal specimens	454	BLASTN	1e-40	NA
(Ng *et al.*, 2009)	Sea turtle fibropapilloma	Lake Worth Lagoon, Florida, USA	Identify DNA viruses associated with fibropapillomatosis in a sea turtle	Sanger	T/BLASTX	NA	NA
(Rosario *et al.*, 2009b)	Reclaimed water	Florida, USA	Characterise viruses in reclaimed water and compare it with viruses in potable water	454	BLASTX	1e-03	>57%
(Willner *et al.*, 2009)	Human respiratory tract	San Diego, USA	Characterise and compare DNA viral communities from individuals with and without cystic fibrosis	454	TBLASTX	1e-05	>86%
(Coetzee *et al.*, 2010)	Vines from a diseased vineyard	South Africa	Identify RNA viruses infecting grapevines	Illumina	BLASTN/X (contigs)	1e-05	59%

Reference	Sample type	Sampling location	Goal	Sequencing method	Classification method	Cut off (E-value)	Unknown hits
(Li *et al.*, 2010a)	Bat guano	San Saba, Texas; Point Reyes, California	Characterise DNA and RNA viruses present in guano from bats	454	BLASTX	1e-03	39%
(Parsley *et al.*, 2010b)	Activated sludge	Auburn, Alabama, USA	Characterise viral diversity in activated sludge and compare to bacterial diversity	Sanger	BLASTX	1e-03	40%
(Reyes *et al.*, 2010)	Human faeces	San Diego, USA	Characterise faecal DNA viromes of monozygotic twins and their mothers and compare to faecal microbiomes	454	TBLASTX	1e-03	>75%
(Allen *et al.*, 2011)	Swine faeces	NA	Characterise fecal viromes over time in swine that were fed the common antibiotics and compare to viromes of nonmedicated swine	454	BLASTX	NA	21-58%
(Cantalupo *et al.*, 2011)	Sewage	Addis Ababa (Ethiopia), Pittsburgh (USA), Barcelona (Spain)	Characterise DNA and RNA viruses obtained from raw sewage	454	T/BLASTX	1e-05	66%
(Li *et al.*, 2011b)	Dogs faeces	California	Characterise the faecal viruses in diarrhoea specimens from dogs	454	BLASTN/X (contigs)	1e-03	NA
(Minot *et al.*, 2011)	Human faeces	Philadelphia, USA	Investigated the dynamics of the gut virome during perturbations to diet	454	BLASTX (contigs)	1e-03	55%
(Ng *et al.*, 2011)	Mosquitoes	San Diego, USA	Characterise the diversity of DNA viruses present in three mosquito samples	454	TBLASTX	1e-03	48-80%
(Park *et al.*, 2011)	Fermented shrimp, Chinese cabbage, sauerkraut	Korea	Characterise DNA viral communities in the fermented foods	454	BLASTX	1e-03	37-50%
(Phan *et al.*, 2011)	Rodents faeces	California, Virginia	Characterise the faecal DNA and RNA viruses from rodents	454	BLASTN/X	1e-05	45%
(Steward & Preston, 2011)	Marine water	Monterey Bay	Characterise viruses from deep ocean, with no amplification prior to cloning	Sanger	BLASTX (contigs)	1e-03	74%
(Willner *et al.*, 2011)	Human oropharyngeal swabs	San Diego, USA	Characterise DNA viral communities in the human oral cavity	454	TBLASTX	1e-05	51-64%
(Boujelben *et al.*, 2012)	Solar salterns	Tunisia	Characterise changes in the viral communities over time and across a salinity gradient	Sanger	BLASTX (contigs)	1e-03	>40%
(Cassman *et al.*, 2012)	Marine water	Eastern Tropical South Pacific off Iquique, Chile	Characterise viral communities isolated from different depths of oceanic water	454	BLASTX	1e-03	>91%
(Emerson *et al.*, 2012)	Hypersaline lake	Lake Tyrrell, Victoria, Australia	Identify the dominant viral populations in hypersaline lake	454, Illumina	BLASTP (contigs)	NA	>85%

Reference	**Sample type**	**Sampling location**	**Goal**	**Sequencing method**	**Classification method**	**Cut off (E-value)**	**Unknown hits**
(Foulongne *et al.*, 2012)	Human skin swabs	France	Characterise viromes of the surface of the skin of five healthy individuals and one patient with Merkel cell carcinoma	Illumina	BLASTN/X	1e-03	90-99%
(Lysholm *et al.*, 2012)	Human nasopharyngeal aspirates	Stockholm, Sweden	Characterise DNA and RNA viruses in patients with severe lower respiratory tract infections	454	BLASTN/X	1e-03	12%
(Masembe *et al.*, 2012)	Pig serum	Uganda	Characterise DNA and RNA viruses in domestic pig serum	454	BLASTN	1e-03	62%
(Ng *et al.*, 2012)	Sewage	Thailand, Nepal, Nigeria, USA	Characterise DNA and RNA viruses from untreated sewage collected from 4 locations	454	BLASTX	1e-04	37%
(Roux *et al.*, 2012a)	Freshwater lake	Lake Bourget, Lake Pavin, France	Characterise two freshwater lakes viromes	454	BLASTX	1e-03	73-85%
(Tamaki *et al.*, 2012)	Wastewater	Singapore	Characterise and compare viral communities in the different stages of wastewater treatment	454	BLASTX	1e-05	79-95%
(Whon *et al.*, 2012)	Air and rainwater	Korea	Characterise airborne viral diversity and its composition in the near-surface atmosphere	454	BLASTX	1e-03	49-80%
(Williamson *et al.*, 2012)	Marine water	17 sites from the Indian Ocean	Examine the Indian Ocean virome	Sanger and 454	BLASTP	1e-10	>88%
(Willner *et al.*, 2012)	CF lung tissue	California	Characterise viral communities in CF lung tissue from spatially distinct areas	454	TBLASTX	1e-05	36-88%
(Baker *et al.*, 2013)	Bat urine, throat swabs, lung tissue	Ghana	Identify viruses of African straw-coloured fruit bats	Illumina	BLASTN, T/BLASTX	1e-04	>99%
(Bibby & Peccia, 2013)	Sludge	USA	Identify viral DNA and RNA viral diversity in mesophilic anaerobic digester from five wastewater treatment plants	Illumina	T/BLASTX (contigs)	1e-03	>80%
(Fancello *et al.*, 2013)	Freshwater ponds	Sahara	Characterise viral communities from freshwater ponds in the Sahara desert	454	BLASTX	1e-05	70-83%
(Mokili *et al.*, 2013)	Human nasopharyngeal and oropharyngeal swabs	USA	Characterise DNA and RNA viruses from patients with febrile respiratory illness	454	BLASTN	1e-05	92%
(Yoshida *et al.*, 2013)	Marine sediment	Northwest Pacific	Characterise viral communities of deep-sea sediments	454	BLASTX	1e-03	>70%

* NA data not available

1.2.2. Specific applications of viral metagenomics

Viral metagenomics enables ecological studies to address questions on what types of viruses are on Earth, how they are distributed and how they interact with their host (Rosario & Breitbart, 2011), in medicine for pathogenic virus discovery and characterisation (Fancello *et al.*, 2012; Mokili *et al.*, 2011) and in biotechnology as a tool for the discovery of novel enzymes.

1.2.2.1. Ecology

Data generated through metagenomic sequencing projects indicates that viral communities are probably dominated by phages, with a few exceptions, such as the viral community from an Antarctic lake that is dominated by eukaryotic viruses (Lopez-Bueno *et al.*, 2009). The viral communities in marine surface waters are abundant in phages, in particular cyanophages (Angly *et al.*, 2006; Williamson *et al.*, 2008), while communities from the marine sediments are dominated by ssDNA viruses (mainly microphages) (Yoshida *et al.*, 2013). Single-stranded DNA microphages have been also found to be very abundant in marine water (Angly *et al.*, 2006) and marine microbialites (Desnues *et al.*, 2008). Phages dominated DNA viral community, of reclaimed water while the RNA viral community was dominated by eukaryotic viruses (Rosario *et al.*, 2009b).

Characterizations of the human gut viromes have revealed predominance of temperate phages in this environment (Minot *et al.*, 2011; Reyes *et al.*, 2010). The viral communities of the human oropharynx were also dominated by phages, some of them encoding streptococcal virulence factors (Willner *et al.*, 2011). The distribution of eukaryotic and prokaryotic DNA viruses in the cystic fibrosis lung was shown to differ across different areas of the lungs, with some lung sections where no phage could be detected (Willner *et al.*, 2012). Recently it has been suggested that lytic phage augment the human immune system in its action against bacteria by their concentration in mucus. Phage binding to mucin glycoproteins is hypothesized to account for this (Barr *et al.*, 2013).

Comparisons between marine viral metagenomes from different oceanic regions indicated that large fraction of marine viruses are globally distributed, but their relative abundance differs between locations (Angly *et al.*, 2006). Similarly, freshwater viral metagenomes showed significant genetic similarity despite the vast geographical distances between sample locations (Roux *et al.*, 2012a). In contrast, little or no phylogenetic overlap was observed between viral groups from different soils (Fierer *et al.*, 2007). Limited global distribution have been also showed for viral populations from geographically diverse hypersaline environments (Emerson *et al.*, 2012).

Metagenomics can also provide valuable insights into understanding of virus-host interactions and co-evolution. In marine environments, significant portions of aquatic viral communities carry genes of host origin involved in host metabolic processes, such as photosynthesis (Angly *et al.*, 2006; Williamson *et al.*, 2008). Phage-encoded photosynthesis genes are expressed during infection and enable their hosts to maintain photosynthesis even in intense sunlight, which normally cause photo-inhibition (Mann *et al.*, 2003), with the end point of increasing phage titres.

Another virus-host interaction explored through metagenomics is analysis of the clustered regularly interspaced short palindromic repeat (CRISPR) system, a defence mechanism against phage infection, widespread in archaea and bacteria (Horvath & Barrangou, 2010; Sorek *et al.*, 2008; Sorek *et al.*, 2013). In this system, short sequences derived from invading phages are integrated as CRISPR spacers between the conserved repeats in the host genome and provide resistance to this phage by targeting invading phage DNA for lysis by the Cas enzyme complex (Figure 1.7). A Cas complex usually contains between 4 and 20 different *cas* genes (Deveau *et al.*, 2010), many of them have been shown to function as helicases or nucleases and are required for new spacer sequence acquisition and degradation of the invading DNA (Sorek *et al.*, 2013). Recently, it have been shown that phage encode its own CRISPR/Cas system, which is used to inhibit host immunity and thereby permit lytic infection (Seed *et al.*, 2013). CRISPR spacer sequences constitute a direct link between phages and their hosts. Comparison of a database of CRISPR spacers (Grissa *et al.*, 2007b) from different bacteria with reads from a viral metagenome can be used as a reverse map to suggest the possible bacterial host for uncharacterised viral metagenomic sequences (Anderson *et al.*, 2011; Stern *et al.*, 2012), or to determine if a bacterial population has been previously infected with a specific phage (Kunin *et al.*, 2008).

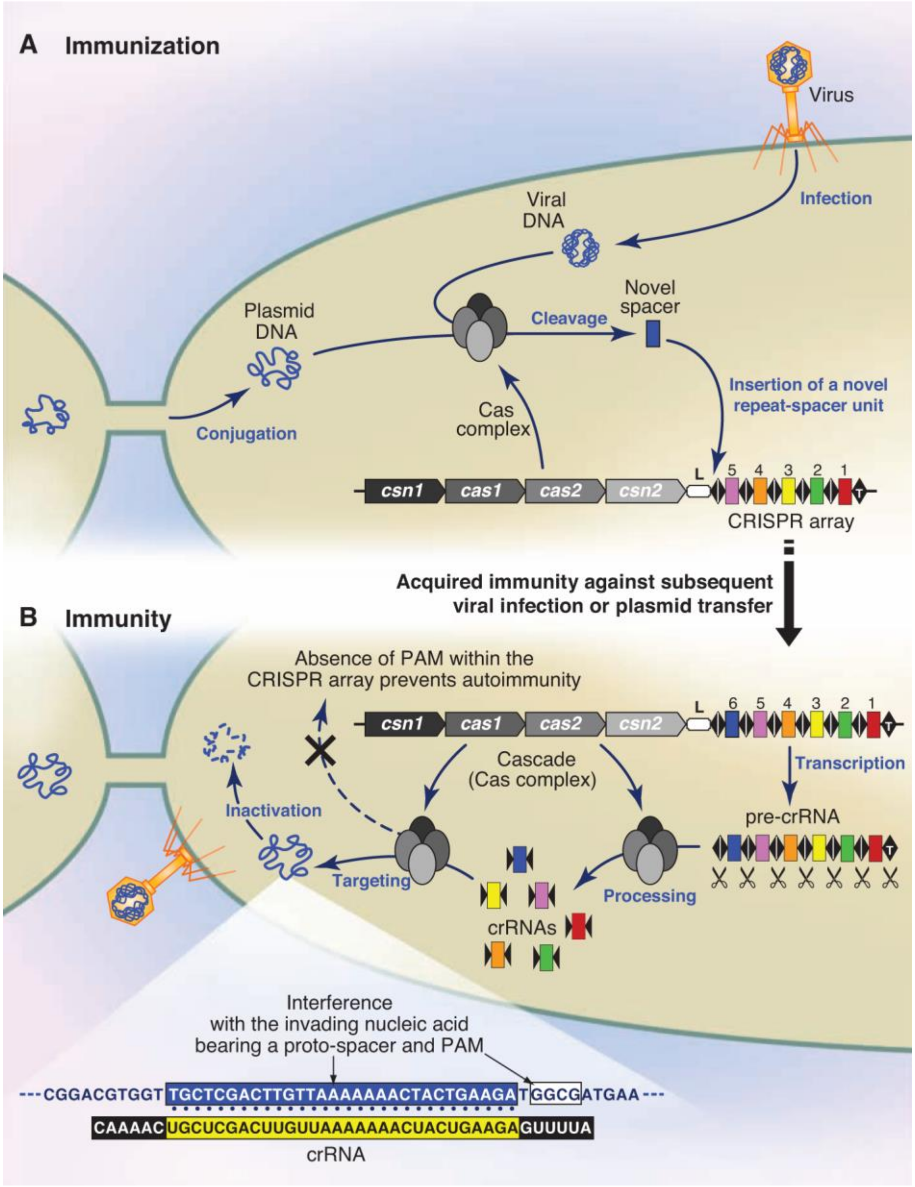

A Immunization
Virus
Infection
Viral DNA
Plasmid DNA
Cleavage
Novel spacer
Insertion of a novel repeat-spacer unit
Conjugation
Cas complex
csn1
cas1
cas2
csn2
L
5 4 3 2 1
CRISPR array
Acquired immunity against subsequent viral infection or plasmid transfer
B Immunity
Absence of PAM within the CRISPR array prevents autoimmunity
6 5 4 3 2 1
Transcription
Cascade (Cas complex)
pre-crRNA
Inactivation
Targeting
Processing
crRNAs
Interference with the invading nucleic acid bearing a proto-spacer and PAM
---CGGACGTGGT TGCTCGACTTGTTAAAAAAACTACTGAAGA T GGCG ATGAA---
CAAAAC UGCUCGACUUGUUAAAAAAACUACUGAAGA GUUUUA
crRNA

Figure 1.7. Mechanism of action of the CRISPR/Cas immune system. (A) Immunization process: Following an exposure to invading DNA from phage or plasmid, a Cas complex recognises and cleaves a short fragment of foreign DNA (proto-spacer) and incorporates it as a novel spacer-repeat unit between conserved short sequence repeats at the leader (L) end of a CRISPR sequence (repeats are represented as diamonds, spacers as rectangles). (B) Immunity process: The CRISPR repeat-spacer array is transcribed into a long RNA that is processed into small RNAs, which subsequently direct a Cas complex to target and inactivate viral genomes that correspond to the viral spacer sequence (Horvath & Barrangou, 2010). Phages can escape CRISPR/Cas-mediated resistance by mutating their proto-spacers or by mutating nearby proto-spacer adjacent motifs (PAMs), regions which probably act as recognition sites for the CRISPR/Cas system (Weinberger *et al.*, 2012). From (Horvath & Barrangou, 2010).

1.2.2.2. Virus discovery

Metagenomics can also be applied as a diagnostic tool to identify potential viral pathogens. It has practical application to discover viruses from humans, animals and plants with diseases of unknown etiology, especially when conventional testing laboratory techniques are unsuccessful. Application of the metagenomic technology for routine identification of infection agents could also help develop treatment or preventive vaccine (Mokili *et al.*, 2011). For example, Palacios *et al.* used high-throughput sequencing method to find the cause of death of two patients following organ transplant and identified a novel arenavirus (Palacios *et al.*, 2008). A similar approach was used to detect viral pathogens by sequencing cDNA from the human nasal and faecal samples during during seasonal influenza virus infections and norovirus outbreaks (Nakamura *et al.*, 2009). In another study, Ng et al. demonstrated the potential of viral metagenomics to establish the aetiology of a new disease in sea turtles (Ng *et al.*, 2009). More recently 454 pyrosequencing has been used to identify viral pathogens causing diarrhoea in dogs (Li *et al.*, 2011b). Coetzee et al. used a deep sequencing analysis to identify RNA viruses causing disease in grapevines (Coetzee *et al.*, 2010). These studies show that metagenomics is promising tool to identify viral candidates to establish aetiology a wide range of different diseases in a wide range of hosts.

1.2.2.3. Biotechnology

Viral metagenomes represent an unexplored source of novel proteins with biotechnological value. In functional metagenomics, DNA from different environments is used to construct a library in a bacterial host which is subsequently screened for clones expressing a desired enzymatic activity (Henne *et al.*, 1999). For example, metagenomic sequencing and functional screening have been used to identify active phage lytic enzymes from animal faeces (Schmitz *et al.*, 2010) or to identify a novel thermostable DNA polymerase from a hot spring (Moser *et al.*, 2012).

1.2.3. Methods for generating viral metagenomes

Metagenomic analysis of uncultured viruses typically involves three main steps: sample preparation, high-throughput sequencing and bioinformatic analysis. Methods used to obtain viral nucleic acids differ slightly depending on the sample type (liquid, solid). The general strategy for obtaining viral nucleic acids from different samples is outlined in Figure 1.8. The main goal is to concentrate viral particles and remove any unwanted background of prokaryotic and eukaryotic DNA by filtration of bacterial and eukaryotic cells, density ultracentrifugation in caesium chloride to purify intact virions and enzymatic digestion of non-capsid protected nucleic acids before nucleic acid extraction and random amplification of viral genomes (Delwart, 2007).

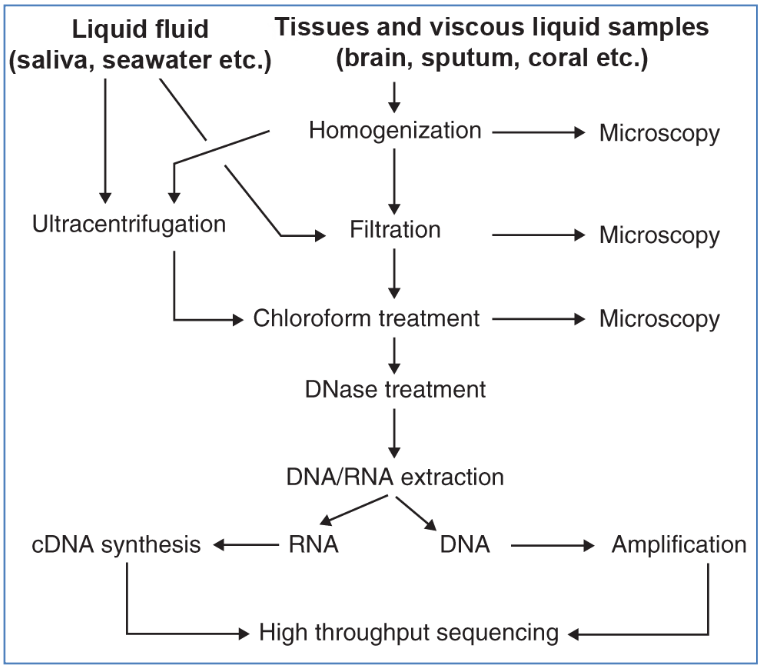


Figure 1.8. Overall protocol for isolating viruses from various samples. Adapted from (Mokili *et al.*, 2011).

1.2.3.1. Purification of viral particles

Samples collected from the environment are initially filtered, typically through a 0.2-µm pore size impact filter, in order to physically separate viruses from cellular organisms. However, filtering through the 0.2-µm filter may result in the loss of some viruses having larger capsids and longer tails (Brum & Steward, 2010), and some giant viruses e.g. mimivirus which is 0.4-µm in diameter (La Scola *et al.*, 2003). Large volume water samples (e.g. seawater, faeces) are filtered and concentrated by tangential flow filtration (Bench *et al.*, 2007; Breitbart *et al.*, 2002; Breitbart *et al.*, 2003). In this method liquid is recirculated across the filter (0.2-0.45 µm) to minimize filter clogging. A backpressure is applied to push virus particles out of the filter pores and the filtrate is collected and concentrated using a TFF (Tangential Flow Filtration) filter with a cutoff of Mr 100 000 (Casas & Rohwer, 2007; Thurber *et al.*, 2009). Some samples require pre-treatment prior to filtration, e.g. solid samples (such as tissue, faeces) require mechanical homogenization (Ng *et al.*, 2009; Zhang *et al.*, 2006), whereas liquid viscous samples such as sputum are treated with a solution of dithiothreitol (DTT) (Willner *et al.*, 2009), which is a reducing agent able to break the disulphide bonds present in mucus glycoproteins, commonly used to liquefy sputum before homogenization (Hammerschlag *et al.*, 1980). Viral particles can be purified using caesium chloride (CsCl) gradient centrifugation, in which viruses are separated based on their buoyant density (Casas & Rohwer, 2007; Thurber *et al.*, 2009). However, the main disadvantage of using CsCl gradients is that some types of viruses do not band in the selected density range or they are sensitive to CsCl (Thurber *et al.*, 2009). Chloroform treatment followed by DNase digestion is used to remove residual microbial contamination. The chloroform disrupts the membranes of bacterial and eukaryotic cells and results in the release of chromosomal DNA, which can be subsequently digested with nuclease (Mokili *et al.*, 2011; Willner *et al.*, 2011). The chloroform treatment however will also disrupt some lipid enveloped viruses (Breitbart & Rohwer, 2005b), therefore this step is sometimes omitted. Epifluorescence microscopy with SYBR Gold staining of nucleic acids enclosed within the capsid is used to verify that viral samples do not contain contaminating nuclei or microbial cells and to ensure that viruses are not lost during the sample processing (Thurber *et al.*, 2009). A commonly used method for the extraction of viral DNA is formamide/SDS-CTAB (Sambrook *et al.*, 1989; Thurber *et al.*, 2009), in which formamide is used in combination with SDS to achieve virus lysis. CTAB cationic detergent is used to remove polysaccharides and DNA is extracted by phenol/chloroform extraction followed by isopropanol precipitation (Sambrook *et al.*, 1989; Thurber *et al.*, 2009). Viral nucleic acids can also be extracted using commercial kits e.g. Qiagen which use QIAamp MinElute Virus Spin Kit to extract DNA/RNA. Once virus particles have been isolated, the viral DNA or RNA must be often amplified before sequencing.

1.2.3.2. Viral DNA amplification

Depending on the sample type, the yield of nucleic acids directly isolated from viral particles is usually very small (Table 1.4), often below the required minimum for standard Illumina or 454 next-generation sequencing (1-5 µg) (Thurber *et al.*, 2009). A variety of methods have been developed to amplify viral DNA such as linker-amplified shotgun library (LASL) or multiple displacement amplification (MDA) (Figure 1.9). Total viral RNA can be amplified using the Whole Transcriptome Amplification (WTA) method (Nakamura *et al.*, 2009) or converted to cDNA and amplified using the method described below. In the LASL method, a single linker is ligated to the fragmented viral DNA or cDNA and a primer complementary to the linker is used for PCR amplification (Breitbart *et al.*, 2002; Breitbart *et al.*, 2003; Culley *et al.*, 2006; Schoenfeld *et al.*, 2008). This method however is time consuming, requires relatively high initial DNA concentration and is limited to dsDNA viruses (Kim & Bae, 2011; Polson *et al.*, 2011).

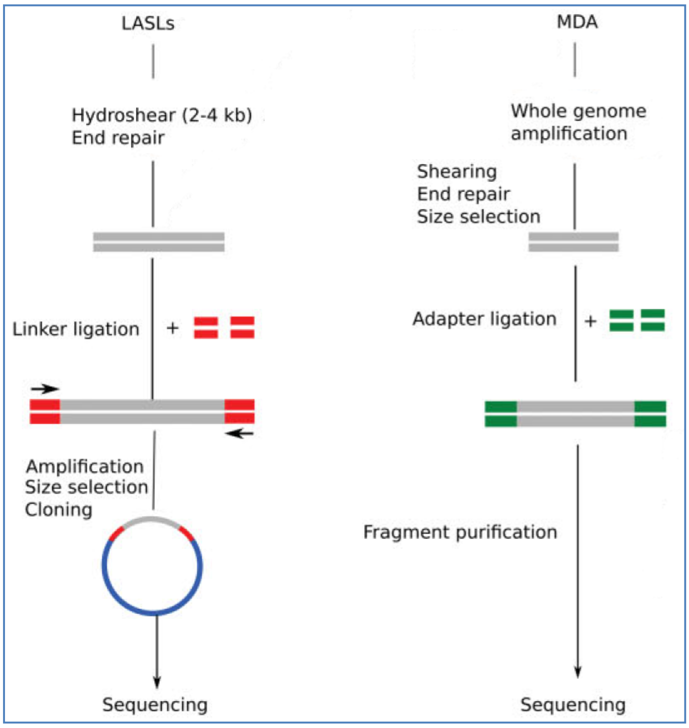


Figure 1.9. Comparison between LASLs and MDA methods. In LASLs (linker-amplified shotgun libraries) method a double-stranded linker is ligated to the randomly sheared metagenomic DNA. A primer complementary to the linker is used for PCR amplification and resulting PCR fragments are gel purified and inserted into a vector, cloned and then sequenced. In MDA (multiple displacement amplification) method, metagenomic DNA is amplified using Phi29 polymerase and sheared, followed by blunt end-ligation and sequencing. Figure adapted from (Willner & Hugenholtz, 2013).

The MDA method (Figure 1.10) uses Phi29 DNA polymerase, a highly processive (>70,000 base insertions per binding event) enzyme with strand displacement activity that enables amplification of complete viral genomes using random primers without the need for thermal cycling (Dean *et al.*, 2001; Pinard *et al.*, 2006; Thurber *et al.*, 2009). Phi29 polymerase can amplify viruses with circular or linear genomes and produce micrograms of products from nanogram DNA quantities (Dean *et al.*, 2001; Spits *et al.*, 2006). Although MDA is associated with biases and artefacts, such as formation of chimeras (Lasken & Stockwell, 2007), more efficient amplification of short circular templates (e.g. ssDNA viruses) than linear DNA (Kim *et al.*, 2008) and introduction of quantitative bias (Yilmaz *et al.*, 2010), it is currently the most widely used technique for viral DNA amplification.

Table 1.4. Amount of viral nucleic acid typically recovered from different samples.

Sample type	**Nucleic acid type**	**Volume or weight**	**Amount of nucleic acids extracted (ng)**	**Reference**
Marine sediments	DNA	1 g	50-100	(Breitbart *et al.*, 2004a; Thurber *et al.*, 2009)
Horse faeces	DNA	1 g	100	(Cann *et al.*, 2005)
Human faeces	RNA	500 g	140-200	(Zhang *et al.*, 2006)
Marine water	DNA	100 l	20-200	(Angly *et al.*, 2006; Thurber *et al.*, 2009)
Marine water	DNA	1,190 l	8000	(Steward & Preston, 2011)
Human diarrhoea	RNA	0.1 ml	100-300	(Finkbeiner *et al.*, 2008)
Hot springs	DNA	400-600 l	< 100	(Schoenfeld *et al.*, 2008)
Coral tissue	DNA	1 g	60-100	(Thurber *et al.*, 2009; Vega Thurber *et al.*, 2008)
Microbialites	DNA	1 g	50-100	(Desnues *et al.*, 2008; Thurber *et al.*, 2009)
Freshwater lake	DNA	1 l	50-80	(Lopez-Bueno *et al.*, 2009)

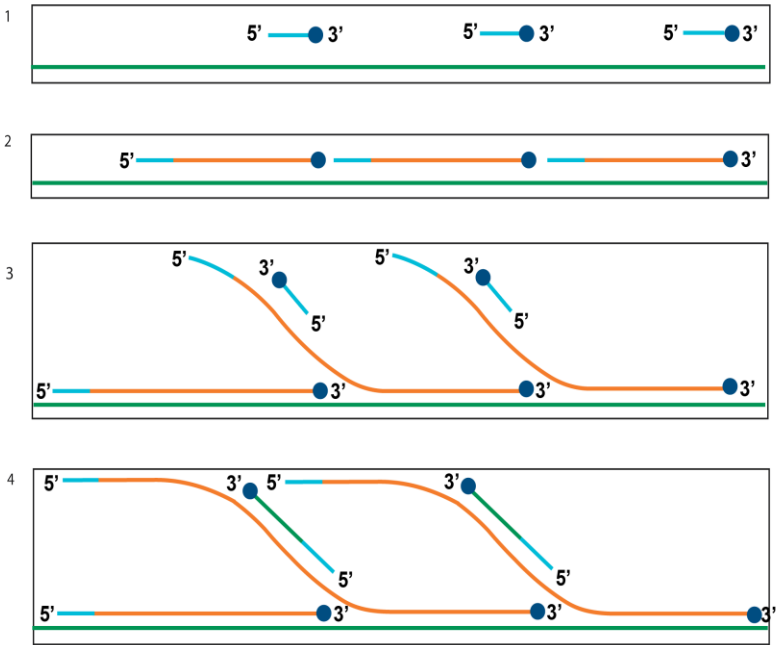


Figure 1.10. Mechanism of multiple displacement amplification. (1) Random hexamer primers (blue line) anneal to the denatured DNA (green line). (2) The phi29 DNA polymerase (blue circle) carries out DNA synthesis (orange line). (3) Following the strand displacement new primers bind to newly formed DNA and (4) replication from the new strand continues, resulting in formation of hyperbranched DNA structures. From (Spits *et al.*, 2006).

1.3. Introduction to next-generation sequencing and bioinformatics

1.3.1. Sequencing strategies

Early applications of metagenomic techniques to viral communities involved shearing of DNA, PCR amplification and random cloning into a vector for subsequent Sanger sequencing (Breitbart *et al.*, 2002; Breitbart *et al.*, 2003). Mechanical DNA shearing and PCR amplification prior to library construction overcame known difficulties of cloning viral DNA, which often contains modified nucleotides and bactericidal genes (Wang *et al.*, 2000; Warren, 1980), in bacterial hosts. Sanger sequencing produces long reads (up to 900 bp), but labour-intensive cloning and relatively high sequencing costs are disadvantages of this method (Table 1.5). Development of ‘‘next-generation’’ sequencing (NGS) technologies enabled high-throughput and less expensive methods for sequencing compared to the Sanger method (Table 1.5). NGS platforms are capable of sequencing hundreds of viral genomes without cloning by producing millions of sequence reads in a single run. The Roche 454 and the Illumina next-generation sequencing platforms have been used most frequently in viral metagenomics studies (Table 1.3). These methods rely on sequencing by synthesis technology (Fuller *et al.*, 2009) and, unlike the Sanger technique (Sanger *et al.*, 1977b), do not use chain-termination chemistry and capillary electrophoresis.

The 454 technology (Margulies *et al.*, 2005) commercialized by Roche (http://www.roche.com) produces the longest reads (up to 1000 bp) amongst NGS platforms (Table 1.5). The latest 454 instrument (454 GS FLX Titanium XL+) produce approximately 1 million sequences comprising 700 Mb of data per run (Table 1.5). For sequencing, DNA is fragmented, ligated to 454-specific adapters and immobilized on magnetic beads. The surfaces of beads carry oligonucleotides that are complementary to the 454-specific adapter sequences and one library fragment is attached per bead. The beads and PCR reagents are emulsified in an oil-water mixture and DNA is amplified on the surface of each bead (emulsion PCR). Next, the beads are loaded into picotiter plate wells, where the single bead enters one of the several hundred thousand individual wells (Mardis, 2008). The enzymes catalyzing the pyrosequencing reaction are added and a sequencing primer is hybridized to the 454-specific adapter. Nucelotides are sequentially added and, if complementary to the template strand, are incorporated by DNA polymerase. As a result of the incorporation, a pyrophosphate is released and subsequently converted via two enzymatic reactions into light detected by a charge coupled device camera (Ronaghi, 2001).

Table 1.5. Comparison of next-generation sequencing platforms. Information obtained from respective websites: Roche (http://454.com/), Illumina (http://www.illumina.com), SOLiD and Sanger (http://www.appliedbiosystems.com).

Company	**Sequencing platform**	**Read length (bp)**	**Number of reads per run**	**Bases per run**	**Run time**	**Cost per Mb (Glenn, 2011)**
Illumina	Genome Analyzer IIx	35 (SE)* 50/75/100/150 (PE)*	320 million (SE)	10-95 Gb*	2-14 days	$0.12
	HiSeq 2500	36 (SE) 50/100 (PE)	3 billion (SE)	95-600 Gb	2-11 days	NA*
	MiSeq	36 (SE) 25/100/150/ 250/300 (PE)	12-15 million (SE)	0.54-15 Gb	4-48 hours	$0.74
Roche	454 GS FLX Titanium XL+	1000	1 million	0.7 Gb	23 hours	$7
	454 GS Junior	400	0.1 million	0.035 Gb	10 hours	$22
ABI Life Technologies	SOLiD 5500xl	75+35	2.8 billion	180 Gb	7 days	< $0.07
ABI Life Technologies	Sanger 3730xl	500-900	96	0.08 Mb*	34 min - 3 hours	$2400

* SE Single read, PE Paired-end read, Mb Megabases, Gb Gigabases, NA data not available

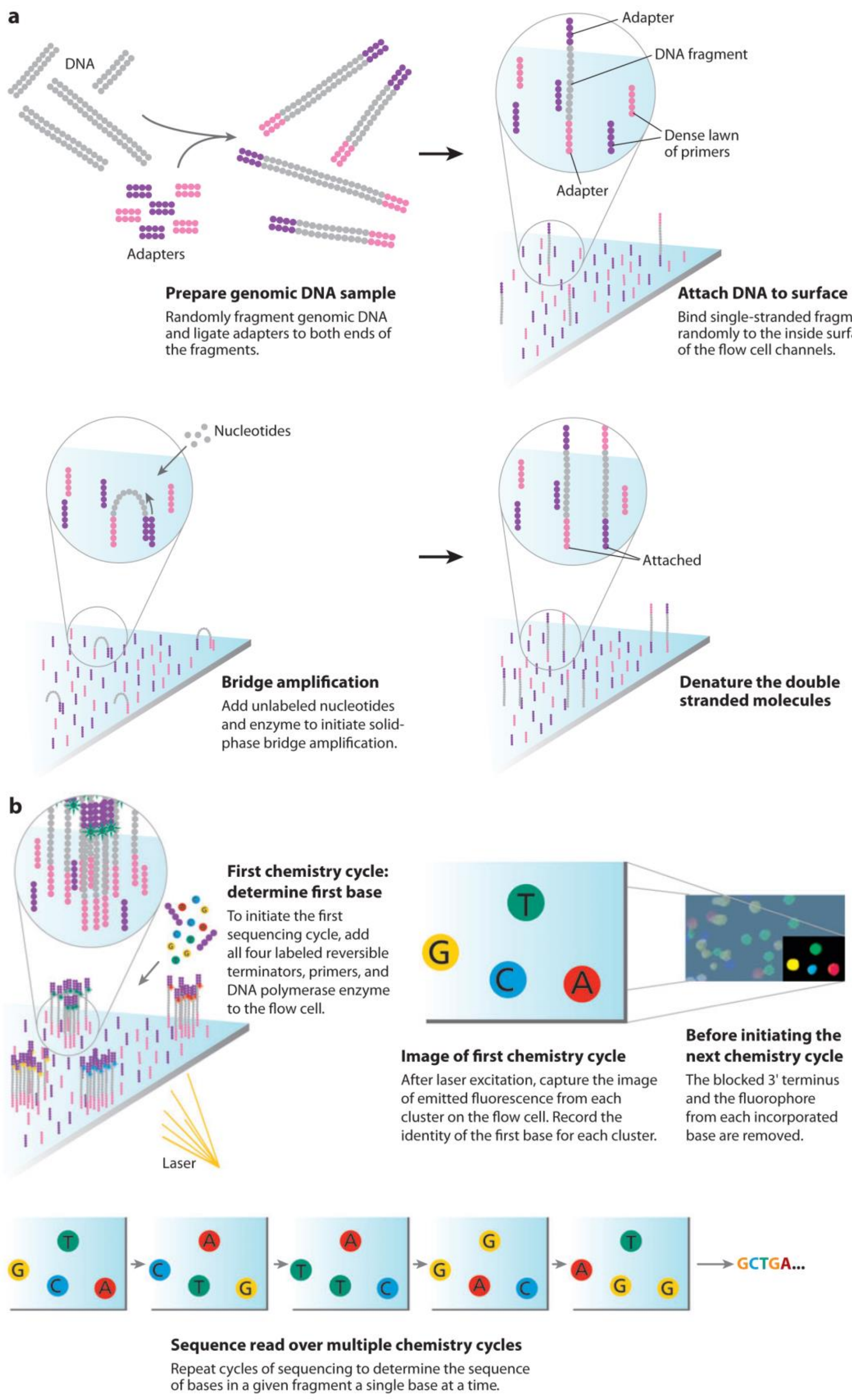

a
DNA
Adapters
Adapter
DNA fragment
Dense lawn of primers
Adapter
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Attached
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Denature the double stranded molecules
b
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.
T
G
C
A
Laser
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
GCTGA...
Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Figure 1.11. The principle of Illumina sequencing process. (A) DNA is converted into an Illumina adapter library and amplified by “bridge amplification” on the surface of the flow cell. (B) Amplified molecules are sequenced by the cycle reversible termination chemistry. Modified from (Mardis, 2008).

Illumina systems produce shorter reads (with single read length up to 300 bp), but have much higher throughput than Roche 454 and can generate up to 600 Gb per machine run (Table 1.5). The sequencing process begins with random fragmentation of starting DNA and ligation of Illumina-specific adapters to both ends of DNA fragments (Figure 1.11). DNA fragments are separated on the agarose gel and fragments between 200-300 bp are selected. Purified single-stranded DNA molecules are immobilized on a surface of a flow cell (a glass slide), which contains eight channels. Each channel can sequence eight independent libraries in parallel during the same instrument run (Shendure & Ji, 2008). The bridge PCR is used for amplification, resulting in clusters of identical DNA fragments. Reverse strands are removed and sequencing primers and modified nucleotides are added to each channel of the flow cell. The nucleotides are fluorescently labelled and carry a terminator group which allows only a single-base incorporation to occur in each cycle (Ju *et al.*, 2006). Sequencing is carried out by annealing primers to the adapter, followed by extension of the sequencing primers by DNA polymerase. The DNA polymerase reaction terminates after the first base incorporation and camera records the fluorescent signal emitted by the cluster. After each imaging step, the terminator group is removed allowing another incorporation cycle to start (Mardis, 2008). Depending on the library construction protocol used, sequences can be derived from one end (single read) or both ends of the library fragments (paired-end read). Paired end reads carry extra information useful for assembly.

1.3.2. Computation approaches for characterizing sequenced viral metagenomes

NGS platforms produce millions of reads from a single sample, which may contain hundreds to thousands of different species. The analysis of the viral metagenomes is challenging because the large amount of data generated by high-throughput sequencing requires significant computational resources. Some challenges in virus identification and functional identification from genome sequences arise from the fact that distantly related viruses share common genomic regions (Hendrix *et al.*, 1999), and that most viruses contain genes without any known homologues (Yin & Fischer, 2008). In addition, viruses can stably integrate into host genomes and have the ability to carry genes of host origin within their own genomes, for example photosynthetic (Lindell *et al.*, 2004) or 16S ribosomal DNA genes (Del Casale *et al.*, 2011a), which can make it difficult to distinguish viral sequences from host sequences. Bioinformatic analyses of viral metagenomes address questions about their diversity (how many different viruses are present), taxonomy (what viruses are present), function (what genes they encode), and how they differ from other metagenomes.

A number of computational tools available as standalone and web-based versions have been developed to analyse metagenomic data (Table 1.6). Data analysis begins with quality control and the pre-processing of the raw reads. Sequence processing tools easily run from a web interface such as PrinSeq can be used to filter out unsatisfactory reads which are too short, or contain ambiguous bases (N), as well as low quality reads and read duplicates (Schmieder & Edwards, 2011a). A necessary step of sequence analysis is removing host-associated (e.g. human, mosquito) and other non-viral sequences that had not been removed by the pre-sequencing filtration, chloroform and DNase treatment. For example viral metagenomes generated from the clinical samples may contain a high background of human reads (Lysholm *et al.*, 2012; Nakamura *et al.*, 2009; Willner *et al.*, 2009). Those sequences slow down the downstream analysis and may result in misassembly of sequence contigs (Schmieder & Edwards, 2011b). Contaminating sequences can be removed with the help of DeconSeq software, which aligns data for removal against a contaminant database (for example the human genome) based on a similarity search (Schmieder & Edwards, 2011b). This is a very fast moving field with rapid obsolescence and rapid development of new analysis tools.

Table 1.6. Computational tools and methods used in viral metagenomics.

Analysis	Tool	Version	Reference	Description
Quality control	PrinSeq	Standalone Web	(Schmieder & Edwards, 2011a)	Provide summary statistics for read length, GC content, sequence complexity and quality score distributions, number of read duplicates, occurrence of Ns and more
Duplicates removal	PrinSeq	Standalone Web	(Schmieder & Edwards, 2011a)	Remove sequence duplicates
	CD-HIT-454	Standalone	(Li & Godzik, 2006)	
Filtering	PrinSeq	Standalone Web	(Schmieder & Edwards, 2011a)	Remove short or long sequences, sequences with N's, low-quality sequences
Decontamination	DeconSeq	Standalone Web	(Schmieder & Edwards, 2011b)	Remove sequence contamination by aligning reads against available reference database (web version) or custom database (standalone version)
Assembly	MetaVelvet	Standalone	(Namiki *et al.*, 2012)	Assemble metagenomic reads based on construction of de Bruijn graph
	Meta-IDBA	Standalone	(Peng *et al.*, 2011)	
	IDBA-UD	Standalone	(Peng *et al.*, 2012)	
	Genovo	Standalone	(Laserson *et al.*, 2011)	Assemble metagenomic reads based on construction of a Bayesian probabilistic model
Annotation	BLAST	Web Standalone	(Altschul *et al.*, 1990)	Sequence similarity search program
	MEGAN	Standalone	(Huson *et al.*, 2007) (Huson *et al.*, 2011)	Import BLAST outputs and automatically calculate a taxonomic and functional classification of the reads
	MG-RAST	Web	(Meyer *et al.*, 2008)	Metagenomic database and automated analysis platform for annotating metagenomes
	Camera	Web	(Seshadri *et al.*, 2007)	Metagenomic database and web server that provide a wide range of tools for metagenomic data analysis
	IMG/M	Web	(Markowitz *et al.*, 2008)	Metagenomic database and web server that provide comparative metagenome data analysis tools
	WebMGA	Web	(Wu *et al.*, 2011a)	Web server that provide a wide range of tools for metagenomic data analysis
Estimation of abundance and diversity	GAAS	Standalone	(Angly *et al.*, 2009)	Estimate relative species abundances based on significant BLAST similarities
	Circonspect	Standalone	(Angly *et al.*, 2006)	Generate contig spectra
	PHACCS	Standalone	(Angly *et al.*, 2005b)	Estimate the structure and diversity of viral metagenomes based on contig spectra information

1.3.2.1. Assembly

The Illumina Genome Analyzer platform used in this study produces reads up to 150 bp long. This read length is generally too short to detect sequence homologs using similarity searches (Wommack *et al.*, 2008). Assembly of short read fragments is performed to generate longer sequences called contigs. Longer reads show increased taxonomic and functional assignment, and enable reconstruction of complete genomic sequences of the dominant species within a metagenome (Fancello *et al.*, 2012). Two strategies are used: mapping assembly, in which short reads are mapped against a reference genome, or *de novo* assembly, in which overlapping reads are assembled into contigs, and no information about reference sequence is required (Thomas *et al.*, 2012). A typical viral metagenome contains 60-99% sequences with no similarity to known sequences (Table 1.3) (Mokili *et al.*, 2011). Therefore *de novo* assembly is usually performed prior to classification. A variety of *de novo* assemblers that assemble metagenomic short sequences are publicly available, including meta-IDBA (Peng *et al.*, 2011), Genovo (Laserson *et al.*, 2011), MetaVelvet (Namiki *et al.*, 2012) and IDBA-UD (Peng *et al.*, 2012). Most *de novo* assemblers are based on the “de Bruijn graph” approach. In the de Bruijn graph method, short reads are split into shorter fragments (called *k*-mers). The graph representing the assembly is built by connecting the *k*-mers with the overlapping fragments (called edges) into nodes using an algorithm (Compeau *et al.*, 2011). The de Bruijn graph programs which generate single genome *de novo* assemblies such as Velvet (Zerbino & Birney, 2008) or SOAPdenovo (Li *et al.*, 2010b) do not work well for metagenomic datasets which contain multiple genomes from different species. Metagenomic *de novo* assemblers such as meta-IDBA or meta-VELVET identify and separate within the entire de Bruijn graph subgraphs that represent very similar regions in subspecies. The sequences subgraphs are then aligned to produce a consensus sequence which represents a contig of subspecies from a single species and is merged into a larger component using paired-end reads. Meta-IDBA can generate longer and more accurate contigs compared to traditional assemblers (Peng *et al.*, 2011). Recently developed new assembler IDBA-UD is capable to assemble short reads with highly uneven sequencing depths, which results in longer sequence assembly and higher accuracy compared to existing assemblers such as (Velvet, SOAPdenovo and Meta-IDBA) (Peng *et al.*, 2012).

1.3.2.2. Classification

Metagenomic sequences are typically assigned to taxonomic groups based on a BLAST comparison (Altschul *et al.*, 1990) against reference databases that contain sequences of known origin such NCBI nucleotide and protein databases (Sayers *et al.*, 2009). Metabolic functions are assigned based on a BLAST comparison against databases containing functionally annotated sequences such as NCBI non-redundant (nr) protein database (Sayers *et al.*, 2009), NCBI Clusters of Orthologous Groups (COGs) (Tatusov *et al.*, 2000), Pfam (Bateman *et al.*, 2002) or SEED (Overbeek *et al.*, 2005). The most commonly used BLAST programs for classification of viral sequences are BLASTN, BLASTX and TBLASTX (Table 1.3). BLASTN compares nucleotide sequences to a nucleotide sequence database, BLASTX compares translated nucleotide sequences to a protein sequence database, whereas TBLASTX compares translated nucleotide sequences to a nucleotide sequence database translated in all possible reading frames (Altschul *et al.*, 1997). The similarity thresholds (E-value) cutoff 10^{-3} (less stringent search) or 10^{-5} (more stringent search) have been frequently used to classify viral sequences (Table 1.3). The BLAST cutoffs have to be chosen carefully, because less stringent searches may result in inaccurate reads classification, while more stringent searches may result in higher number of unannotated sequences (Weng *et al.*, 2010).

High-throughput BLAST output files can be viewed using MEGAN (MEtaGenome ANalyzer) software (Huson *et al.*, 2007). MEGAN uses significant (i.e. with high bit-scores) BLAST hits for each read and assigns them to the lowest node (e.g. species) in the NCBI taxonomy using a lowest common ancestor (LCA) algorithm. Sequences that cannot be assigned with a reasonable level of confidence at e.g. a species level are assigned to a higher taxonomical level (Huson *et al.*, 2007; Huson *et al.*, 2011). GAAS (Genome relative Abundance and Average Size), is another tool used for visualizing annotation results derived from BLAST searches. GAAS runs BLAST against completely sequenced genomes and considers all significant similarities to assign taxonomy (Angly *et al.*, 2009). Then it normalizes the number of reads assigned to a specific genome by the length of that genome. Normalization allows for more accurate estimation of species relative abundance, detecting small genomes which could be missed completely using standard analysis (Angly *et al.*, 2009).

To classify high-throughput datasets BLAST is often run from the command line, however this requires access to high-performance computing clusters and prolonged computational time. Recently, web-based metagenomic annotation platforms, such as CAMERA (Seshadri *et al.*, 2007), MG-RAST (Meyer *et al.*, 2008), IMG/M (Markowitz *et al.*, 2008) and WebMGA (Wu *et al.*, 2011a) have been developed for the remote automatic phylogenetic and functional analysis of metagenomes. The MG-RAST web interface allows data repository, annotation and comparative analysis. Environmental sequences are assigned to taxonomic and functional categories based on similarities to protein databases (e.g. NCBI nr database, SEED). The first step of MG-RAST analysis is open reading frame (ORF) prediction followed by BLAST-like comparison using the BLAT alignment tool (Kent, 2002). Long reads can contain multiple ORFs, and each of these will be annotated separately. The MG-RAST pipeline is therefore more suitable for annotation of short reads, rather than contigs (http://metagenomics.anl.gov/). More than 12,000 public metagenomes are freely available for comparison within MG-RAST (http://metagenomics.anl.gov/). The MG-RAST web interface incorporates PCA (Principal Component Analysis) for metagenome comparison. In PCA, datasets that exhibit similar abundance profiles (taxonomic or functional) are clustered together with respect to components of variation extracted from their normalized abundance profiles (http://metagenomics.anl.gov/).

1.3.2.3. Diversity

Viral community structure and diversity can be estimated using the Circonspect (Angly *et al.*, 2006) and PHACCS (Phage Communities from Contig Spectrum) (Angly *et al.*, 2005b) tools. Circonspect (Control In Research on CONtig SPECTra) uses an external assembly program and a bootstrap technique to compute a contig spectrum. A contig spectrum is the count of the number of contigs of different size generated by assembly. The resulting contig spectra are then mathematically modelled to predict community structure and diversity using PHACCS. The diversity is estimated by measuring the community richness (the total number of different species), evenness (the relative abundance of species) and the Shannon-Wiener index. The Shannon–Weaver index considers both the total number of species and the relative distribution of these species (Angly *et al.*, 2005b). More diverse communities have a higher index (Table 1.2).

1.4. Objectives of this study

The objectives of this study were:

1. To characterise the viral communities from a clinical (cystic fibrosis patient's sputum) and an environmental (dairy food wastewater sludge) sample by sequencing metagenomic DNA using Illumina Genome Analyzer II technology and analysing the sequence data using a variety of bioinformatics tools.

2. To explore functional aspects of metagenomic viral DNA in an *E coli* host by promoter trap approach applied to a metagenomic clone library derived from a dairy food wastewater viral metagenome.

3. To characterise the diversity of T4-type bacteriophages in dairy food wastewater samples using a PCR-based approach of conserved gene.

Chapter 2:

Metagenomic sequencing of DNA viruses in cystic fibrosis sputum

Abstract

In this study metagenomic short read (Illumina) sequencing was used to explore the DNA virus community in sputum from a cystic fibrosis (CF) patient infected with *Pseudomonas aeruginosa*. Bacteriophages were isolated from a sputum sample using a sequential filtration technique with DNase treatment. Phage DNA was extracted and amplified by a rolling circle technique and then sheared before sequencing as a paired-end sequencing run using Illumina technology. Although 97% of reads could not be matched to any GenBank sequence, 53% of the reads were assembled into contigs greater than 100 bp using the meta-IDBA assembler. Forty four contigs were over 10 kb and the largest contig was 61 kb. Four specimen contig assemblies were verified by PCR and sequencing from the original DNA extract. BLASTX assigned reads and contigs were mainly of phage or prophage origin, but the majority of the sequence data obtained was classified as “unassigned”. Phages from bacteria of the genera *Streptococcus*, *Veillonella*, *Pseudomonas*, *Actinobacillus* and *Prevotella* comprised the most frequently assembled contigs. A complete human Torque Teno-like virus was detected in the metagenome assembly and by PCR. A substantial number of assigned sequences could be affiliated with species typically found in the oral cavity, as well as CF lower respiratory tract samples, reflecting the nature of sputum. Viral diversity was low (89 species), which presumably contributed to the successful assembly. Numerous antimicrobial resistance genes were detected flanked by phage sequences, suggesting sputum of cystic fibrosis patients is a source for transmissible antimicrobial resistance. The metagenome assembly contained matches to CRISPR sequences in bacteria known to infect CF patients in other locations, but no matches to locally detected bacterial CRISPR sequences. This is compatible with local phage selection by bacterial CRISPR systems and a limited global diversity of phages in the respiratory tract of cystic fibrosis patients.

2.1. Introduction

Cystic fibrosis is one of the commonest autosomal recessive diseases in Caucasian people, resulting from various mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) (Cheung & Deber, 2008). A mutation in this gene is carried by 1 in 19 Irish individuals, the highest frequency in the world (Farrell *et al.*, 2007). CF patients suffer recurrent lung infections and their respiratory tract becomes colonized with pathogenic bacteria, the most common of which are *Pseudomonas aeruginosa, Staphylococcus aureus* and *Haemophilus influenzae* (Lyczak *et al.*, 2002). In addition, recent culture-independent studies indicate the presence of other bacterial species in CF sputum including *Neisseria, Streptococcus* and anaerobic bacteria such as *Prevotella* and *Veillonella (van der Gast et al., 2011).* Human respiratory viruses may also play a significant role in CF exacerbations (Wat & Doull, 2003). The co-occurrence of multiple infections with multi drug-resistant biofilm-forming bacteria makes infections difficult to treat. Chronic pulmonary infections and the consequent host inflammatory response cause irreversible lung damage which eventually leads to respiratory failure (Ratjen & Doring, 2003). Diagnosis and treatment of agents causing respiratory infection in cystic fibrosis is of high importance. There has been a recent revival of interest in bacteriophage therapy as an alternative to antibiotics in intractable infections such as cystic fibrosis (Harper & Enright, 2011). Bacteriophages have been shown to encode enzymes that can penetrate CF biofilms and access susceptible bacteria (Glonti *et al.*, 2010). Recently, phage therapy has been successfully used to treat staphylococcal infection in a patient with cystic fibrosis (Kvachadze *et al.*, 2011) and in an animal model of *Pseudomonas* lung infection (Alemayehu *et al.*, 2012a). However, lytic bacteriophages have also been shown to select for the mucoid phenotype in *Pseudomonas aeruginosa* in vitro (an adaptive response increasing immune evasion and antimicrobial resistance properties which is characteristic of *Pseudomonas aeruginosa* strains infecting the respiratory tract of CF patients) (Scanlan & Buckling, 2012). Because of this potential application of phages or phage genes in the treatment of microbial infections in cystic fibrosis, the apparent adverse consequences of phage-mediated bacterial selection in the CF disease process, and because phages can encode and transfer harmful toxin and antibiotic resistance genes to bacteria (Rolain *et al.*, 2011), it is important to determine the pre-existing variety of genes encoded in respiratory tract phages in cystic fibrosis.

Most culture-independent studies carried out on CF sputum have mainly focused on analysis of microbial populations. One centre in the USA has carried out comparative metagenomic analysis of viruses in sputum samples from healthy subjects and CF individuals (Willner *et al.*, 2009) and in lung tissue from patients with late-stage CF (Willner *et al.*, 2012). The metabolic functions of the viral community in the sputum of CF patients were different from non-diseased individuals, suggesting that, rather than targeting dominant taxa, changing the environment of the airways may be a way to treat infections in CF individuals, particularly as antimicrobial treatment lose efficacy. Genes encoding antimicrobial resistance were found in both viromes indicating that phages are responsible for spread of antibiotic resistance in CF (Fancello *et al.*, 2011; Willner *et al.*, 2009; Willner *et al.*, 2012). Both of these studies used Roche 454 pyrosequencing technology. There is evidence that Illumina sequencing technology, which provides reads which are shorter, but contain fewer frameshifts potentially truncating read assemblies, may yield longer and more accurate contigs than 454 on assembly of sequence reads from the same metagenomic samples (Luo *et al.*, 2012).

This chapter describes the use of next-generation sequencing technology to characterize the taxonomic and metabolic profile of viral communities present in a sputum sample of a cystic fibrosis patient from Cork, Ireland. A DNA viral metagenome was obtained by filtration, CsCl purification DNase treatment, DNA isolation and whole genome amplification. The resulting DNA was sequenced using Illumina short read technology. Bioinformatic analyses achieved assembly of large contigs demonstrating the presence of a small human DNA virus (TT virus) and numerous phages similar to recognised phages of *Streptococcus*, *Veillonella*, *Pseudomonas*, *Actinobacillus* and *Prevotella*. Many hits to a CRISPR variable region database were obtained. Functional assignment showed high abundance of genes involved in phage lytic and lysogenic growth. Multiple antibiotic resistance genes were identified, including β-lactamases.

2.2. Materials and Methods

2.2.1. Sample collection

A sputum sample of approximately 10 ml was collected from a male Cystic Fibrosis adult patient during a physiotherapy session on 24 March 2010 at Cork University Hospital, Cork, Ireland. The study was approved by the Clinical Research Ethics Committee of the Cork Teaching Hospital and written informed consent was obtained from the patient.

2.2.2. Viral particle purification

The sputum was homogenized in 10 ml of 0.02 µm pre-filtered SM buffer (100 mM NaCl, 8 mM $MgSO_4 \cdot 7H_2O$, 50 mM Tris-HCl pH 7.5, 0.002% (w/v) gelatin) using 10 ml syringe. To reduce viscosity and background human DNA the homogenized sputum was treated with 10 ml of 6.5 mM dithiothreitol (DTT; Sigma) and 1 ml of DNase I (Pulmozyme), homogenized again, and incubated at 37°C for 30 minutes. After incubation the sample was centrifuged at low speed (2000 rpm) for 10 minutes and the supernatant was filtered through a 0.45 µm syringe filter (Millipore) to remove most of bacteria.

Viruses in the filtrate were purified using a published CsCl density gradient centrifugation method (Thurber *et al.*, 2009). Briefly, CsCl was added to the filtrate to a density of 1.15 g/ml and and loaded onto a caesium chloride step gradient consisting of 1 ml each of 1.7, 1.5, and 1.35 g/ml. The gradient was centrifuged in a SW-41 (Beckman) rotor at 22,000 rpm for 2 h at 4°C. A 21-gauge needle on a 1 ml syrine was used to collect1 ml containing concentrated viral particles from the interface between the 1.35- and1.5-g/ml layers. Collected CsCl fraction was additionally filtered through a 0.22 µm impact filter (Millipore) and further concentrated and washed twice with 1 ml of SM buffer on a Millipore Amicon Ultra-15 Centrifugal Filter Unit (30 kDa). The final volume of concentrated CsCl fraction was 400 µl and was used for DNA extraction.

2.2.3. DNA isolation and sequencing

Prior to DNA extraction, concentrated CsCl fraction was incubated for 1 h at 37°C with 125 U/ml of DNase I (New England Biolabs) and 3 µl of RNA (DNeasy Blood & Tissue Kit, Qiagen) to eliminate any remaining free nucleic acids. DNase was inactivated by the addition of EDTA to the final concentration of 20 mM, followed by heat-inactivation at 80°C for 10 minutes. Viral DNA was extracted using CTAB extraction method (Sambrook *et al.*, 1989; Thurber *et al.*, 2009). Briefly, the sample was mixed with 15 µl of 20% (w/v) SDS and 20 µl of proteinase K (DNeasy Blood & Tissue Kit, Qiagen) and incubated for 1 h at 56^{o}C. After protease treatment 100 µl of 5 M NaCl and 80 µl of pre-heated CTAB/NaCl solution (10% (w/v) CTAB in 0.7 M NaCl) was added and incubated for 10 minutes at 65°C. DNA was recovered by phenol/chloroform extraction and isopropanol precipitation and the resulting DNA pellet was resuspended in 30 µl of sterile water. DNA concentration was estimated using NanoDrop (Thermo Scientific), giving a 30 ng/µl concentration and a 260/280 ratio of 1.8. DNA was used for 16S rRNA PCR amplification using universal primers 27F and 907R (Lane, 1991) and no PCR products were obtained, suggesting that filtration, density-gradient ultracentrifugation and DNase treatment steps resulted in successful removal of bacteria.

Extracted metagenomic DNA was used for whole genome amplification using GenomiPhi V2 DNA Amplification Kit (GE Healthcare). Briefly, 1µl (30 ng) of DNA was amplified in 20 µl reaction volumes in triplicate reactions at 30°C for 2 h. The amplified products from each reaction were pooled, purified using DNeasy Blood & Tissue Kit (Qiagen) and resuspended in 100 µl of sterile water. Approximately 3 µg of DNA was used to make libraries with inserts of between 150 and 250 bp at the Wellcome Trust Sanger Institute (UK) which were sequenced in a paired 76 cycle run using an Illumina Genome Analyzer IIx. 6.4 million paired-end 76 bp reads were generated.

2.2.4. Sequence processing

All large-scale computational analyses were performed on The Boole Centre for Research in Informatics (BCRI) compute cluster at University College Cork. The metagenomic library (approximately 6.4 million reads) was filtered using PRINSEQ website http://edwards.sdsu.edu/prinseq_beta/#_(Schmieder & Edwards, 2011a) to remove exact sequence duplicates and reverse complement exact duplicates (1,633,820), reads containing more than one ambiguous bases (N) and low-complexity sequences (DUST score <32) (11,455). Human sequences (94,786) were removed using DeconSeq website http://edwards.sdsu.edu/cgi-bin/deconseq/deconseq.cgi using 90% coverage and 90% identity filtering options. The metagenomic sequences in CF metagenome were compared to another metagenome that was prepared at the same time using DeconSeq standalone (Schmieder & Edwards, 2011b) and any shared sequences that could be the results of contamination during the sample manipulation were removed using 90% coverage and 90% identity options (2,917,648 sequences). These pre-processing steps resulted in 1,715,261 high-quality sequences (130,359,836 total bases).

2.2.5. Assembly of sequence reads

For contigs assembly, sequence reads were prepared as described above, except duplicate reads were not removed, as this resulted in higher N50 (median contig size) of 4335. Sequence reads were pair-end assembled using Meta-IDBA (Peng *et al.*, 2011) into 9,788 contigs (3,131,945 total bases). Sequences less than 100 bp were discarded, leaving a total of 2,859 contigs for analysis (2,637,642 total bases). To calculate number of reads that were recruited into contigs assembly, reads were aligned to the contigs using FR-HIT (Niu *et al.*, 2011) using high-stringency cutoff of 100% identity over 100% of the entire read length.

Coding DNA Sequence (CDS) prediction was performed on the assembled sequences (>100 bp) using MetaGeneMark Heuristic Approach version 1.0 (Zhu *et al.*, 2010) at http://exon.biology.gatech.edu/metagenome/Prediction/index.cgi.

2.2.6. Sequence annotation

Individual reads were automatically annotated against the GenBank database (e-value cutoff of 1e-03 and minimum alignment length of 20 bp) using MG-RAST (Meyer *et al.*, 2008). Contigs with minimum length greater than 1 kb were annotated against the GenBank protein (nr) database (downloaded on 25 October 2011) using BLASTX (version 2.2.24) and an e-value 1e-05. Top BLASTX hit with minimum identity of >80% and highest bit score was used to classify the sequences. The matches to each taxon were normalized by counting number of bases in contigs assigned to each taxon and dividing by the total number of bases (3,131,945 bp). DNA sequences can be accessed through MG-RAST website under Project IDs 4476065.3 (reads) and 4476037.3 (contigs). Filtered reads were also submitted to the Sequence Read Archive (SRA) under accession number SRA057676.

Functional annotation was performed on contigs and CDS using the MG-RAST Subsystems (e-value 1e-05 and minimum alignment length 50) and WebMGA (Wu *et al.*, 2011b) COG (Clusters of Orthologous Groups of proteins) (Tatusov *et al.*, 2003) (e-value 1e-05) databases.

Selected contigs were compared to the reference genome that matched most closely (best hit from BLASTX search). Protein-coding regions and GenBank files of contigs were generated using the phage genome annotation tool ArtAnnoPipe (http://athena.bioc.uvic.ca/node/541). Annotated reference genomes used in this study were downloaded from NCBI. TBLASTX was used to compare the contig sequences with the corresponding reference genome (with minimum alignment length of 50 and e-value 1e-05) and drawn in Easyfig (Sullivan *et al.*, 2011).

2.2.7. Metagenome diversity

Diversity of the metagenome was estimated using PHACCS version 1.1.3 (Angly *et al.*, 2005a) (http://sourceforge.net/projects/phaccs/). Circonspect version 0.2.5 (http://sourceforge.net/projects/circonspect/) implemented with the Octave version 3.6.0 was used to calculate contig spectra based on metagenome assembly using Minimo (98% identity over at least 35 bp overlap). The contig spectra were used as an input for PHACCS, using a logarithmic model and an average genome size of 50 kb.

2.2.8. Metagenome comparison

A multiple comparison based on organism and functional gene abundance between CF sputum contigs (>100bp) and other viromes: CF sputum (MG-RAST ID 4441908.3), human lower respiratory tract (NCBI GenomeProject ID 64629), healthy human saliva (4445731.3) oropharyngeal viromes of healthy individuals (MG-RAST ID 4444195.3 and 4444196.3), and terrestrial hot springs (4441096.3) was performed using MG-RAST Principal Component Analysis. The data was compared to GenBank and SEED databases using a maximum e-value of 1e-05, a minimum identity of 80 %, and a minimum alignment length of 50. The data has been normalized to values between 0 and 1 and drawn using a Bray-Curtis distance.

2.2.9. Antibiotic resistance genes

To identify genes potentially conferring resistance to antibiotics, reads and predicted CDS from contigs were compared to 3,375 antibiotic resistance associated genes downloaded from The Comprehensive Antibiotic Resistance Database (CARD) http://arpcard.mcmaster.ca/ using BLAST and e-value 1e-03. Sequences associated with antibiotic resistance genes according to CARD database BLAST results were extracted for further BLAST against NCBI nr database. Only BLAST hits with 90% identity over 90% of the sequence read length were analysed. Hits to genes encoding integrases, efflux pumps, as well as *gyrA*, *gyrB*, *parC*, *rpoB*, *rpsL*, *bacA*, in which a point mutation in a bacterial gene confers antimicrobial resistance, were discarded. Easyfig was used to visualize CDS of two contigs (contig 22 and contig 71) encoding metallo-β-lactamase genes. CDS were annotated using BLASTP against NCBI nr database. Phylogenetic analysis was conducted using MEGA version 5.04. Contigs used for phylogenetic comparison were submitted to GenBank under accession numbers JX157235 (contig 22) and JX157236 (contig 71).

2.2.10. Virulence genes

Virulence factors including bacterial toxin protein sequences were downloaded from the VFDB (Virulence Factors of Pathogenic Bacteria Database) (Chen *et al.*, 2012) http://mvirdb.llnl.gov/ and compared to the CDS predicted from contigs using BLASTP and e-value 1e-03.

2.2.11. CRISPR spacer analyses

52,511 spacers from CRISPRdb (Grissa *et al.*, 2007b) http://crispr.u-psud.fr/crispr/CRISPRUtilitiesPage.html were downloaded on 29/01/12 and used to search for sequence similarity with metagenomic reads and contigs (>100 bp) using

BLASTN (e-value 1e-03 and word size 7). Only matches that had 100% identity over 20 bp were analysed, as previously described (Anderson *et al.*, 2011).

To identify CRISPR spacers of different *Pseudomonas aeruginosa* strains present in CF sputum samples of local patients, conserved repeat sequences (5’-GTTCACTGCCGT(A/G)TAGGCAGCTAAGAAA-3‘) of complete genomic sequences of *Pseudomonas aeruginosa* strains were retrieved from the CRISPRdb database. The repeat sequence-specific PCR primers CRISPR-PAF 5‘-CTAGTTCACTGCCGT-3‘ and CRISPR-PAR 5‘-TTTCTTAGCTGCCTAY-3‘ were designed manually. The forward primer contained at its 5’ end a three-base pair sequence CTA complement to the leader sequence, directly adjacent to the repeat array. CRISPR spacers and repeats were amplified from total bacterial DNA extracted from 1ml of sputum sample (4 samples in total, including specimen sequenced in this study) using The Wizard® Genomic DNA Purification Kit (Promega). Amplified PCR products were cloned and sequenced at GATC Biotech. CRISPRFinder (Grissa *et al.*, 2007a) was used to identify spacer sequences. All retrieved spacers were subjected to BLASTN analysis against the GenBank databases (nt, HTGS, wgs). Sequenced PCR products were submitted to GenBank under accession numbers JX157240 – JX157254.

2.2.12. PCR validation of contigs

Contigs of interest were selected for PCR verification of the assembly. Primer 3 (Rozen & Skaletsky, 2000) was used to design primers spanning 3 genes (Figure 2.2). Inward-pointing and outward-pointing PCR primers were designed based on the DNA sequence of the longest contig (contig 195) having similarity to TTV virus. PCR reaction (50 µl reaction volume) contained 1 × GoTaq Buffer, $MgCl_2$, 0.2 mM dNTPs, 25 pmol primer (Table 2.1), 1 U GoTaq polymerase (Promega) and 1 µl GenomiPhi amplified metagenomic DNA. The PCR conditions were 2 min at 95 °C; 35 cycles of 1 min at 95 °C, 1 min at 52-55 °C and 2 min at 72 °C; followed by incubation for 10 min at 72 °C. Obtained PCR products were sequenced to verify the accuracy of the assemblies. See Table 2.1 for details on the primers used in this study.

Table 2.1. Primer sequences used for PCR amplification and sequencing.

Contig name	Contig size (bp)	Primer	Sequence (5'→3')	PCR product size (bp)	Application
Contig0	60782	8103F	AACTCCCCGATATTGTAGGC	3574	PCR, sequencing
		11657R	AGCGTGAAAGTGCATACTCC		
		9482F	GAGAGAGCTAGAGCCGATG		Sequencing
		LysAF	CACTGGAGGCAAAGTCAACA		Sequencing
		LysAR	CTCCGTCCGTAATTCTCAGC		Sequencing
Contig17	15213	7401F	TGCAACTGGTGATAGCTTGA	3914	PCR, sequencing
		11295R	GCAATTTCTTGAGCCGAATA		
		8637F	CCAAACCGTAATTGATGCAG		Sequencing
		10526R	CCTGAACAGGCTTCACAGAA		Sequencing
Contig36	11019	2927F	CATTCCAGGAGGACATGAA	3897	PCR, sequencing
		6805R	CAGTGATCGGTACGACGTT		
		3511F	GAACTGGAAAAGACCAAGC		Sequencing
		4059F	AGATGCCCTGACCATCAGT		Sequencing
		4795F	GATCTGGACGAACTGAACG		Sequencing
		5956R	ATGTCGAAAGCCTTCTTCG		Sequencing
Contig195	2847	764F	TAACGGAACGGGCAAGATAC	1837	PCR, sequencing
		2600R	TCTGGGACTGGAAAAGTTGG		
		2478outF	GGGACCCATTAAATCAGCA	1255	PCR, sequencing
		885outR	TTGCATTTTAGGTCGTGGA		

2.2.13. TTV phylogenetic analysis

ORF (Open Reading Frames) were identified using NCBI ORF Finder http://www.ncbi.nlm.nih.gov/projects/gorf/. Predicted ORF1 sequences of three longest contigs were compared to the nr protein database using BLASTP and aligned with the top 5 hits using ClustalW (Larkin *et al.*, 2007). The multiple sequence alignments were trimmed to 220 aa (between position 77 and 297 of multiple sequence alignments) using Jalview (Waterhouse *et al.*, 2009). The phylogenetic tree was constructed using MEGA version 5.04 (Tamura *et al.*) by applying p-distance model and the Neighbor-Joining method with 1000 bootstrap replications. The complete circular genome of contig 195 was drawn using PlasMapper Version 2.0 (Dong *et al.*, 2004). A stem-loop (transcription terminator) was identified manually within the GC rich region. The sequences of TTV contigs used for phylogenetic comparison were submitted to GenBank (JX157237 – JX157239).

2.3. Results

2.3.1. Taxonomic classification of reads

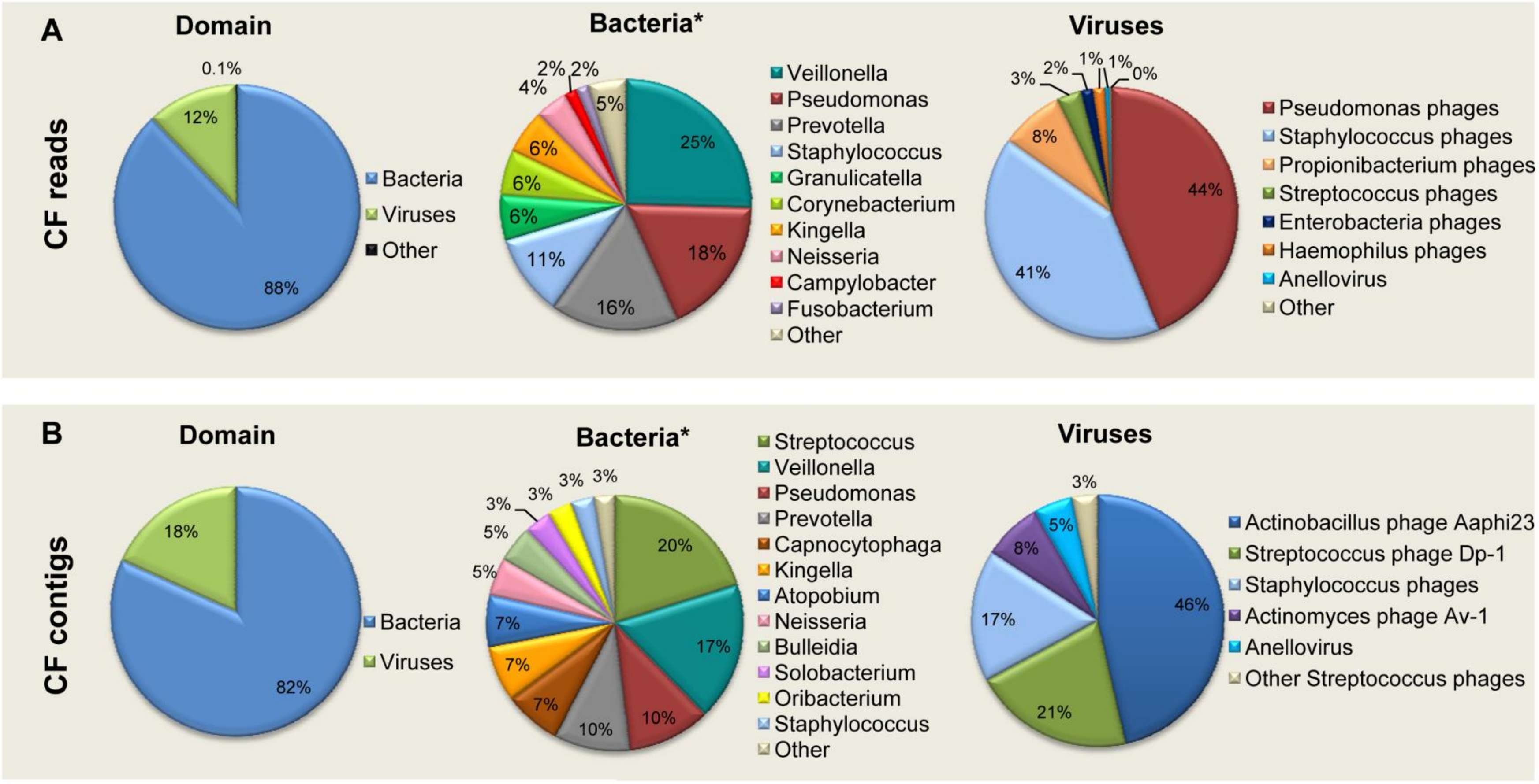
The DNA virus community in sputum from a cystic fibrosis patient was sequenced using Illumina technology and classified using the MG-RAST annotation server (Meyer *et al.*, 2008). Out of 1,715,261 76-bp sequence reads obtained, the vast majority (97%) had no hits to any known sequence in GenBank database (minimum alignment length of 20 bp and e-value of 1e-03) implemented in the MG-RAST server. Of the reads that were assigned (54,490), 87.87% matched bacterial sequences, 11.98% matched viral sequences, 0.13% matched mobile genetic elements and 0.01% matched archaeal and eukaryotic sequences (Figure 2.1A). The most abundant bacterial genera (presumably prophages) were *Veillonella* (25%), followed by *Pseudomonas* (18%), *Prevotella* (16%) and *Staphylococcus* (11%) (Figure 2.1A, Table 2.2). Analysis of the viral matches at species level revealed that phages from *Pseudomonas* (44%) and *Staphylococcus* (41%) were the most abundant GenBank hits in the CF metagenome (Figure 2.1A, Table 2.2). A small number of hits to human viruses were obtained: 41 (0.08%) reads to *Anelloviridae* and 1 (0.002%) read to *Herpesviridae* (Table 2.2).

2.3.2. Assembly characteristics

The sequence reads were assembled into 2,859 contigs (>100 bp), with N50 length of 4.3 kb and the total contig length of 2.6 Mbp. 53% of the reads, as determined by read mapping back onto contigs, were recruited into contigs of length greater than 100 bp. The longest contig was 61 kb, 44 contigs were longer than 10 kb, 102 contigs were longer than 5 kb, and 467 contigs were longer than 1 kb. CDS prediction using MetaGene resulted in prediction of a total of 5150 coding DNA sequences (CDS) in contigs larger than 100 bp.

Contigs (>1 kb) were annotated using standalone BLASTX against GenBank nr database. Because contigs size ranged from 100 bp to 61 kb, similarities to each taxon (>80% identity) for the taxonomic breakdown (Figure 2.1B, Table 2.2) were normalized by counting the number of bases in the contigs matching each taxon and dividing by the total number of bases in all contigs. 21.5% of the contigs with minimum identity of 1 kb (108 contigs totalling 426,201 bases) could be assigned to bacteria (82%) and viruses (18%) (Figure 2.1B). *Streptococcus* (20%), *Veillonella* (17%), *Pseudomonas* (10%) and *Prevotella* (10%) were the most abundant bacterial genera (presumably phage) detected in contigs (Figure 2.1B, Table 2.2). *Actinobacillus* phage Aaphi23 (46%), *Streptococcus pneumoniae* phage Dp-1 (21%) and *Staphylococcus* phages (17%) were dominant viral hits (Figure 1B, Table S2). Apart from *Pseudomonas*, most of the sequences found in CF sputum can be attributed to phages infecting bacteria associated with the upper respiratory tract according to the Human Oral Microbiome Database (http://www.homd.org/index.php). BLASTX comparisons of the 44 largest (>10 kb) contigs against GenBank protein database showed significant (e-value 1e-05) matches to phage genes (Table 2.3). Among large (>5 kb) contigs some had high (>90%) sequence similarity to prophages from *Haemophilus influenzae* (contig 1), *Capnocytophaga sputigena* (contig15, contig 45), *Streptococcus infantis* (contig 17), *Atopobium parvulum* (contig18), *Pseudomonas aeruginosa* (contig 27, contig 36, contig 62, contig 76), *Streptococcus mitis* (contig 29, contig 101), *Prevotella oris* (contig 40), *Kingella oralis* (contig 41), *Solobacterium moorei* (contig 42), *Streptococcus sanguinis* (contig 47), *Bulleidia extructa* (contig 68), *Veillonella sp*. (contig 74), *Staphylococcus aureus* (contig 79) and *Veillonella parvula* (contig 97). TBLASTX comparison of the largest contig (contig 0, 60.7 kb) to the *Mycobacterium* phage Myrna (*Myoviridae* family, environmental phage recovered using an *M. smegmatis* host (Hatfull *et al*., 2010)) showed regions of synteny with some genomic rearrangements (Figure 2.2). The second largest contig (contig 1, 37.3 kb) was syntenic to the genomic region of *Actinobacillus* phage Aaphi23. Four contigs (contig 27, 36, 62 and 76) totalling 39 kb, were ordered and aligned with a prophage in the *Pseudomonas aeruginosa* 39016 genome.

Assembly accuracy was sought by carrying out PCR over multiple CDS using primers designed from these contigs (*Mycobacterium* phage Myrna, *P. aeruginosa* 39016 prophage) and one smaller contig (contig 17, a 15.2 kb contig resembling *Streptococcus infantis* SK1076 prophage) (Figure 2.2). PCR products of the designed size (Table 1.1) were cloned and sequenced yielding nearly identical (>99% identity) sequence to the assembly.



*Bacteria category represents presumed phage sequences

Figure 2. 1. Taxonomic breakdown of assigned reads (n = 53,357) (panel A) and normalized contigs (108 contigs totalling 426,201 bases) (panel B) at domain level, bacteria at genus level and viruses at species level.

Table 2.2. Most abundant bacterial and viral species found in CF sputum sample as classified by MG-RAST (reads) and BLASTX (contigs).

Species	**Reads**	**Species**	**Contigs >1kb, >80% identity (normalized)**
Veillonella sp.	18,93%	*Pseudomonas aeruginosa* 39016	9,12%
Pseudomonas aeruginosa	15,39%	*Actinobacillus* phage Aaphi23	8,77%
Prevotella oris	14,20%	*Prevotella oris*	7,40%
Staphylococcus aureus	9,15%	*Streptococcus infantis*	6,64%
Pseudomonas phages	5,76%	*Capnocytophaga sputigena*	6,26%
Staphylococcus phages	5,37%	*Atopobium parvulum*	5,71%
Granulicatella adiacens	5,05%	*Veillonella* sp.	5,66%
Corynebacterium matruchotii	5,04%	*Veillonella dispar*	4,43%
Kingella oralis	4,88%	*Bulleidia extructa*	4,12%
Veillonella parvula	2,77%	*Streptococcus* phage Dp-1	3,91%
Neisseria gonorrhoeae	1,90%	*Neisseria bacilliformis*	3,86%
Neisseria meningitidis	1,76%	*Streptococcus sanguinis*	3,79%
Campylobacter gracilis	1,39%	*Streptococcus mitis*	3,78%
Haemophilus influenzae	1,18%	*Kingella oralis*	3,67%
Streptococcus pneumoniae	1,15%	*Staphylococcus aureus*	3,30%
Propionibacterium phages	1,07%	*Veillonella parvula*	2,95%
Veillonella atypica	0,82%	*Solobacterium moorei*	2,95%
Fusobacterium sp. 7_1	0,82%	*Staphylococcus aureus* phage 77	2,49%
Escherichia coli	0,67%	*Streptococcus pseudopneumoniae*	2,27%
Fusobacterium sp. D11	0,53%	*Veillonella atypica*	2,18%
Streptococcus phages	0,41%	*Actinomyces* phage Av-1	1,43%
Shigella sonnei	0,35%	*Streptococcus pneumoniae*	1,13%
Bacteroidetes oral taxon 274	0,32%	*Haemophilus influenzae*	1,05%
Rothia dentocariosa	0,24%	Anellovirus	0,94%
Propionibacterium acnes	0,22%	*Prevotella salivae*	0,68%
Enterobacteria phages	0,21%	*Corynebacterium matruchotii*	0,52%
Haemophilus phages	0,20%	candidate division TM7	0,36%
Anellovirus	0,08%	*Gemella haemolysans*	0,34%
Mycobacterium phage Pacc40	0,02%	*Neisseria subflava*	0,29%
Shigella boydii	0,01%		
Pseudomonas sp. M18	0,01%		
Streptococcus suis	0,01%		
Peptoniphilus sp. oral taxon 836	0,01%		
Human *herpesvirus* 6	0,002%		
Other	0,078%		
Total	**100%**		**100%**

Table 2.3. Summary of BLASTX (e-value >1e-05) results of the 44 largest (>10kb) contigs assembled from CF metagenome (sorted by bit score). Hits with BLASTX similarity >80% are shown in bold.

Contig	Contig size (bp)	Phage/Organism containing prophage	Best BLAST hit	Accession number	E-value	% identity	Hit length (aa)
contig0	60,782	*Mycobacterium* phage Myrna	gp192 (DNA polymerase III, alpha subunit)	YP_002225071	2e-178	35	1211
contig1	**37,377**	***Actinobacillus* phage Aaphi23**	**Hypothetical protein**	**NP_852765**	**0.0**	**80**	**502**
contig2	32,643	*Clostridium botulinum*	Phage terminase	CBZ04420	1e-126	50	455
contig3	27,140	*Arthrobacter gangotriensis*	Phage terminase	WP_007271553	2e-166	53	492
contig4	25,518	*Actinomyces turicensis*	Hypothetical protein	WP_006681903	1e-129	38	738
contig5	24,466	*Treponema denticola*	Phage minor structural protein	WP_010699111	0.0	63	1353
contig6	24,071	*Bacillus amyloliquefaciens*	Prophage-derived protein YonE	YP_005421079	3e-40	29	462
contig7	22,729	*Thermosinus carboxydivorans*	Lytic transglycosylase	ZP_01666661	0.0	50	884
contig8	22,651	*Actinomyces turicensis*	Terminase	WP_006681889	2e-118	51	414
contig9	22,112	*Methanobrevibacter ruminantium*	RNA ligase	YP_003423768	1e-95	49	416
contig10	21,200	*Lachnospiraceae* bacterium	Phage tail tape measure protein	WP_009320577	3e-171	36	1372
contig11	19,385	*Prevotella multisaccharivorax*	Phage tail tape measure protein, TP901 family	ZP_08578267	2e-180	32	1493
contig12	17,863	*Coprococcus* sp. HPP007	Anaerobic ribonucleoside-triphosphate reductase	WP_016438838	4e-162	64	417
contig13	17,760	*Rhodococcus* phage ReqiPoco6	Phage-related terminase	ADD81014	0.0	63	578
contig14	17,370	*Acidaminococcus intestini*	Phage portal protein	YP_004896747	4e-135	54	450
contig15	**16,709**	***Capnocytophaga sputigena***	**Hypothetical protein**	**ZP_03392197**	**5e-141**	**92**	**280**
contig16	16,491	*Lactococcus* phage 1706	Putative terminase large subunit	YP_001828656	0.0	65	581
contig17	**15,213**	***Streptococcus infantis***	**Phage tail tape measure protein**	**ZP_08523200**	**0.0**	**92**	**1007**
contig18	14,978	*Atopobium parvulum*	Phage tail tape measure protein	YP_003179609	0.0	76	959
contig19	14,621	*Ruminococcus torques*	Phage primase	ZP_01968459	0.0	50	1365
contig20	13,844	*Clostridium leptum*	Phage terminase	ZP_02079960	0.0	71	580
contig21	13,698	*Rhodococcus* phage E3	Terminase large subunit	YP_008061043	0.0	54	642
contig22	13,687	*Veillonella parvula*	Chromosome segregation protein	EGL77600	0.0	60	633
contig23	13,512	*Rhodococcus* phage ReqiPepy6	Phage primase	ADD80963	0.0	50	1298
contig24	13,410	*Solobacterium moorei*	Hypothetical protein	ZP_08029139	2e-90	69	213

Contig	Contig size (bp)	Phage/Organism containing prophage	Best BLAST hit	Accession number	E-value	% identity	Hit length (aa)
contig25	13,204	*Ruminococcus* sp. SR1/5	Phage tail tape measure protein	CBL20009	6e-22	33	231
contig26	**13,166**	***Streptococcus* sp. F0442**	**Phage tail tape measure protein**	**WP_009732633**	**0.0**	**81**	**1442**
contig27	**12,806**	***Pseudomonas aeruginosa* 39016**	**Phage tail tape measure protein**	**ZP_07793213**	**0.0**	**97**	**1281**
contig28	12,766	*Flavonifractor plautii*	Phage terminase, PBSX family	ZP_09386037	3e-109	47	400
contig29	**12,549**	***Streptococcus mitis***	**Helicase**	**EGU67659**	**0.0**	**97**	**526**
contig30	12,371	*Clostridium bolteae*	Hypothetical protein	WP_002573180	2e-121	40	545
contig31	12,206	*Streptococcus* phage YMC-2011	Collagen binding domain protein	AEJ54387	2e-66	41	381
contig32	11,815	*Lactococcus* phage 1706	Phage primase	YP_001828703	0.0	52	1284
contig33	**11,362**	***Oribacterium* sp.**	**C-5 cytosine-specific DNA methylase**	**ZP_09323670**	**0.0**	**85**	**411**
contig34	11,357	*Clostridium symbiosum*	Hypothetical protein	ZP_08105582	1e-17	45	106
contig35	11,280	*Streptococcus parasanguinis*	Phage tail tape measure protein	WP_003017707	0.0	73	1565
contig36	**11,019**	***Pseudomonas aeruginosa* 39016**	**Integrase**	**ZP_07793182**	**0.0**	**98**	**594**
contig37	10,939	*Coprococcus catus*	1,4-beta-N-acetylmuramidase	CBK80688	3e-74	44	363
contig38	10,910	*Streptococcus anginosus*	Phage tail protein	WP_003038238	0.0	51	620
contig39	10,903	*Micrococcus luteus*	Phage tail tape measure protein	WP_002856287	3e-106	35	1045
contig40	**10,857**	***Prevotella oris***	**ClpP protease**	**ZP_07035944**	**0.0**	**99**	**365**
contig41	**10,724**	***Kingella oralis***	**Baseplate J-like protein**	**ZP_04601766**	**6e-146**	**96**	**299**
contig42	**10,665**	***Solobacterium moorei***	**Site-specific recombinase**	**ZP_08029163**	**4e-119**	**96**	**214**
contig43	10,644	*Dysgonomonas mossii*	Phage terminase	ZP_08471641	7e-51	48	212

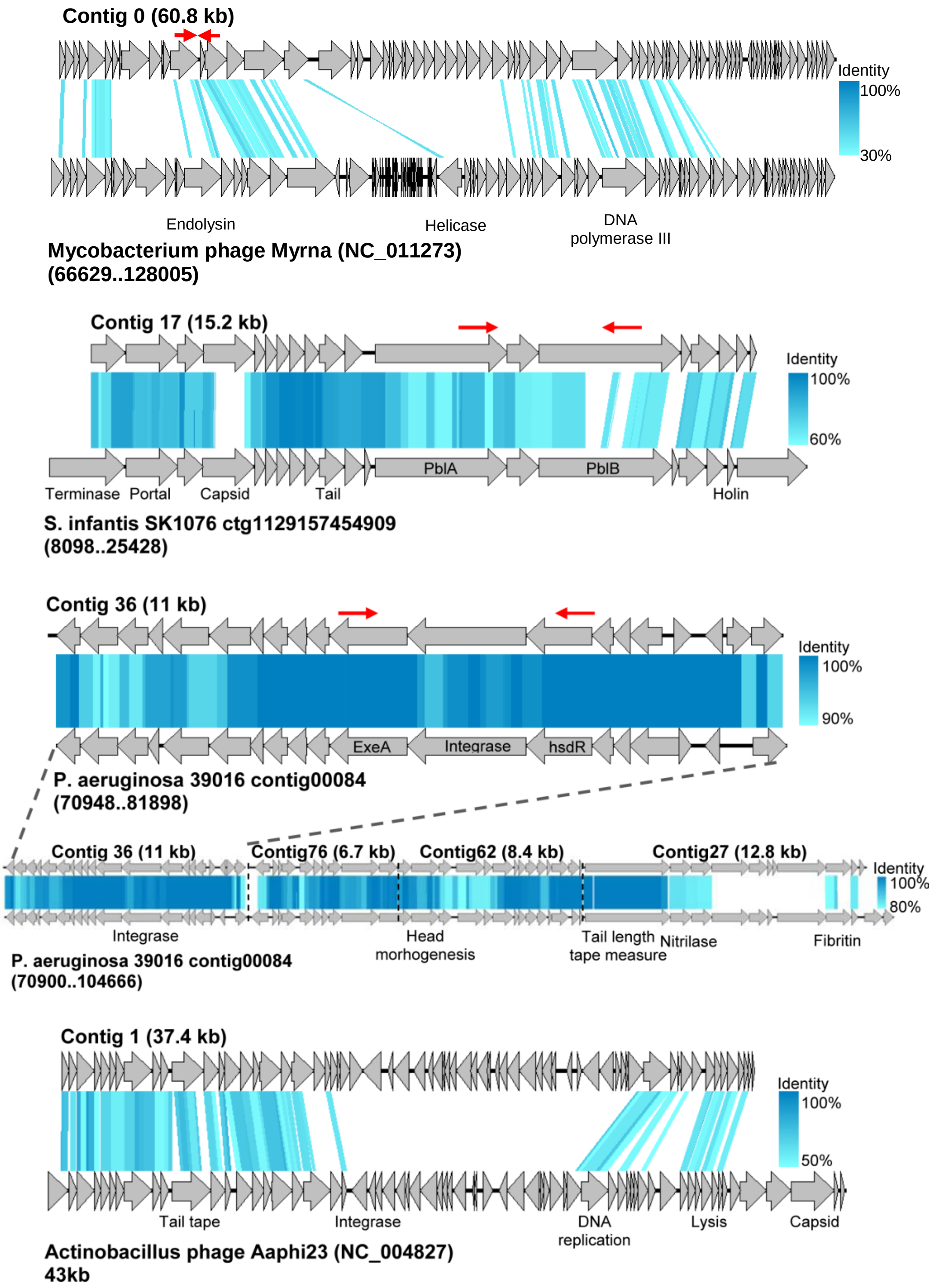


Figure 2. 2. TBLASTX comparison of seven selected contigs assembled from the cystic fibrosis sputum metagenomic short reads to the most similar sequences in GenBank. The level of amino acid identity is shown in the gradient scale. Red arrows indicate position of the PCR primers used for multigene assembly verification. The figure was drawn using Easyfig.

2.3.3. Diversity of viruses in CF sputum

PHACCS analysis showed that the diversity was low with 89 species (Table 2.4). This is comparable to that previously reported from CF sputum (69-154 species) (Willner *et al.*, 2009) and CF lung (14-105 species) (Willner *et al.*, 2012) using a different sequencing methodology (454). The abundance of the most predominant viral genotype identified by PHACCS in the CF virome was 4.7%.

Table 2.4. Diversity analysis of the CF sputum metagenome in this study compared with previously published CF sputum metagenome.

Sample	**Richness**	**Evenness**	**Shannon Index**	**Model**
CF sputum (this study)	89 genotypes	0.98	4.41	logarithmic
CF sputum (Willner *et al.*, 2009)	105 genotypes	0.85	4.17	logarithmic
CF lung (post-mortem, different lobes) (Willner *et al.*, 2012)	14-105 genotypes	NA*	NA	logarithmic

* Data not available

2.3.4. Functional classification

The contig functional annotation based on SEED Subsystem classification was carried out using MG-RAST. 556 genes predicted by MG-RAST were assigned and the dominant subsystem was Phages, Prophages (Figure 2.3A, Table 2.5) comprising 83% of total assigned sequences. This category contained phage-related proteins e.g. phage tail, capsid, lysin, terminase, integrase, portal protein and virulence associated platelet-binding proteins. The other CDS with a functional prediction belonged mostly to the DNA metabolism, including genes encoding DNA polymerase, ssDNA-binding protein and DNA helicase.

Coding DNA sequences (CDS) predicted from contigs were annotated using the COG database. 11% of CDS could be assigned to COG database functional categories (Figure 2.3B, Table 2.6). The three most abundant COG functional categories were replication, recombination and repair (28%), general function (21%) and function unknown (20%). Most of these annotated CDS were predicted to encode genes involved in lysogeny (e.g. site-specific recombinase XerD), assembly (e.g. terminase), replication (e.g. helicase, DNA polymerase, ssDNA-binding protein), transcription (e.g. transcriptional regulator), virulence (e.g. glucan-binding domain), stress response (e.g. SOS-response transcriptional repressor RecA), infection immunity (e.g. phage antirepressor protein) and host lysis (e.g. 1,4-beta-N-acetylmuramidase).

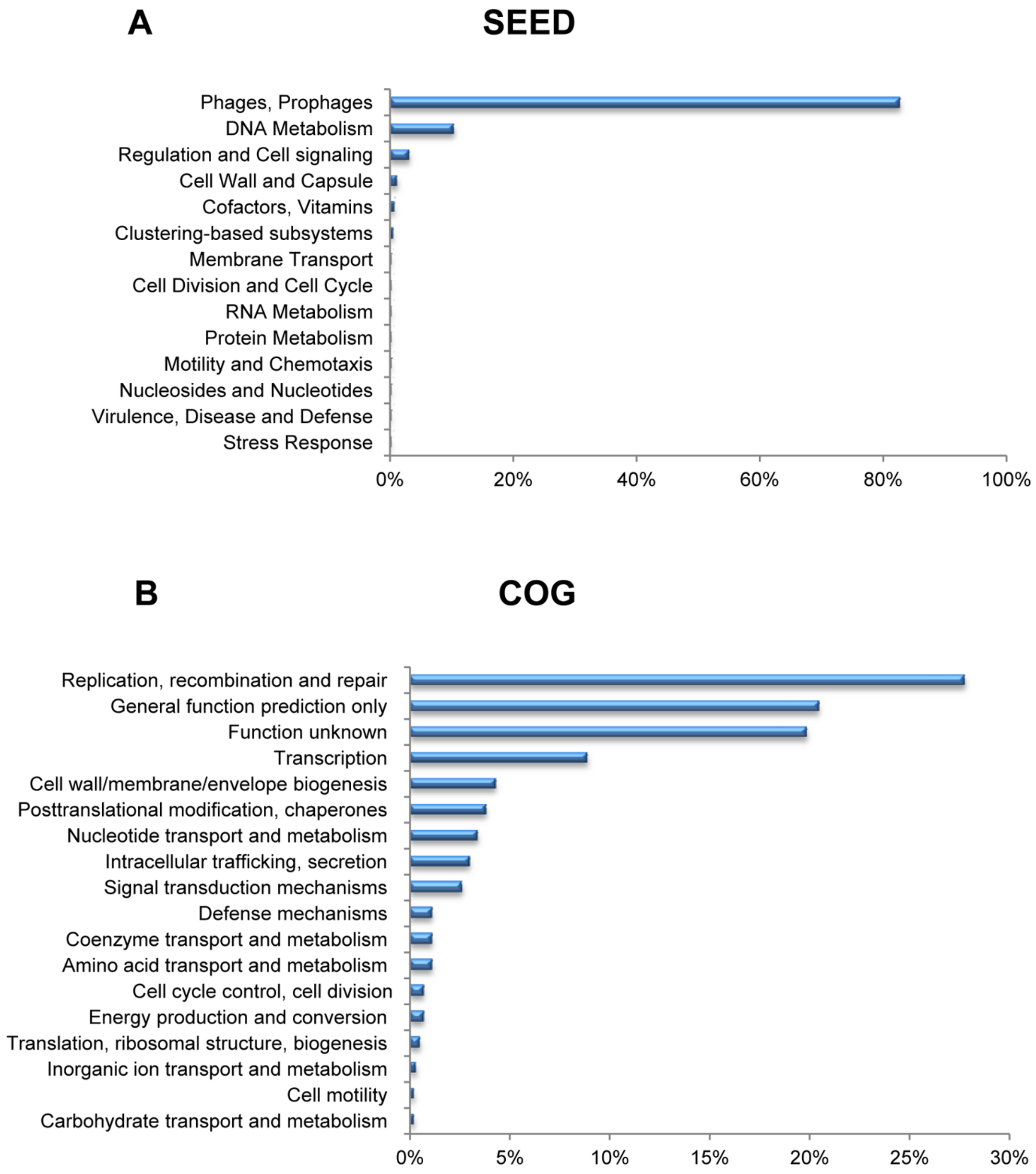


Figure 2. 3. Functional annotation of contigs assembled from CF sputum metagenome to SEED-Subsystem using MG-RAST (A) and contigs CDS to COG database using WebMGA (B).

Table 2.5. Functional classification of the contigs assembled from the CF sputum viral metagenome to the SEED database, based on the MG-RAST analysis (e-value 1e-05 and minimum alignment length 50).

SEED level 1 categories	**Function**	**ORF count**
Phages, Prophages	Phage protein	132
Phages, Prophages	Phage tail length tape-measure protein	37
Phages, Prophages	Phage major capsid protein	34
Phages, Prophages	Phage terminase	34
Phages, Prophages	Phage lysin	26
Phages, Prophages	Phage tail fibers	23
Phages, Prophages	Human platelet-binding protein	22
Phages, Prophages	Phage minor tail protein	20
Phages, Prophages	Phage integrase	19
Phages, Prophages	Phage portal protein	16
DNA Metabolism	DNA polymerase	13
Phages, Prophages	Phage capsid and scaffold	13
DNA Metabolism	Single-stranded DNA-binding protein	12
Regulation and Cell signaling	Prophage Clp protease-like protein	11
DNA Metabolism	DNA helicase	10
Phages, Prophages	Structural protein	9
Phages, Prophages	Phage holin	9
Phages, Prophages	Phage major tail protein	8
Phages, Prophages	Phage antirepressor	7
Phages, Prophages	Phage tape measure	7
Regulation and Cell signaling	Cell wall-associated murein hydrolase LytA	6
Cell Wall and Capsule	N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)	6
Phages, Prophages	Phage collar	6
Phages, Prophages	Phage endopeptidase	5
Phages, Prophages	Phage repressor	5
DNA Metabolism	DNA-cytosine methyltransferase (EC 2.1.1.37)	4
Phages, Prophages	Site-specific recombinase	4
DNA Metabolism	DNA polymerase III alpha subunit (EC 2.7.7.7)	3
DNA Metabolism	DNA polymerase III subunits gamma and tau (EC 2.7.7.7)	3
Cofactors, Vitamins	GTP cyclohydrolase I (EC 3.5.4.16) type 1	3
Phages, Prophages	Phage baseplate	3
Phages, Prophages	Phage minor capsid protein	3
Phages, Prophages	Phage tail assembly	3
Phages, Prophages	Hypothetical homolog in superantigen-encoding pathogenicity islands SaPI	3
DNA Metabolism	DinG family ATP-dependent helicase YoaA	2
DNA Metabolism	DNA primase (EC 2.7.7.-)	2
Phages, Prophages	Phage major tail shaft protein	2
Phages, Prophages	Phage tail fiber protein	2
DNA Metabolism	Recombinational DNA repair protein RecT	2
Clustering-based subsystems	Superfamily II DNA/RNA helicases, SNF2 family	2
DNA Metabolism	ATP-dependent DNA helicase UvrD/PcrA	1
Cell Division and Cell Cycle	Cell division protein FtsK	1
Stress Response	Choline binding protein A	1
Phages, Prophages	CI-like repressor, superantigen-encoding pathogenicity islands SaPI	1
Virulence, Disease and Defense	Collagen-like surface protein	1

Nucleosides and Nucleotides	Deoxycytidine triphosphate deaminase (EC 3.5.4.13)	1
Phages, Prophages	DNA adenine methyltransferase, phage-associated	1
DNA Metabolism	DNA gyrase subunit A (EC 5.99.1.3)	1
Phages, Prophages	DNA methyl transferase, phage-associated	1
DNA Metabolism	DNA topoisomerase I (EC 5.99.1.2)	1
DNA Metabolism	DNA-binding protein HBsu	1
DNA Metabolism	Exodeoxyribonuclease III (EC 3.1.11.2)	1
Motility and Chemotaxis	Flagellar hook-length control protein FliK	1
DNA Metabolism	Helicase loader DnaI	1
Phages, Prophages	Integrase, superantigen-encoding pathogenicity islands SaPI	1
Protein Metabolism	Methionyl-tRNA formyltransferase (EC 2.1.2.9)	1
Phages, Prophages	Phage DNA and RNA binding protein	1
Phages, Prophages	Phage DNA-binding protein	1
Phages, Prophages	Phage endolysin	1
Phages, Prophages	Phage NinG rap recombination	1
Phages, Prophages	Phage tail sheath monomer	1
Clustering-based subsystems	Protein NinG	1
RNA Metabolism	PUA-PAPS reductase like fusion	1
Cofactors, Vitamins	Queuosine Biosynthesis QueC ATPase	1
Membrane Transport	Shufflon-specific DNA recombinase	1
Total		**556**

Table2.6. Functional annotation of CDS predicted on contigs assembled from the CF sputum viral metagenome, based on the BLASTP analysis to the COG database.

Class description	**Description**	**CDS count**
Function unknown	Phage-related protein	63
Replication, recombination and repair	Site-specific recombinase XerD	37
General function prediction only	Phage terminase large subunit	28
Function unknown	Uncharacterized proteins	24
Transcription	Predicted transcriptional regulators	19
Replication, recombination and repair	Single-stranded DNA-binding protein	16
Multiple classes	Protease subunit of ATP-dependent Clp proteases	14
Multiple classes	Superfamily II DNA/RNA helicases, SNF2 family	13
General function prediction only	FOG: Glucan-binding domain (YG repeat)	13
Nucleotide transport and metabolism	dUTPase	12
General function prediction only	Predicted ATPase	9
Replication, recombination and repair	DNA polymerase elongation subunit (family B)	8
Multiple classes	SOS-response transcriptional repressors (RecA-mediated autopeptidases)	8
Function unknown	Small integral membrane protein	8
Replication, recombination and repair	Replicative DNA helicase	7
Replication, recombination and repair	DNA polymerase I - 3'-5' exonuclease and polymerase domains	7
Replication, recombination and repair	Holliday junction resolvase	7
Cell wall/membrane/envelope biogenesis	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)	6
General function prediction only	Bacteriophage capsid protein	6
General function prediction only	Predicted P-loop ATPase and inactivated derivatives	6
Replication, recombination and repair	Site-specific DNA methylase	5
Replication, recombination and repair	DNA primase (bacterial type)	5

Replication, recombination and repair	DNA replication protein	5
General function prediction only	Phage tail sheath protein FI	5
General function prediction only	Phage-related baseplate assembly protein	5
Replication, recombination and repair	RecA/RadA recombinase	4
Replication, recombination and repair	DNA polymerase III, alpha subunit	4
Posttranslational modification, protein turnover, chaperones	Organic radical activating enzymes	4
Multiple classes	DNA or RNA helicases of superfamily II	4
Amino acid transport and metabolism	Predicted Zn peptidase	4
General function prediction only	Phage tail tube protein FII	4
General function prediction only	Phage protein D	4
Replication, recombination and repair	RecA-family ATPase	4
Transcription	Prophage antirepressor	4
General function prediction only	Bacteriophage P2-related tail formation protein	4
General function prediction only	Phage P2 baseplate assembly protein gpV	4
Function unknown	Phage-related minor tail protein	4
Replication, recombination and repair	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), B subunit	3
Replication, recombination and repair	NAD-dependent DNA ligase(contains BRCT domain typeII)	3
Cell wall/membrane/envelope biogenesis	Cell wall-associated hydrolases (invasion-associated proteins)	3
Function unknown	Uncharacterized protein, homolog of phage Mu protein gp30	3
Replication, recombination and repair	Recombinational DNA repair protein (RecE pathway)	3
General function prediction only	Putative secretion activating protein	3
Defense mechanisms	Abortive infection bacteriophage resistance protein	3
Cell wall/membrane/envelope biogenesis	Predicted outer membrane protein	3
Function unknown	Phage-related tail protein	3
General function prediction only	Bacteriophage tail assembly protein	3
Function unknown	Uncharacterized protein conserved in bacteria	3
Cell wall/membrane/envelope biogenesis	N-acetylmuramoyl-L-alanine amidase	3
Multiple classes	3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes	2
Nucleotide transport and metabolism	Guanylate kinase	2
Replication, recombination and repair	Superfamily I DNA and RNA helicases	2
Coenzyme transport and metabolism	GTP cyclohydrolase I	2
Replication, recombination and repair	ATPase involved in DNA repair	2
Replication, recombination and repair	ATP-dependent exoDNAse (exonuclease V), alpha subunit - helicase superfamily I member	2
Energy production and conversion	Glycerophosphoryl diester phosphodiesterase	2
Replication, recombination and repair	Transposase and inactivated derivatives	2
Posttranslational modification, protein turnover, chaperones	Glutaredoxin and related proteins	2
Coenzyme transport and metabolism	6-pyruvoyl-tetrahydropterin synthase	2
Cell wall/membrane/envelope biogenesis	Membrane proteins related to metalloendopeptidases	2
Replication, recombination and repair	DNA modification methylase	2
Cell cycle control, cell division, chromosome partitioning	Chromosome segregation ATPases	2
Nucleotide transport and metabolism	Oxygen-sensitive ribonucleoside-triphosphate reductase	2
Nucleotide transport and metabolism	Predicted alternative thymidylate synthase	2

Signal transduction mechanisms	Predicted ATPase related to phosphate starvation-inducible protein PhoH	2
Function unknown	Uncharacterized protein with SCP/PR1 domains	2
Replication, recombination and repair	DNA polymerase III, gamma/tau subunits	2
Defense mechanisms	Negative regulator of beta-lactamase expression	2
Function unknown	Uncharacterized homolog of phage Mu protein gp47	2
General function prediction only	Phage protein U	2
Transcription	Phage anti-repressor protein	2
General function prediction only	Phage baseplate assembly protein W	2
Replication, recombination and repair	Phage terminase, small subunit	2
General function prediction only	Phage head maturation protease	2
Replication, recombination and repair	Putative primosome component and related proteins	2
General function prediction only	Mu-like prophage protein	2
General function prediction only	SLT domain proteins	2
General function prediction only	ABC-type uncharacterized transport system, ATPase component	2
General function prediction only	Predicted phosphoesterase or phosphohydrolase	2
General function prediction only	Mu-like prophage FluMu protein gp28	2
Function unknown	Mu-like prophage protein gp16	2
General function prediction only	Phage-related holin (Lysis protein)	2
General function prediction only	Phage-related tail fibre protein	2
Replication, recombination and repair	Phage-related protein, predicted endonuclease	2
Cell wall/membrane/envelope biogenesis	Endopolygalacturonase	2
Signal transduction mechanisms	NAD+--asparagine ADP-ribosyltransferase	2
Function unknown	Predicted membrane protein	2
Carbohydrate transport and metabolism	Predicted sugar kinase	1
Transcription	DNA-directed RNA polymerase, beta subunit	1
Replication, recombination and repair	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit	1
Nucleotide transport and metabolism	Thymidylate synthase	1
Translation, ribosomal structure and biogenesis	Cysteinyl-tRNA synthetase	1
Translation, ribosomal structure and biogenesis	Methionyl-tRNA formyltransferase	1
Replication, recombination and repair	5'-3' exonuclease (including N-terminal domain of PolI)	1
Defense mechanisms	Type I restriction-modification system methyltransferase subunit	1
Replication, recombination and repair	Ribonuclease HI	1
Replication, recombination and repair	Site-specific DNA methylase	1
Energy production and conversion	Uncharacterized flavoproteins	1
Posttranslational modification, protein turnover, chaperones	Chaperonin GroEL (HSP60 family)	1
Replication, recombination and repair	ATPase involved in DNA replication	1
Coenzyme transport and metabolism	Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2	1
Replication, recombination and repair	DNA polymerase sliding clamp subunit (PCNA homolog)	1
General function prediction only	Predicted PP-loop superfamily ATPase	1
Replication, recombination and repair	Single-stranded DNA-specific exonuclease	1
Multiple classes	Periplasmic serine proteases (ClpP class)	1
General function prediction only	Predicted permease	1
Signal transduction mechanisms	Signal transduction histidine kinase	1
Replication, recombination and repair	Exonuclease III	1

Nucleotide transport and metabolism	Deoxycytidine deaminase	1
Multiple classes	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	1
Replication, recombination and repair	Bacterial nucleoid DNA-binding protein	1
Transcription	Transcription elongation factor	1
Cell wall/membrane/envelope biogenesis	Periplasmic protein TonB, links inner and outer membranes	1
Replication, recombination and repair	Holliday junction resolvasome, endonuclease subunit	1
Replication, recombination and repair	DNA polymerase III, epsilon subunit and related 3'-5' exonucleases	1
Posttranslational modification, protein turnover, chaperones	Predicted ATP-dependent serine protease	1
Cell cycle control, cell division, chromosome partitioning	ATPases involved in chromosome partitioning	1
Replication, recombination and repair	Recombinational DNA repair ATPase (RecF pathway)	1
Multiple classes	Nucleoside-diphosphate-sugar pyrophosphorylase involved in lipopolysaccharide biosynthesis/translation initiation factor 2B, gamma/epsilon subunits (eIF-2Bgamma/eIF-2Bepsilon)	1
General function prediction only	Metal-dependent hydrolases of the beta-lactamase superfamily I	1
Replication, recombination and repair	Predicted ATPase involved in replication control, Cdc46/Mcm family	1
Replication, recombination and repair	Intein/homing endonuclease	1
Cell wall/membrane/envelope biogenesis	FOG: LysM repeat	1
Replication, recombination and repair	Micrococcal nuclease (thermonuclease) homologs	1
Multiple classes	ABC-type Na+ efflux pump, permease component	1
Cell cycle control, cell division, chromosome partitioning	DNA segregation ATPase FtsK/SpoIIIE	1
Multiple classes	Muramidase (flagellum-specific)	1
Signal transduction mechanisms	DnaK suppressor protein	1
General function prediction only	C-terminal domain of topoisomerase IA	1
Replication, recombination and repair	Site-specific recombinases, DNA invertase Pin homologs	1
Nucleotide transport and metabolism	Deoxycytidylate deaminase	1
Multiple classes	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	1
General function prediction only	ABC-type ATPase fused to a predicted acetyltransferase domain	1
Defense mechanisms	Predicted type IV restriction endonuclease	1
Cell wall/membrane/envelope biogenesis	Membrane protein involved in colicin uptake	1
General function prediction only	Predicted chitinase	1
Intracellular trafficking, secretion, and vesicular transport	Type II secretory pathway, component ExeA (predicted ATPase)	1
Cell wall/membrane/envelope biogenesis	Putative peptidoglycan-binding domain-containing protein	1
Cell wall/membrane/envelope biogenesis	Sortase (surface protein transpeptidase)	1
Intracellular trafficking, secretion, and vesicular transport	Type IV secretory pathway, TrbL components	1
Inorganic ion transport and metabolism	Uncharacterized protein involved in tellurite resistance	1
General function prediction only	Predicted glycosyl hydrolase	1

Function unknown	Mu-like prophage protein gp29	1
Function unknown	Mu-like prophage protein gp36	1
Cell wall/membrane/envelope biogenesis	Predicted soluble lytic transglycosylase fused to an ABC-type amino acid-binding protein	1
General function prediction only	Predicted kinase	1
Function unknown	Microcystin-dependent protein	1
Function unknown	Phage-related protein, tail component	1
General function prediction only	Antirestriction protein	1
Amino acid transport and metabolism	Ornithine/acetylornithine aminotransferase	1
General function prediction only	P2-like prophage tail protein X	1
General function prediction only	Mu-like prophage protein gpG	1
Function unknown	Phage-related minor tail protein	1
Total		**562**

2.3.5. Resistance to antibiotics

To characterize potential antibiotic resistance genes in the CF metagenome, both the unassembled reads and CDS predicted from assembled data were compared to the Comprehensive Antibiotic Resistance Database (CARD). Forty out of 495 assigned reads had significant similarity (>90% reads coverage, >90% identity) to known antibiotic resistance genes (Table 2.7). CARD-represented genes potentially conferring resistance to β–lactams, tetracycline, macrolide and penicillin antibiotics, as well as toxic compounds like cadmium and arsenic. The majority of these genes could be attributed to bacteria abundant in the respiratory tract of CF patients and individuals without CF. Two contigs contained complete metallo-β-lactamase gene sequences along with phage recombinases (Figure 2.4, Table 2.8). BLASTP analysis of CDS predicted on those contigs indicated that multiple CDS from contig 22 were related to *Veillonella* species. Contig 71 had multiple hits with identities to different bacteria; the highest BLASTP identity was 72% to *Gemella haemolysins*. Phylogenetic analysis of translated metallo-β-lactamase sequences is shown in Figure 2.5.

2.3.6. Toxin genes

A staphylococcal complement inhibitor gene *scn* was detected on a 3 kb contig (Table 2.8). An *exeA* gene sometimes associated with type II secretion of toxins was detected on an 11 kb contig (Table 2.8) downstream of an integrase gene (Figure 2.2). This arrangement is seen in a wide variety of *Pseudomonas* and *Salmonella* strains (Stavrinides & Guttman, 2004) and various virulence-associated mobile elements (Stavrinides *et al.*, 2012) where the *exeA* gene is believed to be associated with regulation of transposition, not secretion. Although flanking inverted repeats were not detected on the contig containing the *exeA* gene, two direct 20 bp DNA repeats separated by 14 bp characteristic of E622/Tn*E622* mobile element inverted repeats (Stavrinides *et al.*, 2012) were present upstream of the integrase gene (data not shown).

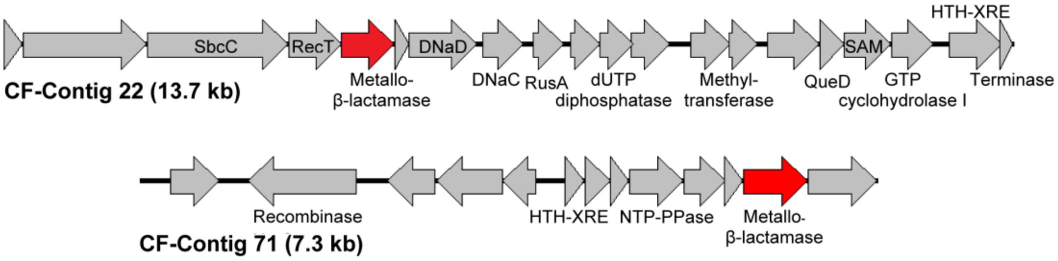


Figure 2. 4. Analysis of the gene order in two contigs containing metallo-β-lactamase and phage recombinase genes. Predicted protein coding genes were annotated using NCBI protein database. CDS of contig 22 displayed similarities to different *Veillonella* species. CDS of contig 71 had multiple hits to different bacteria, and the highest BLASTP identity was 72% to *Gemella haemolysans,* while metallo-β-lactamase CDS showed 46% identity to *Listeria* phage. Metallo-β-lactamase CDS of contig 22 is directly adjacent to phage recombinase suggesting this is an integron. Metallo-β-lactamases are coloured in red. The figure was drawn using Easyfig and edited in Photoshop CS4.

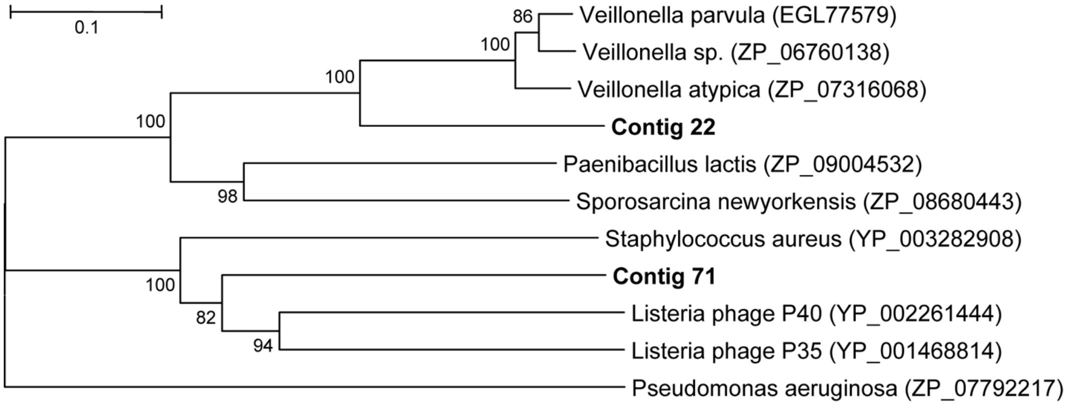

0.1
86
100
100
100
98
100
82
94
Veillonella parvula (EGL77579)
Veillonella sp. (ZP_06760138)
Veillonella atypica (ZP_07316068)
Contig 22
Paenibacillus lactis (ZP_09004532)
Sporosarcina newyorkensis (ZP_08680443)
Staphylococcus aureus (YP_003282908)
Contig 71
Listeria phage P40 (YP_002261444)
Listeria phage P35 (YP_001468814)
Pseudomonas aeruginosa (ZP_07792217)

Figure 2. 5. Neighbour-joining phylogenetic tree showing the relationship between the predicted metallo-β-lactamase sequences assembled from the cystic fibrosis sputum metagenomic short reads and most related sequences in GenBank. Metallo-β-lactamase sequences detected in this study are shown in bold.

Table 2.7. Antibiotic and toxic compound resistance gene hits found in CF sputum metagenome reads. Only hits to reads having >90% identity and 90% coverage are shown.

Resistance gene function	**Resistance gene**	**Resistance conferred**	**Best-match organism**	**No. of hits to reads**
Class B β-lactamase	*MBL*	penicillin, cephalosporin, carbapenem	*Granulicatella adiacens*	6
Class A β-lactamase	*blaTEM-1*	ampicillin	*Streptomyces ghanaensis*	5
Ribosomal protection protein	*tetW, tetM*	tetracycline	*Bifidobacterium longum*	4
Multidrug resistance efflux pump	*tolC*	acriflavin, aminoglycoside, beta-lactam, glycylcycline, macrolide	*Acinetobacter lwoffii*	2
RND multidrug efflux pump	*mexF*	chloramphenicol, fluoroquinolone	*Pseudomonas aeruginosa*	2
Cadmium resistance protein	*cadD*	cadmium	*Streptococcus mitis*	2
RND multidrug efflux pump	*acrB*	acriflavin, aminoglycoside, beta-lactam, glycylcycline, macrolide	*Veillonella parvula*	2
RND multidrug efflux pump	*acrB*	acriflavin, aminoglycoside, beta-lactam, glycylcycline, macrolide	*Escherichia coli*	1
Multidrug resistance efflux pump	*mdtE*	rhodamine, erythromycin, doxorubicin	*Escherichia coli*	1
Multidrug resistance efflux pump	*macB*	macrolide	*Haemophilus haemolyticus*	1
Ribosomal protection protein	*tetM*	tetracycline	multiple hits	1
Glycosyl transferase	*ponA*	penicillin	*Neisseria gonorrhoeae*	1
Arsenical resistance protein	*arsH*	arsenic	*Pseudomonas aeruginosa*	1
RND multidrug efflux pump	*mexN*	chloramphenicol, fluoroquinolone	*Pseudomonas aeruginosa*	1
Class D beta-lactamase	*blaOXA-50a*	oxazolylpenicillins	*Pseudomonas aeruginosa*	1
Class B beta-lactamase	*MBL*	penicillin, cephalosporin, carbapenem	*Selenomonas flueggei*	1
Ribosomal protection protein	*tetM*	tetracycline	*Streptococcus agalactiae*	1
Multidrug resistance efflux pump	*macB*	macrolide	*Streptococcus infantis*	1
Glycopeptide	*vanZ*	vancomycin, teicoplanin	*Streptococcus parasanguinis*	1
Multidrug resistance efflux pump	*macB*	macrolide	*Streptococcus pneumoniae*	1
Peptidoglycan binding protein	*vanW*	vancomycin	*Veillonella atypica*	1
Transglycosylase	*mrcA*	penicillin	*Veillonella atypica*	1
Multidrug resistance efflux pump	*macAB*	macrolide	*Veillonella atypica*	1
Class B beta-lactamase	*MBL*	penicillin, cephalosporin, carbapenem	*Veillonella atypica*	1

Table 2.8. Virulence genes, antibiotic and toxic compound resistance genes found in CDS predicted on the assembled contigs.

Resistance/Virulence gene function	Gene	Resistance conferred	Best-match organism	BLAST identity (%)	E-value	Query coverage (%)	ORF length (aa)	Contig length (kb)	Contig name
metallo-β-lactamase	*MBL*	β-lactam	*Listeria* phage P40	46	1e-43	100	192	7.3	c71
metallo-β-lactamase	*MBL*	β-lactam	*Veillonella atypica*	70	8e-124	100	238	13.7	c22
N-acetyl-anhydromuramil-l-alanine amidase	*ampD*	β-lactam	*Capnocytophaga sputigena*	93	4e-85	100	144	16.7	c15
N-acetyl-anhydromuramil-l-alanine amidase	*ampD*	β-lactam	*Atopobium rimae*	39	3e-68	90	348	21.2	c10
Two-component response regulator	*ompR*	Copper	TM7 (oral taxon)	58	7e-88	99	222	0.9	c490
Tellurite resistance protein	*telA*	Tellurite	*Lactococcus garvieae*	33	4e-08	86	94	2.1	c253
Type II secretory pathway	*exeA*	Virulence gene	*Pseudomonas aeruginosa*	100	0.0	100	388	11	c36
Collagen binding domain	*cbsA*	Virulence gene	*Streptococcus salivarius*	39	3e-80	93	478	12.2	c31
Collagen binding domain	*cbsA*	Virulence gene	*Streptococcus salivarius*	40	7e-65	80	471	9.4	c51
Staphylococcal complement inhibitor SCIN	*scn*	Virulence gene	*Staphylococcus aureus*	100	2e-74	100	116	3.4	c154

2.3.7. Comparison with other viral metagenomes

The comparative analysis was based on Principal Component Analysis (PCA) using MG-RAST. The viral metagenomes used for comparison with assembled CF sputum metagenome were: CF sputum (Willner *et al.*, 2009), oropharyngeal swabs from healthy individuals (Willner *et al.*, 2010), human lower respiratory tract (Lysholm *et al.*, 2012), human saliva (Pride *et al.*, 2012) and terrestrial hot springs used as an outgroup (Pride & Schoenfeld, 2008). PCA analysis indicated that viral communities found in the CF sputum in this study were more similar to salivary viruses (Figure 2.6A), while functional genes were distinct (Figure 2.6B).

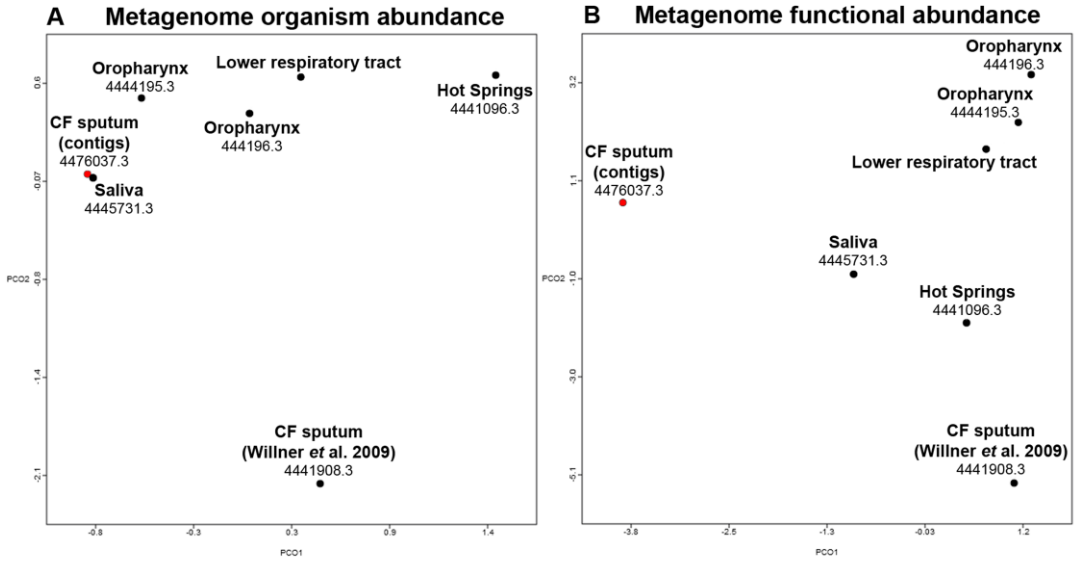


Figure 2. 6. Principal component analysis (PCA) based on organism abundance (A) and functional abundance (B) using MG-RAST. Metagenome comparisons were calculated for CF sputum metagenome (contigs >100bp) (red circle) and other published viral metagenomes (black circles): CF sputum (Willner *et al.*, 2009), human oropharynx (Willner *et al.*, 2010), human lower respiratory tract (Lysholm *et al.*, 2012), human saliva (Pride *et al.*, 2012) and hot springs (Pride & Schoenfeld, 2008). The data was compared to GenBank (organism abundance) and Subsystems (functional abundance) using a maximum e-value of 1e-05, a minimum identity of 80 %, and a minimum alignment length of 50.

2.3.8. CRISPR sequences in CF viral metagenome

Analysis of bacterial CRISPR spacers matching the CF metagenome revealed that there were 54 different spacers matching CF reads and 15 different spacers matching assembled contigs (Table 2.9). Multiple metagenome sequence matches were seen with CRISPR sequences from the following genera: *Clostridium* (5), *Streptococcus* (5), *Corynebacterium* (2), *Leptotrichia* (2), *Methanothermococcus* (2), *Pelobacter* (2), *Pseudomonas* (2), *Rothia* (2) and *Veillonella* (2). The most common identified contig CDS containing CRISPR hits were phage structural genes followed by phage enzymes. Functional assignment of these CDS is shown in Table 2.10.

CRISPR repeat-based PCR amplification on bacterial DNA in sputum from local CF patients and CRISPRfinder analysis of sequenced clones resulted in identification of 40 unique spacers (Table 2.11). BLASTN analysis revealed that 29 (72.5%) spacers had identity to known *Pseudomonas aeruginosa* isolates and phages, while remaining spacers were novel. None of the spacers matched the viral metagenome contigs, including the spacers identified in the bacterial DNA extracted from the same sputum as the metagenomic phage DNA used in this study. The most frequent CDS identity of the CRISPR hits were tail proteins and P-loop NTPase.

2.3.9. Identification of novel Torque Teno viruses in CF sputum

Eleven contigs analysed had similarity to human Torque Teno virus (TTV). Divergent PCR using primers designed from the largest (2.8 kb) assembled contig followed by product sequencing confirmed it formed closed circle in the phage DNA sample (Figure 2.7). Phylogenetic analysis based on the translated capsid protein (ORF1) from the three largest contigs indicated co-infection with at least three highly divergent anelloviruses (Figure 2.8).

Table 2.9. Bacterial CRISPRdb (Grissa *et al.*, 2007b) spacers matching CF reads and contigs.

Bacterial genome	**No. of unique spacers matching CF reads**	**No. of unique spacers matching CF contigs**
Acidianus hospitalis W1	1	0
Arcanobacterium haemolyticum DSM 20595	1	0
Bacillus coagulans 2-6	1	0
Bifidobacterium dentium Bd1	1	0
Caldicellulosiruptor obsidiansis OB47	1	0
Caldicellulosiruptor owensensis OL	0	1
Candidatus Arthromitus sp. SFB-mouse-Japan	1	0
Clostridium botulinum A3 str. Loch Maree	1	0
Clostridium difficile CD196	1	0
Clostridium novyi NT	2	1
Clostridium thermocellum ATCC 27405	1	0
Corynebacterium resistens DSM 45100	1	0
Corynebacterium urealyticum DSM 7109	1	0
Dictyoglomus turgidum DSM 6724	1	0
Erwinia tasmaniensis Et1/99	1	0
Fusobacterium nucleatum subsp. nucleatum 25586	1	0
Isosphaera pallida ATCC 43644	1	0
Leptotrichia buccalis C-1013-b	2	0
Meiothermus ruber DSM 1279	1	0
Metallosphaera sedula DSM 5348	1	1
Methanocaldococcus jannaschii DSM 2661	1	0
Methanococcus voltae A3	1	0
Methanosarcina mazei Go1	1	0
Methanosphaera stadtmanae DSM 3091	1	1
Methanothermobacter thermautotrophicus Delta H	1	0
Methanothermococcus okinawensis IH1	2	0
Neisseria lactamica 020-06	1	0
Pelobacter carbinolicus DSM 2380	2	1
Pseudomonas aeruginosa LESB58	1	1
Pseudomonas aeruginosa UCBPP-PA14	1	1
Roseiflexus castenholzii DSM 13941	1	1
Rothia dentocariosa ATCC 17931	2	2
Salmonella enterica subsp. enterica Typhimurium LT2	1	0
Streptococcus agalactiae NEM316	1	0
Streptococcus gordonii str. Challis substr. CH1	2	1
Streptococcus salivarius CCHSS3	1	1
Streptococcus sanguinis SK36	1	0
Sulfolobus islandicus L.D.8.5	1	1
Syntrophomonas wolfei subsp. wolfei str. Goettingen	1	0
Thermincola potens JR	1	0
Thermoanaerobacter pseudethanolicus ATCC 33223	1	0
Thermobispora bispora DSM 43833	1	0
Thermodesulfobium narugense DSM 14796	1	0
Thermomonospora curvata DSM 43183	1	0
Thermosipho melanesiensis BI429	1	0
Thermotoga neapolitana DSM 4359	1	0
Veillonella parvula DSM 2008	2	2
Xenorhabdus nematophila ATCC 19061	1	0
Total	**54**	**15**

Table 2.10. Analysis of CRISPR spacer matches to CF phage metagenome contigs.

CF phage contig	Bacterial genome CRISPR matches to CF phage contigs	Bacterial source	CF phage CDS assignment
contig3	*Rothia dentocariosa* ATCC 17931	Human oral cavity	N-acetylmuramoyl-L-alanine amidase
contig7	*Veillonella parvula* DSM 2008	Human gastrointestinal tract (Gronow *et al.*, 2010)	ArpU family transcriptional regulator
contig27	*Pseudomonas aeruginosa* LESB58	CF patient (Winstanley *et al.*, 2009)	Prophage tail length tape measure protein
contig63	*Rothia dentocariosa* ATCC 17931	Human oral cavity	XRE family transcriptional regulator
contig76	*Pseudomonas aeruginosa* UCBPP-PA14	Pathogenic isolate from human (unspecified site) (Rahme *et al.*, 1995)	Hypothetical protein
contig121	*Caldicellulosiruptor owensensis* OL	Lake sediment (Huang *et al.*, 1998)	No CDS in match region
contig132	*Clostridium novyi* NT	Soil	Phage structural protein
contig144	*Methanosphaera stadtmanae* DSM 3091	Archaeal commensal of human intestine (Fricke *et al.*, 2006)	DNA polymerase
contig176	*Pelobacter carbinolicus* DSM 2380	Aquatic environments	Phage portal protein
contig298	*Streptococcus gordonii* str. Challis	Human oral cavity (Vickerman *et al.*, 2007)	Phage tail protein
contig334	*Veillonella parvula* DSM 2008	Human dental plaque	Hypothetical protein
contig1149	*Metallosphaera sedula* DSM 5348	Hot springs	No match
contig1490	*Streptococcus salivarius* CCHSS3	Human oral cavity	No match
contig1758	*Roseiflexus castenholzii* DSM 13941	Hot springs	Endolysin
contig2439	*Sulfolobus islandicus* L.D.8.5	Volcanic field, Hot springs	No match

Table 2.11. BLASTN analysis of CRISPRs amplified from bacterial DNA in CF sputum samples.

Spacer name	Sputum sample No.	Closest database match	Database phage CDS assignment	% coverage	% identity	E -value
CFspacer1	1	*Pseudomonas aeruginosa* strain SMC4396 CRISPR repeat sequence	-	100	100	3e-08
CFspacer2		*Pseudomonas aeruginosa* strain SMC4396 CRISPR repeat sequence	-	100	100	3e-08
CFspacer3		*Pseudomonas aeruginosa* strain SMC4512 CRISPR repeat sequence	-	100	100	3e-08
CFspacer4		*Pseudomonas aeruginosa* strain PACS458 clone fa1376	Hypothetical protein	100	100	3e-08
CFspacer6		*Pseudomonas aeruginosa* strain PACS171b clone fa1386	Hypothetical protein	100	97	7e-06
CFspacer7		*Pseudomonas* phage phi297	Hypothetical protein	100	100	3e-08
CFspacer10		*Pseudomonas aeruginosa* NCGM2.S1	Tail fiber assembly protein	100	97	7e-06
CFspacer12		*Pseudomonas aeruginosa* strain PACS458 clone fa1376	Hypothetical protein	100	97	7e-06
CFspacer13		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	Hypothetical protein	100	94	0.010
CFspacer14		*Pseudomonas* phage MP38	Hypothetical protein	90	97	4e-04
CFspacer16		*Pseudomonas aeruginosa* strain PACS171b clone fa1386	P-loop NTPase	96	100	1e-07
CFspacer19		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	Hypothetical protein	100	100	2e-07
CFspacer20		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	Tail tape measure protein	100	100	2e-07
CFspacer21		*Pseudomonas aeruginosa* PADK2_CF510 contig009	Hypothetical protein	100	100	2e-07
CFspacer24	2	*Pseudomonas* phage PAJU2	Replication protein P	100	100	3e-08
CFspacer25		Bacteriophage D3	Hypothetical protein	100	100	3e-08
CFspacer27		*Pseudomonas aeruginosa* strain PACS458 clone fa1374	Hypothetical protein	100	97	7e-06
CFspacer28	3	*Pseudomonas aeruginosa* strain PACS171b clone fa1386	No CDS in match region	100	100	3e-08
CFspacer29		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	No CDS in match region	100	100	2e-07
CFspacer30		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	Endolysin	100	97	4e-05
CFspacer31		*Pseudomonas aeruginosa* strain PACS458 clone fa1376	Hypothetical protein	96	100	1e-07
CFspacer32		*Pseudomonas aeruginosa* strain PACS171b clone fa1386	No CDS in match region	100	100	3e-08
CFspacer33		*Pseudomonas aeruginosa* NCGM2.S1	Hypothetical protein	100	100	3e-08
CFspacer34		*Pseudomonas aeruginosa* PA7	Hypothetical protein	100	100	3e-08
CFspacer35		*Pseudomonas aeruginosa* strain PACS171b clone fa1386	P-loop NTPase	100	100	3e-08
CFspacer36		*Pseudomonas aeruginosa* 9BR 9BRScaffold1	Tail fibre protein	100	100	2e-07
CFspacer37		*Pseudomonas aeruginosa* NCGM2.S1	Hypothetical protein	100	100	3e-08
CFspacer39		*Pseudomonas aeruginosa* strain PACS458 clone fa1374	P-loop NTPase	93	97	1e-04
CFspacer40	4*	*Pseudomonas* phage MP38	Hypothetical protein	90	100	2e-06

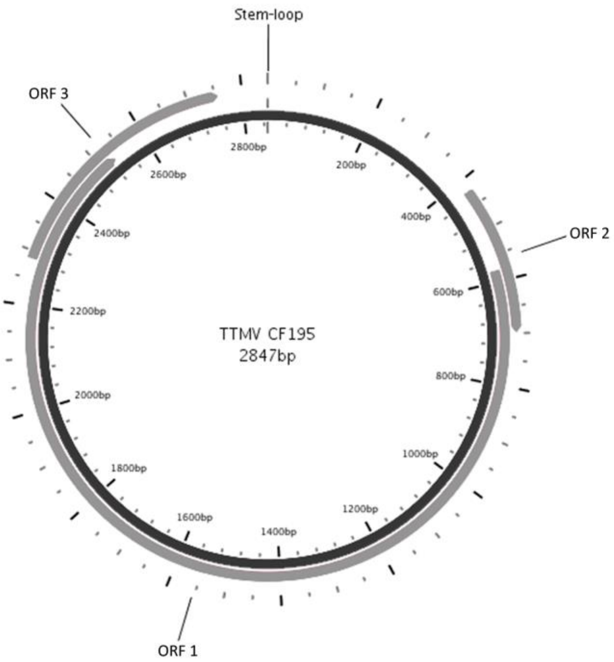


Figure 2. 7. Genomic organization of the TTMV (Torque Teno Mini Virus) assembled from CF sputum metagenome. ORF1 encode the capsid protein, and ORF2 and ORF are uncharacterized proteins.

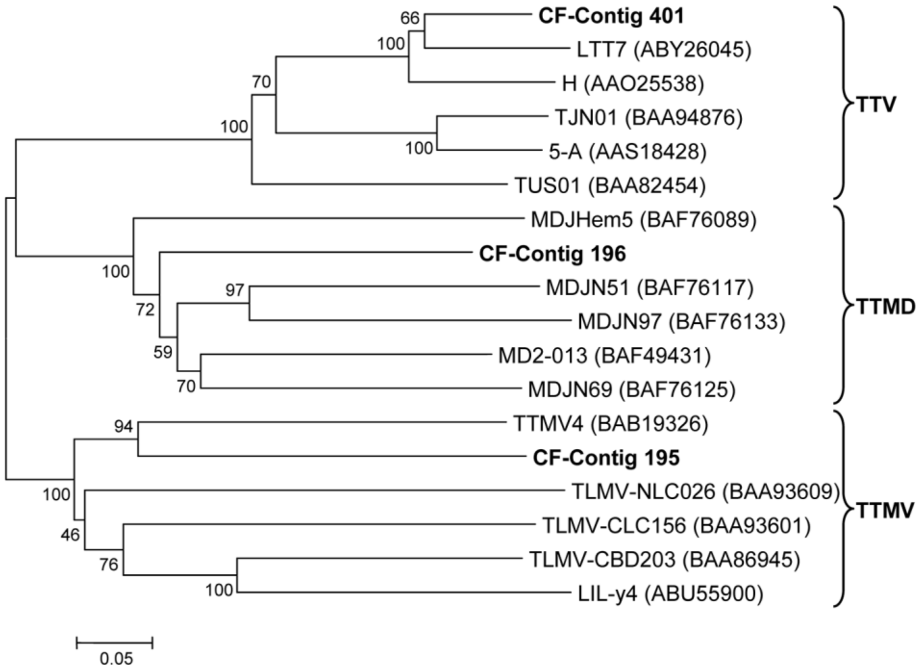


Figure 2.8. Neighbor-joining phylogenetic tree based on the partial amino acid sequences of the ORF1 (capsid protein) of the three anelloviruses identified in this study and most similar anellovirus sequences in GenBank. TTV sequences isolated in this study are shown in bold. GenBank accession numbers are shown in parentheses. TTV (Torque Teno Virus), TTMDV (Torque Teno Midi Virus) and TTMV (Torque Teno Mini Virus) indicate phylogenetic groups of TT viruses.

2.4. Discussion

The respiratory tract of cystic fibrosis patients is heavily colonised with biofilm-forming bacteria which play a major role in the disease process. Environments containing large numbers of bacteria generally support bacteriophage predation. The phage population in the respiratory tract of CF patients is of relevance to the disease process because of the potential for adaptive selection of bacteria by phage, the acquisition of novel traits by lysogeny and the future use of phage therapy or phage effectors in cystic fibrosis treatment. This study used Illumina short-read-based DNA sequencing of filtered sputum to examine the diversity of phages in the potentially transmissible form of sputum from a cystic fibrosis patient. We showed that expectorated sputum from a cystic fibrosis patient contains phages parasitizing both the oral and lower respiratory tract microbiota, and incorporating antimicrobial resistance genes.

Although the majority of database hits both from reads and assembled contigs were to bacterial sequences, functional analysis showed that the metagenome (from a sample filtered and enriched for phage content) was in fact dominated by phage-related genes, specifying proteins involved in phage integration, assembly, replication and host lysis. The abundance of bacterial sequences in viral metagenome is consistent with findings of other studies, including a study of viral communities in cystic fibrosis sputum (Willner *et al.*, 2009). High abundance of bacterial-like sequences in viral metagenomes are likely a result of unannotated prophage regions in bacterial genome sequences deposited in public databases (Fouts, 2006) Phages similar to those known to infect different species of *Streptococcus*, *Veillonella*, *Pseudomonas*, *Actinobacillus* and *Prevotella* were the most abundant source for these sequences (Figure 2.1, Table 2.2). These results mirror findings of culture-independent studies of bacterial diversity in CF sputum, in which these genera were the most common detected, including the core CF pathogens like *Streptococcus pneumoniae, Streptococcus mitis*, *Veillonella dispar*, *Veillonella atypica*, *Veillonella parvula*, *Pseudomonas aeruginosa*, *Prevotella oris* and *Staphylococcus aureus* (Bittar *et al.*, 2008; Guss *et al.*, 2011; van der Gast *et al.*, 2011).

The vast majority of 76 bp Illumina reads could not be assigned to any known database sequences. This could be because sensitivity in the detection of sequence homology by BLAST decreases with short read lengths (Wommack *et al.*, 2008). As the *de novo* assembly generated longer contigs, the number of assigned sequences improved, with many assignments found that were not detected by read classification. The most frequent phage CDS detected in assembled data resembled phages infecting different species of *Streptococcus* (with large contigs having similarity to *Streptococcus pneumoniae*, *Streptococcus infantis*, *Streptococcus mitis* and *Streptococcus sanguinis*), constituting about 21% of the assigned sequences (Figure 2.1B, Table 2.2). Most of the other contigs resembled phages from bacteria present in the oropharynx including a temperate phage of the periodontal pathogens *Actinobacillus actinomycetemcomitans* (Resch *et al.*, 2004), and *Capnocytophaga sputigena* (Stevens *et al.*, 1980), or an organism associated with halitosis, *Atopobium parvulum* (Copeland *et al.*, 2009). Sequence analysis of some of the large contigs showed also similarity to the phages of known CF pathogens (Figure 2.2). The largest contig had weak similarity to phage isolated by environmental screening with a *Mycobacterium smegmatis* host (Hatfull *et al.*, 2010). Non-tuberculosis mycobacteria are known to be prevalent in airways of CF patients (Levy *et al.*, 2008). Multiple contigs were found with a marked similarity to a prophage from a *Pseudomonas aeruginosa* genome sequence of an organism associated with ocular keratitis isolated in Manchester, UK (Stewart *et al.*, 2011). Smaller contigs corresponding to sequences from oral bacteria *Oribacterium* sp., *Neisseria bacilliformis*, *Kingella oralis*, *Solobacterium moorei* and *Bulleidia extructa* were also detected (Carlier *et al.*, 2004; Chen, 1996; Downes *et al.*, 2000; Han *et al.*, 2006; Haraszthy *et al.*, 2007).

The CF sputum sample described in this study is different from CF sputum described by (Willner et al. 2009) at taxonomic and functional level (Figure 2.6). For example, Willner *et al*. study, that used 454 technology, did not report assembly of contigs containing apparent *Pseudomonas* phage sequences, despite the presence of culturable *Pseudomonas aeruginosa* in the sputum of the CF patient used for metagenome assembly. The contrasting assembly of multiple *Pseudomonas* contigs in our short read study is compatible with an enhanced assembly capacity of Illumina short reads compared with 454 reads (Luo *et al.*, 2012). Illumina sequencing is also cheaper than 454 at present (approximately 25% for a similar metagenome coverage) (Luo *et al.*, 2012), and the combination of Illumina sequencing and de novo assembly seems suitable for future applications of phage metagenomics such as monitoring the effects of phage therapy in cystic fibrosis.

The only eukaryotic virus detected in the contig assembly was Torque Teno virus (TTV), a small circular ssDNA virus from the *Anelloviridae* family, highly prevalent in the blood of healthy individuals (Vasilyev *et al.*, 2009). TTV has been associated with lower respiratory disease in nasal secretions and bronchoalveolar lavage fluid (Maggi *et al.*, 2003; Wootton *et al.*, 2011), although its etiological role in respiratory disease is uncertain. TTV has recently been demonstrated in a viral metagenome recovered from CF lung sections (Willner *et al.*, 2012), with herpes simplex virus and human papilloma virus, which were not detected in our assembly, (there was a single read hit to a herpesvirus in our dataset). In this study, we identified at least three contigs that were closely related to different groups of TTV (Figure 2.8), suggesting coinfection with multiple genogroups as previously noted in faeces by PCR (Pinho-Nascimento *et al.*, 2011). Potential transmission of this virus by cystic fibrosis sputum is therefore possible.

De novo assembly with Meta-IDBA (Peng *et al.*, 2011) recruited the majority of short reads and produced some surprisingly large contigs, with the largest being over 60 kb long. The assembly accuracy was confirmable by PCR and sequencing on the original amplified DNA over multiple CDS in three contigs (Figure 2.2). The use of Phi 29 polymerase multiple strand displacement amplification as in this study is known to preferentially amplify small circular DNA molecules (Pinard *et al.*, 2006), but this did not prevent contig assembly up to 60 kb from a mycobacteriophage likely to have a capsid diameter of over 80 nm. The fact that 59% of reads assembled into 2,859 contigs >100 bp suggests that the overall viral diversity was limited. This is in agreement with the PHACCS analysis that showed that the viral diversity in CF sputum was low (Table 2.4) and comparable with viral diversity previously found within cystic fibrosis sputum and lung tissue using a different sequencing technique (454) (Willner *et al.*, 2009; Willner *et al.*, 2012).

Some prophages in the CF metagenome described here may result from antibiotic induction, as antibiotic treatment of CF patients contributes to phage induction in bacteria (Fothergill *et al.*, 2010). Phages contribute to the development of antimicrobial resistance in environmental bacteria by transferring genes encoding antibiotic resistance (Colomer-Lluch *et al.*, 2011). β-lactamase genes are highly abundant in phages in bacteria-rich environments (Colomer-Lluch *et al.*, 2011; Muniesa *et al.*, 2004). In this study two metallo-β-lactamase gene sequences (class B β-lactamases) potentially conferring resistance to β-lactam antibiotics were detected in the CF sputum viral metagenome (Figure 2.5, Table 2.8). Metallo-β-lactamase gene sequences were previously reported in two publications involving 454 sequencing of CF respiratory tract-associated metagenomic viral DNA from the same centre in the USA. One of these reports involved sequence data from lung tissue, (Willner *et al.*, 2012), a site with limited potential for transmission - it is well established from bronchoalveolar lavage studies that lower airways and lungs in cystic fibrosis and other conditions may contain bacteria (and presumably their phage predators) that are not detectable in sputum (Hilliard *et al.*, 2007), However the other report used data from sputum samples (Willner et al. 2009), in which β-lactamase sequences were subsequently detected (Fancello *et al.*, 2011; Willner *et al.*, 2009). The consistent presence of metallo-β-lactamase genes in phage sequences from CF sputum in different continents obtained using different sequencing methodology suggests CF sputum is a potential transmission source for transducible antimicrobial resistance genes.

The only toxin gene detected was *cin* (Table 2.8). This encodes an antiopsonic protein which acts on complement C3 convertase and is present in the vast majority of *Staphylococcus aureus* strains isolated from humans (van Wamel *et al.*, 2006). It is located in the *S. aureus* genome on a β-haemolysin converting prophage in an 8 kb region containing several other genes acting on the human innate immune system, an innate immune evasion cluster (IEC). These prophages are readily inducible as infective phages *in vitro* by mitomycin-C treatment, and it is not surprising to find evidence of them in CF sputum, because of the common presence of *S. aureus* in the upper and lower respiratory tract of CF patients.

Many bacteria contain clustered, regularly interspaced short palindromic repeats (CRISPR elements). These elements are thought to function as defences against bacteriophage attack by an RNA interference (RNAi)-like mechanism, inactivating extrachromosomal DNA using a short phage sequences placed between conserved direct repeats (DR) (Barrangou *et al.*, 2007; Sorek *et al.*, 2008). The presence of CRISPR spacer matches (protospacers) from database searches (Bhaya *et al.*, 2011) in the CF phage contigs (Table 2.9) suggest that these spacers are derived from phages, and the phage matches could be used as a tool to determine possible origins of these sequences. Among CRISPR spacers matching our phage contigs (Table 2.10) were those identified in *Pseudomonas aeruginosa* genome sequences (including a strain isolated from a cystic fibrosis patient), and genomes of other bacteria associated with CF including *Rothia dentocariosa* (Tunney *et al.*, 2008), and others whose role in the pathogenesis of cystic fibrosis is uncertain e.g. *Leptotrichia buccalis*. However, locally identified CRISPR spacers (Table 2.11) amplified from bacterial DNA in CF sputum with *Pseudomonas*-specific primers (including bacterial DNA from the sputum sample used for the phage metagenome) were not matched to the metagenome. This is compatible with local effectiveness of bacterial CRISPR systems in resisting phage attack and selecting for phages not represented in CRISPR spacer sequences, and also suggests a limited global diversity of phages in the respiratory tract of cystic fibrosis patients.

Chapter 3:

Metagenomic sequencing of DNA viruses in dairy wastewater

Abstract

In this chapter short read (Illumina) sequencing was used to investigate the diversity of the DNA virus community in the activated sludge of a dairy food wastewater treatment plant. Bacteriophages were isolated from a wastewater sample using a tangential flow filtration (TFF) technique with DNase treatment. Phage DNA was extracted and amplified by a rolling circle technique and then sheared before sequencing as a paired-end sequencing run using Illumina technology. Although over 99% of reads could not be matched to any GenBank sequence, 47% of the reads were assembled into contigs greater than 100 bp using the meta-IDBA assembler. Fifty eight contigs were over 10 kb and the largest contig was 114 kb. Most of the assigned sequences were mainly of phage or prophage origin, but the majority of the sequence data obtained was classified as "unassigned". Phages from bacteria of the genera *Vibrio*, *Mycobacterium*, *Synechococcus*, *Pseudomonas* and *Burkholderia* comprised the most frequently assembled contigs. Fifteen complete single-stranded eukaryotic and prokaryotic viruses were detected in the metagenome assembly representing diverse members of *Circoviridae*, *Nanoviridae*, *Geminiviridae* and *Microviridae* families. A substantial number of assigned sequences could be affiliated with species typically found in the soil and aquatic environments, including wastewater. Viral diversity was moderate, with a total of 20,278 contigs originating from 409 species. Genes involved in phages, prophages, DNA metabolism and sulphate metabolism were prevalent. Antimicrobial resistance genes detected were flanked by phage sequences, suggesting dairy wastewater is a source for transmissible antimicrobial resistance.

3.1. Introduction

Wastewater generated by the dairy food industry is an extremely eutrophic environment due to high concentrations of milk products, which render it rich in organic and inorganic matter. The main components of dairy wastewater are proteins, carbohydrates, lipids and minerals such as ammonia, nitrogen and phosphorus (Britz *et al.*, 2005). This environment is inhabited by many different prokaryotic and eukaryotic microorganisms specialized in the degradation of organic pollutants and nutrients through nitrification, denitrification, ammonification, sulphur and sulphate reduction (McGarvey *et al.*, 2007; Tocchi *et al.*, 2012; Yu & Zhang, 2012). In addition, dairy wastewater may contain many zoonotic pathogens, including *Escherichia coli*, *Campylobacter*, *Listeria*, *Mycobacterium* and *Salmonella* (Dungan *et al.*, 2012).

The process of wastewater treatment can be affected by bacteriophage addition to control the growth of bacteria responsible for foaming and bulking in activated sludge (Choi *et al.*, 2011; Withey *et al.*, 2005). Phages can be also used to control pathogenic bacteria population and currently have application in controlling of bacterial pathogens in food including dairy products (Greer, 2005). Bacteriophages in wastewater may also support microorganisms in turnover of organic matter. In other environments it is known that phages contribute to nutrient cycling through the lysis of bacteria (Fuhrman, 1999). Phages can also influence their host pathogenicity and adaptation by horizontal transfer of toxin and antimicrobial resistance genes (Boyd & Brussow, 2002). Antibiotic resistance amongst bacterial pathogens is an increasing problem, and several studies indicate that antibiotic resistance genes are highly abundant in bacteriophages detected in the environment (Colomer-Lluch *et al.*, 2011; Parsley *et al.*, 2010a; Stalder *et al.*, 2012).

Few studies to date have employed metagenomics techniques to characterize viral communities from wastewater samples. A metagenomic analysis of reclaimed water viruses from USA using pyrosequencing revealed that bacteriophages dominated the DNA viral community, while eukaryotic viral sequences were dominated by viruses containing single-stranded DNA circular genomes, including plant pathogens from the *Geminiviridae* and *Nanoviridae* families, and animal pathogens from the

Circoviridae family (Rosario *et al.*, 2009b). A metagenomic analysis of viral sequences derived from the activated sludge from USA showed that 60% of sequenced reads were of bacterial origin, while reads of viral origin were dominated by bacteriophages (95%) (Parsley *et al.*, 2010b). Cantalupo (Cantalupo *et al.*, 2011) used pyrosequencing to analyse raw sewage from three locations (USA, Spain and Ethiopia) and detected 51 virus families. Bacteriophages from families *Microviridae*, *Siphoviridae*, *Myoviridae*, *Podoviridae*, and *Inoviridae* dominated viral fraction (Cantalupo *et al.*, 2011). A metagenomic analysis of viruses in untreated sewage samples from four locations (USA, Thailand, Nepal and Nigeria) revealed sequences related to 29 eukaryotic viral families, including many known human pathogens (Ng *et al.*, 2012). Another study used pyrosequencing to analyse the DNA viruses in the influent, activated sludge, effluent, and anaerobic digester of a wastewater treatment plant in Singapore and found that > 80% of viral reads had similarities to bacteria and the remaining sequences were mainly classified as bacteriophages (Tamaki *et al.*, 2012). One study used Illumina sequencing to characterize viral communities from the influent and effluent sludge from anaerobic digesters of five domestic wastewater treatment plants from USA (Bibby & Peccia, 2013). The majority of assembled contigs were of bacteriophage origin, followed by eukaryotic viruses, including human pathogens (Herpesvirus most abundant) (Bibby & Peccia, 2013). A recent study used pyrosequencing to characterize viruses in a dairy waste treatment lagoons (Alhamlan *et al.*, 2013). Viral communities were dominated by bacteriophages, the majority of them belonging to the *Siphoviridae* family, whereas among eukaryotic viruses were known animal pathogens such as those from *Circoviridae* and *Herpesviridae* family and plant pathogens from *Geminiviridae* and *Nanoviridae* family (Alhamlan *et al.*, 2013).

High-throughput sequencing have been applied to investigate the diversity of viruses in wasters from municipal wastewater treatment plants and treatment lagoons receiving dairy manure, however no metagenomic work have been conducted on viral communities from a dairy food wastewater. The aim of this chapter was to investigate taxonomic and functional diversity of the viral community in a water sample obtained from the dairy wastewater treatment plant receiving milk product-polluted wastewater using metagenomic Illumina-based high throughput sequencing.

The resulted sequences have been analysed using different bioinformatics tools and compared with metagenomic sequences from other wastewater environments.

3.2. Material and Methods

3.2.1. Sample collection

An activated sludge sample (15 litre volume) was collected in January 2010 from an open steel aeration tank raised off the ground receiving milk product-polluted wastewater treated by dissolved air flotation, and anaerobic digestion. The site was Kerry Ingredients Wastewater Treatment Plant in Listowel, Co. Kerry, Ireland (52°26’20” N; 9°29’7” W).

3.2.2. Viral particles purification

Viruses were purified using a published combination of filtration and density gradient centrifugation (Thurber *et al.*, 2009). Initially, the wastewater sample was passed through a 100 µm mesh to remove large particles and filtered using a 0.45 µm Tangential Flow Filter (TFF) (QuixStand, Amersham Bioscience) to remove prokaryotes and eukaryotes. Viruses in the filtrate were then concentrated with the 100kD TFF filter (QuixStand, Amersham Bioscience) to a final volume of ~50 ml. Viral concentrate was overlaid onto a CsCl gradient (1 ml of CsCl of densities at 1.7, 1.5 and 1.35 g/ml) and ultracentrifuged in a SW-41 (Beckman) rotor at 22,000 rpm for 2 h at 4°C. 1 ml of the fraction between 1.35 g/ml and 1.5 g/ml density was collected, filtered through a 0.22 µm syringe filter (Millipore) and further concentrated and washed twice with 1 ml of SM buffer on a Millipore Amicon Ultra-15 Centrifugal Filter Unit (30 kDa). The final volume of the concentrated CsCl fraction was 500 µl and was used for DNA extraction.

3.2.3. DNA isolation and sequencing

Prior to DNA extraction, the concentrated CsCl fraction was incubated for 30 min at 37°C with 20 U/ml of DNase I (New England Biolabs) to remove free nucleic acids. DNase was inactivated by the addition of EDTA to the final concentration of 20 mM, followed by heat-inactivation at 75°C for 10 minutes. Viral DNA was extracted using formamide/CTAB extraction protocols (Sambrook *et al.*, 1989; Thurber *et al.*, 2009). Briefly, the sample was mixed with 0.1 volumes of 2M Tris-HCl/0.2M EDTA, 0.01 volumes of 0.5 M EDTA, 1.0 volume of deionized formamide, 10 µl/ml sample of glycogen and incubated for 30 minutes at 37°C. The phage DNA was precipitated for 30 min at 4°C with 2 volumes of 100% ethanol and washed with 70% ethanol. The phage DNA pellet was resuspended in 500 µl of TE buffer and used for CTAB extraction protocol. The sample was incubated for 1 h at 56°C with 15 µl of 20% (w/v) SDS and 20 µl of proteinase K (DNeasy Blood & Tissue Kit, Qiagen). After protease treatment 100 µl of 5 M NaCl and 80 µl of pre-heated CTAB/NaCl solution (10% (w/v) CTAB in 0.7 M NaCl) was added and incubated for 10 minutes at 65°C. DNA was recovered by phenol/chloroform extraction and isopropanol precipitation and the resulting DNA pellet was resuspended in 30 µl of sterile water. DNA concentration was estimated using NanoDrop (Thermo Scientific), giving a 35 ng/µl concentration and a 260/280 ratio of 0.96. To check for bacterial contamination, the extracted DNA was screened for the presence of 16S rRNA genes by PCR using primers 27F and 907R (Lane, 1991).

Extracted metagenomic DNA was used for whole genome amplification using GenomiPhi V2 DNA Amplification Kit (GE Healthcare). Briefly, 1µl (35 ng) of DNA was amplified in 20 µl reaction volumes in triplicate reactions at 30°C for 2 h. The amplified products from each reaction were pooled, purified using DNeasy Blood & Tissue Kit (Qiagen) and resuspended in 100 µl of sterile water. Approximately 5 µg of DNA was used to make libraries with inserts of between 150 and 250 bp at the Wellcome Trust Sanger Institute (UK) which were sequenced in a paired 76 cycle run using an Illumina Genome Analyzer IIx.. 10.3 million paired-end 76 bp reads were generated.

3.2.4. Confocal microscopy

Aliquots (100 µl) of the viral concentrate and CsCl gradient fractions were preserved with an equal volume of 4% formaldehyde (0.02 µm pre-filtered), diluted in 5 ml of (0.02 µm pre-filtered) water and filtered onto 0.02 µm Anodiscs filters (Whatman). Filters were stained with 2.5x SYBR Gold (Invitrogen), viewed at 630x under a Confocal Laser Scanning Microscope (Zeiss LSM 5 exciter CLSM) and images were acquired using Zen software. Argon laser was used at <2% power at 488nm with emission filtered at 530 nm.

3.2.5. Electron microscopy

An aliquot (100 µl) of the CsCl gradient fraction was fixed with 1 ml of EM Buffer (0.5% Glutaraldehyde in 20 mM HEPES buffer, pH 7.0) and send to Dr. Heinrich Lünsdorf (Germany) for electron microscopy analysis. Viruses were adsorbed to the carbon-coated Butvar grids for 2 minutes and stained negatively with 4% uranyl acetate. The grids were viewed under a Transmission Electron Microscope (Zeiss CEM 902; Zeiss, Oberkochen, Germany). Images were obained using a charge-coupled-device camera (Proscan Electronic Systems, Scheuring, Germany).

3.2.6. Sequence processing

All large-scale computational analyses were performed on the BRCI cluster at University College Cork unless otherwise indicated. The metagenomic library was filtered using the PRINSEQ website (Schmieder & Edwards, 2011a) to remove exact sequence duplicates and reverse complement exact duplicates (2,144,351), reads containing more than one ambiguous bases (N) and low-complexity sequences (DUST score <32) (18,766). Human sequences (637) were removed using DeconSeq website, using 90% coverage and 90% identity filtering options. The metagenomic sequences in the activated sludge metagenome were compared using DeconSeq standalone (Schmieder & Edwards, 2011b) to another metagenome based on different source DNA that was prepared in the laboratory at the same time and any shared sequences were removed using 90% coverage and 90% identity options (3,702,737 sequences). These pre-processing steps resulted in 4,474,959 high-quality sequences (340,096,884 total bases).

3.2.7. Assembly of sequence reads

For contigs assembly, sequence reads were prepared as described above, except duplicate reads were retained, as this resulted in a higher N50 of 1470. Sequence reads were pair-end assembled using Meta-IDBA (Peng *et al.*, 2011) into 34,636 contigs (11,432,658 total bases). Sequences less than 100 bp were discarded, leaving a total of 20,278 contigs for analysis (10,412,900 total bases). To calculate number of reads that were recruited into contigs assembly, reads were aligned to the contigs using FR-HIT (Niu *et al.*, 2011) using stringent parameters of 100% identity over 100% of the entire read length.

Putative open reading frames (ORFs) were identified on the assembled sequences (>100 bp) using MetaGeneMark Heuristic Approach version 1.0 (Zhu *et al.*, 2010) at http://exon.biology.gatech.edu/metagenome/Prediction/index.cgi.

3.2.8. Sequence annotation

Individual reads were automatically annotated against the GenBank database (e-value cutoff of 1e-03 and minimum alignment length of 20 bp) using the 'Representative Hit Classification' of the MG-RAST version 3.2 (Meyer *et al.*, 2008). Contigs were compared to the NCBI non-redundant (nr) protein database with an e-value < 1e-05 using BLASTX implemented in CAMERA 2.0 server (Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis) (Sun *et al.*, 2011). The data from the blast output was analysed using MEGAN version 4.69.4 (MEtaGenome ANalyzer) (Huson *et al.*, 2011) using Min Score of 100, a Top Percent value 10% and Min Support of 1. Classification to the taxon was based on counting number of bases in contigs assigned to each taxon. DNA sequences can be accessed through MG-RAST website under Project IDs 4476066.3 (reads) and 4476039.3 (contigs).

Functional annotation was performed on contigs (>100 bp) using the MG-RAST SEED Subsystems database (with an e-value cutoff of 1e-05 and minimum alignment length 50 bp) and ORFs using WebMGA (Wu *et al.*, 2011b) COG database (Clusters of Orthologous Groups of proteins) (Tatusov *et al.*, 2003) (using an e-value cutoff of 1e-05).

3.2.9. Analysis of complete genomes

Putative open reading frames (ORFs) of contig 0 and small circular genomes of ssDNA phages (*Microviridae*) were predicted using GeneMark (Besemer & Borodovsky, 1999). Each ORF was manually annotated against the NCBI nr protein database (nr) using BLASTP. Complete genomes were visualized using CG View (Grant & Stothard, 2008) or SnapGene Viewer free software http://www.snapgene.com/products/snapgene_viewer/ and edited in Photoshop.

3.2.10. Phylogenetic analyses

Translated sequences of 15 complete circular genomes of ssDNA viruses assembled from the metagenomic reads were aligned with selected reference sequences using ClustalW (Larkin *et al.*, 2007). The multiple sequence alignment of full-length Rep (*Circo*-, *Nano*-, *Geminiviridae* family) or major capsid proteins (*Microviridae*) was used for phylogenetic tree construction using MEGA version 5.04 (Tamura *et al.*, 2011) by applying p-distance model and the Neighbour-Joining method with 1000 bootstrap replications.

3.2.11. Metagenome diversity

Diversity of the metagenome was estimated using PHACCS version 1.1.3 (Angly *et al.*, 2005a) (http://sourceforge.net/projects/phaccs/). Circonspect version 0.2.5 (http://sourceforge.net/projects/circonspect/) implemented with the Octave version 3.6.0 was used to calculate contig spectra based on metagenome assembly using Minimo (98% identity over at least 35 bp overlap). The contig spectra were used as an input for PHACCS, using a logarithmic model and an average genome size of 50 kb.

3.2.12. Metagenome comparison

A multiple comparison based on organism and functional gene abundance between dairy wastewater contigs (>100bp) and other viromes: wastewater (Bibby & Peccia, 2013; Ng *et al.*, 2012; Parsley *et al.*, 2010b; Rosario *et al.*, 2009b; Tamaki *et al.*, 2012), lake (Roux *et al.*, 2012b) and CF sputum (Willner *et al.*, 2009) was performed using MG-RAST Principal Component Analysis (PCA). The data was compared to GenBank and SEED databases using a maximum e-value of 1e-05 and a minimum alignment length of 50. The data has been normalized to values between 0 and 1 and drawn using a Bray-Curtis distance.

3.2.13. Antibiotic resistance genes

To identify genes potentially conferring resistance to antibiotics, predicted ORFs from contigs were compared to 3,375 antibiotic resistance associated genes downloaded from The Comprehensive Antibiotic Resistance Database (CARD) http://arpcard.mcmaster.ca/ using BLASTP and e-value 1e-03. Hits to genes encoding integrases, efflux pumps, as well as *gyrA, gyrB, rpsL*, *bacA*, in which a point mutation in a bacterial gene confers antimicrobial resistance, were discarded. Easyfig was used to visualize ORFs of contig 50, encoding two antibiotic resistance genes. Phylogenetic analysis was conducted using MEGA version 5.04.

3.2.14. CRISPR spacer analyses

CRISPR spacer database containing 52,511 spacers was downloaded from http://crispr.u-psud.fr/crispr/CRISPRUtilitiesPage.html (Grissa *et al.*, 2007b) on 29/01/12 and used to search for sequence similarity with metagenomic reads and contigs (>100 bp) using BLASTN (e-value 1e-03 and word size 7). Only matches that had 100% identity over 20 bp were analysed, as previously described (Anderson *et al.*, 2011).

3.2.15. Microbial diversity in activated sludge

Total microbial DNA was extracted from 2 ml of raw wastewater. Prior to DNA extraction the sample was split into two aliquots that were centrifuged at 12,500 rpm for 5 min to pellet microbial cells. One tube was processed according to the protocol of the Promega Wizard Genomic DNA Purification Kit (Promega) to isolate DNA from Gram-positive and another to isolate DNA from Gram-negative bacteria. Extracted DNA was pooled from both tubes.

16S ribosomal RNA sequences were PCR amplified from the purified microbial DNA using universal bacterial primers 27F and 907R (Lane, 1991) and *Accumulibacter*-specific 16S rRNA primers cap438f and cap846r (Kunin et al., 2008). The PCR mixtures contained, in a total volume of 50 µl, 1x GoTaq® Green Flexi Reaction Buffer (Promega), 15 pmol of each of the primers, 2.5mM $MgCl_2$ (Promega), 0.1mM dNTPmix (New England Biolabs), 1U GoTaq® Flexi DNA Polymerase (Pomega) and 50 ng template DNA. For PCR amplifications an initial denaturation step of 5 min at 95°C was followed by 26 cycles of 30 sec at 94°C for denaturation, 30 sec at 55°C (for cap438f and cap846r) or 56°C (for 27F and 907R) for annealing, and 1 min at 72°C for extension. A final extension was carried out at

72°C for 5 min. The resulting PCR products were purified using QIAquick PCR Purification Kit (Qiagen) and cloned into pCR2.1 using TOPO TA Cloning Kit (Invitrogen) according to protocol instructions. Plasmid DNA was isolated from randomly selected bacterial colonies using QIAprep Spin Miniprep Kit (Qiagen) and digested with EcoRI (New England Biolabs) to identify clones containing insert. Ten clones showing different restriction pattern from each 16S rRNA library were sequenced at GATC Biotech (Germany).

Sequences were checked for chimera formation using the Bellerophon server (Huber et al., 2004) and no chimeras were detected. BLASTN was used to identify the closest match in GenBank against reference genomic sequences (refseq_genomic) database. The ribosomal RNA gene sequences have been submitted to the GenBank under accession numbers JN393605–JN393623.

3.3. Results

3.3.1. Visualization of viruses in activated sludge

Viruses present in the activated sludge tank of a dairy wastewater treatment plant were concentrated and extracted metagenomic DNA purified and sequenced using Illumina technology. Viruses in the concentrate were visualized prior to DNA extraction by confocal and electron microscopy. SYBR-Gold staining of nucleic acids within the virus capsid showed presence of numerous virus particles (Figure 3.1A). Transmission electron microscopy revealed different virion morphotypes, including tailed bacteriophages and enveloped viruses (Figure 3.1B to 3.1E).

Although no bacterial cells were seen by microscopy in the filtered concentrate, bacterial ribosomal RNA genes were detected by PCR on the extracted DNA (data not shown). The level of bacterial 'contamination' in the metagenomic phage sequence library was further checked by BLAST search against the RDP (Ribosomal Database Project) database using MG-RAST (with minimum alignment length of 20 bp and e-value cutoff of 1e-03). No 16S rRNA sequences were identified on contigs and only a small number of reads (30 reads) were found to be similar to ribosomal proteins (genera: *Roseateles*, *Mannheimia*, *Bacillus*, *Rubrivivax*, *Shewanella*, *Burkholderia*, *Cupriavidus*, *Pandoraea*, *Ralstonia*, *Acidovorax*, *Pseudacidovorax*, *Pelomonas*, *Leptothrix*, *Thiobacillus*, *Azovibrio* and *Bdellovibrio*). Several studies indicate that phage-mediated transduction of bacterial 16s rRNA genes naturally occurs within the viral communities in the environment (Del Casale *et al.*, 2011b; Harrington *et al.*, 2012; Sander & Schmieger, 2001). Therefore, the PCR result could represent transduction of bacterial DNA through horizontal gene transfer, and the assumption has been made that these sequences were of phage origin.

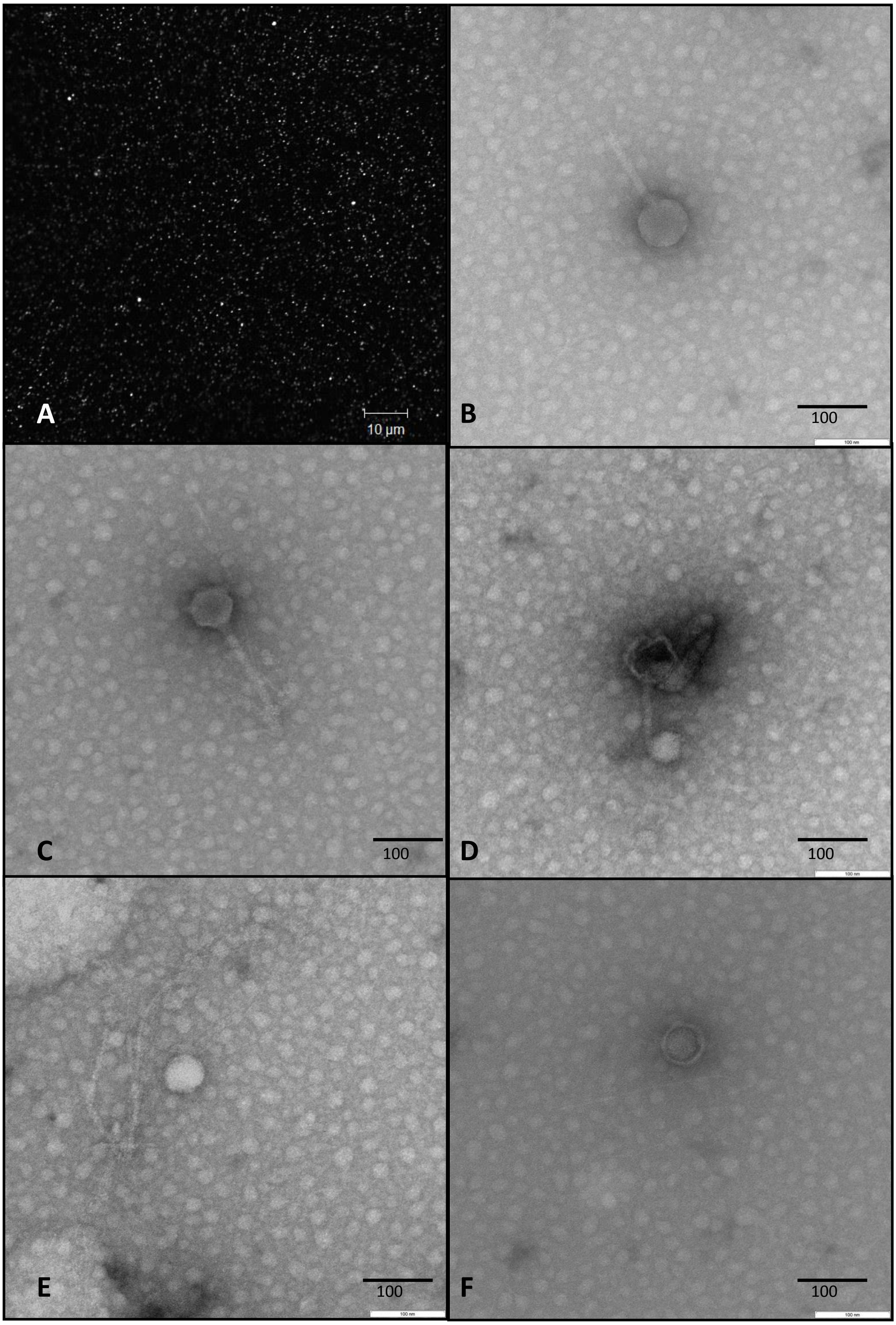


Figure 3. 1. Confocal (A) and electron (B-F) micrographs showing purified virus particles found in activated sludge.Confocal micrograph taken at 600× magnification.

3.3.2. Sequence reads taxonomic classification

Data filtering resulted in 4,474,959 high-quality 76-bp reads. A total of 6223 (0.14%) reads had hits to known sequences in the GenBank database (minimum alignment length of 20 bp and e-value cutoff of 1e-03) implemented in the MG-RAST server. Among these sequences, 96% were assigned to cellular organisms (mainly bacteria) and 4% to viruses (Figure 3.2A). Taxonomic assignment was further split according to the phylum (bacteria) or family (viruses) and species.

Reads classified as bacteria-like sequences were mainly affiliated to *Proteobacteria* (Figure 3.2B). The most abundant genera (data not shown) were *Neisseria* (6% of all assigned reads), *Pseudomonas* (3%), *Escherichia* (2%), *Salmonella* (1%) and *Acinetobacter* (1%). Bacterial species assignments of genes found in the sample included those associated with human and animal pathogens common in dairy wastewater such as *Neisseria* spp., *Escherichia coli, Salmonella enterica, Acinetobacter baumannii***,** *Campylobacter jejuni* or bacteria typically found in sludge e.g. *Pelobacter propionicus* and soil e.g. *Pseudomonas putida* (see table Table 3.2 for the top 50 bacteria species detected).

Reads assigned to viruses could be divided into 9 viral families, with the dsDNA bacterial viruses from order *Caudovirales*: *Siphoviridae* (35.4%), *Myoviridae* (35.4%) and *Podoviridae* (24.3%) most abundant (Figure 3.1C). In addition, members of ssDNA bacteriophages from *Microviridae* (1.3%) were also identified. The most prevalent phage host species for the assigned virus sequences were *Enterobacteria* spp., *Pseudomonas* spp., *Burkholderia* spp. and others (Figure 3.2D). The most 5 abundant viral species were Iodobacteriophage phiPLPE, *Escherichia* phage phiV10, *Salmonella* phage SETP3, uncultured myovirus, and *Burkholderia* phage phiE12-2 (see table Table 3.3 for list of top 50 virus species detected). Amongst eukaryotic viruses, ssDNA viruses (2.2%) from *Parvoviridae* and *Circoviridae* family, and dsDNA viruses (1.2%) from *Herpesviridae*, *Iridoviridae* and *Phycodnaviridae,* were found. A small portion (1%) of the reads was classified to the Archaea domain (Figure 3.2A). A further split into phylum (data not shown) showed that *Euryarchaeota* (97%) dominated amongst Archaea, with archeons within the class *Methanomicrobia* (72%) *Halobacteria* (13%) and *Methanococci* (10%) most abundant.

Reads classification

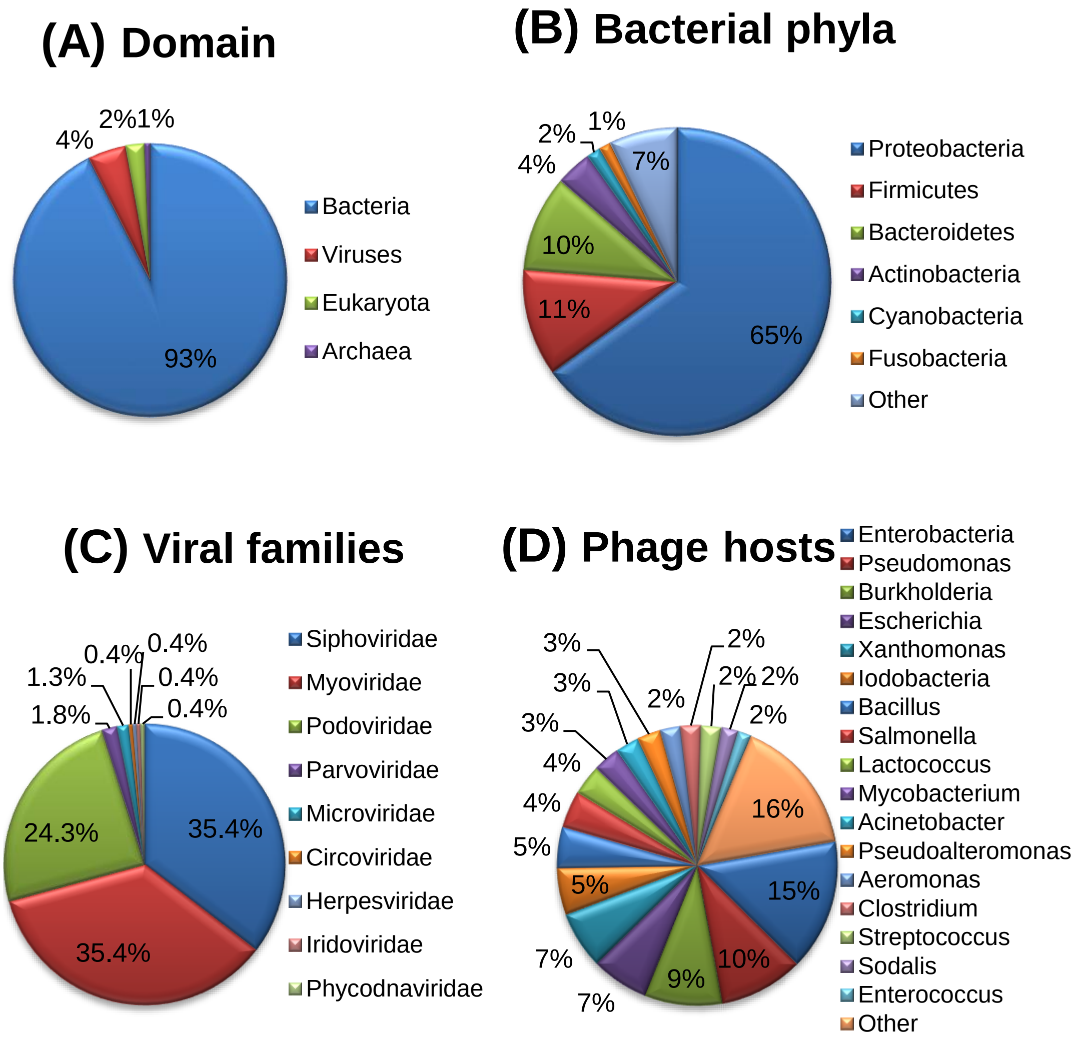


Figure 3. 2. Taxonomic affiliation of the assigned (6223) viral metagenomic reads from the activated sludge to the GenBank protein database based on the MG-RAST analysis (e-value < 1e-03). (A) Distribution of the assigned reads to the main taxonomic groups. (B) The proportion of the most abundant bacterial phyla. (C) The proportion of the most abundant viral families. (D) Phage host distributon.

Table 3. 1. Summary of best BLASTX (e-value >1e-05) hits from the 58 largest (>10kb) contigs assembled from the activated sludge viral metagenome (sorted by bit score). Hits with BLASTX similarity >50% are shown in bold.

Contig	Contig size (bp)	Phage/Organism containing prophage	Best BLAST hit	Accession number	E-value	% identity	Hit length (aa)
contig0	114,109	*Pelagibaca bermudensis*	Ribonucleoside-diphosphate reductase	ZP_01444584	5e-97	40	550
contig1	56,729	*Acinetobacter brisouii*	Chaperonin GroEL	WP_004900257	7e-79	35	533
contig2	**53,095**	***Chthoniobacter flavus***	**TROVE domain protein**	**ZP_03130643**	**2e-90**	**52**	**344**
contig3	46,440	*Vibrio* phage VP16C	DEAD-like helicase	AAQ96576	3e-75	35	590
contig4	44,123	*Mycobacterium* phage PLot	DNA polymerase III	YP_655445	3e-144	37	878
contig5	40,448	*Desulfotomaculum ruminis*	Hypothetical protein	AEG59378	3e-103	32	895
contig6	38,311	*Paenibacillus polymyxa*	Phage terminase large subunit	CCC85706	3e-64	34	470
contig7	**38,095**	***Pedobacter agri***	**Site-specific DNA methylase**	**WP_010603268**	**0.0**	**51**	**604**
contig8	35,905	*Lachnospiraceae bacterium*	Phage terminase	ZP_08334821	1e-96	40	508
contig9	31,782	*Thermocrinis albus*	Chaperonin GroEL	YP_003474201	1e-94	40	532
contig10	29,580	*Roseobacter* sp.	S-adenosylmethionine-dependent methyltransferase	ZP_01903545	0.0	34	1657
contig11	27,700	*Clostridium thermocellum*	ParB-like nuclease	YP_001038054	2e-81	43	400
contig12	**26,127**	***Nitrosomonas eutropha***	**Phage terminase**	**YP_747683**	**0.0**	**71**	**458**
contig13	24,156	*Bacillus cereus*	GIY-YIG endonuclease	YP_085011	2e-17	41	141
contig14	**23,737**	***Pseudomonas* sp. GM67**	**P22 coat protein**	**WP_008036983**	**8e-177**	**73**	**376**
contig15	22,887	*Bradyrhizobium* sp. ORS 285	Hypothetical protein	ZP_09477388	1e-83	38	482
contig16	**22,766**	***Oxalobacter formigenes***	**Hypothetical protein**	**ZP_04576510**	**9e-69**	**58**	**227**
contig17	22,220	*Roseobacter* sp.	S-adenosylmethionine-dependent methyltransferase	ZP_01903545	0.0	34	1615
contig18	21,443	Uncultured bacterium from groundwater metagenome	Cell division protein FtsK	EKD92820	2e-89	48	348
contig19	21,017	*Robiginitalea biformata*	Hypothetical protein	YP_003196495	2e-61	27	544
contig20	18,509	*Bilophila wadsworthia*	Phage terminase	ZP_07944391	2e-92	41	426
contig21	17,762	*Wolbachia* phage WOVitA1	Phage terminase	ADW80145	1e-142	45	624
contig22	17,163	*Bacillus cereus*	Chromosome segregation ATPase Smc	NP_976737	5e-85	37	582

Contig	Contig size (bp)	Phage/Organism containing prophage	Best BLAST hit	Accession number	E-value	% identity	Hit length (aa)
contig23	**17,064**	***Photobacterium profundum***	**Phosphoadenosine phosphosulfate reductase**	**YP_129561**	**5e-105**	**50**	**390**
contig24	**16,911**	***Thermoanaerobacterium saccharolyticum***	**Phage tail tape measure protein**	**YP_006393013**	**1e-37**	**59**	**153**
contig25	16,629	*Deftia* phage phiW-14	DNA helicase	YP_003358924	4e-26	26	516
contig26	16,443	*Burkholderia pseudomallei*	Bacteriophage replication protein A	ZP_02495847	7e-37	30	344
contig27	**16,294**	***Variovorax* sp. HH01**	**Hypothetical protein**	**AER23951**	**2e-77**	**69**	**193**
contig28	16,292	*Bacillus halodurans*	S-adenosylmethionine-dependent methyltransferase	NP_244402	4e-73	39	467
contig29	15,951	*Desulfatibacillum alkenivorans*	Hypothetical protein	YP_002433692	1e-73	27	913
contig30	15,901	*Polysphondylium pallidum*	IPT domain of Plexins and Cell Surface Receptors	EFA78206	4e-10	23	549
contig31	15,107	*Aquifex aeolicus*	Chaperonin GroEL	NP_214512	1e-94	39	529
contig32	14,451	*Denitrovibrio acetiphilus*	Phage terminase	YP_003504938	2e-79	37	459
contig33	14,194	*Sinorhizobium meliloti*	Stage 0 sporulation protein J	YP_004549199	5e-92	45	398
contig34	14,129	*Capnocytophaga sp.*	Phage terminase	ZP_08202667	1e-36	30	396
contig35	14,005	*Brachyspira intermedia*	Phage terminase large subunit	AEM22154	6e-50	28	651
contig36	**13,789**	***Niastella koreensis***	**Replicative DNA helicase**	**YP_005011225**	**5e-121**	**50**	**529**
contig37	13,475	*Cellulophaga* phage phiSM	Phage tail protein	AGH07768	2e-130	48	487
contig38	**13,283**	***Oxalobacter formigenes***	**Phage tail protein**	**ZP_04578334**	**0.0**	**50**	**867**
contig39	13,076	*Streptococcus mitis*	C-5 cytosine-specific DNA methylase	CBJ21706	4e-71	45	341
contig40	12,715	*Mucilaginibacter paludis*	Hypothetical protein	ZP_07745716	2e-106	44	505
contig41	12,295	*Marinobacter* sp.	Superfamily II DNA/RNA helicase	ZP_01738782	2e-44	28	877
contig42	11,725	*Aurantimonas manganoxydans*	Hypothetical protein	ZP_01226938	3e-76	35	579
contig43	11,663	*Sphingobium yanoikuyae*	Type III restriction protein, res subunit	ZP_09908736	1e-97	44	469
contig44	**11,475**	***Acinetobacter baumannii***	**Phage terminase**	**WP_001136863**	**7e-171**	**58**	**452**
contig45	**11,384**	***Flavobacterium columnare***	**AdoMet-dependent methyltransferase**	**YP_004941706**	**1e-98**	**76**	**201**
contig46	**11,274**	***Bacteroides vulgatus***	**HNH endonuclease**	**WP_005852289**	**1e-19**	**51**	**91**

Contig	**Contig size (bp)**	**Phage/Organism containing prophage**	**Best BLAST hit**	**Accession number**	**E-value**	**% identity**	**Hit length (aa)**
contig47	11,229	*Pirellula staleyi*	Phage terminase	YP_003369204	1e-54	34	498
contig48	11,027	*Methyloversatilis universalis*	Coat protein	ZP_08504648	2e-60	38	383
contig49	10,940	*Filifactor alocis*	Chaperonin GroEL	YP_005055286	2e-44	29	540
contig50	**10,820**	***Flavobacterium* sp. CF136**	**Recombinational DNA repair protein (RecT)**	**WP_007804118**	**9e-86**	**52**	**302**
contig51	10,740	*Flavobacterium* sp. CF136	Hypothetical protein	WP_007804358	2e-38	43	187
contig52	**10,684**	***Thauera* sp. 27**	**PBSX family phage terminase large subunit**	**WP_002937346**	**0.0**	**72**	**459**
contig53	10,498	*Arcticibacter svalbardensis*	Metallophosphoesterase	WP_016197049	6e-47	43	209
contig54	**10,221**	***Nitrosomonas eutropha***	**Phage terminase**	**YP_747683**	**9e-179**	**70**	**557**
contig55	10,056	*Rhodococcus erythropolis*	Phage terminase	YP_002765956	1e-73	38	402
contig56	10,054	*Bordetella pertussis*	Hypothetical protein	NP_881912	5e-69	32	472
contig57	10,027	Delta proteobacterium NaphS2	Integron integrase	ZP_07198461	3e-62	41	326

3.3.3. Contig assembly and taxonomic classification

Reads were assembled into 20,278 contigs (>100 bp), with N50 length of 1.5 kb and the total contig length of 10.4 Mbp. 47% of the reads, as determined by read mapping back onto contigs, were recruited into contigs of length greater than 100 bp. The longest contig was 114 kb, 58 contigs were longer than 10 kb, 184 contigs were longer than 5 kb, and 1876 contigs were longer than 1 kb. Coding DNA sequence (CDS) prediction using MetaGeneMark resulted in prediction of a total of 27,363 CDS in contigs larger than 100 bp.

Contigs with minimum length of 100 bp were classified based on the ‘best’ BLASTX homology (i.e. hits with the highest BLAST bit score) to the NCBI non-redundant (nr) protein database implemented in CAMERA server. Table 3.1 show the summary of the BLAST results for contigs longer than 10 kb. Some contig CDS showed >70% (but no more than 86%) similarity at amino acid level with bacterial or phage protein sequences e.g. *Nitrosomonas eutropha* (contig 12, contig 54), *Pseudomonas* sp. GM67 isolated from plant roots (contig 14), *Flavobacterium columnare* (contig 45*),* *Thauera* sp. (contig 52), *Dechloromonas aromatica* (contig 95), *Burkholderia cenocepacia* phage Bcep1 (contig 165), *Nitrobacter winogradskyi* (contig 221), *Novosphingobium aromaticivorans* (contig 525), *Achromobacter piechaudii* (contig 780), *Legionella pneumophila* (contig 845), *Stenotrophomonas* sp. SKA14 (contig 901). The majority of predicted genes exhibited less than 70% similarity at amino acid level, indicating that the viral metagenome was mostly novel.

BLASTX results with an e-value of <1e-05 and bit score >100 were analyzed using MEGAN software. Because contig size distribution varied from 100 bp to 114 kb, similarities were calculated based on the count of the total number of bases assigned to each taxon, rather than counting the number of contigs. The majority (65% based on the total base count) of the assembled contigs showed no sequence homology to any known sequences in the database. Among known sequences (1557 contigs, total of 3,676,618 bp), 15% (281 contigs) were of viral origin, and the remaining sequences (85%) were classified as cellular organisms (bacteria, archaea and eukaryotes) (Figure 3.2A).

Contigs classified as bacteria were further split into phyla and, similarly to the reads classification, the dominant bacterial phylum was *Proteobacteria*, with *Bacteroidetes* and *Firmicutes* as the two other leading phyla. The minority phyla differed between the read and contig classifications (Figure 3.2B). A large portion of the assigned contigs, which was entirely composed of contig 0, was originally assigned to marine bacterium *Pelagibaca bermudensis* (Table 3.2), however further analysis (described in section 3.3.5) suggests that this contig represents novel *Bacillus*-like phage from the *Myoviridae* family. In general, many bacterial species identified (Table 3.2) are known to be involved in metabolic processes observed in sludge, such as nitrification (*Nitrosomonas eutropha*, *Nitrobacter winogradskyi*), nitrogen fixation (*Bradyrhizobium* sp.*, Paenibacillus polymyxa*), hydrocarbon oxidation (*Methyloversatilis universalis, Desulfatibacillum alkenivorans, Dechloromonas aromatica*) and sulphate reduction (*Desulfatibacillum alkenivorans, Desulfotomaculum ruminis*). Some sequences were similar to pathogenic species associated with diseases in humans and/or animals e.g. *Flavobacterium columnare, Legionella pneumophila*, *Achromobacter piechaudii* and *Acinetobacter baumannii* (Table 3.2).

Contigs classified as viral sequences split into 8 viral families (Figure 3.3C). Bacterial viruses from *Siphoviridae* (26.1%), *Myoviridae* (23.9%), *Podoviridae* (14.3%) and *Microviridae* (13.5%) comprised 78% of these viral sequences (Figure 3.3C). Markedly more *Microviridae* and *Circoviridae* were assigned than in the read-based classification. The dominant bacteriophage host genera were *Vibrio* (13%), *Mycobacterium* (12%)*, Synechococcus* (10%), *Pseudomonas* (9%) and *Burkholderia* (5%) (Figure 3.3D). The top 50 viral species found in the activated sludge metagenome are presented in Table 3.3. In general, the most abundant virus species assignment was not in agreement with the reads classification, with a *Vibrio parahaemolyticus* phage phi16, uncultured microphage and *Mycobacterium smegmatis* phage PLot most abundant compared with Iodobacteriophage phiPLPE, *Escherichia* phage phiV10, *Salmonella* phage SETP3, and uncultured microphage in reads. Among eukaryotic viruses, 9.2% of the ‘viral’ contigs were related to ssDNA circular viruses (*Geminiviridae*, *Nanoviridae* and *Circoviridae*) that infect plants and animals. A small portion (0.6%) of the contigs was classified to *Phycodnaviridae*, dsDNA viruses infecting eukaryotic algae. The remaining viral contigs (13%) were uclassified. These unclassified sequences included, among others, hits to unclassified 'Circo-like' viruses; ssDNA virus that infects fungus *Sclerotinia sclerotiorum*; ssRNA virus that infects the fungus *Sclerophthora macrospora*; and phage TJE1 infecting *Tetrasphaera* - bacteria involved in phosphorus removal in the activated sludge.

As for the read-based classification approximately 1% of the assigned contigs (0.9%) were assigned to the domain Archaea, (14 contigs, totalling 33089 bp) (Figure 3.3A) and affiliated with the phylum *Euryarchaeota*, particularly with the classes *Methanomicrobia* (56%) and *Halobacteria* (26%) (data not shown). Members of *Methanomicrobia* include the methanogens, with produce methane and are abundant in sludge e.g. (Narihiro *et al.*, 2009).

3.3.4. Identification of novel ssDNA circular viruses

Single-stranded DNA viruses with a small circular genome (1-6 kb) (81 contigs totalling 136,097 bases) accounted for 23.3% of the contigs assigned to viruses or 3.7% of the total assigned contigs (Figure 3.4). Fifteen out of 81 contigs contained complete circular sequence and these were further investigated.

3.3.4.1. *Circoviridae, Nanoviridae* and *Geminiviridae* families

BLASTP analysis of the CDS predicted on these genomes revealed that 11 contigs with circular sequence contained CDS with 30-50% amino acid identity to the replication-associated protein (Rep) of animal viruses from *Circoviridae*, and plant pathogens from the *Nanoviridae* and *Geminiviridae*. A phylogenetic tree of the eleven complete Rep protein sequences with nearest neighbours by BLAST demonstrated that assembled ssDNA viruses represent novel members within these families (Figure 3.5). Four wastewater Rep sequences clustered with different circovirus-like genomes derived from sewage (Blinkova *et al.*, 2009) and mammal faeces (Ge *et al.*, 2011). One Rep sequence clustered with an algal nanovirus. The long tree branch lengths supported by low bootstrap values indicates large genetic distances between these sequences.

Contigs classification

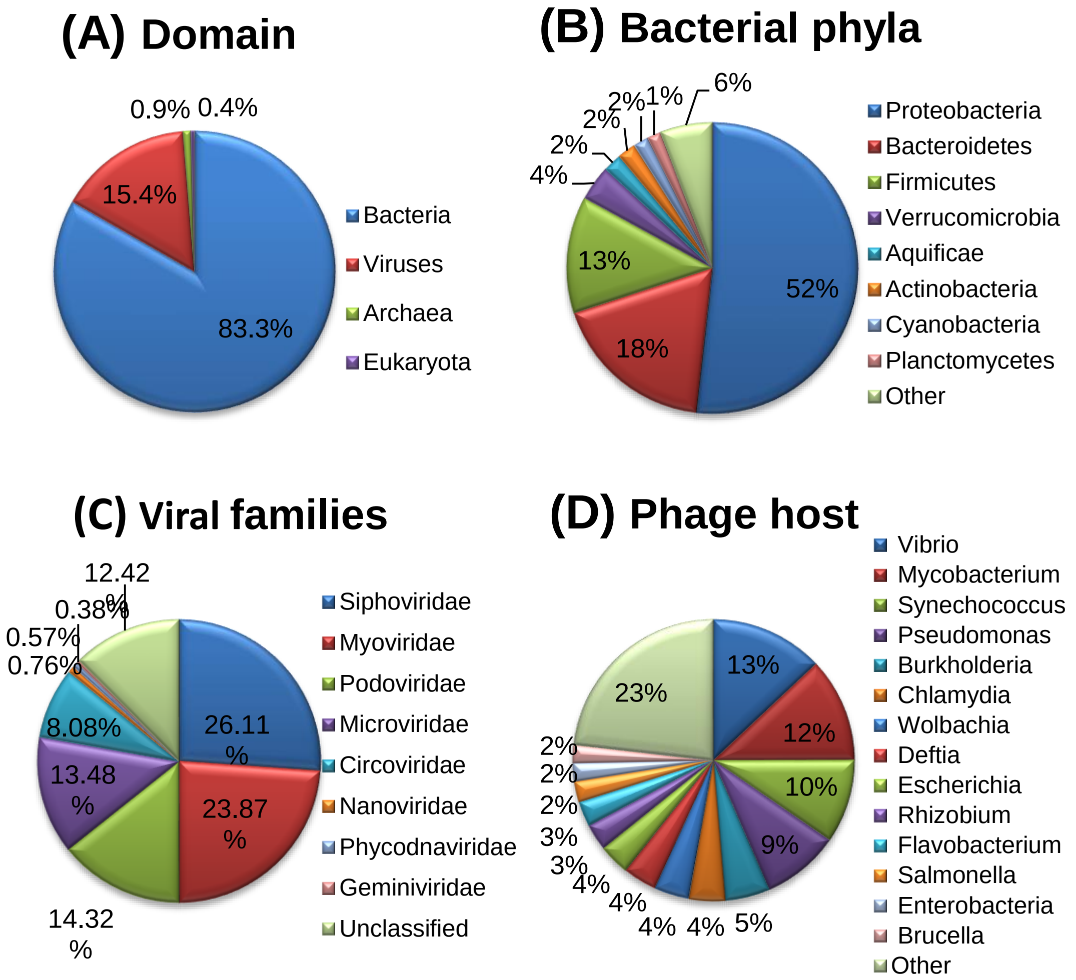


Figure 3. 3. Taxonomic affiliation of the assigned contigs (1557) assembled from the activated sludge to the GenBank protein database based on the BLASTX analysis (e-value < 1e-05). Classification was based on counting number of bases in contigs assigned to each taxon. (A) Distribution of the assigned contigs to the main taxonomic groups. (B) The proportion of the most abundant bacterial phyla. (C) The proportion of the most abundant viral families. (D) Phage host distribution.

Table 3. 2. Top 50 uncharacterized prophage species present in the activated sludge viral metagenome based on MG-RAST BLAT (Reads) and MEGAN BLASTX (Contigs) assignment. Information about habitat was retrieved from NCBI Genome project organism overview http://www.ncbi.nlm.nih.gov/genome. Aquatic (A), animal associated (AN), human associated (H), plant associated (P), marine (M), sediment (SD), **sludge (SL)**, soil (S).

Bacteria species (habitat)	Reads count	Bacteria species (habitat)	Contigs (total bases)
Escherichia coli (A, H, AN)	163	~~*Pelagibaca bermudensis* (M)~~ (reclassified as *Myoviridae* phage)	114109
Pseudomonas putida (S, A)	133	*Acinetobacter brisouii* (A)	56729
***Salmonella enterica* (H, AN, SL)**	104	*Chthoniobacter flavus* (S)	54351
Neisseria gonorrhoeae (H)	102	*Roseobacter* sp. (M)	53060
Neisseria meningitides (H)	88	*Bradyrhizobium* sp. (P)	46994
Haemophilus influenzae (H, AN)	86	***Nitrosomonas eutropha* (SL)**	43486
Achromobacter piechaudii (H, S)	80	*Desulfotomaculum ruminis* (AN)	40448
Bacteroides sp. (H, AN)	75	*Paenibacillus polymyxa* (S)	38311
***Pelobacter propionicus* (M, A, SL)**	74	*Pedobacter agri* (S)	38095
Pseudomonas fluorescens (S, A, P)	72	*Acinetobacter baumannii* (H, A)	37205
Pseudomonas aeruginosa (S, A, H)	64	*Oxalobacter formigenes* (H, AN)	36707
***Campylobacter jejuni* (H, AN, SL)**	54	*Verrucomicrobium spinosum* (A, S, H)	36099
Rhodopseudomonas palustris (S, A)	54	*Pseudomonas* sp. (P)	36009
Nitrosococcus halophilus (A)	50	*Clostridium sporogenes* (S)	35905
***Comamonas testosteroni* (SL)**	44	*Desulfatibacillum alkenivorans* (SD)	31897
Pseudomonas syringae (P)	44	*Thermocrinis albus* (A)	31782
Actinobacillus pleuropneumoniae (AN)	43	*Clostridium thermocellum* (S, A, H)	31119
Pseudomonas mendocina (S, A, H)	43	*Burkholderia pseudomallei* (S, H, AN)	29769
Acidovorax delafieldii (S, A)	36	***Comamonas testosteroni* (SL)**	26598
Enterococcus faecalis (H, AN)	35	***Pirellula staleyi* (M, SL)**	26010
Vibrio cholera (M, A, H)	35	*Bilophila wadsworthia* (H)	25412
Acinetobacter baumannii (A, H)	34	*Dysgonomonas mossii* (H)	24984
Enterobacter sp. 638 (P)	33	*Robiginitalea biformata* (M)	24483
Flavobacterium johnsoniae (S, A)	32	*Myroides odoratus* (AN)	24116
Brucella pinnipedialis (AN)	31	*Chitinophaga pinensis* (P)	23470
Cytophaga hutchinsonii (S, A, M)	28	*Denitrovibrio acetiphilus* (A)	22173
Fusobacterium sp. (H, AN)	28	*Flavobacterium* sp. (A)	21560
Novosphingobium aromaticivorans (A)	28	Uncultured bacterium (metagenome) (A)	21443
Shewanella sp. (M)	28	*Escherichia coli* (A) (H)	20018
Geobacter lovleyi (SD)	27	*Niastella koreensis* (S)	19918
Pseudomonas savastanoi (P)	26	*Mucilaginibacter paludis* (P)	19416
Fusobacterium mortiferum (H)	25	*Methyloversatilis universalis* (SD)	18576
***Delftia acidovorans* (S, SD, SL, A)**	24	*Clostridium* sp. BNL1100	18328
Edwardsiella tarda (H, AN)	24	***Dechloromonas aromatica* (A, S, SL)**	**17448**
Laribacter hongkongensis (AN)	24	*Pseudomonas aeruginosa* (S, A, H)	17048
***Clostridium perfringens* (S, SL, AN)**	23	*Photobacterium profundum* (M)	17064
Xylella fastidiosa (P)	23	*T. saccharolyticum* (A)	16911
Acidovorax citrulli (P)	22	*Legionella pneumophila* (A, S, H)	16475
Arcobacter butzleri (A, AN, H)	22	*Variovorax* sp. HH01 (S)	16294
Burkholderia pseudomallei (S,H,AN)	22	*Bacillus halodurans* (S)	16292
Cyanothece sp. (M)	22	*Flavobacterium columnare* (A)	16174
Helicobacter pylori (H)	22	*Dysgonomonas gadei* (H)	15900
Chitinophaga pinensis (P)	21	*Sinorhizobium meliloti* (P)	15681
Paenibacillus polymyxa (S)	21	*Capnocytophaga* sp. (H)	15675
Rhizobium etli (P)	21	*Alishewanella jeotgali*	15107
Saccharophagus degradans (M)	21	*Aquifex aeolicus* (A)	15107

Asticcacaulis excentricus (A)	20	*Brachyspira intermedia* (AN)	14754
Bacteroides vulgates (H, AN)	20	*Bacteroides clarus* (H, AN)	14664
***Geobacter sulfurreducens* (A, SL)**	20	*Cyclobacterium marinum* (M)	14654
***Rhodobacter sphaeroides* (S, SL, A)**	20	***Gemmata obscuriglobus* (A, SL)**	14487

Table 3.3. Top 50 virus species present in the activated sludge viral metagenome based on MG-RAST BLAT (Reads) and BLASTX (Contigs) assignment.

Virus species	**Reads count**	**Virus species**	**Contigs (total bases)**
Iodobacteriophage phiPLPE	14	*Vibrio* phage phi16	48575
Escherichia phage phiV10	13	Uncultured *Microviridae*	47094
Salmonella phage SETP3	10	*Mycobacterium* phage PLot	44123
uncultured *Myoviridae*	10	Bat circovirus ZS/Yunnan-China/2009	22112
Burkholderia phage phiE12-2	9	*Synechococcus* phage S-CBS3	19524
Bacillus phage phi29	8	*Wolbachia* endosymbiont wVitA of *Nasonia vitripennis* phage WOVitA1	17762
Burkholderia phage BcepIL02	8	*Deftia* phage phiW-14	17160
Xanthomonas phage Xp15	8	*Rhizobium* phage 16-3	13585
Pseudoalteromonas phage H105/1	7	*Flavobacterium* phage 11b	12739
Pseudomonas phage D3	7	*Escherichia* phage vB_EcoM_ECO1230-10	12070
Xanthomonas phage phiL7	7	*Pseudomonas* phage PAK_P1	11294
Enterobacteria phage N4	5	EBPR siphovirus 2	10630
Enterobacteria phage P2	5	EBPR podovirus 3	10429
Pseudomonas phage PAK_P1	5	*Brucella* phage Pr	9150
Sodalis phage phiSG1	5	*Ircinia* phage phiJL001	8925
Aeromonas phage phiAS5	4	*Salmonella* phage epsilon15	8886
Clostridium phage phiCD27	4	*Chlamydia* phage CPAR39	8669
Enterobacteria phage RB49	4	*Synechococcus* phage S-CBS1	8359
Enterobacteria phage T7	4	*Pseudomonas* phage PA11	8315
Lactococcus phage 936	4	EBPR podovirus 1	6612
Pseudomonas phage B3	4	uncultured phage MedDCMOCTS04C1161	6558
Pseudomonas phage F116	4	*Bdellovibrio* phage phiMH2K	6193
Acinetobacter phage 133	3	Rodent stool-associated circular virus	6150
Bacteroides phage B40-8	3	*Synechococcus* phage S-CBS4	5912
Enterobacteria phage T4	3	*Enterobacteria* phage Phieco32	5716
Enterobacteria phage T5	3	*Burkholderia* phage Bcep1	5543
Enterococcus phage phiFL3A	3	Porcine circovirus 1	5309
Lactococcus phage ul36	3	*Pseudomonas* phage PaP2	5212
Mycobacterium phage Corndog	3	*Tetrasphaera* phage TJE1	5168
Acinetobacter phage AB1	2	*Mycobacterium* phage Gladiator	5102
Acinetobacter phage Ac42	2	Circovirus-like genome CB-A	4399
Aeromonas phage 65	2	*Myxococcus* phage Mx8	4384
Bacillus phage BCJA1c	2	*Picobiliphyte* sp. MS584-5 nanovirus	4231
Brochothrix phage NF5	2	*Synechococcus* phage S-SM2	4198
Burkholderia phage Bcep176	2	*Burkholderia* phage Bcep781	4164
Clostridium phage phiC2	2	*Burkholderia* phage phi1026b	4093
Cronobacter phage ESSI-2	2	Mosquito VEM virus SDRBAJ	4080
Enterobacteria phage phiEcoM-GJ1	2	*Yersinia* phage PY100	3937
Enterobacteria phage WA13	2	*Chlamydia* phage phiCPG1	3919
Enterobacteria phage YYZ-2008	2	*Escherichia* phage phiKT	3715

Escherichia phage K1ind3	2	*Pseudomonas* phage 119X	3708
Feline panleukopenia virus	2	*Burkholderia* phage BcepNazgul	3592
Kluyvera phage Kvp1	2	*Bordetella* phage BPP-1	3578
Leptospira phage LE1	2	Enterobacteria phage alpha3	3564
Mycobacterium phage Omega	2	*Chlamydia* phage 3	3551
Pseudomonas phage PaP2	2	Circovirus-like genome RW-B	3375
Pseudomonas phage PaP3	2	Raven circovirus	3332
Rhizobium phage 16-3	2	Deep-sea thermophilic phage D6E	3283
Rhodococcus phage ReqiPepy6	2	*Sclerophthora macrospora* virus A	3245
Staphylococcus phage 53	2	Meles meles circovirus-like virus	3226

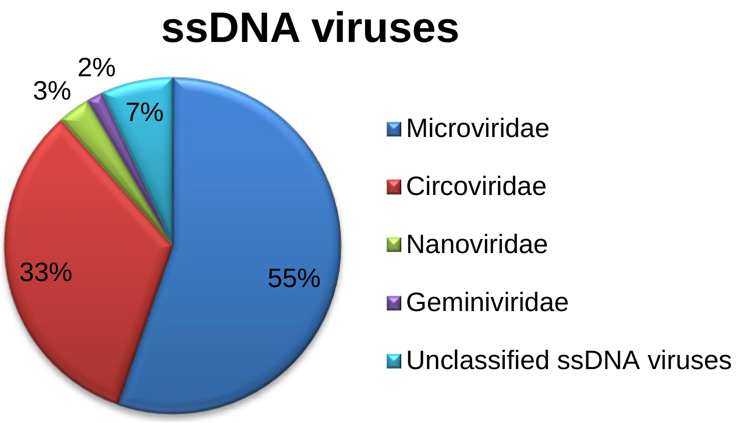


Figure 3.4. Taxonomic classification of contigs assembled from the activated sludge (81 contigs totalling 136,097 bases, which represent 3.7% of all assigned contigs) with similarity to ssDNA viruses of prokaryotes (*Microviridae*) and eukaryotes (*Circoviridae*, *Nanoviridae* and *Geminiviridae*). Contigs were compared to the GenBank protein database using BLASTX (e-value < 1e-05) and analysed using MEGAN. Classification was based on counting number of bases in contigs assigned to each family.

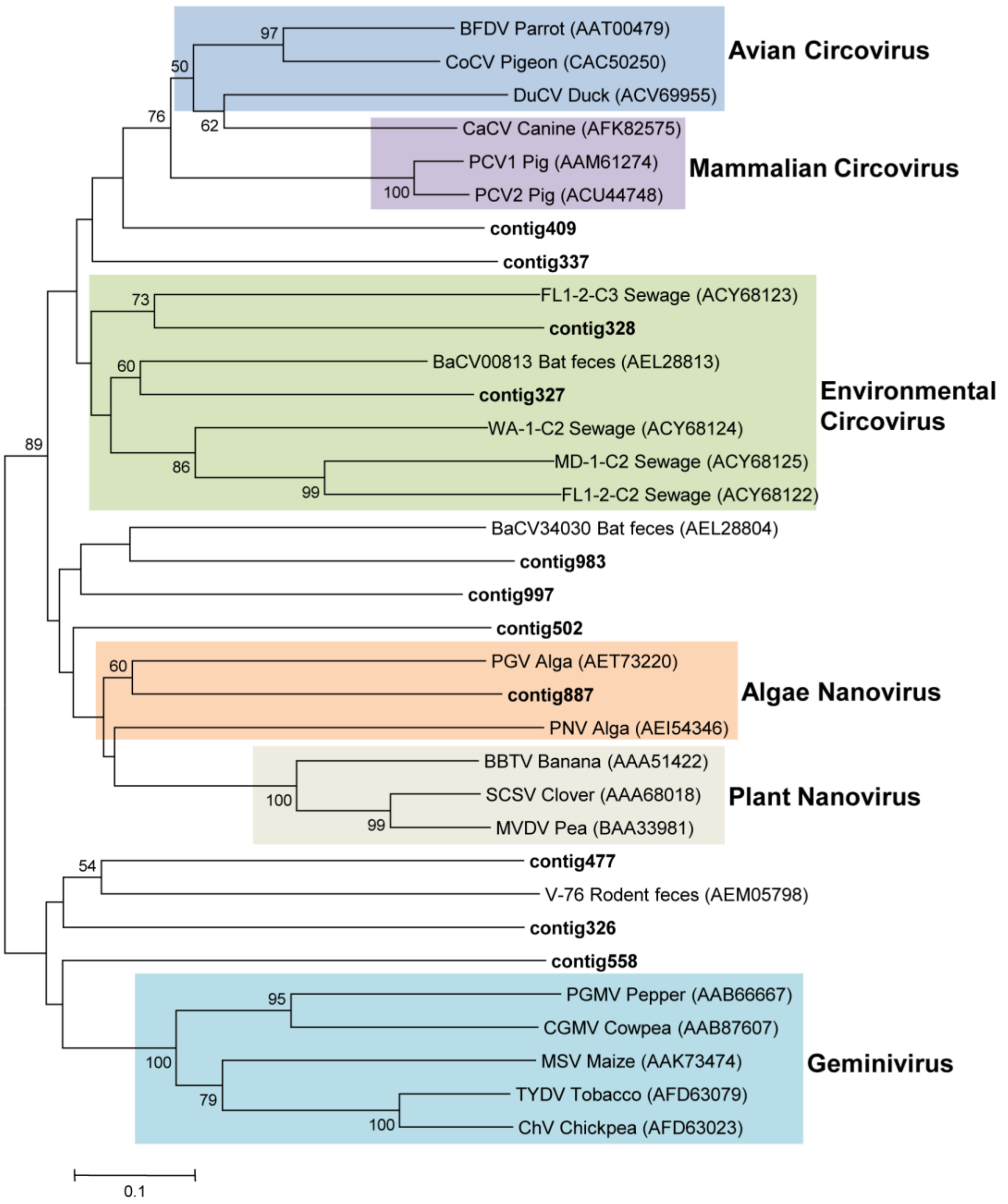


Figure 3.5. Neighbour-joining phylogenetic tree of the activated sludge ssDNA viruses (shown in bold) and known members of the ssDNA viruses from *Gemini*-, *Nano*-, and *Circoviridae* family based on the complete amino acid sequences of the Rep protein. Reference sequences were selected based on nearest BLASTP matches. GenBank accession numbers of the reference sequences are shown in parentheses. Only bootstrap values >50 are shown.

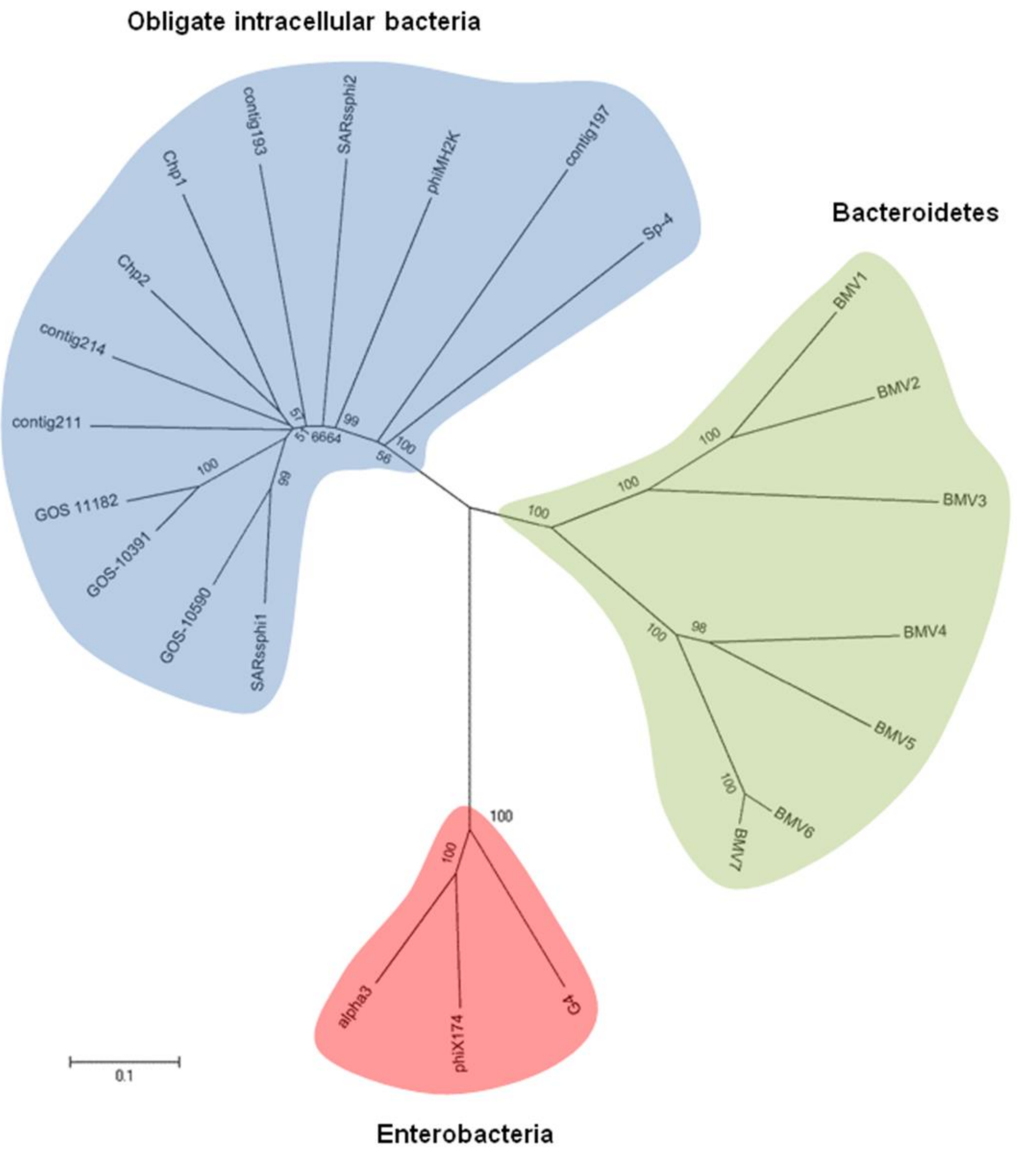


Figure 3.6. Unrooted neighbour-joining phylogenetic tree based on major capsid proteins identified from the activated sludge contigs and representative phages from the *Microviridae* family, grouped according to the host they infect. The reference sequences and their accession numbers used in the phylogenetic analysis are: Enterobacteria group – phiX174 (AAA32578), alpha3 (CAA42881), G4 (AAA32323); obligate intracellular bacteria group – *Chlamydia psittaci* Chp1 (BAA00515), *Chlamydophila abortus* Chp2 (CAB85589), uncultured ocean phage SARssphi1 (ADR80653), uncultured ocean phage SARssphi2 (ADR80650), *Bdellovibrio bacteriovorus* phage phiMH2K (AAG45340), *Spiroplasma* phage Sp-4 (AAA72621), uncultured marine phage GOS_10590 (ECU79385), uncultured marine phage GOS_10391 (ECU79568), uncultured marine phage GOS_11182 (ECU78785); *Bacteroidetes*: *Bacteroides sp*. BMV1 (EEO54684), *Bacteroides eggerthii* BMV2 (EEC52970), *Bacteroides plebeius* BMV3 (EDY96368), *Prevotella sp.* BMV4 (EFC69154), *Prevotella buccalis* BMV5 (EFA93043), *Prevotella bergensis* BMV6 (EFA43821), *Prevotella bergensis* BMV7 (EFA44667). The tree was drawn using MEGA software, with p-distance model and 1,000 bootstrap replications. Only bootstrap values >50 are shown.

3.3.4.2. *Microviridae* family

Four contigs containing circular sequence showed sequence similarities to members of the *Microviridae* family, bacteriophages with genomes of 4-6 kb infecting a wide range of hosts, including Enterobacteria, some *Bacteroidetes* and obligate parasitic bacteria such as *Chlamydiae*, *Bdellovibrio* and *Spiroplasma*. Phylogenetic analysis of the major capsid proteins revealed that these four sequences belong exclusively to a group of viruses infecting obligate intracellular parasitic bacteria, but are distinct from known members of this group (Figure 3.6). BLASTP search of the major capsid proteins encoded by these genomes revealed 41-57% sequence identity to the uncultured virus SARssphi1 isolated from the Atlantic ocean (Tucker *et al.*, 2011). The second BLAST hit was to *Chlamydiae* phages, with the highest protein sequence identity 57% to *Chlamydophila abortus* phage Chp4.

3.3.5. Sequence analysis of the largest assembled contig resembling *Myoviridae*

The largest contig assembled from the activated sludge viral metagenome was contig 0. Sequence analysis showed that it formed a closed circle, 114,060 bp long, suggesting the completion of this phage genome sequence. 157 Coding DNA Sequences (CDS) were predicted, of which 30 CDS had a putative function (Figure 3.7). Six of the predicted genes were found to encode for phage genes, including terminase, prohead protease, capsid, portal, tail sheath and baseplate proteins. The other CDS with a functional prediction belonged mostly to the phage structural genes, nucleotide metabolism, DNA replication and recombination genes.

The best BLASTX hit of this genome was a CDS of a prophage of the marine bacterium *Pelagibaca bermudensis* (showing 40% amino acid identity) (Table 3.1), however no other CDS from the contig matched with the same prophage. Sequence analysis of other CDS revealed that 7 predicted CDS matched with the proteins of *Bacillus subtilis* phage SP10 from the *Myoviridae* family (with identity ranging from 21 to 38%). With a similarity less than 38% at the protein level, these phages are only weakly related, suggesting that contig0 may represent a novel species within the *Myoviridae*.

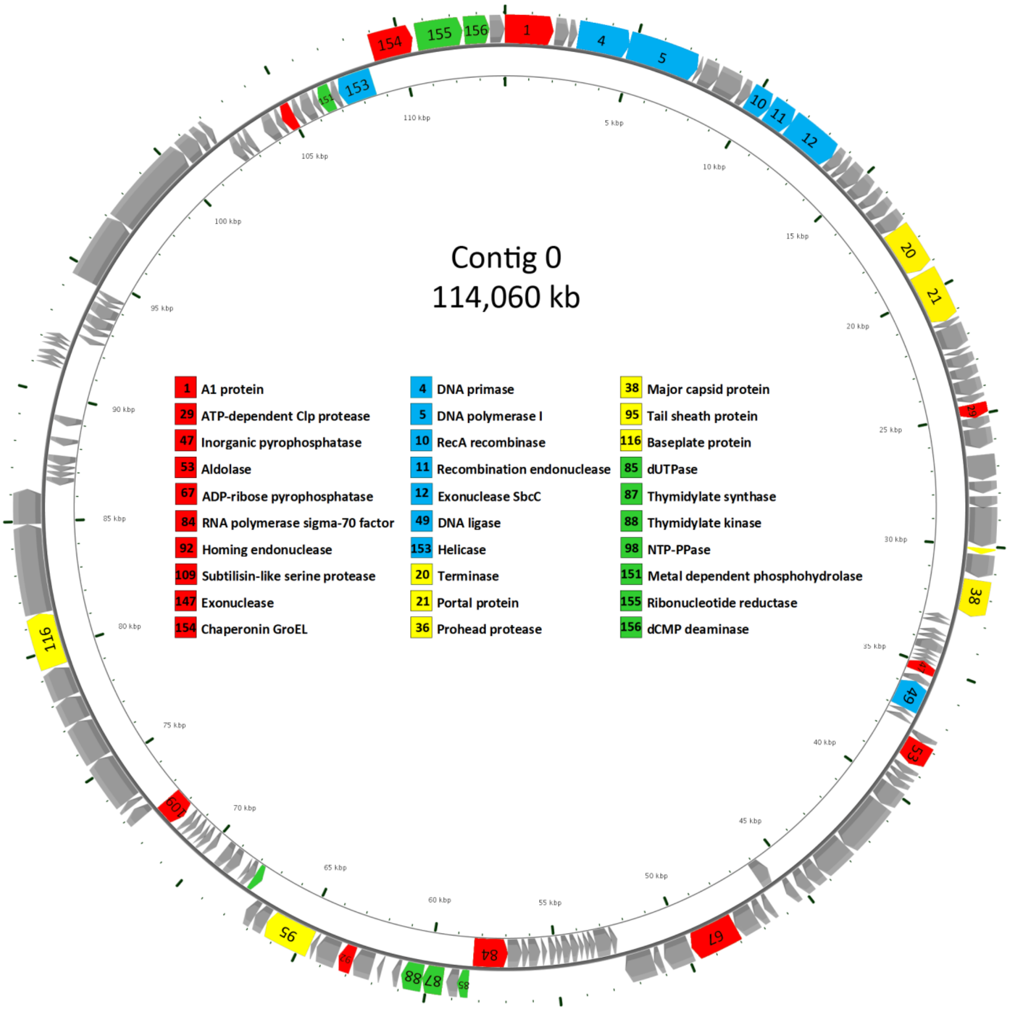


Figure 3.7. Genomic organization of the largest complete contig assembled from dairy wastewater. CDS (Coding DNA Sequence) were predicted using Genemark and annotated using NCBI protein database. CDS were categorized according to their predicted functions: DNA packing and structural proteins (yellow), replication and recombination (blue), nucleotide metabolism (green), other genes (red), and function unknown (grey). A total of 34200 reads (0.76%) were aligned to ‘contig 0’ with FR-HIT (at 100% identity and 100% read coverage). Figure was drawn in CG View and edited in Photoshop.

3.3.6. Diversity of viruses in activated sludge

Circonspect and PHACCS analyses were used to estimate the viral diversity in activated sludge sample. The viral metagenome showed slightly lower (409 species) but comparable degree of diversity when compared with another activated sludge viral community (511 species) that used a different sequencing methodology (454) (Table 3.4). The abundance of the most predominant viral genotype identified by PHACCS was 2.35%, strikingly less than the previous study.

Table 3.4. Diversity analysis of the activated sludge (AS) metagenomes.

Sample	**Richness**	**Most abundant genotype (%)**	**Evenness**	**Shannon Index**	**Model**
AS (This study)	409 genotypes	2.35	0.98	5.9	logarithmic
AS (Tamaki *et al.*, 2012)	511 genotypes	19.0	0.81	5.1	logarithmic

3.3.7. Functional analysis of the activated sludge viral genes

The putative proteins were assigned functional annotations using MG-RAST against the SEED Subsystems database and WebGMA against the COG (Clusters of Orthologous Groups of proteins) database (e-value cutoff of 1e-05 for both methods of annotation). MG-RAST functionally assigned 398 genes into 13 functional categories (SEED Level 1). The dominant subsystem was “Phages, prophages, transposable elements, and plasmids”, comprising 64% of total assigned sequences (Figure 3.8A, Table 3.5). The fact that most of the sequences fell into this category indicates that the metagenome was enriched for bacteriophage content. Indeed, this category contained mainly phage-related proteins such as terminase, phage capsid, portal protein and adenine-specific DNA methyltransferase. Other sequences with a predicted function belonged mostly to DNA metabolism, accounting for 19% of total assigned genes in the SEED database, including genes encoding for replicative DNA helicase, cytosine-specific DNA methyltransferase and single-stranded DNA-binding protein.

Annotation using the COG database showed that 4.5% (1248) of CDS had predicted functions, mainly associated with replication, recombination and repair (39%);

general function prediction (18%); and function unknown (12%) (Figure 3.8B, Table 3.6). Genes encoding DNA modification methylase and site-specific DNA methylase were particularly abundant in the activated sludge viral metagenome accounting for 12.5% (9% and 3.5% respectively) of total assigned CDS in the COG database. Other assigned genes were involved in replication and recombination (e.g. DNA helicases, DNA polymerase, site-specific recombinase XerD, DNA primase, ssDNA-binding protein), phage assembly (e.g. terminase), transcription (e.g. transcriptional regulator), stress response (e.g. groEL, nicotinamide mononucleotide adenylyltransferase), host lysis (e.g chitinase, lysozyme), virulence (cysteine protease C1 family) and immunity to infection (e.g. phage antirepressor protein).

Of particular interest were genes associated with bacterial metabolism genes. These included genes involved in sulphur metabolism such as 3'-phosphoadenosine 5'-phosphosulphate sulphotransferase and adenylylsulphate kinase, enzymes used by sulphate-reducing bacteria to assimilate sulphate, and genes involved in iron metabolism such as ferritin, an iron storage protein.

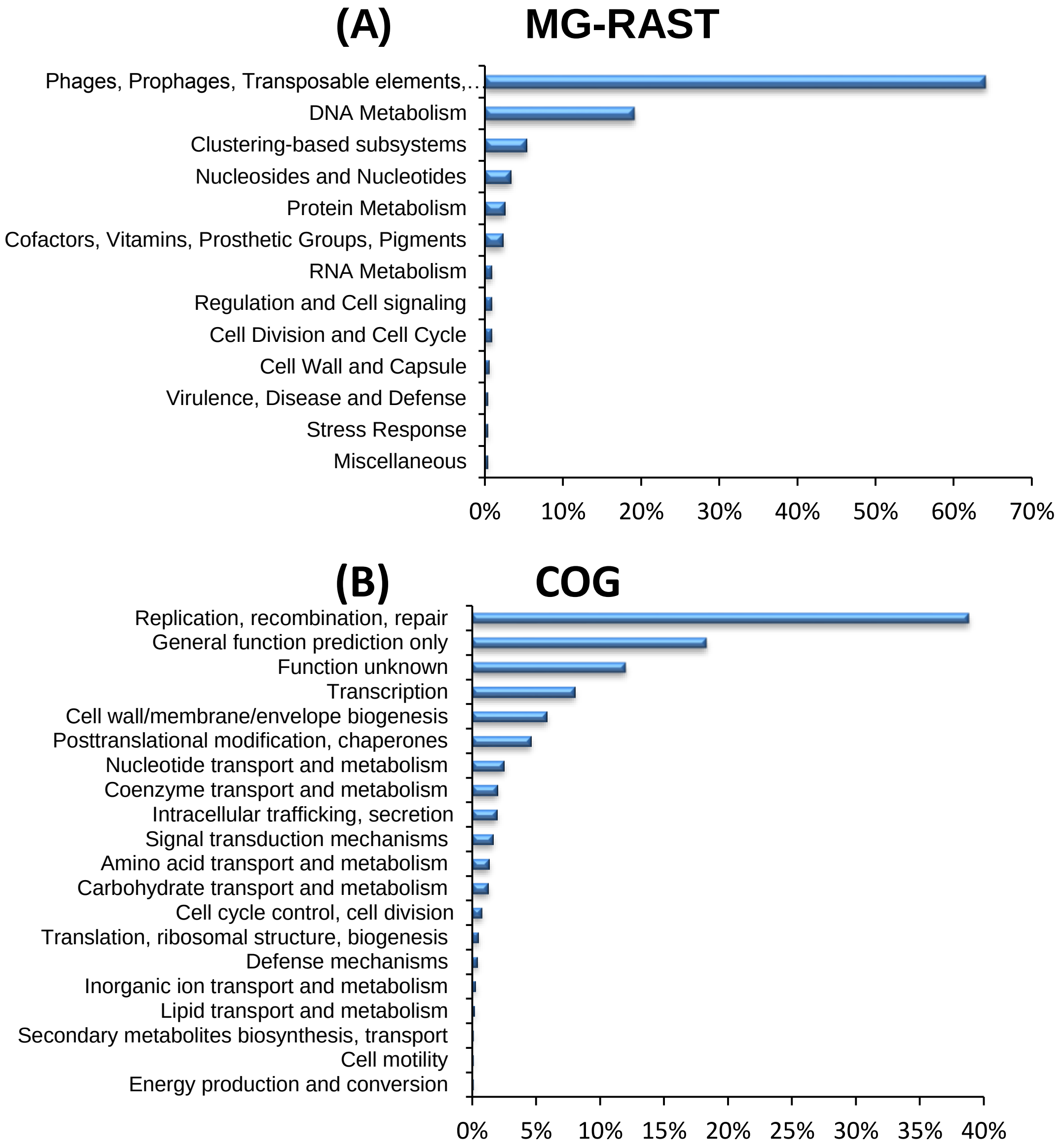


Figure 3.8**.** Functional annotation of contigs assembled from activated sludge viral metagenome to SEED-Subsystem using MG-RAST (A) and contigs CDS to COG database using WebMGA (B).

Table 3.5. Functional classification of the contigs assembled from the activated sludge viral metagenome to the SEED database, based on the MG-RAST analysis (e-value 1e-05).

Subsystem (Level 1)	Function	Abundance
Phages, Prophages	Phage terminase	109
Phages, Prophages	Phage capsid protein	83
DNA Metabolism	Replicative DNA helicase (EC 3.6.1.-)	26
DNA Metabolism	DNA-cytosine methyltransferase (EC 2.1.1.37)	20
Phages, Prophages	DNA-adenine methyltransferase, phage-associated	17
Clustering-based subsystems	Superfamily II DNA/RNA helicases, SNF2 family	16
Phages, Prophages	Portal protein	14
DNA Metabolism	Single-stranded DNA-binding protein	14
Phages, Prophages	DNA helicase, phage-associated	12
Protein Metabolism	Heat shock protein 60 family chaperone GroEL	10
DNA Metabolism	Recombinational DNA repair protein RecT	8
Nucleosides and Nucleotides	Deoxycytidine triphosphate deaminase (EC 3.5.4.13)	6
Phages, Prophages	Phage integrase	6
Cofactors, Vitamins, Pigments	Queuosine biosynthesis QueD, PTPS-I	6
Nucleosides and Nucleotides	Deoxyuridine 5'-triphosphate nucleotidohydrolase	4
Cell Division and Cell Cycle	Cell division protein FtsK	3
Clustering-based subsystems	DNA packaging	3
Cofactors, Vitamins, Pigments	GTP cyclohydrolase I (EC 3.5.4.16) type 1	2
RNA Metabolism	NADPH-dependent 7-cyano-7-deazaguanine reductase	2
Phages, Prophages	Phage NinB DNA recombination	2
Phages, Prophages	Phage NinC	2
Regulation and Cell signaling	Prophage Clp protease-like protein	2
DNA Metabolism	RecA protein	2
Nucleosides and Nucleotides	Ribonucleotide reductase of class Ia (aerobic), beta subunit	2
DNA Metabolism	Uracil-DNA glycosylase, family 4	2
DNA Metabolism	Chromosomal replication initiator protein DnaA	1
Stress Response	Cold shock protein CspA	1
DNA Metabolism	DNA polymerase III epsilon subunit (EC 2.7.7.7)	1
Clustering-based subsystems	DNA primase (EC 2.7.7.-), Phage P4-associated	1
Phages, Prophages	DNA replication protein, phage-associated	1
DNA Metabolism	DNA-binding protein HU	1
Virulence, Disease and Defense	Fibronectin-binding protein	1
Cell Wall and Capsule	GDP-mannose 4,6-dehydratase (EC 4.2.1.47)	1
Phages, Prophages	Gene Transfer Agent capsid protein	1
Regulation and Cell signaling	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	1
Phages, Prophages	Integron integrase IntI4	1
DNA Metabolism	Late competence protein ComEB	1
Cell Wall and Capsule	N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)	1
Cofactors, Vitamins, Pigments	Nicotinamide phosphoribosyltransferase (EC 2.4.2.12)	1
Miscellaneous	Peptidase, M23/M37 family	1
Phages, Prophages	Phage baseplate	1
Phages, Prophages	Phage EaA protein	1
Phages, Prophages	Phage exonuclease (EC 3.1.11.3)	1
Phages, Prophages	Phage NinG rap recombination	1

Phages, Prophages	Phage repressor	1
Phages, Prophages	Phage rIIB lysis inhibitor	1
Phages, Prophages	Protein gp47, recombination-related [Bacteriophage A118]	1
Nucleosides and Nucleotides	Ribonucleotide reductase of class Ia, alpha subunit	1
RNA Metabolism	tRNAHis-5'-guanylyltransferase	1
Clustering-based subsystems	unknown protein encoded within prophage CP-933V	1
Total		**398**

Table 3.6. Functional annotation of CDS predicted on contigs assembled from the activated sludge viral metagenome, based on the BLASTP analysis to the COG database.

Class description	**Annotation**	**ORF scount**
Replication, recombination and repair	DNA modification methylase	116
Function unknown	Uncharacterized proteins	96
Replication, recombination and repair	Replicative DNA helicase	69
Multiple classes	DNA or RNA helicases of superfamily II	63
General function prediction only	Phage terminase	58
Replication, recombination and repair	Site-specific DNA methylase	44
Replication, recombination and repair	DNA polymerase I	32
Function unknown	Phage-related protein	31
Replication, recombination and repair	Site-specific recombinase XerD	30
Replication, recombination and repair	DNA primase	25
General function prediction only	Predicted chitinase	24
General function prediction only	Phage-related lysozyme (muraminidase)	21
General function prediction only	Predicted ATPase	21
Replication, recombination and repair	Single-stranded DNA-binding protein	18
Multiple classes	Protease subunit of ATP-dependent Clp proteases	17
Replication, recombination and repair	Transposase and inactivated derivatives	16
Multiple classes	**3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)**	13
Cell wall/membrane/envelope biogenesis	Glycosyltransferases involved in cell wall biogenesis	12
Replication, recombination and repair	Recombinational DNA repair protein (RecE pathway)	12
Posttranslational modification, chaperones	Chaperonin GroEL (HSP60 family)	11
Function unknown	Bacteriophage protein gp37	11
General function prediction only	Bacteriophage capsid protein	11
Replication, recombination and repair	ATPase involved in DNA replication initiation	10
Cell wall/membrane/envelope biogenesis	Membrane proteins related to metalloendopeptidases	10
Multiple classes	Transcriptional activator, adenine-specific DNA methyltransferase	10
Replication, recombination and repair	Phage-related protein, predicted endonuclease	10
Transcription	Predicted transcriptional regulators	10
General function prediction only	Putative secretion activating protein	9
Carbohydrate transport and metabolism	Muramidase (phage lambda lysozyme)	9
Posttranslational modification, chaperones	Co-chaperonin GroES (HSP10)	8
Cell wall/membrane/envelope biogenesis	Glycosyltransferase	8
Function unknown	Uncharacterized homolog of phage Mu protein gp30	8
General function prediction only	Predicted phage phi-C31 gp36 major capsid-like protein	8

General function prediction only	Bacteriophage tail assembly protein	8
Replication, recombination and repair	ATPase involved in DNA repair	7
Replication, recombination and repair	RecA/RadA recombinase	7
Multiple classes	Periplasmic serine proteases (ClpP class)	7
Nucleotide transport and metabolism	Deoxycytidine deaminase	7
Replication, recombination and repair	DNA polymerase III, epsilon subunit	7
General function prediction only	Bacteriophage head-tail adaptor	7
Nucleotide transport and metabolism	Ribonucleotide reductase, alpha subunit	6
Coenzyme transport and metabolism	6-pyruvoyl-tetrahydropterin synthase	6
Replication, recombination and repair	Site-specific recombinases, DNA invertase Pin homologs	6
General function prediction only	Phage head maturation protease	6
Replication, recombination and repair	Holliday junction resolvase	6
Cell wall/membrane/envelope biogenesis	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	5
Multiple classes	Nucleoside-diphosphate-sugar epimerases	5
Nucleotide transport and metabolism	dUTPase	5
Cell wall/membrane/envelope biogenesis	Acyl-[acyl carrier protein]--UDP-N-acetylglucosamine O-acyltransferase	5
Cell wall/membrane/envelope biogenesis	dTDP-D-glucose 4,6-dehydratase	5
Nucleotide transport and metabolism	**Predicted alternative thymidylate synthase**	5
Function unknown	Uncharacterized homolog of phage Mu protein gp47	5
Replication, recombination and repair	RecA-family ATPase	5
General function prediction only	Predicted phosphoesterase or phosphohydrolase	5
General function prediction only	Mu-like prophage FluMu protein gp28	5
General function prediction only	Predicted P-loop ATPase and inactivated derivatives	5
Transcription	Transcriptional regulators	5
Posttranslational modification, chaperones	Membrane protease, stomatin/prohibitin homologs	4
Replication, recombination and repair	ATP-dependent exoDNAse (exonuclease V), alpha subunit - helicase superfamily I member	4
Cell cycle control, cell division, chromosome partitioning	Chromosome segregation ATPases	4
General function prediction only	Predicted glycosyltransferases	4
Replication, recombination and repair	ATP-dependent DNA ligase	4
General function prediction only	Predicted hydrolases of HD superfamily	4
General function prediction only	Predicted O-methyltransferase	4
Function unknown	Mu-like prophage protein gp29	4
Function unknown	Phage-related protein, tail component	4
Function unknown	Phage-related minor tail protein	4
Transcription	Prophage antirepressor	4
Coenzyme transport and metabolism	GTP cyclohydrolase I	3
Replication, recombination and repair	Superfamily I DNA and RNA helicases	3
Multiple classes	Guanosine polyphosphate pyrophosphohydrolases/synthetases	3
Replication, recombination and repair	Ribonuclease HI	3
Amino acid transport and metabolism	Asparagine synthase (glutamine-hydrolyzing)	3
Replication, recombination and repair	DNA polymerase elongation subunit (family B)	3
Cell wall/membrane/envelope biogenesis	Glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains	3
Transcription	DNA-directed RNA polymerase, sigma subunit	3
Posttranslational modification, chaperones	Organic radical activating enzymes	3
Replication, recombination and repair	Bacterial nucleoid DNA-binding protein	3
Cell wall/membrane/envelope biogenesis	ADP-heptose:LPS heptosyltransferase	3
Replication, recombination and repair	Uracil-DNA glycosylase	3

Cell cycle control, cell division, chromosome partitioning	DNA segregation ATPase FtsK/SpoIIIE and related proteins	3
Amino acid transport and metabolism	Lysophospholipase L1 and related esterases	3
Defense mechanisms	Negative regulator of beta-lactamase expression	3
General function prediction only	Mu-like prophage protein	3
General function prediction only	Mu-like prophage major head subunit gpT	3
General function prediction only	AAA ATPase containing von Willebrand factor type A	3
Cell wall/membrane/envelope biogenesis	Endopolygalacturonase	3
General function prediction only	Phage baseplate assembly protein	3
Nucleotide transport and metabolism	Adenylosuccinate synthase	2
Multiple classes	Glutathione synthase/Ribosomal protein S6 modification enzyme (glutaminyl transferase)	2
Nucleotide transport and metabolism	Ribonucleotide reductase, beta subunit	2
Posttranslational modification, chaperones	ATPases of the AAA+ class	2
Signal transduction mechanisms	RecA-family ATPases implicated in signal transduction	2
Coenzyme transport and metabolism	Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2	2
Inorganic ion transport and metabolism	**Adenylylsulfate kinase**	2
General function prediction only	Predicted Fe-S oxidoreductases	2
Replication, recombination and repair	DNA polymerase sliding clamp subunit	2
Replication, recombination and repair	Uracil DNA glycosylase	2
Cell wall/membrane/envelope biogenesis	Soluble lytic murein transglycosylase	2
Cell wall/membrane/envelope biogenesis	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	2
Nucleotide transport and metabolism	ADP-ribose pyrophosphatase	2
Posttranslational modification, chaperones	Predicted ATP-dependent serine protease	2
General function prediction only	Metal-dependent beta-lactamase superfamily I	2
Transcription	Cold shock proteins	2
Defense mechanisms	Restriction endonuclease	2
Transcription	DNA-directed RNA polymerase, sigma subunit	2
Multiple classes	SOS-response transcriptional repressors (RecA-mediated autopeptidases)	2
Coenzyme transport and metabolism	2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-1,4-benzoquinol methylase	2
Cell wall/membrane/envelope biogenesis	Glycosyltransferase involved in LPS biosynthesis	2
Cell wall/membrane/envelope biogenesis	Putative peptidoglycan-binding protein	2
General function prediction only	Phage tail sheath protein FI	2
General function prediction only	Phage tail tube protein FII	2
Cell wall/membrane/envelope biogenesis	Mannosyltransferase OCH1 and related enzymes	2
Replication, recombination and repair	Antirestriction protein	2
General function prediction only	Bacteriophage P2-related tail formation protein	2
Cell wall/membrane/envelope biogenesis	Predicted soluble lytic transglycosylase fused to an ABC-type amino acid-binding protein	2
General function prediction only	Phage-related holin (Lysis protein)	2
Posttranslational modification, chaperones	**Cysteine protease**	2
Cell cycle control, cell division, chromosome partitioning	Membrane-bound metallopeptidase	2
General function prediction only	Predicted methyltransferase (contains TPR repeat)	2
General function prediction only	Mu-like prophage protein gpG	2

General function prediction only	Predicted kinase	2
Nucleotide transport and metabolism	Glutamine phosphoribosylpyrophosphate amidotransferase	1
Cell cycle control, cell division, chromosome partitioning	Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	1
Translation, ribosomal structure	Histidyl-tRNA synthetase	1
Nucleotide transport and metabolism	Thymidylate kinase	1
Replication, recombination and repair	Ribonuclease HII	1
Nucleotide transport and metabolism	Thymidylate synthase	1
Energy production and conversion	**Inorganic pyrophosphatase**	1
Carbohydrate transport and metabolism	Ribulose-5-phosphate 4-epimerase	1
Translation, ribosomal structure	N-formylmethionyl-tRNA deformylase	1
Posttranslational modification, protein turnover, chaperones	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain	1
Replication, recombination and repair	NAD-dependent DNA ligase	1
Amino acid transport and metabolism	Spermidine synthase	1
Multiple classes	Phosphoribosylpyrophosphate synthetase	1
Posttranslational modification, chaperones	DnaJ-class molecular chaperone with C-terminal Zn finger domain	1
Replication, recombination and repair	DNA polymerase III, alpha subunit	1
General function prediction only	Predicted PP-loop superfamily ATPase	1
Multiple classes	Cytidylyltransferase	1
General function prediction only	Predicted kinase	1
General function prediction only	Predicted dehydrogenases and related proteins	1
General function prediction only	MoxR-like ATPases	1
Carbohydrate transport and metabolism	Predicted xylanase/chitin deacetylase	1
Multiple classes	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	1
Cell wall/membrane/envelope biogenesis	Cell wall-associated hydrolase	1
Cell wall/membrane/envelope biogenesis	Lipoproteins	1
Cell wall/membrane/envelope biogenesis	N-acetylmuramoyl-L-alanine amidase	1
Multiple classes	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	1
Coenzyme transport and metabolism	Nicotinamide mononucleotide adenylyltransferase	1
Replication, recombination and repair	ATP-dependent exoDNAse (exonuclease V) beta subunit	1
Cell wall/membrane/envelope biogenesis	UDP-glucose 4-epimerase	1
Cell wall/membrane/envelope biogenesis	GDP-D-mannose dehydratase	1
Posttranslational modification, chaperones	Pyruvate-formate lyase-activating enzyme	1
Transcription	DNA-directed RNA polymerase, sigma subunit	1
Multiple classes	Rad3-related DNA helicases	1
Posttranslational modification, chaperones	ATP-dependent 26S proteasome regulatory subunit	1
Replication, recombination and repair	DNA topoisomerase VI, subunit B	1
Posttranslational modification, chaperones	Subtilisin-like serine proteases	1
Replication, recombination and repair	ATP-dependent DNA ligase	1
Replication, recombination and repair	DNA replication protein	1
Coenzyme transport and metabolism	Nicotinic acid phosphoribosyltransferase	1
Inorganic ion transport and metabolism	**Ferritin-like protein**	1

Signal transduction mechanisms	Carbon storage regulator	1
Replication, recombination and repair	Holliday junction resolvase - archaeal type	1
Translation, ribosomal structure and biogenesis	Acetyltransferases, including N-acetylases of ribosomal proteins	1
Signal transduction mechanisms	DnaK suppressor protein	1
Replication, recombination and repair	DNA polymerase IV (family X)	1
Cell wall/membrane/envelope biogenesis	Spore coat polysaccharide biosynthesis protein F	1
Signal transduction mechanisms	Predicted ATPase related to phosphate starvation-inducible protein PhoH	1
Replication, recombination and repair	Transposase and inactivated derivatives	1
Replication, recombination and repair	ERCC4-type nuclease	1
Replication, recombination and repair	DNA repair proteins	1
General function prediction only	Predicted glutamine amidotransferases	1
Cell wall/membrane/envelope biogenesis	Sialic acid synthase	1
General function prediction only	Predicted phosphoesterases, related to the Icc protein	1
Nucleotide transport and metabolism	Deoxycytidylate deaminase	1
Carbohydrate transport and metabolism	Predicted glycosylase	1
Multiple classes	Type II secretory pathway, pseudopilin PulG	1
Replication, recombination and repair	DNA polymerase III, alpha subunit (gram-positive type)	1
Replication, recombination and repair	Adenine specific DNA methylase Mod	1
General function prediction only	Predicted Zn-dependent hydrolases of the beta-lactamase	1
Coenzyme transport and metabolism	Methylase involved in ubiquinone/menaquinone biosynthesis	1
General function prediction only	Predicted phosphohydrolase (DHH superfamily)	1
General function prediction only	Predicted archaeal methyltransferase	1
General function prediction only	Predicted Zn-dependent protease	1
Amino acid transport and metabolism	3-deoxy-D-arabino-heptulosonate 7-phosphate synthase	1
Coenzyme transport and metabolism	Molybdenum cofactor biosynthesis enzyme	1
Replication, recombination and repair	DNA recombination-dependent growth factor C	1
Cell wall/membrane/envelope biogenesis	Membrane protein involved in colicin uptake	1
Function unknown	Uncharacterized iron-regulated protein	1
General function prediction only	Predicted double-stranded RNA/RNA-DNA hybrid binding protein	1
General function prediction only	Phage protein U	1
General function prediction only	Phage protein D	1
Function unknown	Predicted periplasmic protein	1
Posttranslational modification, chaperones	Predicted proline hydroxylase	1
Intracellular trafficking, secretion	Type IV secretory pathway, TrbL components	1
General function prediction only	Surface antigen	1
General function prediction only	Predicted O-methyltransferase	1
General function prediction only	ABC-type uncharacterized transport system, permease and ATPase components	1
General function prediction only	Mu-like prophage tail protein gpP	1
General function prediction only	Mu-like prophage tail sheath protein gpL	1
Function unknown	Mu-like prophage protein gp36	1
General function prediction only	Competence protein	1
Function unknown	Microcystin-dependent protein	1
Posttranslational modification, chaperones	Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins	1
Replication, recombination and repair	Recombination DNA repair protein (RAD52 pathway)	1
General function prediction only	Pyocin large subunit	1
Total		**1248**

3.3.8. Antibiotic resistance genes in activated sludge viral metagenome

To characterize the potential antibiotic and toxic compound resistance genes assembled from activated sludge viruses, predicted ORFs were compared to the Comprehensive Antibiotic Resistance Database (CARD). Fourteen ORFs encoded on 13 contigs resembled known antibiotic resistance genes (Table 3.7). This included genes conferring resistance to vancomycin and penicillin (D-alanyl-D-alanine carboxypeptidase), β–lactam (metallo-beta-lactamase and N-acetyl-anhydromuramil-l-alanine amidase), trimethoprim (dihydrofolate reductase) and arsenate (ArsR family transcriptional regulator). The low sequence identities of genes identified in the wastewater sample (Table 3.7) with known proteins indicate the novelty of these proteins. Phylogenetic analysis of the three complete metallo-β-lactamase sequences showed that two genes were distantly related with *Bacteroidetes* sequences and one grouped with *Betaproteobacterium* (Figure 3.9A). A close inspection of contig50 (10.8 kb in length) demonstrated that it encodes two antibiotic resistance genes (vancomycin and β–lactam) along with a phage recombinase (RecT) (Figure 3.9B). Predicted ORFs matched with proteins affiliated with different bacterial species, making taxonomic assignment impossible. None of these ORFs matched known phage proteins. Low amino-acid identities (42-67%, with identity of 67% to *Clostridium difficile*), suggests that contig50 represents a novel phage genome.

3.3.9. Comparison with other viral metagenomes

The comparative analysis was based on Principal Component Analysis (PCA) using MG-RAST. Three municipal wastewater viral metagenomes (Parsley *et al.*, 2010b; Rosario *et al.*, 2009b; Tamaki *et al.*, 2012), along with a freshwater (lake) viral metagenome (Roux *et al.*, 2012b), were used for comparison with the assembled viral metagenome from dairy wastewater. A cystic fibrosis (CF) sputum viral metagenome was used as an outgroup (Willner *et al.*, 2009). PCA analysis indicated that taxonomic and functional profile of the dairy wastewater virus community was different from those of other environments, although of all compared viral metagenomes, it was most similar to another activated sludge viral metagenome from a municipal treatment plant in Singapore (Figure 3.10A and B). These two metagenomes both contain overrepresentation of DNA methylase genes, and have many phage hits in common e.g. to *Pseudomonas aeruginosa* bacteriophage PaP2, *Bordetella* phage BPP-1, *Burkholderia cenocepacia* phage Bcep1 and *Flavobacterium* phage 11b, however at different relative proportions. Phage communities in the dairy wastewater metagenome were most distant from CF sputum and freshwater communities.

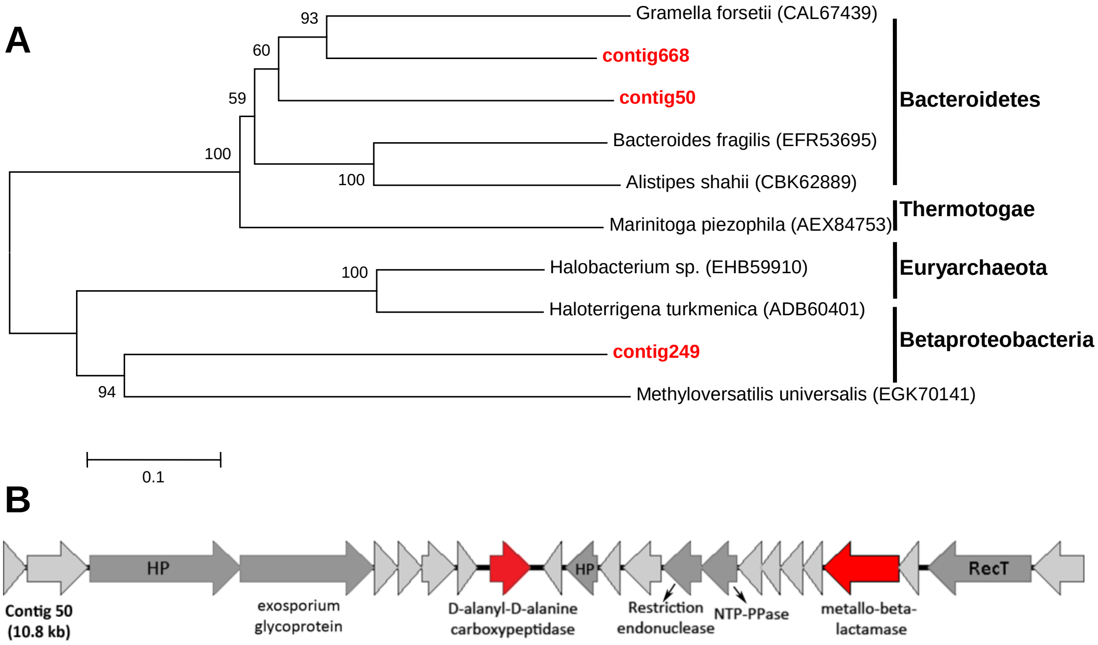
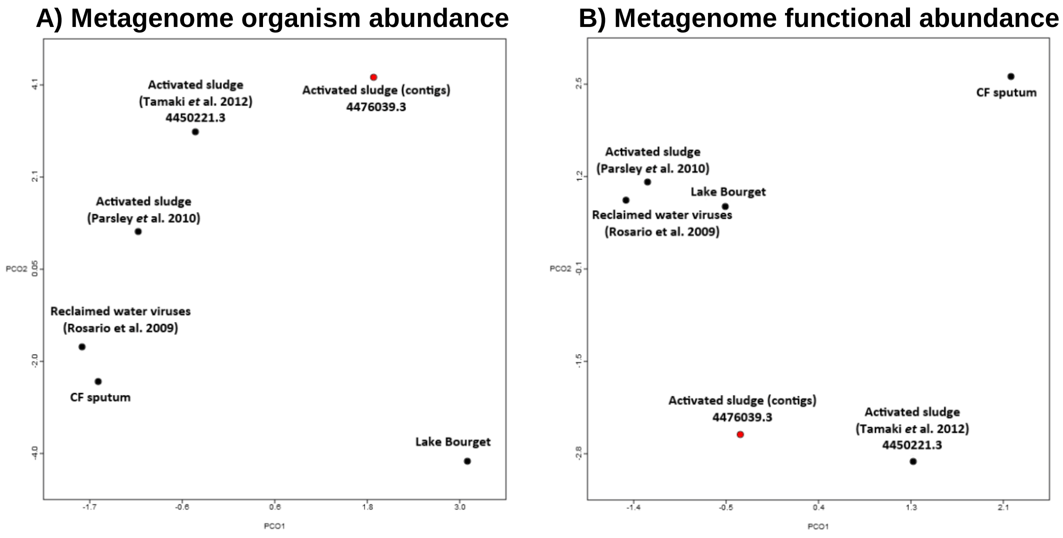


Figure 3.9. (A) A neighbour-joining phylogenetic tree showing the relationship between the predicted metallo-β-lactamase sequences assembled from the activated sludge viral metagenome to the most related sequences in NCBI protein database. Metallo-β-lactamase sequences detected in this study are shown in bold red. (B) Analysis of the gene order in contig 50 containing two antibiotic resistance genes (coloured in red), known genes including hypothetical proteins (HP) (dark grey) and unknown ORFs (light grey). Predicted protein coding genes were annotated using NCBI protein database. ORFs displayed low amino-acid similarities (42-67%) to different bacteria species. The figure was drawn using Easyfig.



A) Metagenome organism abundance

B) Metagenome functional abundance

Figure 3.10. Principal component analysis (PCoA) based on organism abundance (A) and functional abundance (B) using MG-RAST. Metagenome comparisons were calculated for dairy wastewater viral metagenome (contigs >100bp) (red circle) and other published viral metagenomes (black circles): wastewater (Parsley *et al.*, 2010b; Rosario *et al.*, 2009b; Tamaki *et al.*, 2012), lake (Roux *et al.*, 2012b) and CF sputum (Willner *et al.*, 2009). The data was compared to GenBank (organism abundance) and Subsystems (functional abundance) using a maximum e-value of 1e-05 and a minimum alignment length of 50.

3.3.10. CRISPR sequences in activated sludge viral metagenome

To identify matches between the bacterial and archaeal CRISPR spacers and activated sludge phage sequences, spacers downloaded from the CRISPR database (Grissa *et al.*, 2007b) were compared to the metagenomic reads and contigs using BLASTN. Spacers having 100% identity over 20 bp were retrieved. A total of 75 different spacers matching activated sludge reads and 22 different spacers matching assembled contigs were identified (Table 3.8). These spacers belonged to 77 different bacteria species, including species typically found in wastewater environment such us *Candidatus Accumulibacter phosphatis* (Garcia Martin et al., 2006) and *Nakamurella multipartita* (Tice et al., 2010), or pathogenic species found dairy wastewater such as *Campylobacter jejuni* (Dungan *et al.*, 2012). The most common identified contig CDS containing CRISPR hits were DNA methylase followed by phage enzymes. Phylogenetic and functional assignment of these CDS is shown in Table 3.9.

Table 3.7. Antibiotic and toxic compound resistance genes found in ORFs predicted on the assembled contigs from the activated sludge virome.

Resistance gene function	Gene	Resistance conferred	Best-match organism	BLAST identity (%)	E-value	ORF coverage (%)	ORF length (aa)	Contig length (kb)	Contig name
Transcriptional regulator	*ArsR*	arsenate	*Thermoanaerobacter pseudethanolicus*	43	6e-11	83	92	2.6	c473
D-alanyl-D-alanine carboxypeptidase	*VanY*	vancomycin penicillin	*Herbaspirillum seropedicae*	45	3e-32	78	180	31.8	c9
D-alanyl-D-alanine carboxypeptidase	*VanY*	vancomycin penicillin	*Solitalea canadensis*	50	1e-37	95	136	10.8	c50
Metallo-beta-lactamase	*MBL*	β-lactam	*Gramella forsetii*	49	3e-71	96	252	10.8	c50
Metallo-beta-lactamase	*MBL*	β-lactam	*Halobacterium* sp.	29	6e-51	100	416	4.0	c249
Metallo-beta-lactamase	*MBL*	β-lactam	*Gramella forsetii*	57	1e-92	100	246	2.1	c668
N-acetyl-anhydromuramil-l-alanine amidase	*AmpD*	β-lactam	*Anaerophaga thermohalophila*	34	3e-32	90	292	5.6	c163
N-acetyl-anhydromuramil-l-alanine amidase	*AmpD*	β-lactam	*Brevundimonas diminuta*	56	5e-59	83	180	4.0	c247
N-acetyl-anhydromuramil-l-alanine amidase	*AmpD*	β-lactam	*Gemmatimonas aurantiaca*	41	2e-20	97	127	3.2	c346
R67 dihydrofolate reductase	*DHFR*	trimethoprim	*Burkholderiales* bacterium	70	4e-11	20	193	1.1	c1606
R67 dihydrofolate reductase	*DHFR*	trimethoprim	*Burkholderiales* bacterium	66	2e-04	32	87	0.5	c4075
R67 dihydrofolate reductase	*DHFR*	trimethoprim	*Salmonella* enterica	75	6e-23	91	62	0.3	c6659
R67 dihydrofolate reductase	*DHFR*	trimethoprim	*Salmonella* enterica	74	3e-22	91	62	0.3	c7130
R67 dihydrofolate reductase	*DHFR*	trimethoprim	*Burkholderiales* bacterium	76	2e-13	75	54	0.2	c12808

Table 3.8. Bacterial CRISPRdb spacers matching AS reads and contigs. Information about habitat was retrieved from NCBI Genome project organism overview http://www.ncbi.nlm.nih.gov/genome. Aquatic (A), animal associated (AN), human associated (H), plant associated (P), marine (M), sediment (SD), **sludge (SL)**, soil (S).

Bacterial genome (habitat)	**No. of unique spacers matching AS reads**	**No. of unique spacers matching AS contigs**
Acinetobacter baumannii (A, H)	1	0
Acidovorax sp. JS42 (SD)	1	0
Actinobacillus pleuropneumoniae (AN)	1	0
Ammonifex degensii (A)	1	0
Anaeromyxobacter dehalogenans (S)	1	0
Anoxybacillus flavithermus (A)	1	0
Azospirillum sp. B510 (P)	1	1
Azotobacter vinelandii (S)	1	0
Caldicellulosiruptor hydrothermalis (A)	1	0
Caldicellulosiruptor kristjanssonii (A)	1	1
***Campylobacter jejuni* (H, AN, SL)**	1	1
***Candidatus Accumulibacter phosphatis* (SL)**	1	2
Candidatus Korarchaeum cryptofilum (A)	0	1
Cellulomonas fimi (S)	1	0
Chlorobium chlorochromatii (A)	1	0
Chlorobium limicola (A)	1	0
Chloroflexus aurantiacus (A)	1	0
Chloroherpeton thalassium (A)	2	0
Clostridium botulinum (H, AN)	1	0
Clostridium thermocellum (P)	0	1
Cyanothece sp. (M)	1	0
Desulfarculus baarsii (SD)	1	0
***Desulfitobacterium hafniense* (S, SL)**	1	0
Desulfotomaculum kuznetsovii (A)	1	0
Desulfovibrio vulgaris (S) (A)	1	0
Dickeya dadantii (P)	1	0
Escherichia coli (A) (H) (AN)	1	0
Fluviicola taffensis (A)	1	0
Haliangium ochraceum (M)	1	0
Halorubrum lacusprofundi (M)	1	0
Isosphaera pallida (A)	1	0
Kyrpidia tusciae (A)	2	0
Leptotrichia buccalis (H)	1	1
Magnetococcus marinus (M)	1	0
Marinomonas mediterranea (M)	1	1
Metallosphaera sedula (A)	1	0
Methanobacterium sp. (A)	1	0
Methanosphaera stadtmanae (H)	1	0
Methanothermococcus okinawensis (M)	2	0
Methanotorris igneus (M)	1	0
Methylomonas methanica (A)	0	2

Bacterial genome (habitat)	**No. of unique spacers matching AS reads**	**No. of unique spacers matching AS contigs**
Microcystis aeruginosa (A)	1	0
***Nakamurella multipartita* (SL)**	0	1
Natrialba magadii (A)	1	0
***Nitrosomonas europaea* (SL)**	1	0
Pectobacterium wasabiae (P)	1	0
***Pelotomaculum thermopropionicum* (SL)**	1	0
Prevotella denticola (H)	1	0
*Propionibacterium freudenreichii**	1	0
Pseudomonas mendocina (S, A, H)	2	0
Pyrococcus abyssi (M)	1	0
Rhodoferax ferrireducens (SD)	1	1
***Rhodopseudomonas palustris* (S, A, SD, SL)**	1	1
Rhodothermus marinus (M)	1	1
Riemerella anatipestifer (AN)	1	0
Runella slithyformis (A)	0	1
Salinispora arenicola (M)	1	0
Shewanella baltica (M)	1	0
Streptococcus gallolyticus (H, AN)	0	1
*Streptococcus thermophilus**	1	0
Streptomyces avermitilis (S)	1	0
Sulfolobus tokodaii (A)	1	0
Sulfurihydrogenibium azorense (A)	1	0
Sulfurihydrogenibium sp. (A)	2	0
***Syntrophothermus lipocalidus* (SL)**	1	0
***Tepidanaerobacter* sp. Re1 (SL)**	1	0
Thermoanaerobacter sp. (A)	1	1
Thermoanaerobacter tengcongensis (M)	1	0
Thermobifida fusca (isolated from waste)	1	0
Thermobispora bispora (isolated from manure)	1	0
Thermococcus onnurineus (SD)	0	1
Thermodesulfovibrio yellowstonii (A)	0	1
Thermosipho africanus (M)	1	1
Thermotoga neapolitana (M)	1	0
Treponema brennaborense (AN)	1	0
Verminephrobacter eiseniae (AN)	2	1
Xanthomonas oryzae (P)	1	0
Total	**75**	**22**

* Commonly isolated from cheese and other dairy products

Table 3.9. Analysis of CRISPR spacer matches to dairy wastewater phage contigs. CDS predicted on contigs that matched CRISPR spacers were further analysed using BLASTP.

WW phage contig	Bacterial genome CRISPR matches to WW phage contigs	Bacterial source	BLASTP organism assignment (CDS % coverage, % identity)	WW phage CDS functional assignment
contig4299	*Azospirillum* sp. B510	Rhizosphere	*Comamonas testosteroni* (96%, 53%)	Phage integrase
contig7938	*Caldicellulosiruptor kristjanssonii*	Hot springs	*Elusimicrobium minutum* (83%, 53%)	Phage terminase
contig1395	*Campylobacter jejuni* RM1221	Animal feces	*Burkholderia* phage Bcep43 (94%, 36%)	DNA methylase
contig38	Candidatus *Accumulibacter phosphatis*	Wastewater	No BLASTP similarity	No BLASTP similarity
contig6615	Candidatus *Accumulibacter phosphatis*	Wastewater	Spacer not within the CDS	-
contig5700	Candidatus *Korarchaeum cryptofilum*	Hot springs	*Escherichia* phage vB_EcoM_ECO1230-10 (98%, 46%)	Helicase
contig8363	*Clostridium thermocellum* ATCC 27407	Multiple habitats (cellulose rich)	No BLASTP similarity	No BLASTP similarity
contig8573	*Leptotrichia buccalis* C-1013-b	Human oral cavity	*Bacillus* sp. BT1B_CT2 (93%, 49%)	Sensor protein, YopX family
contig12688	*Marinomonas mediterranea* MMB-1	Marine	No BLASTP similarity	No BLASTP similarity
contig36	*Methylomonas methanica* MC09	Freshwater	*Capnocytophaga* sp. CM59 (82%, 41%)	Hypothetical protein
contig2198	*Methylomonas methanica* MC09	Freshwater	*Collimonas fungivorans* (90%, 49%)	Peptidase M15 family
contig17029	*Nakamurella multipartita* DSM 44233	Wastewater	*Chlamydia* phage 4 (87%, 44%)	Replication initiation protein
contig7327	*Rhodoferax ferrireducens* T118	Marine sediment	*Herbaspirillum* sp. GW103 (98%, 42%)	Hypothetical protein
contig8879	*Rhodopseudomonas palustris* DX-1	Freshwater, soil	No BLASTP similarity	No BLASTP similarity
contig2734	*Rhodothermus marinus* DSM 4252	Hot springs	*Elizabethkingia anopheles* (88%, 37%)	Antirestriction protein (ArdA)
contig3786	*Runella slithyformis* DSM 19594	Freshwater	*Bacteroides* sp. D20 (95%, 48%)	DNA methylase
contig13361	*Streptococcus gallolyticus*	Human blood	No BLASTP similarity	No BLASTP similarity
contig7347	*Thermoanaerobacter* sp. X514	Freshwater	Spacer not within the CDS	-
contig10721	*Thermococcus onnurineus* NA1	Marine	*Frateuria aurantia* (100%, 59%)	Restriction endonuclease
contig13271	*Thermodesulfovibrio yellowstonii*	Hot springs	No BLASTP similarity	No BLASTP similarity
contig7188	*Thermosipho africanus* TCF52B	Marine oil reservoir	*Odoribacter laneus* (97%, 34%)	DNA methylase
contig866	*Verminephrobacter eiseniae* EF01-2	Earthworm nephridia	*Pseudomonas fluorescens* (100%, 52%)	DNA methylase

3.3.11. Microbial diversity in activated sludge

To identify the putative phage host species, bacterial DNA was extracted from the same sample that was used for the viral diversity study and amplified using 16S ribosomal RNA universal primers (Lane, 1991) and *Accumulibacter*-specific primers (Kunin *et al.*, 2008) targeting *Candidatus Accumulibacter phosphatis*, a bacterium widespread in sludge samples. Nineteen clones that were successfully sequenced from the clone libraries were affiliated with the *Proteobacteria*, *Bacteroidetes* and *Gemmatimonadetes* (Table 3.10). *Proteobacteria* dominated the 16S rRNA clone libraries, and the identified genera were typical for a soil and wastewater environment. Some species identified such as *Novosphingobium* sp., *Dechloromonas aromatica, Niastella koreensis* and *Gemmatimonas aurantiaca* were also detected in the viral metagenome (Table 3.2, Table 3.7).

Table 3.10. Sequence similarity of the 16S rRNA clones amplified from activated sludge to the NCBI reference genome database.

No. of clones	Closest known match in BLAST analysis (accession number)	Identity (%)	Isolation source of the match	Phylum
1	*Novosphingobium* sp. PP1Y (NC_015580)	96	Seawater	α-Proteobacteria
1	*Beijerinckia indica* (NC_010581)	95	Soil	
1	*Nitrosospira multiformis* (NC_007614)	89	Soil	β-Proteobacteria
1	*Methylibium petroleiphilum* (NC_008825)	89	Sewage	
1	*Ralstonia solanacearum* (NC_014311)	90	Soil	
1	*Azoarcus* sp. BH72 (NC_008702)	93	Soil	
4*	*Candidatus Accumulibacter phosphatis* (NC_013194)	99	EBPR bioreactor	
5*	*Dechloromonas aromatica* (NC_007298)	97-98	Sediment	
1	*Niastella koreensis* (NC_016609)	88	Soil	Bacteroidetes
2	*Owenweeksia hongkongensis* (NC_016599)	85-86	Seawater	
1	*Gemmatimonas aurantiaca* (NC_012489)	83	EBPR bioreactor	Gemmatimonadetes

*Sequenced using *Accumulibacter*-specific primers

3.4. Discussion

This chapter describes the application of Illumina short-read-based DNA sequencing to characterise taxonomic and functional diversity of viruses in activated sludge sample collected from a dairy food wastewater treatment plant. Previous metagenomic analysis of viruses associated with wastewater environment (raw sewage, activated sludge, dairy waste lagoon) revealed that a large fraction of metagenomic reads had no significant similarity to any known sequences, indicating the high proportion of unknown viruses in this environment (Alhamlan *et al.*, 2013; Bibby & Peccia, 2013; Cantalupo *et al.*, 2011; Ng *et al.*, 2012; Rosario *et al.*, 2009b; Tamaki *et al.*, 2012). Majority of these studies reported that bacteriophages dominated the known fraction of the viral community.

The majority of reads (99.86%) and assembled contigs (65%) had no hits to any sequences in databases, suggesting that the dairy food wastewater viral community was mostly novel. This is comparable with findings of previous studies of viral diversity in various other environments, including an activated sludge from a municipal wastewater treatment plant (Tamaki *et al.*, 2012), reclaimed water (Rosario *et al.*, 2009b), Antarctic lake (Lopez-Bueno *et al.*, 2009) or marine water (Angly *et al.*, 2006).

Despite the stringent virus purification method applied, a large proportion of assigned sequences showed similarity to bacterial sequences. However, analyses of the genes in assembled contigs confirmed that these sequences were dominated by phage proteins (e.g. presence of high number of terminases or phage structural genes) (Table 3.1, Figure 3.7, Table 3.5, and Table 3.6). A high abundance of prokaryotic sequences in an analysis of a viral metagenome is consistent with other studies (Angly *et al.*, 2006; Lopez-Bueno *et al.*, 2009; Tamaki *et al.*, 2012), and it may be explained by the common inclusion of uncharacterised prophage sequences in databases containing bacterial genome sequences (Fouts, 2006; Rosario *et al.*, 2009b). Previous metagenomic studies of viral communities have shown that the majority of sequences initially classified as bacteria can be re-classified as plasmids or phages by further consideration of gene context and associations (Rosario *et al.*, 2009b; Roux *et al.*, 2012b).

The majority of these uncharacterized prophage sequences resembled phage sequences from bacteria commonly found in soil and aquatic environments, including sludge (Table 3.2), and the majority of assignments were to bacteria of the *Proteobacteria* phylum. Based on the reads classification, top five (13%) bacterial hits could be assigned to genus *Neisseria*, *Pseudomonas*, *Escherichia*, *Salmonella* and *Acinetobacter*. Ten contigs (0.6% of all assigned contigs) contained genes related to *Neisseria* species, however when re-BLASTed these contigs had the best BLAST hits to other species, suggesting that the presence of *Neisseria* species in the sample was low and might be the results of contamination during the sample manipulation. Contigs assigned to *Pseudomonas* species (P. *aeruginosa*, P. *syringae* and P. *putida* most common) accounted for 3%, *Escherichia* species accounted for 1% and *Salmonella* species accounted for 0.6% of the assigned contigs. This is consistent with finding of a previous study that detected high abundance of prophages, including those of *Escherichia*, *Pseudomonas* and *Salmonella* in reclaimed water (Rosario *et al.*, 2009b). *Acinetobacter* species accounted for 3% of the assigned contigs, including two large contigs (Table 3.1), one related to *Acinetobacter brisouii*, a gammaproteobacterium isolated from a wetland (Anandham *et al.*, 2010) and the other to *Acinetobacter baumannii*, a nosocomical pathogen. High abundance of *Acinetobacter* species have been previously detected in freshwater metagenome using pyrosequencing, suggesting that they are common in the environment (Ghai *et al.*, 2011).

Based on the best BLASTX hit some contigs contained genes showing high similarity (>70% of amino acid identity, but no more than 86%) to known sequences. Two large contigs (26 kb and 10 kb), possibly originating from the same phage genome, showed high sequence similarity to a prophage of *Nitrosomonas eutropha* (Table 3.1) (1% of the assigned contigs). *N. eutropha* is an ammonia-oxidizing bacterium converting ammonia to nitrate via nitrification, found in polluted environments like wastewater or eutrophic sediments (Stein *et al.*, 2007). Several small contigs including one larger (~7 kb) showed high sequence similarity to a prophage of *Dechloromonas aromatica* (0.5% of the total assigned contigs), microorganism involved in the degradation of aromatic compounds that have been isolated from sludge (Salinero *et al.*, 2009). Presence of *Dechloromonas aromatica* in dairy wastewater has been also confirmed by 16S rRNA PCR (Table 3.10). Another large contig (> 10 kb) showed similarity to a prophage of *Thauera* sp. 27, a microorganism isolated from a wastewater reactor (Liu et al., 2013). Few contigs, including a large contig (~ 11 kb), had similarity to a prophage of *Flavobacterium columnare* (0.4% of the assigned contigs), a freshwater fish pathogen (Decostere et al., 1997). Among other prophages of potentially pathogenic bacteria was *Legionella pneumophila* (0.4%), an environmental pathogen causing lung infections in humans and animals including cattle (Catalan *et al.*, 1997; Fabbi *et al.*, 1998).

The taxonomic affiliation of most of the contigs could not be ascertained. For example, the largest contig assembled in this study (~ 114 kb) was initially classified as marine bacterium *Pelagibaca bermudensis* (Table 3.1), however further sequence analysis showed that it contained multiple ORFs with weak similarity to phage SP10 infecting *Bacillus subtilis*, a ubiquitous bacterium commonly found in water and soil. This contig therefore might represent a novel *Bacillus* phage.

Viral sequences were classified into 8-9 different viral families, much less than previously reported in raw sewage, which contained 51 viral families (Cantalupo *et al.*, 2011), and comparable to dairy lagoon wastewater (13 families) (Alhamlan *et al.*, 2013). The directly identified viral fraction of dairy wastewater was dominated by double-stranded (ds) DNA bacteriophages from *Siphoviridae*, *Myoviridae* and *Podoviridae* families, which accounted for 64% of the contigs assigned to viruses (Figure 3.3C). These findings are consistent with previous studies, which observed that viral communities in wastewater samples were dominated by bacteriophages (Alhamlan *et al.*, 2013; Cantalupo *et al.*, 2011; Tamaki *et al.*, 2012). Phages infecting *Vibrio* spp., *Mycobacterium* spp., *Synechococcus* spp., *Pseudomonas* spp. and *Burkholderia* spp. were the most abundant among assembled sequences. These phages may be generally common in the wastewater environment as previous studies also demonstrated the presence of phages infecting those bacteria (Rosario *et al.*, 2009b; Tamaki *et al.*, 2012).

Of particular interest were sequences with similarities to uncultured phages from enhanced biological phosphorus removal (EBPR) sludge, which is dominated by *Candidatus Accumulibacter phosphatis* (CAP) (Skennerton et al., 2011) (Table 3.3).

CAP is globally distributed in a substantially identical form and subject to locally variable phage predation (Kunin et al., 2008), including phage from *Siphoviridae* and *Podoviridae* families (Skennerton et al., 2011). Despite dairy wastewater viruses shared less than 76% amino acid sequence identities to known EBPR viruses, suggesting that these viruses are only weakly related, the PCR amplification of 16S rRNA genes confirmed the presence of *Candidatus Accumulibacter phosphatis* in the dairy wastewater sample. Analysis of matches of known CRISPR spacers from *Accumulibacter phosphatis* to viral reads and assembled contigs also suggests that viruses infecting CAP are present in this environment (Table 3.8). This suggests the presence of novel phage infecting CAP in this dairy wastewater sample.

Viruses with small circular single-stranded (ss) DNA genomes infecting eukaryotes (*Circoviridae*, *Nanoviridae* and *Geminiviridae*) and prokaryotes (*Microviridae*) accounted for 23% of the contigs assigned to viruses (Figure 3.3C and Figure 3.4). Phylogenetic analyses showed that fifteen complete ssDNA viruses assembled from dairy metagenome were diverse and novel (Figure 3.5 and Figure 3.6). Single-stranded viruses were dominated by bacteriophages (Figure 3.4) resembling those associated with obligate intracellular bacteria closely related to *Chlamydia* and *Chlamydophila* species (Figure 3.6). The natural hosts of *Chlamydia*-like species in wastewater are most likely free-living amoebae (Corsaro & Greub, 2006). Chlamydiae have been isolated from water samples, including lake (Pizzetti *et al.*, 2012) and wastewater (Corsaro *et al.*, 2009). It has been recently demonstrated that host-free *Chlamydia*-like organisms can survive outside of their host for extended periods and are resistant to heat, which may account for their ubiquitous presence in the environment (Coulon *et al.*, 2012). Some Chlamydia can be pathogenic to humans causing respiratory infections (Corsaro & Greub, 2006; Friedman *et al.*, 2003; Greub, 2009). Therefore, the presence of Chlamydia-like phages demonstrates that dairy wastewater is a source of pathogens that might be involved in human pathogenicity. Other ssDNA viruses that were identified in wastewater were numerous circovirus-like sequences (Figure 3.5). Circoviruses are known to infect a wide range of animal hosts and can be associated with disease in variety of animals including dairy cattle (Delwart & Li, 2012; Li *et al.*, 2011c; Nayar *et al.*, 1999), therefore dairy products may be a source of these viruses. A high abundance of novel ssDNA viruses has been found in metagenomic surveys of many different environments including sewage (Blinkova *et al.*, 2009; Cantalupo *et al.*, 2011), reclaimed water (Rosario *et al.*, 2009b), lake (Lopez-Bueno *et al.*, 2009; Roux *et al.*, 2012b), marine water (Angly *et al.*, 2006), human feces (Kim *et al.*, 2011) and dairy lagoon wastewater (Alhamlan *et al.*, 2013). The extensive representation of ssDNA viruses in all these metagenomic samples may reflect methodological bias, as it may be a result of the amplification step with phi29 polymerase that is known to preferentially amplify small circular DNA (Kim *et al.*, 2008). Although their true abundance cannot currently be assessed with reliability, ssDNA viruses are however certainly present in dairy wastewater.

The diversity of the dairy wastewater viral metagenome (409 species) was lower compared to marine ecosystems (4110) (Bench *et al.*, 2007), higher than that found in human associated samples (105) (Willner *et al.*, 2009) and comparable to that found in another activated sludge from a municipal wastewater treatment plant (511) (Tamaki *et al.*, 2012). Principal component analysis (Figure 3.10) showed that the dairy wastewater metagenome contained genes and species distinct from those previously described, and was most similar to the activated sludge from municipal wastewater treatment plant (Tamaki *et al.*, 2012).

Analysis of the genes present in dairy wastewater viral metagenome revealed a substantial number of phage-related genes including structural proteins and genes involved in nucleic acid metabolism, such as DNA replication and synthesis. High abundance of phage genes involved in DNA metabolism may reflect phage activity in wastewater. Wastewater contains high concentration of nutrients and as a consequence a high bacterial load that favours rapid turnover in phage–host interactions (Shapiro *et al.*, 2010). In particular, genes related to DNA methyltransferases, including DNA adenine methyltransferase (*dam*) and DNA cytosine methyltransferase (*dcm*), were overrepresented in dairy wastewater (Table 3.5 and Table 3.6) as previously noted in municipal wastewater datasets (Tamaki *et al.*, 2012). In prokaryotes, methylation of DNA is primarily involved in restriction-modification system used to protect them from phage infection. Typically, restriction-modification systems are composed of genes that encode a methyltransferase, which introduces a specific methylation that protects the DNA against the restriction enzyme cleavage and a restriction enzyme, which recognize the same sequence (Wilson & Murray, 1991). Once inside the host, phage DNA is restricted because it is not methylated in the DNA sequence recognized by the host restriction endonuclease. Several phage genomes encode methyltransferases that modify their DNA and protects phage genome from cleavage (Chiou *et al.*, 2010; Drozdz *et al.*, 2011; Kruger & Bickle, 1983). Phage encoded DNA methyltransferases may also play a role in switching between phage lytic and lysogenic life cycles. In prokaryotes, expression of certain genes is regulated by DNA adenine methylation (DAM) of the promoter region, which leads to activation or repression of the gene expression (Collier, 2009). Recently, it was suggested that phage *dam* methylates the phage antirepressor gene, which allow the lytic repressor gene to repress the lytic life cycle. Once methylation is removed, the repressor protein becomes repressed and non-functional leading to switching to the lytic cycle (Bochow *et al.*, 2012).

Some viral sequences were apparently associated with enzymes involved in sulphur assimilation (Table 3.6). In bacteria, sulphur is used for tRNA modification, a mechanism used for improvement of reading frame maintenance (Urbonavicius *et al.*, 2001). Viruses use programmed ribosomal frameshifting to synthesize their proteins (Farabaugh, 1996). Recently, it was suggested that internal competition for sulphur stores may increase host resistance to phage infection, because it preferentially affects translation of phage proteins, through an indirect effect on ribosomal frameshifting via tRNA (Maynard *et al.*, 2012). Therefore phage genes that increase host sulphur uptake could aid viral replication.

The presence of antibiotic resistance genes in close proximity to phage integrase genes (Figure 3.9B) indicates that dairy wastewater is a potential source for transmissible antimicrobial resistance. In general, however, the abundance of antibiotic resistance genes was low. Some phage contigs contained genes conferring resistance to β-lactam, vancomycin and trimethoprim antibiotics (Table 3.7). The presence of these genes in dairy wastewater may be a result of extensive use of β-lactams and trimethoprim to treat dairy cattle infections in Ireland (More *et al.*, 2012).

Chapter 4:

Functional screening for bacterial promoter activity in dairy wastewater metagenomic viral DNA using a promoter trap vector

Abstract

Identification of strong viral promoters can be useful in the construction of novel expression vectors. A metagenomic library containing DNA fragments of viruses extracted from dairy wastewater was screened for their ability to drive expression of the promoter-less *gfp* gene in *E.coli*. Metagenomic sequencing of the DNA used for the library was available. Twenty clones (inserts 65- 992 base pairs) that showed constitutive promoter activity were sequenced and further characterized. Eighteen of these were present in the metagenome assembly and identification of the short promoter sequence source was aided by flanking sequence from the assembly. Sequence analysis of trapped DNA fragments showed that ten inserts were related to the ssDNA viruses of eukaryotes (mainly Geminiviruses), with the majority of ORFs associated with the replication initiation protein. Three inserts were characterized by the presence of phage-specific early transcription regulatory cassettes, which play a role in phage-mediated horizontal gene transfer.Two inserts contained an ORF associated with phage structural proteins. The remaining inserts had unannotated ORFs. These results demonstrate that promoter trapping is useful for identifying regulatory sequences from environmental viral DNA.

4.1. Introduction

Promoters are regulatory regions that initiate transcription of genes. In *E. coli*, the main (sigma^{70}) promoter sequence incorporates two sequence motifs approximately -10 (TAATAT consensus sequence) and -35 (TTGACA consensus sequence) nucleotides upstream of the start of transcription (Hawley & McClure, 1983; Sinoquet *et al.*, 2008). These sequence motifs are recognized by the sigma factors of the RNA polymerase, which bind to the promoter region of DNA and direct transcription (Browning & Busby, 2004). Most bacterial sigma factors belong to the sigma^{70} family and are primarily responsible for regulation of bacterial cell growth (Paget & Helmann, 2003). Other sigma factors (such as sigma^{54}or sigma^{37}) regulate expression of genes associated with specific circumstances such as virulence or genes involved in stress response to changes in environmental conditions (Kazmierczak *et al.*, 2005). Viruses, including bacteriophages, use the infected host cell's transcriptional machinery to enable the expression of their own genes. After infection, prokaryotic or eukaryotic virus early gene promoters are recognized by their host's RNA polymerase sigma factors and transcribed (Fassler & Gussin, 1996; Hinton, 2010). Phages such as T4 are known to have regions of extreme plasticity (variability) called hyperplastic regions (HPRs) in their genome, which are small open reading frames flanked by early transcription promoters composed of sigma^{70} bacterial promoters and stem-loopsThese HPRs show wide horizontal gene transfer mediated by recombination between homologous promoter regions (Arbiol *et al.*, 2010; Comeau *et al.*, 2008). Foreign DNA can be transcribed in bacteria, because sigma^{70} sequences are conserved across different species, and there is evidence that prophage DNA is highly transcribed in *E.coli* (Warren *et al.*, 2008). In addition, prophage promoters subject to the host regulation during the bacterial stress response to environmental stresses, which triggers prophage induction (Glinkowska *et al.*, 2010). Therefore, *E.coli* should express phage genes adapted for expression in a wide range of prokaryotic hosts.

One method to identify functional promoters is to use a promoter-trap vector containing a promoter-less reporter gene such as green fluorescence protein (GFP) that is adjacent to a multiple cloning site. A genomic or metagenomic library of random DNA fragments cloned upstream of a promoter-less *gfp* gene is transformed into *E. coli* and screened for bacteria expressing insert-GFP gene fusions. Identification of novel promoters may have practical applications e.g. in the construction of expression vectors (Han *et al.*, 2008). Promoter-trapping has been used for screening of promoters derived from individual phage DNA (Zargar *et al.*, 2001), and bacterial genomic (Chen *et al.*, 2007; Dunn & Handelsman, 1999) or metagenomic DNA (Han *et al.*, 2008; Lee *et al.*, 2011; Park & Kim, 2010). However, this approach was not previously applied for identification of promoters from a viral metagenome.

Bacteriophage DNA may be difficult to clone because of the inherent bacterial toxicity of some of the gene products if expressed, the presence of modified nucleotides in phage genomes and methylase modification of phage DNA preventing standard digestion (Wang *et al.*, 2000; Warren, 1980). Recently, a linker amplification method has been successfully used to obtain clone libraries from viral metagenomes (Schoenfeld *et al.*, 2008). It involves blunt-ending mechanically fragmented viral DNA and attaching double stranded linkers which serve as a target for PCR primers that can be used to amplify the insert with a proofreading polymerase. Because the linkers of known sequence are ligated to unknown DNA fragments no viral sequence homology is required for functional primers.

Metagenomic analyses of viruses isolated from the dairy food wastewater treatment plant described in Chapter 3 demonstrated that assembled sequences contain mainly phage or prophage-like ORFs. This chapter combines linker amplification for construction of a metagenomic library and promoter trapping to identify phage-specific promoter DNA sequences in the dairy wastewater. Because bacteriophages from some environments have been shown to contain bacterial toxin genes (Casas *et al.*, 2006), genetic modification (GM) risk assessment required that short DNA fragments only were cloned (100 -1000 bp) and the vector used should be non-mobilisable. In addition, the sample DNA used for the library was free of recognised toxin genes on metagenomic sequencing (Chapter 3). A previously described promoter trap plasmid was modified to make it non-mobilisable for screening virus-derived DNA.

4.2. Material and Methods

4.2.1. Deletion of mobilization genes from pZEP08

pZEP08 (Cm^R, Kn^R, Amp^R) (Lab stock; gift from David Clarke) (Hautefort *et al.*, 2003), a pBR322-derived vector carrying a promoterless *gfp* was used for this experiment. In order to minimize the possibility of onward phage gene transfer through bacterial conjugation, a 1634 bp region containing genes required for plasmid mobility (*mobA*, *mobB*, *mobC* and *oriT*) was deleted by PCR. The outward pointing PCR primers: mobF (5'- TTATACTAGTCGATGAAGAACGACAGGAC-3') and mobR (5'- TTATACTAGTGCTGAATGATCGACCGAGA-3') were designed to amplify the DNA region flanking the mob sequence (Figure 4.1A). The underlined bases indicate a *Spe*I restriction site incorporated into each primer. After amplification using Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen), PCR fragments were digested with SpeI restriction enzyme (NEB) and self-ligated using the T4 DNA Ligase (Invitrogen) to produce pZEP08Δ*mob*. Successful deletion of *mob* region was confirmed by sequencing of the entire plasmid (Figure 4.1B).

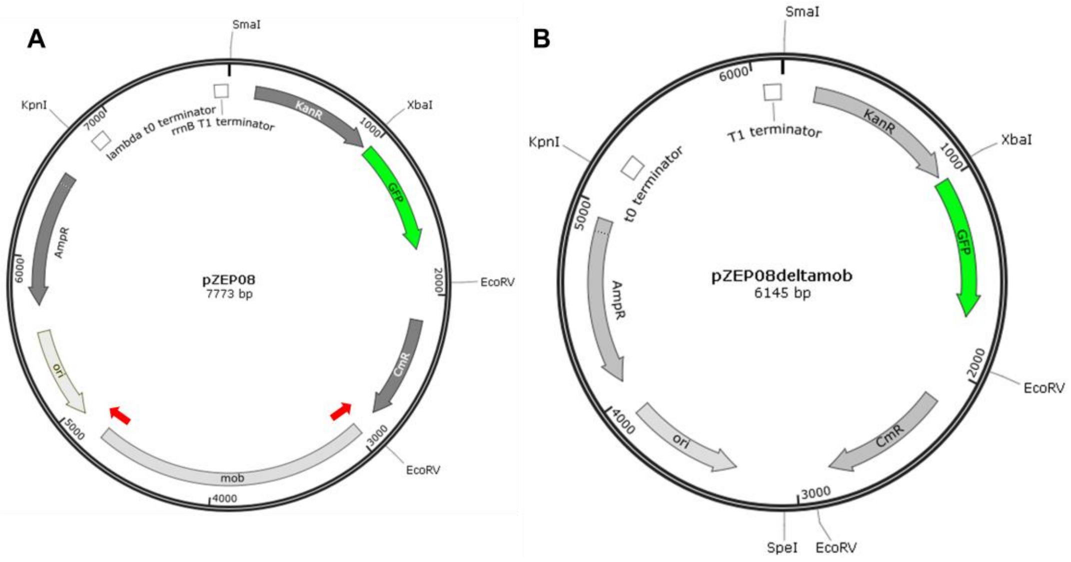


Figure 4. 1. Promoter trap vector used for expression assay. (A) Cloning vector pZEP08gfp+ for promoter trap insertions. Red arrows indicate position of the PCR primers used for the mob region deletion. (B) Cloning vector pZEP08Δmobgfp+ used for construction of metagenomic library.

4.2.2. Construction of metagenomic library

The method for library construction was adapted from a published source (Schoenfeld *et al.*, 2008) and is outlined in Figure 4.2. Briefly, viral DNA isolated from dairy wastewater was amplified in triplicate using GenomiPhi V2 DNA Amplification Kit (GE Healthcare) (the method of DNA isolation and amplification is described in Chapter 3 Section 3.2.3). Approximately 8 µg of the amplified metagenomic DNA was sheared for 3 min at 10 psi using a Nebulizer (Invitrogen), blunt end-repaired using the DNATerminator® End Repair Kit (Lucigen), and fragments between 100 bp and 1 kb were gel purified. A double-stranded DNA linker consisting of forward phosphorylated linker: 5'–p–GATTCTAGATTGTATCTGATACTGCT–3' and reverse nonphosphorylated linker: 5'–GGAGCAGTATCAGATACAATCTAGAATC– 3' was ligated to the sheared metagenomic DNA and PCR-amplified using primer 5'–AGCAGTATCAGATACAATCTAGAATC–3'. The underlined bases indicate XbaI restriction site incorporated into each linker and the primer. The PCR mixture contained, in a total volume of 50 µl, 1× High Fidelity PCR Buffer, 2mM $MgSO_4$, 2U Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen), 0.2mM dNTPmix (NEB), 25 pmol of primer, and 50 pg linker ligated DNA. For PCR amplification, an initial denaturation step of 5 min at 94°C was followed by 25 cycles of 30 sec at 94°C for denaturation, 30 sec at 57°C for annealing, and 1 min at 68°C for extension. A final extension was carried out at 68°C for 10 min. After amplification resulting fragments were digested with *XbaI* restriction enzyme (NEB) and ligated into a linearized pZEP08Δ*mob* vector. One aliquot of the ligation products (10 µl) was transformed into One Shot® TOP10 chemically competent *E. coli* cells (Invitrogen) and plated on LB agar plates containing chloramphenicol (20 µg/ml) and kanamycin (50 µg/ml) (library A). The remaining ligation mix (1 µl) was electroporated into One Shot® TOP10 electrocompetent *E. coli* cells (Invitrogen) and spread across $LB_{Kn/Cm}$ agar Q-Tray plate (22 x 22 cm; Genetix) (library B). A high throughput robotics platform $QPix2^{xt}$ (Genetix) randomly selected 5487 colonies that grew on the Q-Tray and transferred them to 96-well plates (Genetix) containing 125 µl of LB freezing buffer (LB broth containing 36 mM K_2HPO_4, 13.2 mM KH_2PO_4, 1.7 mM sodium citrate, 0.4 mM $MgSO_4 \cdot 7H_20$, 6.8 mM ammonium sulfate, 4.4% (v/v) glycerol) and library was stored at -80°C. Each 96-well microtiter plates in the library was replicated into a plate containing 125 µl of $LB_{Kn/Cm}$ broth, $LB_{Kn/Cm}$ containing TPEN and $LB_{Kn/Cm}$ containing Desferal, as described below.

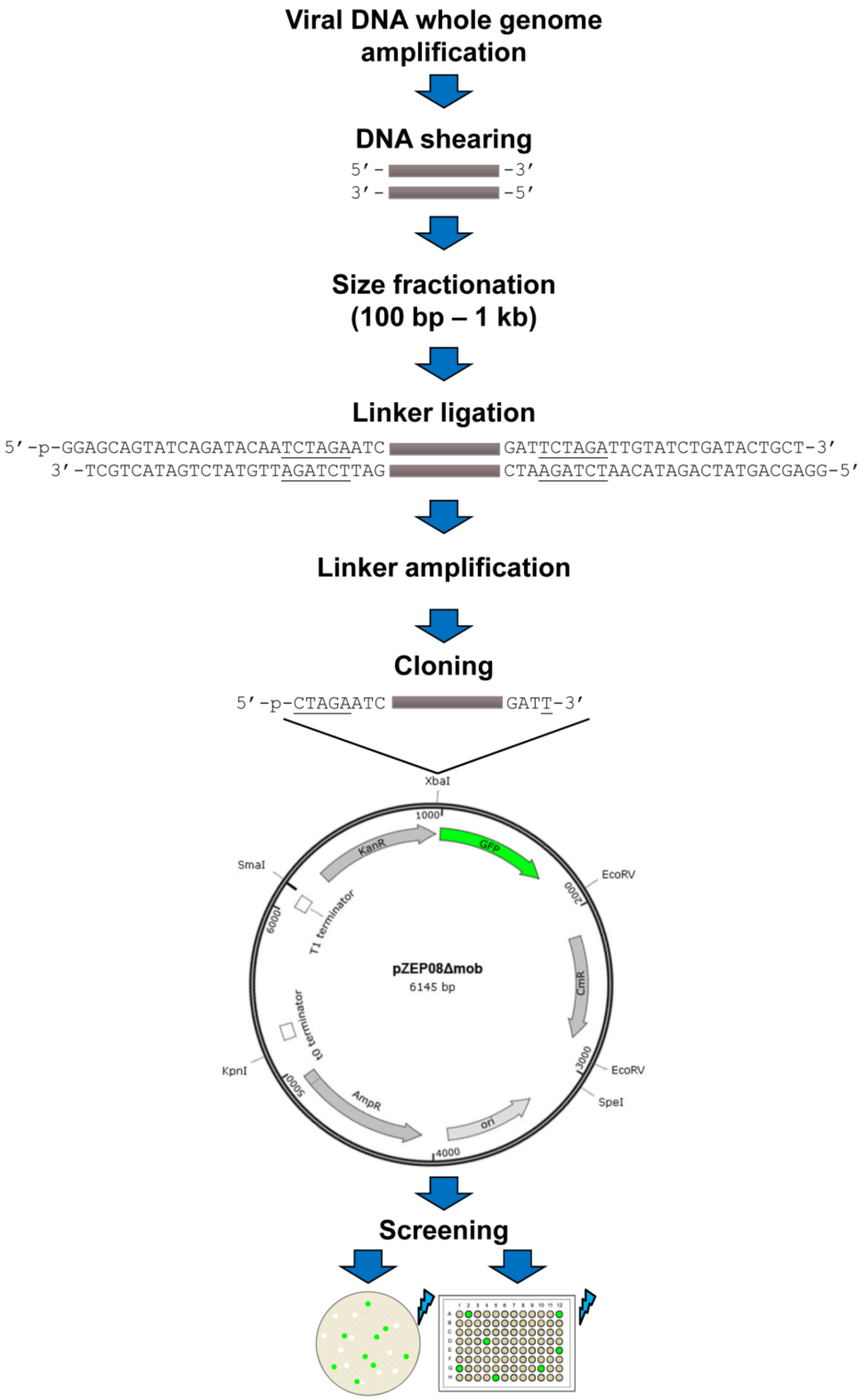


Figure 4.2. A flowchart showing the strategy of promoter trap library construction. A library of random metagenomic DNA fragments was cloned into promoter-less GFP vector and transformed into *E.coli* cells. Screening of the colonies in the presence of UV light allowed selection of clones containing promoter insert.

4.2.3. Screening of the metagenomic library

After 48 h incubation at 37°C colonies from ‘library A’ (hand plated) were viewed on a UV Transilluminator and DNA inserts from fluorescent clones were sequenced. For ‘library B’, fluorescence of clones that grew on LB broth was measured after 24 h of incubation at 37°C using a GENios XFLUOR4 microplate reader (Tecan®) at the 485 nm excitation and the 535 nm emission wavelengths, 3 flashes, gain 60 and 40 µs integration time. Colonies that exhibited fluorescence readings 2× greater than the negative control (*E. coli* containing pZEP08Δ*mob* without an insert) were counted as GFP-expressing. Library B was subsequently screened for transcriptional responses to stress with metal chelators. Metagenomic clones were cultured in $LB_{Kn/Cm}$ liquid media containing 5 µM N,N,N′,N′-Tetrakis(2-pyridylmethyl)ethylenediamine (TPEN; SIGMA), which is a zinc chelator and 15 µM deferoxamine mesylate (Desferal; SIGMA), an iron chelator. After 24 h incubation at 37°C fluorescence was measured as described above and clones that showed an increase or decrease in fluorescence intensity compared with the initial reading in non-chelating medium were selected for sequencing.

Plasmids were purified from 10 ml of overnight culture using QIAprep Spin *MiniPrep Kit* (Qiagen). In total, 21 GFP-positive clones were selected for sequencing. This included 11 clones from library A that showed visible fluorescence and 10 clones from library B that showed strong fluorescence ratio compared to the negative control and different levels of GFP expression when grown on chelating media (see Table 4.1 for more details).

DNA inserts were sequenced by GATC-Biotech (Germany) using the pZEP08-specific primers pZEP08F (5’–CCTTCTTGACGAGTTCTTCTGAGCG–3’) and pZEP08R (5’–TCACCTTCACCCTCTCCACT–3’).

4.2.4. Insert sequence analysis

Open reading frames (ORFs) were identified using GeneMark (Besemer & Borodovsky, 1999) in combination with ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) and annotated using BLASTP against NCBI protein nr database. Putative promoter sequences were predicted using

BPROM (http://linux1.softberry.com/berry.phtml) and GenomeMatScan (http://www.pdg.cnb.uam.es/icases/promscan/). Transcription terminators (stem-loops) were predicted using the mfold Web Server (http://mfold.rna.albany.edu/?q=mfold) (Zuker, 2003) and PePPER (Prediction of Prokaryote Promoter Elements and Regulons) (http://pepper.molgenrug.nl/) (de Jong *et al.*, 2012). Ribosomal binding sites (RBS) were found manually based on the following criteria: close proximity to the codon start and presence of purine bases (A and G). A Fur (ferric uptake regulator) binding motifs for *E. coli* were searched using the Virtual footprint (http://www.prodoric.de/vfp/vfp_promoter.php). Inserts were compared to the metagenomic DNA sequenced from the same sample (described in Chapter 3) using BLASTN and overlapping sequences were manually assembled. Assembled inserts were annotated using BLASTX against protein database. Unannotated protein sequences were further searched for homology detection using HHpred (Homology detection and structure prediction by HMM-HMM comparison) (http://toolkit.tuebingen.mpg.de/hhpred) (Soding *et al.*, 2005) with minimum query coverage 40% and e-value < 1.

4.3. Results

4.3.1. Screening of the promoter library

In order to identify viral promoters in metagenomic DNA derived from an activated sludge of a dairy wastewater treatment plant, a promoter-trap library was constructed. Prior to library construction, a double-stranded linker was ligated to randomly fragmented metagenomic DNA and amplified to facilitate cloning into promoter-less GFP vector. In an initial proof of principle, the ligation products were transformed into chemically competent *E. coli* and hand plated onto LB agar plates (Library 'A'). The resulting colonies were viewed under UV light for fluorescence induction. From 413 clones that grew on LB agar, 11 clones exhibited strong visible fluorescence (Figure 4.3) on inspection and those were sequenced.

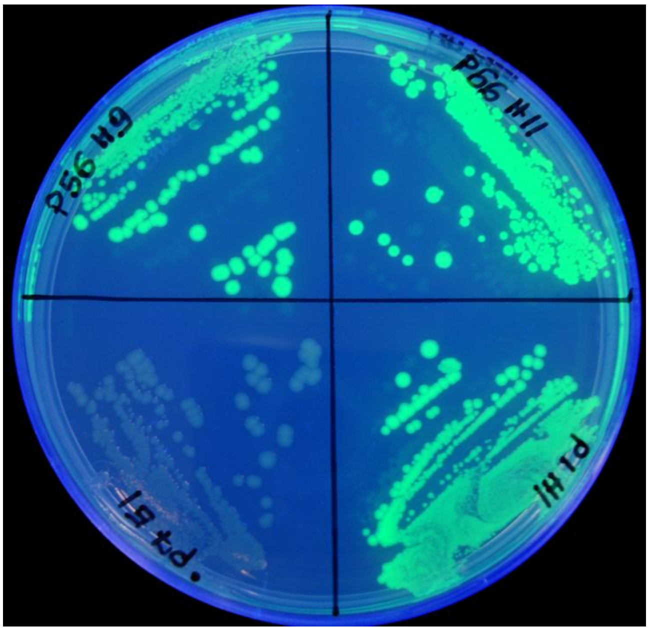


Figure 4.3. Agar plate containing GFP-expressing *E.coli* TOP10 pZEP08Δmob cells with inserts compared to the non-expressing *E.coli* TOP10 (bottom left).

To make a larger library (Library 'B'), ligation products were then cloned into electro-competent *E. coli* and plated on LB Q-tray. Approximately 5500 colonies were randomly picked from the Q-tray and GFP expression was measured fluorometrically after 24 hours of incubation in 96-well plates containing LB broth. In total, 318 clones showed over 2-fold greater GFP expression when compared to control clones not expressing GFP. To screen for clones containing promoters that may be activated under different host stressing conditions, Library B clones were cultured in iron-limiting and zinc-limiting media. No non-fluorescent clones on LB showed any fluorescence on chelating media. Although library screening did not result in detection of ON/OFF phenotypes when changing from rich media to chelating media, some clones appeared to have elevated or reduced levels of GFP expression compared to LB when grown on chelating media. In total, 73 clones on TPEN and 87 clones on Desferal showed up to 2.4-fold fluorescence change relative to LB. Ten clones with the most marked response to either of the stress conditions were selected for sequencing (Table 4.1).

Table 4.1. Change of *gfp* expression in clones selected for sequencing, grown in LB and LB supplemented with metal chelators (Desferal and TPEN).

Clone	**Fluorescence ratio to the negative control (fold)**		
	LB	**Desferal**	**TPEN**
B1F1	7.1	7.4	11.5
B1H1	17.6	14.5	OVER
B6A7*	7.5	4.9	6.0
B22C1	8.1	16.1	11.8
B43H11	14.3	9.7	5.7
B46A2	13.9	13.0	OVER
B50H12	17.3	12.6	9.5
B56H9	12.7	15.5	8.9
B56H11	OVER**	OVER	11.5
B58A11	9.9	OVER	11.2
A142/A229	5.8	N/A***	N/A
A48	10.4	N/A	N/A
A83	8.0	N/A	N/A
A15	7.6	N/A	N/A
A78	8.2	N/A	N/A
A5	10.8	N/A	N/A
A7	OVER	N/A	N/A
A42	9.4	N/A	N/A
A45	9.1	N/A	N/A
A52	5.6	N/A	N/A

* Sequencing of this clone failed.

** Over-expressed. Level of fluorescence intensity could not be determined because it exceeded the maximum recordable by the plate reader.

*** Library ‘A’ clones were not screened on chelating media.

4.3.2. Sequence analysis of the gene fusions

In total, 20 clones expressing GFP (see Material and Methods for selection criteria clones were chosen based on) and their inserts were sequenced and analysed for evidence of prokaryotic transcription and translation signals, such as σ^{70} and σ^{54}

promoters, terminators and translation initiation sequences. The DNA sequences of the sequenced clones and DNA motifs identified are shown in the Appendix. No fur (ferric uptake regulator) binding motifs were detected.

The lengths of the inserts ranged from 65 bp to 992 bp. Open reading frames (ORFs) were predicted to identify genes for which a fragment of the insert may function as a promoter. Inserts were compared to contigs (described in Chapter 3) assembled from the same metagenomic DNA that was used for GFP promoter-trap library construction and overlapping fragments were assembled to increase the chance of ORF identification. The nucleotide and predicted amino acid sequences of inserts and inserts assembled with matching contigs were compared to the GenBank and HHpred databases, and putative identification of the genes encoded was made based on sequence homology (Figure 4.5, Table 4.2). Of the 20 clones sequenced, 14 represented unique fragments. Clones A7, B46A2, B56H9, B56H11 and A142, A229, A48, B1H1 contained overlapping DNA fragments of different lengths (Figure 4.5). BLAST results demonstrated that 9 inserts (A142, A229, A48, B1H1, A7, B46A2, B56H9, B56H11 and A83) showed sequence similarity (28% to 47% identity) to the replication-associated (Rep) protein of single-stranded DNA viruses infecting plants and animals from *Geminiviridae* and *Circoviridae* families. Analysis of the conserved motifs in the N termini of Rep protein of clone A48 (which is identical to clone A142, A229 and B1H1 at the DNA level) indicate that these clones may contain novel Rep protein related to plant virus genus Mastrevirus (Figure 4.4). Additionally, one insert (A5) showed weak homology to Geminivirus coat protein. Two clones (A78 and B58A11) contained inserts with similarity to phage structural proteins and one clone (A15) had a similarity to hypothetical protein of nematode *Wuchereria bancrofti* (Figure 4.5). However, when clone A15 was assembled with an overlapping metagenomic contig, the predicted ORF showed similarity to a hypothetical protein of *Pseudomonas* sp. Ag1 (Table 4.2). Additionally, clone A15 and overlapping contig assembled into a circular DNA molecule of 1.1 kb in length. Assembly and HHpred homology search showed that the ORF predicted on clone A5 had weak homology to the PD-(D/E)XK nuclease superfamily. The remaining nine ORFs could not be annotated, but four of these were part of putative geminivirus Rep sequences in the metagenome.

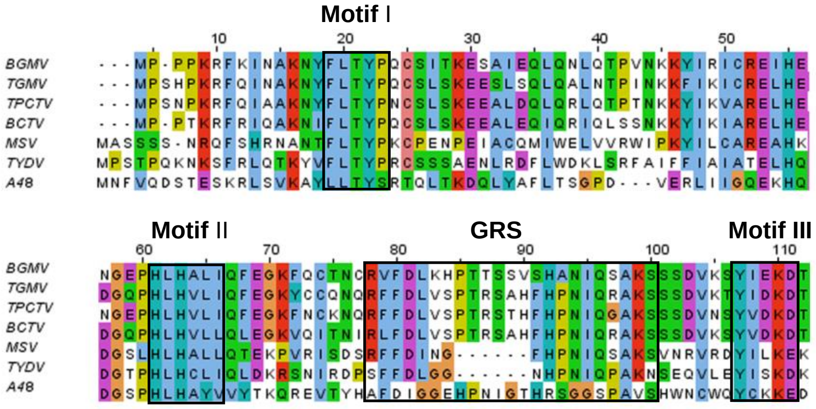
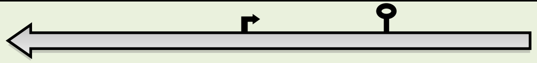
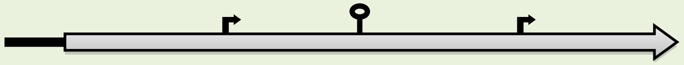
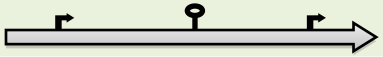
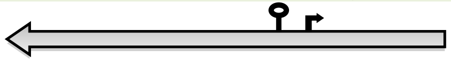
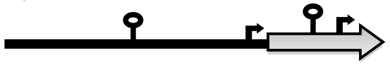
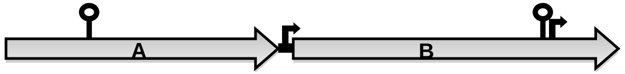
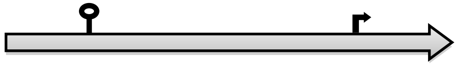
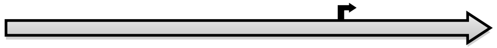
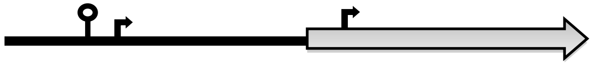
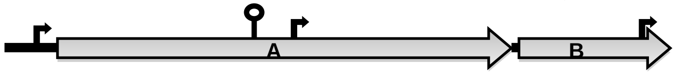
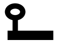
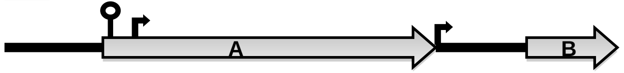
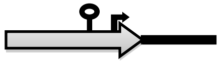
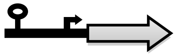
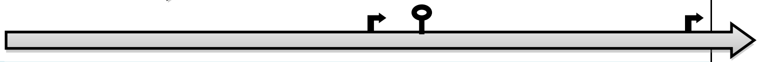
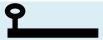
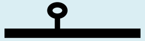
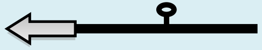
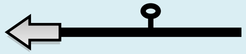

Motif I
10
20
30
40
50
BGMV ---MP-PPKRFKINAKNYFLTYPQCSITKESAIEQLQNLQTPVNKKYIRICREIHE
TGMV ---MPSHPKRFQINAKNYFLTYPQCSLSKEESLSQLQALNTPINKKFIKICRELHE
TPCTV ---MPSNPKRFQIAAKNYFLTYPNCSLSKEEALDQLQRLQTPTNKKYIKVARELHE
BCTV ---MP-PTKRFRIQAKNIFLTYPQCSLSKEEALEQIQRIQLSSNKKYIKIARELHE
MSV MASSSS-NRQFSHRNANTFLTYPKCPENPEIACQMIWELVVRWIPKYILCAREAHK
TYDV MPSTPQKNKSFRLQTKYVFLTYPRCSSSAENLRDFLWDKLSRFAIFFIAIATELHQ
A48 MNFVQDSTESKRLSVKAYLLTYSRTQLTKDQLYAFLTSGPD---VERLIIGQEKHQ
Motif II
GRS
Motif III
60
70
80
90
100
110
BGMV NGEPHLHALIQFEGKFQCTNCRVFDLKHPTTSSVSHANIQSAKSSSDVKSYIEKDT
TGMV DGQPHLHVLIQFEGKYCCQNQRFFDLVSPTRSAHFHPNIQRAKSSSDVKTYIDKDT
TPCTV NGEPHLHVLIQFEGKFNCKNQRFFDLVSPTRSTHFHPNIQGAKSSSDVNSYVDKDT
BCTV DGQPHLHVLLQLEGKVQITNIRLFDLVSPTRSAHFHPNIQRAKSSSDVKSYVDKDT
MSV DGSLHLHALLQTEKPVRISDSRFFDING------FHPNIQSAKSVNRVRDYILKEK
TYDV DGTPHLHCLIQLDKRSNIRDPSFFDLGG------NHPNIQPAKNSEQVLEYISKDK
A48 DGSPHLHAYVVYTKQREVTYHAFDIGGEHPNIGTHRSGGSPAVSHWNCWQYCKKED

Figure 4.4. Amino acid sequence alignment showing the conserved motifs in the N termini of Rep proteins of clone A48 and different members of the *Geminiviridae* family. Begomovirus: BGMV, *Bean golden mosaic virus* (AAA46312), curtovirus: BCTV, *Beet curly top virus* (AAA42751), topocuvirus: TPCTV, *Tomato pseudo-curly top virus* (CAA59223), and mastrevirus: MSV, *Maize streak virus* (AAK73446) and TYDV, *Tobacco yellow dwarf virus* (AFD63094). The amino acid sequences were aligned using ClustalW (Thompson *et al.*, 1994). The conserved motifs are boxed in black. Motif I (FLTY) is required for specific dsDNA binding, motif II (HLH) is a metal-binding site that may be involved in protein conformation and DNA cleavage, GRS motif is required for initiation of rolling-circle replication, motif III (YxxKD/E) is the catalytic site for DNA cleavage (Nash *et al.*, 2011). The Geminivirus sequences used for comparison were as in reference (Nash *et al.*, 2011). The Rep protein of TYDV virus was the closest homolog of clone A48 (sharing 30% of amino acid identity) as determined by BLASTX against NCBI protein database.

4.3.3. Identification of putative promoters

Putative promoters were predicted using a bacterial σ^{70} promoter recognition program (BPROM) and the sequences of the identified -35 and -10 promoter elements are shown in Table 4.3. No σ^{54}-like promoters were predicted using GenomeMatScan. Of the 23 predicted σ^{70} promoters, 17 were found to lie within the ORF and 6 were located upstream of the ORF (Figure 4.5). Five clones had no predicted promoter sequences.

Clone	Insert size (bp)	ORF BLASTP analysis (% coverage; % identity)	Structural organization of the insert
A142/ A229	659	Geminivirus rep protein (85%; 30%)	
A48	889	Geminivirus rep protein (85%; 30%)	
B1H1	490	Geminivirus rep protein (85%; 29%)	
A83	585	Circovirus rep protein (75%; 39%)	
A15	751	*Wuchereria bancrofti*, hypothetical protein (64%; 33%)	
A78	814	A: *Methylophilales* phage HAM624-A, hypothetical protein (96%; 31%) B: *Bradyrhizobium* sp., phage major head protein (98%; 52%)	
B58A11	592	*Rhodovulum* sp., P22 coat protein (98%; 48%)	
A5	640	No hit [Geminivirus coat protein, e-value 0.19, coverage 43%, identity 21%]**	
A42	778	No hit	
A45	883	A: No hit B: No hit	
A52	65	No hit	
B1F1	805	A: No hit B: No hit	
B22C1	280	No hit	
B43H11	224	No hit	
B50H12	992	No hit	
A7	120	No hit *	
B46A2	185	No hit *	
B56H9	341	No hit *	
B56H11	323	No hit *	

* Assembly of the insert with the overlapping contig revealed matches to Geminivirus Rep protein (see Table 4.2)

** ORF homology predicted using HHpred (see material and methods)

Figure 4.5. Sequence analysis of the sequenced clones. ORF were identified with GeneMark and annotated using GenBank protein database. Grey arrows represent the direction of putative ORFs with the putative bacterial promoter(s) (black arrow), and putative transcriptional terminators (lollipop). Non-coding regions are shown as black lines. Groups of inserts that contain identical DNA fragment have been highlighted in blue and green.

Sequence analysis of inserts with no predicted bacterial promoter (clone A7, B46A2, B56H9 and B56H11 containing overlapping fragments of different lengths, and clone A52) revealed the presence of terminators with G+C-rich stems (Table 4.3). Stem-loops of clones A7, B46A2, B56H9 and B56H11 were located in an intergenic region immediately upstream of the Geminivirus *rep* translation start position (Figure 4.5) and contained a nonanucleotide motif TACGGGGTA. Among clones that had a promoter located within the ORF were five clones (A142, A229, A48 and B1H1 containing overlapping fragments of different lengths and orientation, and clone A83) with a putative prokaryotic promoter located within the Rep ORF.

Three clones (A42, A15 and B43H11) contained inserts with potential phage-specific early transcription promoters characterized by presence of the σ^{70}-like promoter upstream of the ORF and stem-loop structure (Figure 4.6). High sequence similarity to the *E. coli* σ^{70} promoter consensus indicates that these promoters should act as phage early promoters, and it is no surprise that they drive expression in *E.coli*. In particular, clone A42 contained a stable terminator with G+C-rich stem, a clear σ^{70}-like promoter (TTGACA-nt17-TATTAT) with a -35 element sequence identical to the TTGACA consensus sequence in *E. coli* and a putative ribosomal binding site (Shine-Dalgarno) identical to the AGGAGG consensus sequence in *E. coli*. The ORF encoded by this clone was incomplete (124 amino acids long) and showed weak homology to PD-(D/E)XK nuclease superfamily protein (Table 4.2).

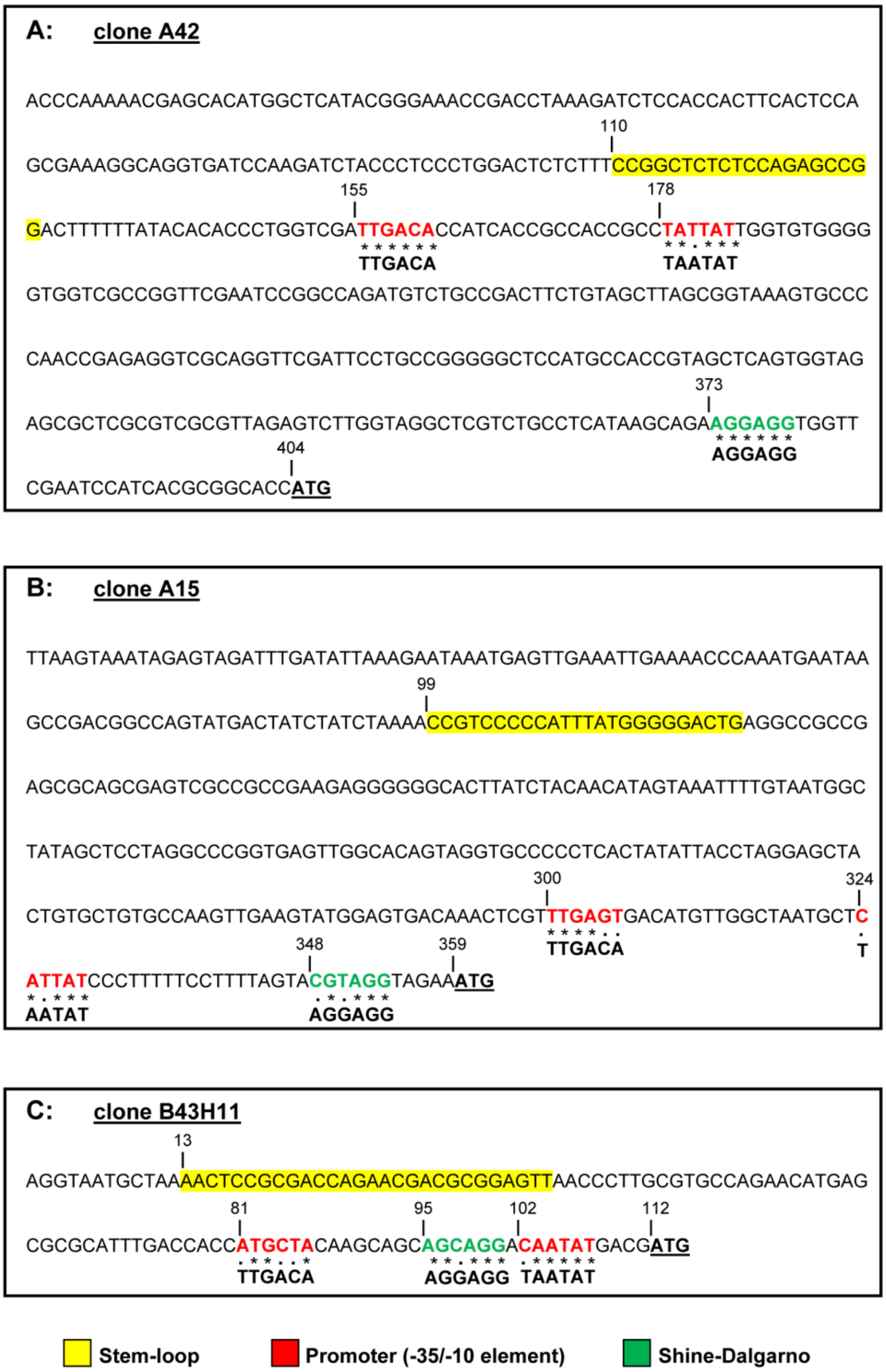

A:
clone A42
ACCCAAAAACGAGCACATGGCTCATACGGGAAACCGACCTAAAGATCTCCACCACTTCACTCCA
110
GCGAAAGGCAGGTGATCCAAGATCTACCCTCCCTGGACTCTCTTTCCGGCTCTCTCCAGAGCCG
155
178
GACTTTTTTATACACACCCTGGTCGATTGACACCATCACCGCCACCGCCTATTATTGGTGTGGGG
* * * * * *
TTGACA
* * . * * *
TAATAT
GTGGTCGCCGGTTCGAATCCGGCCAGATGTCTGCCGACTTCTGTAGCTTAGCGGTAAAGTGCCC
CAACCGAGAGGTCGCAGGTTCGATTCCTGCCGGGGGCTCCATGCCACCGTAGCTCAGTGGTAG
373
AGCGCTCGCGTCGCGTTAGAGTCTTGGTAGGCTCGTCTGCCTCATAAGCAGAAGGAGGTGGTT
* * * * * *
AGGAGG
404
CGAATCCATCACGCGGCACCATG
B:
clone A15
TTAAGTAAATAGAGTAGATTTGATATTAAAGAATAAATGAGTTGAAATTGAAAACCCAAATGAATAA
99
GCCGACGGCCAGTATGACTATCTATCTAAAACCGTCCCCCATTTATGGGGGACTGAGGCCGCCG
AGCGCAGCGAGTCGCCGCCGAAGAGGGGGGCACTTATCTACAACATAGTAAATTTTGTAATGGC
TATAGCTCCTAGGCCCGGTGAGTTGGCACAGTAGGTGCCCCCTCACTATATTACCTAGGAGCTA
300
324
CTGTGCTGTGCCAAGTTGAAGTATGGAGTGACAAACTCGTTTGAGTGACATGTTGGCTAATGCTC
* * * * . .
TTGACA
.
T
348
359
ATTATCCCTTTTTCCTTTTAGTACGTAGGTAGAAATG
* . * * *
AATAT
. * . * * *
AGGAGG
C:
clone B43H11
13
AGGTAATGCTAAAACTCCGCGACCAGAACGACGCGGAGTTAACCCTTGCGTGCCAGAACATGAG
81
95
102
112
CGCGCATTTGACCACCATGCTACAAGCAGCAGCAGGACAATATGACGATG
. * * . . *
TTGACA
* * . * * *
AGGAGG
. * * * * *
TAATAT
🟨 Stem-loop
🟥 Promoter (-35/-10 element)
🟩 Shine-Dalgarno

Figure 4.6. Phage-specific promoter early stem-loop like regulatory cassettes identified in clone A42 (A), A15 (B) and B43H11 (C). The putative -35 and -10 promoter elements (red) and the putative Shine-Dalgarno (green) were aligned with *E. coli* consensus sequences. The predicted terminator is highlighted in yellow and putative start codon is underlined.

Table 4.2. BLASTX (e-value < 0.001) results of assembled insert with overlapping contig assembled from dairy wastewater metagenome.

Clone (bp)	**Contig (bp)**	**Overlapping fragment length (bp)**	**Overlapping fragment identity (%)**	**Insert/Contig BLASTX (bp length, % query coverage, %identity)**
B1F1 (490)	ctg1540 (1143)	290	100	*Candidatus Glomeribacter gigasporarum*, dCTP deaminase* (1658, 18%, 65%)
B1H1 (490)	ctg7701 (299)	299	100	Geminivirus, replication protein (490 bp , 49%, 36%)
A229/A142 (659)	ctg7701 (299)	281	100	Geminivirus, replication protein (677 bp, 85%, 28%)
A48 (889)	ctg7701 (299)	299	100	Geminivirus, replication protein (889 bp, 77%, 28%)
B46A2 (185)	ctg6947 (335)	40	100	Geminivirus, replication protein (480 bp, 34%, 47%)
B56H9 (341)	ctg6947 (335)	196	100	Geminivirus, replication protein (480 bp, 34%, 47%)
B56H11 (323)	ctg6947 (335)	178	100	Geminivirus, replication protein (480 bp, 34%, 47%)
A83 (585)	ctg318 (3332)	585	100	Circovirus, replication protein (3332 bp, 24%, 34%)
B58A11 (592)	ctg1323 (1264)	592	99	*Rhodovulum* sp., P22 coat protein (1264 bp, 88%, 51%)
A78 (814)	ctg8352 (275)	275	100	*Bradyrhizobium* sp., phage major head protein (814 bp, 47%, 52%)
A15 (751)	ctg3362 (648)	213	99	No hit (1110 bp, circular sequence) [251 aa, 61%, 33%, *Pseudomonas* sp. Ag1, hypothetical protein]**
A5 (640)	ctg5657 (410)	410	99	No hit (640 bp) [Geminivirus coat protein, e-value 0.19, coverage 43%, identity 21%]***
A42 (778)	ctg4251 (528)	412	100	No hit (894 bp) [PD-(D/E)XK nuclease superfamily, e-value 0.003, coverage 89%, identity 20%]***
A45 (883)	ctg5347 (431)	305	99	No hit (939 bp)
B22C1 (280)	ctg708 (1995)	280	99	No hit (1995 bp)
B43H11 (224)	ctg5505 (419)	224	99	No hit (419 bp)
B50H12 (992)	ctg8371 (274)	175	100	No hit (1091 bp)

*	ORF not overlapping with the insert sequence
**	Identified using BLASTP
***	ORF homology predicted using HHpred (see material and methods)

Table 4.3. Putative bacterial promoters and transcription terminators predicted in clones expressing GFP. Scores with weights to the predictions are given by BPROM.

Clone	-10 element	Score	-35 element	Score	Stem-loop	ORF annotation
A7	None	-	None	-	GGGGGTACGGGGTACCCCC	Geminivirus, replication protein
B56H9	None	-	None	-	GGGGGTACGGGGTACCCCC	Geminivirus, replication protein
B56H11	None	-	None	-	GGGGGTACGGGGTACCCCC	Geminivirus, replication protein
B46A2	None	-	None	-	GGGGGTACGGGGTACCCCC	Geminivirus, replication protein
A52	None	-	None	-	GGGGGCCGGCAAGGCCCCC	None
A142/A 229	CCTTAGAAC	20	TTGCAG	49	ACCTTTAATCAAAGGT	Geminivirus, replication protein
A48	AGGTAACCT CTCTATGCT	54 51	TTTACA TTTCAG	47 30	ACCTTTGATTAAAGGT	Geminivirus, replication protein
B1H1	AGGTAACCT CTGCATACG	54 9	TTTACA TTTACT	47 42	ACCTTTGATTAAAGGT	Geminivirus, replication protein
A83	TGTTAAGCT	60	GAGCCG	-15	GCGATGTCTGACATCGC	Circovirus, replication protein
A5	CTGCAACCT	22	TTGAAG	54	None	Geminivirus, coat protein
A15	GCTCATTAT TTCCAAGAT	36 32	TTGAGT TTGAAG	38 54	CCGTCCCCCATTTATGGGGGACTG CGAATGAATTCG	*Wuchereria bancrofti*, hypothetical protein
A42	GCCTATTAT CGCTATCTT	57 46	TTGACA TTGAAG	66 54	CCGGCTCTCTCCAGAGCCGG	PD-(D/E)XK nuclease superfamily
A45	GGATAAAAT TGTTGTTAT ATTTATGGT	72 41 41	TTATCG TCGACG TCGATT	24 23 16	GTTGTCGATGCGCTGGACGCAATCGGCAAG	None
A78	GACTAAAGT CGGCAAAGT	40 33	TTTAAG ATGACA	35 36	CGATGCAATACGACCGCATCA GACATGATGTC	*Bradyrhizobium* sp., phage major head protein
B1F1	GGCTATCTT CGCTATGCT	47 58	TTGAGG TGAACA	37 6	GGCGAAACTAGGGCGCC	None
B22C1	GAGGAAGAT	18	TTGACT	61	CCGTTGGTGATCAACGG	None
B43H11	GGACAATAT	32	ATGCTA	23	AACTCCGCGACCAGAACGACGCGGAGTT	None
B50H12	TTGTAGAGT GAATACATT	46 29	TTGCCA TTGATT	61 53	ACATTGCCTACAACAATGT	None
B58A11	CGGCAAACT	47	TTGCCA	61	CACCGCGAATAAGCGGTG	*Rhodovulum* sp., P22 coat protein
***E. coli* consensus**	**TAATAT**	**89**	**TTGACA**	**66**		

4.4. Discussion

Metagenomic samples contain a mixture of different genes derived from hundreds or even thousands of different species, mostly unknown or novel. The functional study of metagenomic DNA can be useful for assessing DNA sequence function. This chapter describes the construction and large-scale screening of a promoter-trap metagenomic library to identify virus-specific regulatory sequences able to drive GFP expression in *E.coli*. The DNA used for promoter screening was obtained from a wastewater sample, which was processed for virus particles isolation through a 0.22-µm filtration, CsCl gradient ultracentrifugation and DNAse treatment. Sample prepared in such a fashion should contain all viruses present in this environment – including prokaryotic as well as eukaryotic viruses. Metagenomic sequencing and sequence analyses described in Chapter 3 demonstrated that indeed wastewater DNA was overrepresented by sequences of phage, prophage and viral origin. Cloning of short fragments of this DNA upstream of the promoter-less *gfp* reporter gene resulted in increased expression of green fluorescent protein in approximately 6% of the screened clones, suggesting they contained functional promoter sequences. Expression of the *gfp* gene in *E.coli* from foreign non-bacterial promoters is possible because viruses and phages contain regulatory sequences that are recognised by *E.coli* biosynthesis machinery (Lewin *et al.*, 2005; Warren *et al.*, 2008). A promoter trapping strategy has been successfully used for detection of promoters, mainly from bacterial DNA (Dunn & Handelsman, 1999; Han *et al.*, 2008; Lee *et al.*, 2011) and no such work has been described on metagenomic viral DNA.

Screening of the small insert metagenomic library revealed that of 20 selected clones that showed strong promoter activity, 10 clones (50%) were derived from single-stranded (ss) DNA viruses (Table 4.3). Nine of these inserts (representing three unique inserts, see Figure 4.5) were identified as plant-infecting ssDNA viruses from the *Geminiviridae* family. Eighteen of the 20 clones were represented in the metagenome assembly, including all the putative Geminivirus sequences (Figure 4.5). Four of the Geminivirus sequences were only identifiable by similarity to longer contigs in the metagenome assembly. Eukaryotic promoters can be recognised by the prokaryote transcription system (Jacob *et al.*, 2002) because RNA polymerase is evolutionarily conserved among prokaryotes and eukaryotes (Allison *et al.*, 1985; Ebright, 2000; Fassler & Gussin, 1996). Previous studies have shown that viral promoters can direct gene transcription in *E. coli* (Jenkins *et al.*, 1983; Lewin *et al.*, 2005; Li *et al.*, 2011a; Mitsialis *et al.*, 1981). Therefore, it is not surprising that viral promoters cloned from the wastewater metagenomic DNA can initiate transcription of the GFP protein in the *E.coli* TOP10 strain used for the expression study. The fact that the majority of identified sequences detected in the promoter screen belong to a group of ssDNA viruses (circoviruses and geminiviruses) may be the effect of amplification bias. Viruses with circular ssDNA genomes replicate via a rolling-circle amplification mechanism initiated by viruses-encoded replication initiation protein (Rep) and are highly detected in metagenomic studies of viral diversity, which use rolling-circle amplification technique to obtain sufficient DNA amount prior to sequencing (Kim *et al.*, 2008; Lopez-Bueno *et al.*, 2009; Rosario *et al.*, 2009a; Rosario *et al.*, 2009b).

Nine inserts contained predicted promoters associated with the replication initiator (Rep) protein of ssDNA viruses (Figure 4.5). There were two predicted promoter locations identified in putative Rep proteins, one inside the gene and the other upstream (Figure 4.5). Four of these inserts (clones with identical sequence of different length: A7, B46A2, B56H9 and B56H11) contained a predicted promoter region derived from the intergenic region of the Rep protein of Geminivirus. Five inserts (clones with identical sequence of different length: A142, A229, A48, B1H1 showing sequence similarity to Geminivirus and clone A83 showing similarity Circovirus) contained a predicted prokaryotic promoter located inside the Rep protein. These promoters may be involved in regulation of the transcription of the capsid protein, as previously shown in Circovirus (Mankertz & Hillenbrand, 2002). Because of technical limitations in virus purification, including inefficacy of DNase I in removing free ssDNA templates, it has been suggested that the presence of bacterial or eukaryotic sequences in viral metagenomic libraries cannot be excluded (Rosario *et al.*, 2012). Rep-like sequences of circoviruses and geminiviruses have recently been shown to be very widely distributed in eukaryotic host genomes, suggesting widespread insertion of genes from circular single stranded DNA viruses into host genomes in evolution, with evidence of expression of viral genes by the host (Liu *et al.*, 2011). However, genes similar to geminivirus Rep sequences have been reported from eubacterial and algal plasmids (Rosario *et al.*, 2012). Some features increase the confidence in the viral origin of small circular genomes detected in metagenome sequence data, such as predicted Rep protein of clone A48, which showed conserved motifs similar to those found in Rep proteins of viruses from *Geminiviridae* family (Figure 4.4). Promoters derived from ssDNA viruses have wide application in study of gene expression in transgenic plants (Dugdale *et al.*, 1998; Shirasawa-Seo *et al.*, 2005) and mammals (Tanzer *et al.*, 2011). Therefore viral promoters identified in this study could be useful in the construction of novel expression vectors, especially as they showed high expression activity (Table 4.1).

At least three clones were characterized by the presence of phage-specific promoter early cassettes (Figure 4.6), similar to those previously identified in T4-type phage genomes (Arbiol *et al.*, 2010). These cassettes are composed of nonessential genes flanked by σ^{70}-like promoters and stem-loop structures, and are thought to be involved in horizontal gene transfer (Arbiol *et al.*, 2010; Cornelissen *et al.*, 2012).

The work described in this chapter show preliminary results on the potential use of promoter trap vector pZEP08Δ*mob* constructed in this thesis to isolate virus-derived regulatory regions. Promoter-trapping approach has not previously been applied for identification of promoters from a viral metagenome. The results provide a proof of principle that metagenomic DNA fragments obtained from the clone library can be functionally screened for virus and phage promoters. Most of these DNA fragments were present in the metagenomic sequence assembly described in previous Chapter 3, in which DNA from the same source was directly sequenced, omitting the cloning step. Availability of the metagenome assembly assisted identification and classification of the cloned inserts. Screening of the clone library on media imitating host environmental stress conditions such as low iron or zinc did not result in detection of sequences containing metal related regulatory sequences or promoters that would selectively induce gene expression. Previous studies demonstrated that these media can be used for identification of bacterial genes that are differentially expressed in response to metal chelators (Himpsl *et al.*, 2010; Sigdel *et al.*, 2006), and the preliminary screening assay used was relatively insensitive. A comparative assay of relative GFP production would be worth applying to this type of media. Future work in needed to screen a larger number of positive clones. Further characterization of viral promoters derived from metagenomic DNA could be directed to detection of sequences whose transcription is regulated in response to different environmental conditions such as anaerobic conditions or the presence of antimicrobials. To characterise the putative promoter sequences experimentally will require further work. To assess a large number of putative promoters in parallel, a high thoughput experimental confirmation method could be devised by the application of ChIP-seq (Mardis, 2007) and RNA-seq (Wang *et al.*, 2009d). This would allow determination of the transcription start site, RNA polymerase binding site and transcription factor binding sites.

Appendix

DNA sequences of clones sequenced from the GFP promoter-trap metagenomic library. The predicted open reading frames (ORF), promoters, terminators (stem-loop) and ribosomal binding sites (RBS) have been highlighted.

Colour codes:

- ■ Putative ORF
- 🟦 Putative -35/-10 promoter
- 🟨 Putative stem-loop
- 🟩 Putative RBS

GFP library:

>**A5**

CGAGCTATATGCGCAGCGCTCGGAGTAGCGCAAAGCTCGCGCCAGAAACGAAGTACTTCGATACTACG
TTCTCGGCGAATGTGGACTCGGCCGAAGACTGGGCCACAACGAGTGTACCTATGACCTCGTACATCAA
TAGTGACGGAGTCACTGTATCTGGTTACACCGACCGTGCCCTCATTCCCAGTGCCGTAGGGTCTGGGT
ACGGGCAGGTCGTAGGCACGAAGTACCTACTCAAGCGGCTTGCGGTCAAAGGCGAGATTTTCTCGAAT
CCCGCCCAGGACCAAGCCGATGTGCTTGGGTCCCGTACAGTCCGATGTGTCCTAGTTATGGACACTCA
GCCCAATGGTGCACAGGCCACAGGAGATCTTGTGTTCACAGATCTCGGCCTTGCGACCAATTGCAACC
ACTCCTTTATAGCCATGGGAGCTGCCGGAAACGGCCGCTTCCGTGTCTTGAAGGACAAGACGTTCCTG
CTGCAACCTGCCGTTGCAGGAACTGATGGTGCGAACACCAACTCGCAGACTCACAATGGGGCTCTTGT
GAAGATGACGTACAGCCCGAAGAAGCCGCTTGCGGTTCGTATTCGAGGCAGTTCTGCCACCCCCACTG
TGGCCAGCTTGACAGATGTGAACATCTT

>**A7**

GCACGGGGGTACGGGGTACCCCCTCCCGTTCATGGGTTGCCGGTGGCACCCCATGGGGGGGCACGGGG
AGGCTGCGCCCCCCCTCTGGGGAAGTCGGTGACGTTCCCCACAGTTGAGTCC

>**A15**

TTAAGTAAATAGAGTAGATTTGATATTAAAGAATAAATGAGTTGAAATTGAAAACCCAAATGAATAAG
CCGACGGCCAGTATGACTATCTATCTAAAACCGTCCCCCATTTATGGGGGACTGAGGCCGCCGAGCGC
AGCGAGTCGCCGCCGAAGAGGGGGGCACTTATCTACAACATAGTAAATTTTGTAATGGCTATAGCTCC
TAGGCCCGGTGAGTTGGCACAGTAGGTGCCCCCTCACTATATTACCTAGGAGCTACTGTGCTGTGCCA
AGTTGAAGTATGGAGTGACAAACTCGTTTGAGTGACATGTTGGCTAATGCTCATTATCCCTTTTTCCT
TTTAGTACGTAGGTAGAAATGGGTGTTTTTCGTCGTAATCCTCCTCGTTCTGCTAAGAGGGCTAAGAA
GTCGCCATACAAGGCGCGTATGTCCGTCTATAAGAAGCCTCCTTCGGATGATTCTATGTATCGGAGTC
CTATTCCGCGTATTAGGGGATCTGATTTCGGATTTCCTGATAAGCTTGTAACTAACTTGCGTTATGTT
GATACGTTTCGATTGACAGGTAACGCAGGTGTTCCTGGTTCAAATGTGTTTCGAATGAATTCGTTGTT
CGATCCGGACCTGTCGGGTATCGGTCATCAACCAATGTATTTCGATCAGTTGTGTGGTGCTGCGGGCA
CTGGTCCATATTTGAAGTATCGTGTTCTTGGTTCCAAGATTACGGTTAAGTATTGTGTCGAGAATGCG
CCT

>**A42**

```
ACCCAAAAACGAGCACATGGCTCATACGGGAAACCGACCTAAAGATCTCCACCACTTCACTCCAGCGA
AAGGCAGGTGATCCAAGATCTACCCTCCCTGGACTCTCTTTCCGGCTCTCTCCAGAGCCGGACTTTTT
TATACACACCCTGGTCGATTGACACCATCACCGCCACCGCCTATTATTGGTGTGGGGGTGGTCGCCGG
TTCGAATCCGGCCAGATGTCTGCCGACTTCTGTAGCTTAGCGGTAAAGTGCCCCAACCGAGAGGTCGC
AGGTTCGATTCCTGCCGGGGGCTCCATGCCACCGTAGCTCAGTGGTAGAGCGCTCGCGTCGCGTTAGA
GTCTTGGTAGGCTCGTCTGCCTCATAAGCAGAAGGAGGTGGTTCGAATCCATCACGCGGCACCATGAA
CACCGAATATCCAAAATTGATTCTTCCGAAAGGGTATCTTTCGTGGTCTCAAATTGATTGCTGGATGA
AAAACCCAGGACGCTATGTGCGTGAATACTTTGAAGCAGGAGAACGACTCGACACCCGCTATCTTCGC
TTTGGATCTAAATTTTCAAAGATGGTCGAGCGGTTGTGTGAACTTATGGATCAGTTTCCGGACCGTGC
AACCGCGGTGATGGAGCTTGCTAAAGAGCACCCAATGGATGAGAACATGCAGAGCGTTCTCATGGAGC
TCGATATTGAGGGGACATCAGAGTTCCAAATTGGAAACTCTGGGCGGAAAGAGGACACCAACCCCGTC
GTTAAGGTGCGCGGAATCGTTCCGATTCTT
```

>**A45**

```
ATCAACATTACCTCTAACAACCGCCTCATTTCGATTGGTCCACAAAGAGGCGGTATTTATGGTTGTGA
GCGCATGGATATGGCTCACGTTTGGCAATACAAAAATAAGTCGGGTCAGCAGGTGCAATTTTTTCCTC
GAGTCGTGCCGATCCCGGAAAGTGTACCGGTGCCTCCTGCTGATGTTGATACAGGATTAGGGCCGGGG
CGGCGTCCTCGTCCGCGCCCGGAACCTTATGAAGACCCGATTATTTCTCGGTGGCCGAAAGATCATCC
CGTTCCGGAACCCTATCATAAACGGGTGTCTCCTGAACCCGGTATGAAGGAGAAAAAGCTTAAGCTTG
GAAAGGGGGGTTTAATTGGTAGTGTGTACGGTGCGGTTACGGAAGTTGTCGATGCGCTGGACGCAATC
GGCAAGGCGCTTGATCCTAAGGTTCGTGGCGGTTATCGCCAACAGAAAACTGTACAGGATAAAATTCG
GTATCTCCTGCGCAATTGGCGTTATATCGATCATAACGCGGCGTTGGGTAATATTATTTTGGATCAAG
TCGAGGATTGGTTCATTGGCAAATCAAATCAACTTGCCAATCGCACCGCTTCTCATCCATACTGGCGC
GGCATGCGAGGTCCGAACGTCTCTCGCATGCCGGGGCGAATTCCCGCCCCTCAACTGTAGAAAGGTAG
GTCAATGGCTTATTACCGCAGGCGCCGTCGCACTAGCTATCGCTCCCGGCGCGGCTACTCCGCACGGT
CCGGTTATCGGATGCGGTCTTATGGGCGACGCTCTTATGGTCGTCGTTATTCGAGGCGTCGCAGGTCG
ACGGCTCGGGCACAGCGTGTTGTTATTCAGGTCATCGGTGGTCCGGGCGGTGTAGCAACGTCGCCCG
```

>**A48**

```
TGTCTGCTGCTCCCCCCTCACGACGAAATGTGCAGACTGATTCATCTGGATGCCATGCAGTGTGCTGA
ACTACATTCAGAATGAATTTTGTTCAGGACTCGACAGAGTCAAAGCGACTGAGTGTGAAGGCGTACCT
TCTCACGTACAGTCGGACCCAGTTGACGAAAGATCAACTGTATGCTTTCCTGACAAGTGGACCAGATG
TGGAGCGTCTGATCATCGGGCAGGAAAAGCATCAGGACGGTAGTCCCCATCTCCATGCTTACGTTGTT
TACACGAAGCAGAGAGAGGTAACCTACCATGCGTTCGACATTGGTGGTGAACACCCCAATATTGGAAC
GCATAGGTCGGGTGGTAGCCCCGCAGTAAGCCACTGGAACTGTTGGCAATACTGCAAGAAGGAGGATT
CGGAACCTTTGATTAAAGGTGATCCTCCTACCGAACCTCCCCCATCTCGCAAGCGGCCTGCGGACGGG
GAGGTATCGAAGGCGAAGCGTAGCAAGAAGGACGACTTAGTCCGCACTTGTATGGAAATCGCAAAGGA
CCCTGAACGCAGTTCTAAGGAAGCCTTCGATTTACTGCAAGACCGTATGCCTGCATACGCGGTCGAAC
GAGCCAATGCATATCGTATGGAGTTTCAGCGCATTCGGAGCGAAGCTCTATGCTATGAAGCTCCGGCA
CGCCCCCTGTCCGAATTTGCGCGGGCTCCGAAGGTGCTCCCAAATTGGCGGACGCTCTACATCTACGG
GCCCACGAAGTTTGGCAAAACAGAGTATGCCCGTGCACTGCTCCCGGGTGCCGAAGTCATTCGCCACC
GGGATCAGCTCAAAGACGCAGACATGTCGAAAGGCCTCATCTTCGATGACTTTGAGACGTGTCACTGG
CCAAT
```

>**A52**

```
GGGGGCCGGCAAGGCCCCCGGCCGCAGGTGACGGTGCGCAACGCGCTGAACAACAAATGGTTCAG
```

>**A78**

```
TTTGTTGCTGATGTTCCTGAACTAACCGACCCCGCAAAGGCCGGTCAGGTTCTGTCGGATATTGTCAA
CTATGCAAAACAGGCGGGCGTTCCAGAAAGCGTATTTGAGGCTGAAAACCTCAACGCGATTACGTCTG
CCGAACTGCATCTCGCGTGGAAAGCGATGCAATACGACCGCATCAAGAGCGCGCAAGGCGAGGTGAAG
AAAACACCCGCGCCGAAGCCCGCCCAGCCAGCAGTAAGGCCCGGTGTAGCCATTCCAAGGTCTGCAAC
CAAAGCAACGGCAGTTCGGAAAGCAAATGAACGATTGGCCGCAGAGGGCAGCATCGAAGCCGGTGCCG
CTGTTTGGAAAAACTTTCTTTAAGGGATTTTTGAAATGACTAAAGTTACTGGCGCGATGGCGACGTAT
GACGTCACCACGAACCGCGAAGATTTGGCAGATGCGGTTTACCGTATTTCGCCCGCAGACACCCCGTT
TATGTCGGCAGTTCCCCGCGTGAAAGCGACGGCTGTTCTGCATGAATGGTCAACCCATGCGCTATCGA
GCATCAACACGACTAACGCCCGCCTTGAAGGTGACGCGCTGACCCGTGTTGCATCGACCGCGCCCGTT
CGCCGTCAAAACTACTGCCAGATTTCAAGCCGTGATGCGACTGTGACTGGCACACAGCGCGCTACCAA
TCCGGCGGGCATTGATGACATGATGTCTTTCCAGATGTCGGCAAAGTCGCTTGAACTGCGCCGCGATA
TGGAAGCCATCCTTCTGGGTAACACCGGACAGACGGCGGGCAACACCACCACTGCACGGACCCTGC
```

```
>A83
TCATGCCGCCAGCGTGATAATCGTCGAATATCACACAGGATTCGTAGTTGTAGTTGTCGAAGCGGATG
CCGTTGGCCTTGGCCTGCGGAGCCCAATAGGTCGGCTGATCGCCAGCGTGCACGAGAGCAGACGAGGT
CTTGCCAGTCCCTGATGGGCCCCAGATCCAAACAACGACCATGCCATCATGGGTCTCGCGGGTGCGGT
GTTGAATCCGGAGATCGCGGTACTTGCTGAAGCCGCGCTCGTACTTGATGAACTCGGAAAAATGATCC
TTCGCGATGTCTGACATCGCTTTGTTGGCATCGAGCTCTTGTTGAACGACGAGAAGGTCTGTGCGACC
TCCGCGCTTGCCAGCGCGTGGCTTAAGTTCGCCCCACATATGGGGTCCCGCAAATGGGTGGCCGACGA
GCCGGGCCGGTGAACCGTCCTTGTTAAGCTCCTTGGTGCAATACGCTTGGTTCTCGAGGTGCGTGCCG
TTGGCAGCCTCGACATGAGCACGGGGAATCAACTTGCGGATGCGTTCACCGAAAGCACGAGTGGTGAA
CTGAACATAGCCCTGGAAGTGCGGCGTGCCGTTCTTCCCGG
```

```
>A142/A229
CAGTGACACGTCTCAAAGTCATCGAAGATGAGGCCTTTCGACATGTCTGCGTCTTTGAGCTGATCCCG
GTGGCGAATGACTTCGGCACCCGGGAGCAGTGCACGGGCATACTCTGTTTTGCCAAACTTCGTGGGCC
CGTAGATGTAGAGCGTCCGCCAATTTGGGAGCACCTTCGGAGCCCGCGCAAATTCGGACAGGGGGCGT
GCCGGAGCTTCATAGCATAGAGCTTCGCTCCGAATGCGCTGAAACTCCATACGATATGCATTGGCTCG
TTCGACCGCGTATGCAGGCATACGGTCTTGCAGTAAATCGAAGGCTTCCTTAGAACTGCGTTCAGGGT
CCTTTGCGATTTCCATACAAGTGCGGACTAAGTCGTCCTTCTTGCTACGCTTCGCCTTCGATACCTCC
CCGTCCGCAGGCCGCTTGCGAGATGGGGGAGGTTCGGTAGGAGGATCACCTTTAATCAAAGGTTCCGA
ATCCTCCTTCTTGCAGTATTGCCAACAGTTCCAGTGGCTTACTGCGGGGCTACCACCCGACCTATGCG
TTCCAATATTGGGGTGTTCACCACCAATGTCGAACGCATGGTAGGTTACCTCTCTCTGCTTCGTGTAA
ACAACGTAAGCATGGAGATGGGGACTACCGTCCTGATGCTTTTCCTG
```

```
>B1F1
GGTGCAGCGAGTCGCCGCCCTCATCGGTCGACCGCCAACCGGCAACGACTGGGCGCAACACAAAGCTT
GGGGATTCACCGAGGCGGTGTCATTAAGGCGCCAAGGCTATCCATGGCCGAAGCTGCTAGTCATGGCA
GGATTCGAAGTCACAGAGCAACAGGCGAAACTAGGGCGCCGTGCCATCTCACCTGAACAAGTCGAAGG
CGCAAGAATCGCTATGCTGGCACTGTGTGACGAAAACGGATTCCTGCCACACCGCCGGTGGAAAGAGA
ACCATGGTGGCCATCCCGGTACGTTTGCTTTGATGAGTTACCACAATACCCGAGACTGGCTGGTTGTT
CTGGAGCGTATGGGGTTCTATGCGCCACCACGGAAGTTGGGCAGCAAAGTCACCGCCAAGGTCGAGAA
GTTCAAGGATATCAACGACTACATCACCAAGAAGAAAGCCCAGCCGGTGACGATTCAGGAAGAACGCA
ACACATCTCAAACCATGGCAGGCTATGTCAACCGGGTAGAGATTATCGAGCGGAACCATGGCGGACAC
CTGGTGCGCACTGAACGCACGTACATTCAATTGAGGTAACAACATGAAGGCTATCTTTTTCCGCACTC
TGTATCGTATCGTTTTCGGTCGTGGTGCCGCCTACCGTGGCTCGCGCTGGTATGCCGTCGCCGACTGG
ATGAAGGCGCAGTGGATGGAGGTTCCATGACAACACATATTTCATTTCCTGATTGGTGGCCCGAATGC
CCGTATCCTACAGATATTTTTCCAATGAGTATGGGTGAATATGCAAAAATCGTACCG
```

```
>B1H1
ACGAAAGATCAACTGTATGCTTTCCTGACAAGTGGACCAGATGTGGAGCGTCTGATCATCGGGCAGGA
AAAGCATCAGGACGGTAGTCCCCATCTCCATGCTTACGTTGTTTACACGAAGCAGAGAGAGGTAACCT
ACCATGCGTTCGACATTGGTGGTGAACACCCCAATATTGGAACGCATAGGTCGGGTGGTAGCCCCGCA
GTAAGCCACTGGAACTGTTGGCAATACTGCAAGAAGGAGGATTCGGAACCTTTGATTAAAGGTGATCC
TCCTACCGAACCTCCCCCATCTCGCAAGCGGCCTGCGGACGGGGAGGTATCGAAGGCGAAGCGTAGCA
AGAAGGACGACTTAGTCCGCACTTGTATGGAAATCGCAAAGGACCCTGAACGCAGTTCTAAGGAAGCC
TTCGATTTACTGCAAGACCGTATGCCTGCATACGCGGTCGAACGAGCCAATGCATATCGTATGGAGTT
TCAGCGCATTCGGA
```

```
>B22C1
CTCGGAGGGCCCGGTAACTTGGGTGTGACGTACATCGAGACGATCAAGGACATGGAGCACGAGAGGGC
CACGGTAGCATTTGAGCTGCTCGTGGGAGAGTATCGTGAGAACCCGTTGGTGATCAACGGAGAAGTGA
TTGACTTAGTGACAACTGAGGAGGAAGATAGTGAGGAGGAATAAGATTACTTCGCTTTAGAGCTACGC
TCACAGCTACATTCATTAAGAAGACTACACTACTAGTTGACTAAGACTAAGATTAGCCTGATGCTTCG
CATCATTC
```

```
>B43H11
AGGTAATGCTAAAACTCCGCGACCAGAACGACGCGGAGTTAACCCTTGCGTGCCAGAACATGAGCGCG
CATTTGACCACCATGCTACAAGCAGCAGCAGGACAATATGACGATGAATAAAACGCAGAAGGAAATGG
AAGCCACAGGCGAGAAGGATCAATCAACCACGCCTGTGACGAAGACCGGCAAAGAAAAACTGCTCACG
GTCTGCAACGATATCATAGC
```

>**B46A2**

```
AGTTGTTCTCTGCCCTGGGCACACCTATGCAACCCGGGGGGTCACATGGGTGCACGGGGCGGGGGTAG
GCACGGGGGTACGGGGTACCCCCTCCCGTTCATGGGTTGCCGGTGGCACCCCATGGGGGGGCACGGGG
AGGCTGCGCCCCCCCTCTGGGGAAGTCGGTGACGTTCCCCACAGTTGAG
```

>**B50H12**

```
CGCCGCCGGATCCAAACTCAACTTCGACAGGTATAGCCGGGTGCAGGGGTATGTGCTGGACGGCGTAT
GGATGCACTGGACGGAGGCGAAGGTCAACGACATGGTAATCGAGGGCAATGAGCCGAGGAAGATAAAT
CTCGAAAAGGCAGCAAACGCCGGCAAGAACCTCGAATATGAGTTGTATGCAGGCTTTGACGGGCAGGG
GCAGTTTGCCAATGGCGACCAGTTTGTGTTCAACGTCAGAAGCCTTGACGGTGCAACGCTCTACGCCA
CGCAGACAATCACGATTCCGAACGGCACATACCTGAGCGACGTGCAGGCTGGTCTTGAGTACCTGTAT
GACCAACTCACGTCGGGAACGATGGATGAGCATTTGACCATAGAACTGTGCGACTGCAAGTTAGGTAT
CACTGCTTTGGTGGCAGAGCGTAAGGTGACAATGACGGTTACTGGCAGTGTCAAGCCCATCCTTGTTC
CCAAGAATCACTATGGTCCGCAGGTGACAGACAAGGAGACGTACCATTTTGCTCTCAAAAAAGAGACA
TTTGATTGCCCGCCAGACCCCGAATACATTGGCGCTGATGACATTGCCTACAACAATGTGCGCGATGA
CTGCTTCCAGTTTCGCGCCCGCTTCGTTTACGATGATGGCGGAGCGAGCCACTGGTCCGGCGTTTCGA
TAGTGCCGCTGAACAATGCTCAGTTTGCCGACCCACAGCCGTCTCTCAATGCCATCAAGATTGATTAC
ACTGATGAGCGCCTCAATACAATCAGTTGGCTATCCATCCTTGATTACGTTGACATCGCTTCTCGATA
CAATGAGAATGATGTGTGGCGCCTCATTCGTCGCATCCCCGTTTGCGAGGTGGGCATTGATGAGCAAT
TCATCATCTTTGCCAATGACAACTCGTACACGGTTGTAGAGTCCGATGACCCTTCGGTGGCAACCGGA
GATACGCAGGTGCTCACAAACTATCATCGTGTCCCGTACA
```

>**B56H9**

```
CGTAAGGGATGTCTGGCTATATGTGAGCAGGTACGCTTTCGCGTTCCGTCTCACTGACGTGGTTGAAT
CAACCAAGTCCGGGATCTCCATCTTCGATGTTGGTCCCACTGCACCCCATTCTCGACTACCTGCGGCA
GCTCGTCAGTTCGTGGGGGCAGTTGTTCTCTGCCCTGGGCACACCTATGCAACCCGGGGGGTCACATG
GGTGCACGGGGCGGGGGTAGGCACGGGGGTACGGGGTACCCCCTCCCGTTCATGGGTTGCCGGTGGCA
CCCCATGGGGGGGCACGGGGAGGCTGCGCCCCCCCTCTGGGGAAGTCGGTGACGTTCCCCACAGTTGA
G
```

>**B56H11**

```
ATATGTGAGCAGGTACGCTTTCGCGTTCCGTCTCACTGACGTGGTTGAATCAACCAAGTCCGGGATCT
CCATCTTCGATGTTGGTCCCACTGCACCCCATTCTCGACTACCTGCGGCAGCTCGTCAGTTCGTGGGG
GCAGTTGTTCTCTGCCCTGGGCACACCTATGCAACCCGGGGGGTCACATGGGTGCACGGGGCGGGGGT
AGGCACGGGGGTACGGGGTACCCCCTCCCGTTCATGGGTTGCCGGTGGCACCCCATGGGGGGGCACGG
GGAGGCTGCGCCCCCCCTCTGGGGAAGTCGGTGACGTTCCCCACAGTTGAG
```

>**B58A11**

```
GACACGTTGGTCGTCACGTTGGACACGATGCGGAACTGACGCAGGAAGTTCAGTGTCGCCTTCGTGAT
CGGGTTGACCGCGTACACACCGGAGATGGTGAACACGTCGCCTTCCTTCAGGCCCGAGTTTGTCGTCC
AGCCATCCGTCACGAGGTCCATCGTGTCGGTTGTGAGGCTCGACGCATAGGTCGTGGACTGCGAGGCA
CCGCGAATAAGCGGTGTGCCGCCGTGAGCACCAACCGTGTGCGTCTGGACGTTCTGGGCCATCCACAC
ATCGGAACCGCCGATCATCGGCAGCTTGGCCTTTTCAATGGCCCCAGCCGCGATGCGTTCCACATAGG
AGCCGGTGAACGAGTTCGCCAGCCCATAGTAGTCCGTAGGCGTCACCATGCCGACGCGGTTGCCATCG
TTTGGAACGGCAAACTCGTCGAGGCGCTCCAGACCGACCGACCAATCCGAGAAGGAGTTGATCGTCTG
GCCTGCAGTTCCTACCCAGTTCGGCACCTGCTTGTAGAGAGCGAGCAAGTCGCTATCCACCTGGTTCG
CCAACTGGATCATGGCGGGCTTGATGTACCGCTCGCTGAACATATCGA
```

Chapter 5:

Wide bacteriophage diversity in a single terrestrial wastewater site

Abstract

T4-type bacteriophages are distributed worldwide in aquatic and terrestrial environments. The Major Capsid Protein sequence (MCP), conserved in T4-type bacteriophages, has been used to document phylogenetically informative variations among bacteriophages from various environments. In this study, we present a protein-based phylogenetic tree of uncultured T4-type phages circulating in a defined freshwater environment (an activated sludge from a dairy wastewater treatment plant), derived using an amino acid sequence alignment of predicted MCP proteins. Thirty distinct (<99% identity) DNA clones were obtained following PCR with degenerate primers for the *g23* gene on bacteriophage DNA prepared from multiple samples from one dairy industry wastewater plant in Kerry, Ireland. The translated sequences showed amino acid identities of 46–99 % with NCBI database sequences. Following alignment with uncultured and cultured phage genomes, neighbor-joining trees were constructed. Wide phylogenetic diversity was seen in multiple samples from the same terrestrial wastewater environment, comparable with that seen in widely geographically-separated samples from the ocean and land. Three novel groups of environmental T4-type capsid sequences were identified. There was little sequence overlap between samples taken at different times from the same site. Among the most similar g23 sequences in the NCBI database were samples from an Antarctic lake. This provides evidence for worldwide distribution of phages with widespread genetic exchange.

5.1. Introduction

T4-type bacteriophages constitute a large group of viruses that belong to the *Myoviridae* family, viruses with a head and a contractile tail (Ackermann & Krisch, 1997). They infect Enterobacteria and other phylogenetically distant bacteria (such as *Vibrio*, *Acinetobacter*, *Aeromonas*, *Prochlorococcus* and *Synechococcus*), and are widely distributed in nature (Kim *et al.*, 2010; Mann *et al.*, 2005; Matsuzaki *et al.*, 1998; Petrov *et al.*, 2006). Cultured T4-type phages can be classified into four subgroups: the T-evens (closely related to T4 phage), pseudoT-evens (distantly related to T4), schizoT-evens (more distantly related to T4), and exo-T-even (the most distantly related to T4) (Desplats & Krisch, 2003; Tetart *et al.*, 2001). The evaluation of the diversity of naturally occurring phage communities in the environment by phage culture is limited because only a small percentage of their hosts are culturable. However, the development of techniques to isolate viral DNA from the environment (Thurber *et al.*, 2009), together with the PCR amplification of conserved genes from environmental DNA has provided information about the diversity of phages present in a particular environment (Chen *et al.*, 1996). One such conserved gene, the major capsid protein (gp23), is part of the core genome of T4-related phages (Petrov *et al.*, 2010b) and has been used to evaluate the genetic diversity of the T4-type family of bacteriophages in different environments. Several studies carried out on aquatic (Butina *et al.*, 2010; Filee *et al.*, 2005; Lopez-Bueno *et al.*, 2009) as well as on terrestrial environments (Cahyani *et al.*, 2009a; Cahyani *et al.*, 2009b; Fujihara *et al.*, 2010; Fujii *et al.*, 2008; Jia *et al.*, 2007; Nakayama *et al.*, 2009a; Nakayama *et al.*, 2009b; Wang *et al.*, 2009a; Wang *et al.*, 2009b; Wang *et al.*, 2009c; Wang *et al.*, 2011), support the hypothesis that culturable phages constitute a small proportion of the entire population in an ecosystem.

Wastewater bacteria include species known to be infected by T4-type viruses, including *E. coli*, *Acinetobacter*, *Klebsiella*, *Aeromonas*, *Sphingomonas*, and *Vibrio* (Dafale *et al.*, 2010; Jorgensen & Pauli, 1995; Loperena *et al.*, 2009). Metagenomic analysis of uncultured phage communities in activated sludge (Parsley *et al.*, 2010b; Tamaki *et al.*, 2012) and reclaimed water (Rosario *et al.*, 2009b) indicate that *Myoviridae* phages are abundant in these environments. Numerous phages with T4 phage morphology have been cultured from this environment (Goyal *et al.*, 1980; Kaliniene *et al.*, 2010; Zuber *et al.*, 2007).

In this chapter, we evaluate the genetic diversity of uncultured T4-type bacteriophages found in dairy plant wastewater at a single location by analysing sequences of major capsid protein genes (*g23*) derived by PCR amplification. The novel wastewater-specific sequences obtained in this study are compared with a wide range of sequences from other locations. Our results show that the dairy plant wastewater contained very diverse and previously uncharacterized phages. Three previously uncharacterized groups of environmental T4-type capsid sequences were identified, while the genetic diversity of other previously described groups has been expanded.

5.2. Material and Methods

5.2.1. Samples

The sludge samples were collected in 10-15 litre volumes from the same location (an open aeration tank receiving milk product polluted wastewater treated by dissolved air flotation, and anaerobic digestion) on four occasions, in May 2006 (sample WL1), November 2006 (sample WL2), June 2007 (sample WL3) and January 2010 (sample WL4), in the Kerry Ingredients Wastewater Treatment Plant in Listowel, Co. Kerry, Ireland (52°26’20” N; 9°29’7” W). For further sample characteristics see Table 5.1.

Table 5.1. Characteristics of wastewater samples. Differences in sample parameters reflect varying input of waste from different processes involving milk, cheese, butter, and milk powder.

Sample	Wastewater collection date	pH	COD* (ppm)	Ortho-P (ppm)	Ammonia (ppm)
W1	30/05/06	7.08	664	60.74	39
W2	28/11/06	7.37	1254	21.1	12.5
W3	27/07/07	6.56	1030	59.8	44.5
W4**	07/01/10	9.35	1175	17.9	12.5

*COD Chemical oxygen demand
**Sample used for metagenomic analysis described in Chapter 3

5.2.2. Purification of phage particles and DNA extraction

Phage particles purification and DNA extraction have been described in Chapter 3 Section 3.2.2 and 3.2.3.

5.2.3. Major capsid protein genes (g23) amplification

The g23 sequences were amplified from the environmental samples with the degenerate primer pair MZIA1bis and MZIA6 and PCR program as described previously (Filee *et al.*, 2005). The PCR mixture contained, in a total volume of 50 µl, 1x GoTaq® Green Flexi Reaction Buffer (Promega), 25 pmol of each of the primers, 2.5 mM $MgCl_2$ (Promega), 0.2 mM dNTPmix (New England Biolabs), 1U of GoTaq® Flexi DNA Polymerase (Promega) and 1µl of phage DNA. Reaction mixtures were heated to 94°C for 90 sec followed by 30 cycles of denaturation at 94°C for 45 sec, annealing at 50°C for 1 min, and extension at 72°C for 45 sec. A final extension was carried out at 72°C for 5 min. PCR products of samples WL1, WL2 and WL3 were extracted from the 1% agarose gel using QIAquick Gel Extraction Kit (Qiagen), while PCR products of sample WL4 were purified using QIAquick PCR Purification Kit (Qiagen). The purified PCR products were cloned into the *E. coli* pCR2.1-TOPO vector using TOPO TA Cloning Kit (Invitrogen). Plasmid DNA was isolated from randomly selected colonies using QIAprep Spin Miniprep Kit (Qiagen) and digested with restriction enzyme (EcoRI) to identify clones with insert. Positive clones were sequenced by GATC Biotech (Germany) using the M13F and M13R universal primers.

5.2.4. g23 phylogenetic analyses

Forward and reverse sequences were assembled and corrected manually using the sequence chromatograms. Consensus nucleotide sequences were translated into amino acids using EMBOSS Transeq (Rice *et al.*, 2000), trimmed to the PCR product length and compared using BLASTP to all protein sequences present in non-redundant database of NCBI website (http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi). Phylogenetic trees were constructed for 30 representative gp23 sequences out of 83 sequences obtained (inclusion criterion: <99% amino acid identity in the same wastewater sample). To construct an aquatic and terrestrial phage tree, the wastewater sequences were aligned using ClustalW (Chenna *et al.*, 2003) with top hits from BLASTP search together with capsid sequences from previously established Marine groups I-V (Filee *et al.*, 2005) and Paddy groups I-IX (Fujii *et al.*, 2008; Jia *et al.*, 2007; Wang *et al.*, 2009b), as well as representative cultured T4-type cultured phages (Tetart *et al.*, 2001). The resulting alignment was manually edited in Jalview (Waterhouse *et al.*, 2009) to remove highly variable regions between amino acids 133-141, 157-221 and 238-245 of the phage T4 gp23 protein (GenBank accession number AAD42428) as described in (Filee *et al.*, 2005) (Figure 5.1). The phylogenetic tree was constructed in MEGA (version 5.04) program (Tamura *et al.*, 2007) using a p-distance model, complete deletion of gaps and the neighbor-joining method with 1000 bootstrap replications. Forty six unique g23 sequences obtained in this study have been submitted to the GenBank database under accession numbers JN393559–JN393604.

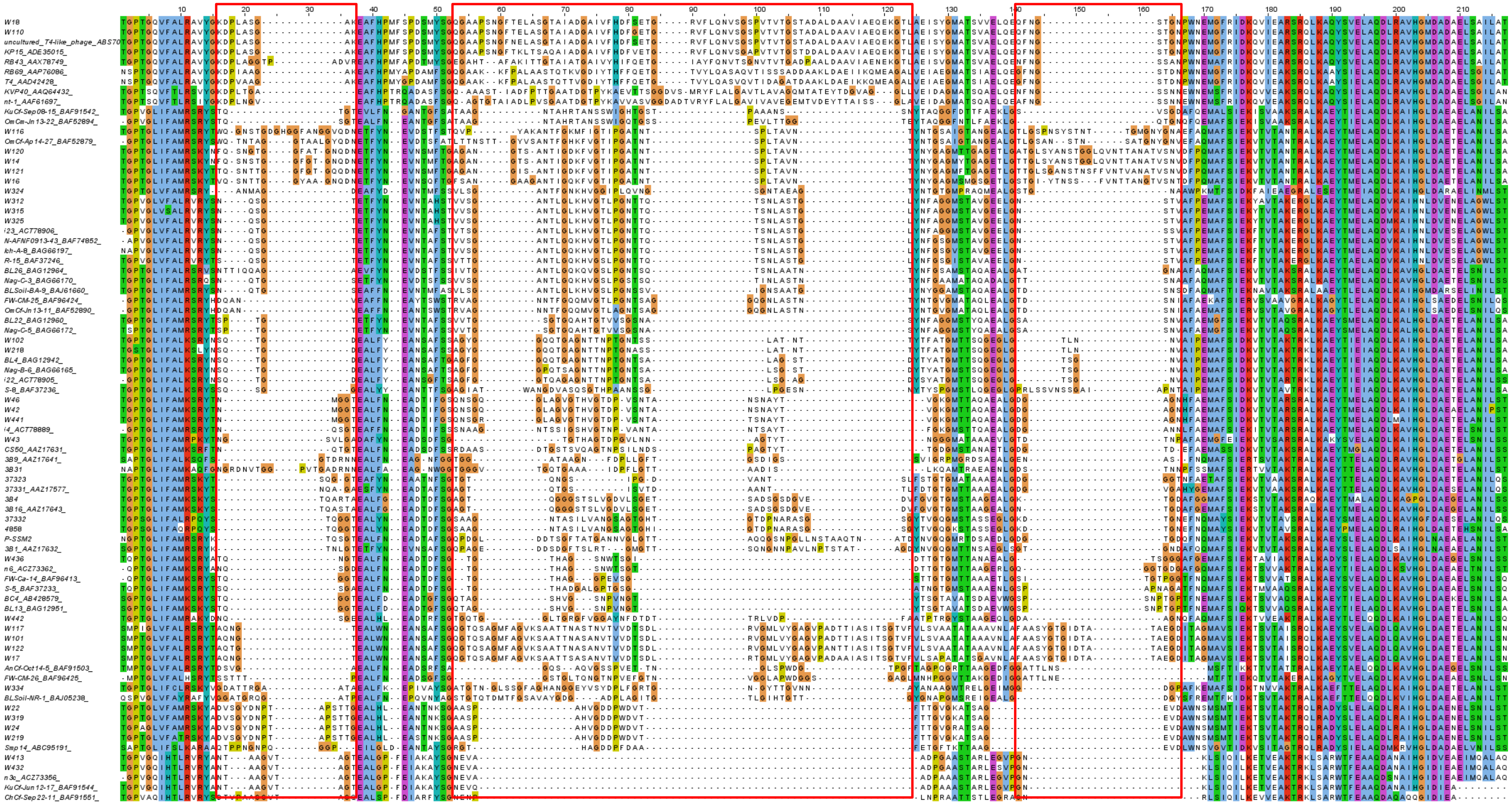


Figure 5.1. Amino acid sequence alignment of deduced major capsid protein sequences obtained from wastewater as well as other capsid sequences showing >70% identity in BLAST analysis, and sequences of representative sequences from previously reported subgroups. Regions in red boxes indicate hypervariable regions and were excluded from the phylogenetic analysis.

For the large ‘global’ phylogenetic tree, the dataset consisted of the same 30 wastewater sequences (<99% amino acid identity), 42 representative gp23 sequences of cultured isolates (<95% amino acid identity), 423 representative gp23 sequences from PCR-amplified environmental clone libraries (<95% amino acid identity) (Table 5.2) and 370 gp23 sequences from metagenomes (<95% amino acid identity). The latter were obtained by BLASTP search against NCBI Environmental samples (env_nr) database using gp23 from cultured representatives of Near T4 and Far T4 groups as queries, as previously described (Comeau & Krisch, 2008): 500 hits against gp23 from T4 and 100 hits against RM378 were initially retained. The sequences were then filtered to remove short sequences and sequences with identity above 95%. Three sequences (CS43, FW-Ca-1 and BLSoil-NR-9) that created very long branches on the tree were also excluded from further analysis. All gp23 sequences used in the comparison, both cultured and environmental phages, were obtained from GenBank. The combined 862 gp23 amino acid sequences were aligned using MUSCLE v3.6 (Edgar, 2004) using a gap opening penalty of -4 and a gap extension penalty of -0.1. The alignment was edited using Jalview software. Only conserved regions between amino acids 118-131 and 247-296 of the phage T4 were used for phylogenetic analysis (Figure 5.2). Phylip v3.67 program protdist was used with the JTT model to calculate a distance matrix from protein sequences. The Phylip neighbor program was used to construct a neighbor-joining tree from the distance matrices. The tree was drawn using iTOL v2.1 software (Letunic & Bork, 2011) and rooted using the RM378 sequence.

Table 5.2. Summary of all PCR-amplified g23 sequences retrieved from GenBank (up to June 2011).

g23 study	**No. of all clones (DNA)**	**No. of unique clones (aa)**	**No. of clones used for phylogenetic analysis <95%**	**References**
Black soil, China	46	42	27	(Wang *et al.*, 2011)
Black soil, China	99	84	51	Liu *et al.*, Unpublished
Lake in China	46	35	27	Huang et al., Unpublished
Lake Baikal, Siberia, Russia	23	22	12	(Butina *et al.*, 2010)
Paddy field, Japan	68	49	21	(Fujihara *et al.*, 2010)
Antarctic lake	63	31	21	(Lopez-Bueno *et al.*, 2009)
Borehole water from gold mine, South Africa	8	5	4	Mabizela and Litthauer, Unpublished
Paddy soil, Japan	97	68	21	(Wang *et al.*, 2009c)
Paddy soil, China	53	49	34	(Wang *et al.*, 2009b)
Rice straw compost, Japan	50	37	8	(Cahyani *et al.*, 2009a)
Paddy floodwater, Japan	40	39	24	(Nakayama *et al.*, 2009a)
Mn nodules in paddy soil, Japan	44	41	24	(Cahyani *et al.*, 2009b)
Paddy soil, Japan	56	54	31	(Wang *et al.*, 2009a)
Paddy field floodwater, Japan	58	41	15	(Nakayama *et al.*, 2009b)
Paddy soil, Japan	44	35	22	(Fujii *et al.*, 2008)
Marine environment	23	17	14	Sandaa and Kristiansen, Unpublished
Paddy field, Japan	17	17	17	(Jia *et al.*, 2007)
Marine environment	85	83	50	(Filee *et al.*, 2005)
Total	**920**	**749**	**423**	

5.2.5. Inter-sample comparison

∫-LIBSHUFF implemented in mothur (Schloss *et al.*, 2009) was used to compare a total of 83 aligned g23 sequences (nucleotide and translated) from the wastewater samples containing a single open reading frame (Table 5.3) assigned to groups WL1-4. A nucleotide distance matrix was calculated using the dist program in mother (http://www.mothur.org/) and an amino acid distance matrix was calculated using protdist at http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::protdist.

5.2.6. Comparison with the viral metagenome

Amplified g23 DNA sequences from the libraries were compared to metagenomic reads and contigs described in Chapter 3. Short reads and contigs were aligned to the g23 nucleotide sequences using BLASTN with e-value cutoff of 0.001.

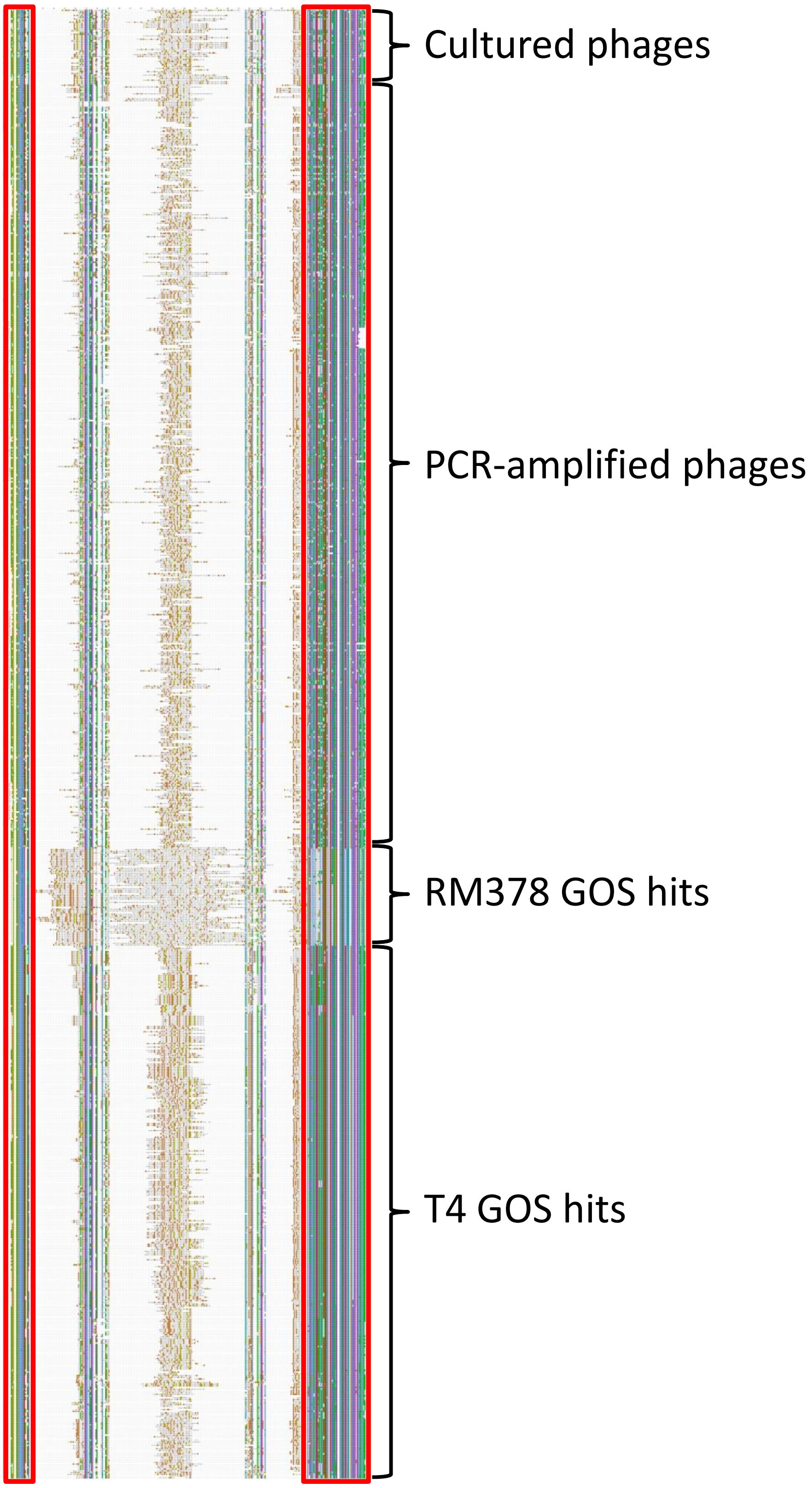


Figure 5.2. Multiple sequence alignment of 862 major capsid protein sequences from cultured phages and environmental samples. Regions in red boxes indicate conserved blocks used for phylogenetic analysis.

5.3. Results

5.3.1. Sequences of g23 clones

Major capsid protein gene sequences were amplified from four wastewater samples collected at four different time points from the same location (activated sludge from dairy wastewater treatment plant). In total, 85 clones were sequenced from clone libraries (WL1, WL2, WL3 and WL4), and 46 unique (only seen once) clones at DNA level and 43 unique clones at amino acid level were obtained (Table 5.3). Two sequences from library WL3 contained non-translatable sequences and were excluded from further analysis. After exclusion of primer sequences, sequences ranged from 323 to 542 bp (108-181 amino acid residues). Amino acid identity ranged from 33.17% (W110 and W116) to 100% (e.g. W21 and 17 other clones).

Table 5.3. Summary of g23 PCR sequencing results from this study.

Wastewater Library	No. of sequenced clones	No. of unique clones (DNA)	No. of unique clones (aa)	No. of distinct (<99% identical) clones used for phylogenetic analysis
WL1	21	12	12	12 (W11,W12,W16,W18,W19,W110, W112,W116,W117,W119,W120,W121)
WL2	23	8	7	4 (W21,W23,W216,W217)
WL3	21	13	11 (+2 clones not in any frame)	6 (W34,W37,W39,W313,W314,W320)
WL4	20	13	13	8 (W41,W42,W43,W46,W411,W415, W419,W420)
Total	**85**	**46**	**43**	**30**

BLAST analysis revealed that all amino acid sequences showed more than 46% similarity to known sequences within GenBank (Table 5.4). The most similar hits (>70% identity) were g23 sequences from sewage, an Antarctic Lake (Lopez-Bueno *et al.*, 2009) and paddy field soil in China (Wang *et al.*, 2009b) and Japan (Fujii *et al.*, 2008). The highest amino acid sequence identity of 99% was observed for clones W110 and W112 with uncultured T4-type phage isolated from sewage in Portugal. These clones showed 97% identity at DNA level (96% at amino acid level) to the cultured *Klebsiella* phage KP15, therefore it is possible that they belong to novel *Klebsiella* phages. Thirteen sequences showed less than 70% identity to known sequences: these hits were from paddy field soil, seawater (Filee *et al.*, 2005), upland black soil in China (Wang *et al.*, 2011), a Chinese lake, and Lake Baikal in Siberia (Butina *et al.*, 2010).

Table 5.4.Sequence similarity of the g23 clones (<99% identity) isolated from wastewater to the closest match in GenBank by BLASTP.

Clone	**Length (aa)**	**Best hit in NCBI**	**Accession number**	**Identity (%)**	**Isolation source**	**Reference**
W110	181	Uncultured T4-like phage	(ABS70720)	99	Sewage, Portugal	(Carvalho *et al.*, 2010)
W112	181	Uncultured T4-like phage	(ABS70720)	99	Sewage, Portugal	(Carvalho *et al.*, 2010)
W44*	108	Uncultured Myoviridae clone n3c	(ACZ73356)	98	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W46	108	Uncultured Myoviridae clone n3c	(ACZ73356)	97	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W411	108	Uncultured Myoviridae clone n3c	(ACZ73356)	97	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W31*	135	Uncultured Myoviridae clone j23	(ACT78906)	96	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W34	135	Uncultured Myoviridae clone j23	(ACT78906)	95	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W37	135	Uncultured Myoviridae clone j23	(ACT78906)	95	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W314	135	Uncultured Myoviridae clone j23	(ACT78906)	95	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W12	133	Uncultured Myoviridae clone BL4	(BAG12942)	92	Paddy soil, China	(Wang *et al.*, 2009b)
W216	133	Uncultured Myoviridae clone BL4	(BAG12942)	89	Paddy soil, China	(Wang *et al.*, 2009b)
W415	118	Uncultured Myoviridae clone n6	(ACZ73362)	82	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W116	169	Uncultured Myoviridae clone OmCf-Ap14-27	(BAF52879)	79	Paddy soil, Japan	(Fujii *et al.*, 2008)
W120	170	Uncultured Myoviridae clone OmCf-Ap14-27	(BAF52879)	71	Paddy soil, Japan	(Fujii *et al.*, 2008)
W119	170	Uncultured Myoviridae clone OmCf-Ap14-27	(BAF52879)	70	Paddy soil, Japan	(Fujii *et al.*, 2008)
W16	170	Uncultured Myoviridae clone OmCf-Ap14-27	(BAF52879)	70	Paddy soil, Japan	(Fujii *et al.*, 2008)
W18	167	Uncultured Myoviridae clone OmCf-Ap14-27	(BAF52879)	69	Paddy soil, Japan	(Fujii *et al.*, 2008)
W43	132	Uncultured Myoviridae clone j4	(ACT78889)	79	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W41	132	Uncultured Myoviridae clone j4	(ACT78889)	77	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W419	132	Uncultured Myoviridae clone j4	(ACT78889)	77	Antarctic lake	(Lopez-Bueno *et al.*, 2009)
W42	124	Uncultured Myoviridae clone 37323	(AAZ17574)	66	Marine environment	(Filee *et al.*, 2005)
W313	136	Uncultured Myoviridae clone BLSoil-DH-15	(BAJ61623)	66	Black soil, China	(Wang *et al.*, 2011)
W420	136	Uncultured Myoviridae clone s10-8	(ADI87648)	61	Lake in China	Huang et al., Unpublished
W320	151	Uncultured Myoviridae clone BLSoil-NR-1	(BAJ05238)	58	Black soil, China	(Wang *et al.*, 2011)
W21	124	Uncultured Myoviridae clone N0508/1-5	(ADA61135)	56	Lake Baikal, Siberia, Russia	(Butina *et al.*, 2010)
W39	124	Uncultured Myoviridae clone N0508/1-5	(ADA61135)	56	Lake Baikal, Siberia, Russia	(Butina *et al.*, 2010)
W23	124	Uncultured Myoviridae clone N0508/1-5	(ADA61135)	55	Lake Baikal, Siberia, Russia	(Butina *et al.*, 2010)
W217	124	Uncultured Myoviridae clone N0508/1-5	(ADA61135)	55	Lake Baikal, Siberia, Russia	(Butina *et al.*, 2010)
W121	175	Uncultured Myoviridae clone AL3	(BAG12936)	47	Paddy soil, China	(Wang *et al.*, 2009b)
W11	175	Uncultured Myoviridae clone AL3	(BAG12936)	46	Paddy soil, China	(Wang *et al.*, 2009b)
W19	175	Uncultured Myoviridae clone AL3	(BAG12936)	46	Paddy soil, China	(Wang *et al.*, 2009b)
W117	175	Uncultured Myoviridae clone AL3	(BAG12936)	46	Paddy soil, China	(Wang *et al.*, 2009b)

* Sequences having more than 99% identity to other wastewater clones in this study

5.3.2. Inter-sample variation

The *P* value calculated by libshuff for all of the 12 individual (bidirectional) nucleotide group comparisons was <0.0001, except for WL2-WL3 (0.0003). *P* value for amino acid group comparisons was 0.001, except for WL2-WL3 (0.2871). WL3-WL2 was 0.001. A Bonferroni correction for intergroup differences at a significance level of *P* <0.05 given 12 comparisons is <0.004166. Therefore, the groups WL1-4 are significantly different (*P* <0.05).

5.3.3. Phylogeny of gp23 sequences

A phylogenetic tree (Figure 5.3) was constructed using MEGA (Tamura *et al.*, 2007) based on the alignment (Figure 5.1) of 30 representative gp23 translated sequences (<99% identity) out of 83 sequences obtained from wastewater with highly related sequences from GenBank. Sequences from previously established groups (Marine groups I-V and Paddy groups I-IX) as well as T-evens, PseudoT-evens, SchizoT-evens and ExoT-evens groups were also included for comparison. In the phylogenetic tree, 2 out of 83 wastewater sequences grouped together with cultured representatives of PseudoT-evens, appearing on the same branch with *Klebsiella* phage KP15. A total of 34 out of 83 sequences were placed into previously established groups (Paddy Groups I, IV, V and Marine Groups I, II, IV), while nearly half (41 out of 83) formed 3 independent deep-branching clusters (novel clusters W1a, W1b and W2). W1a and W1b were only represented in the first sample. Six sequences formed two clusters with other g23 sequences from GenBank not previously assigned to any group. Bootstrap support was better for assignment to the Paddy groups (83-99%) than the Marine Groups (50-93%). Although, apart from the novel groups, most sequences (18) in this study grouped with phages from freshwater or paddy soil (a freshwater influenced environment), sixteen sequences grouped with Marine Groups I, II and IV (Figure 5.3).

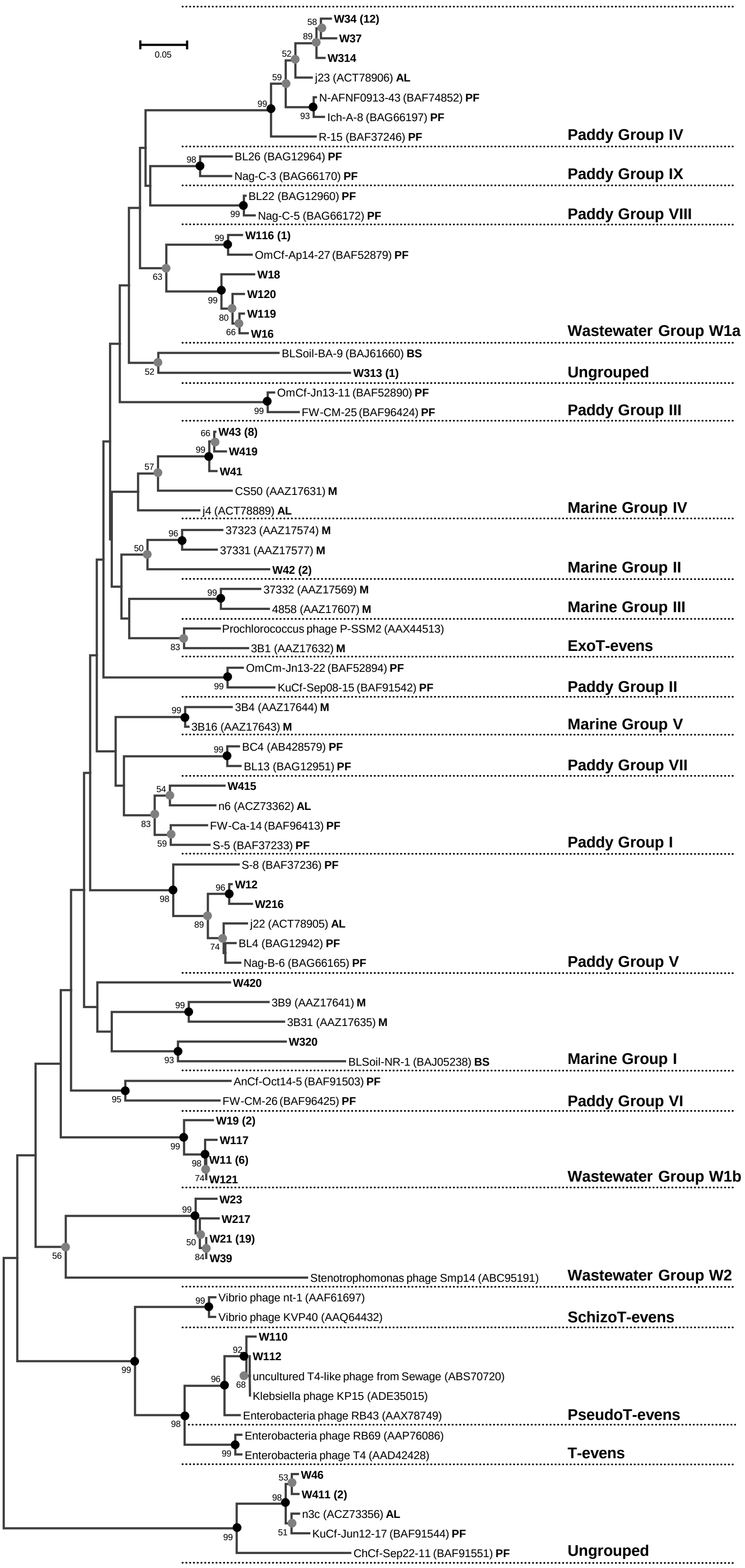

0.05
58
W34 (12)
89
W37
52
W314
59
j23 (ACT78906) AL
99
N-AFNF0913-43 (BAF74852) PF
93
Ich-A-8 (BAG66197) PF
R-15 (BAF37246) PF
Paddy Group IV
98
BL26 (BAG12964) PF
Nag-C-3 (BAG66170) PF
Paddy Group IX
BL22 (BAG12960) PF
99
Nag-C-5 (BAG66172) PF
Paddy Group VIII
99
W116 (1)
OmCf-Ap14-27 (BAF52879) PF
63
W18
99
W120
80
W119
66
W16
Wastewater Group W1a
BLSoil-BA-9 (BAJ61660) BS
52
W313 (1)
Ungrouped
OmCf-Jn13-11 (BAF52890) PF
99
FW-CM-25 (BAF96424) PF
Paddy Group III
66
W43 (8)
99
W419
57
W41
CS50 (AAZ17631) M
j4 (ACT78889) AL
Marine Group IV
96
37323 (AAZ17574) M
50
37331 (AAZ17577) M
W42 (2)
Marine Group II
99
37332 (AAZ17569) M
4858 (AAZ17607) M
Marine Group III
Prochlorococcus phage P-SSM2 (AAX44513)
83
3B1 (AAZ17632) M
ExoT-evens
OmCm-Jn13-22 (BAF52894) PF
99
KuCf-Sep08-15 (BAF91542) PF
Paddy Group II
99
3B4 (AAZ17644) M
3B16 (AAZ17643) M
Marine Group V
99
BC4 (AB428579) PF
BL13 (BAG12951) PF
Paddy Group VII
54
W415
n6 (ACZ73362) AL
83
FW-Ca-14 (BAF96413) PF
59
S-5 (BAF37233) PF
Paddy Group I
S-8 (BAF37236) PF
98
96
W12
W216
89
j22 (ACT78905) AL
74
BL4 (BAG12942) PF
Nag-B-6 (BAG66165) PF
Paddy Group V
W420
99
3B9 (AAZ17641) M
3B31 (AAZ17635) M
W320
93
BLSoil-NR-1 (BAJ05238) BS
Marine Group I
AnCf-Oct14-5 (BAF91503) PF
95
FW-CM-26 (BAF96425) PF
Paddy Group VI
W19 (2)
99
W117
98
W11 (6)
74
W121
Wastewater Group W1b
W23
99
W217
50
W21 (19)
84
W39
56
Stenotrophomonas phage Smp14 (ABC95191)
Wastewater Group W2
99
Vibrio phage nt-1 (AAF61697)
Vibrio phage KVP40 (AAQ64432)
SchizoT-evens
99
W110
92
W112
96
68
uncultured T4-like phage from Sewage (ABS70720)
Klebsiella phage KP15 (ADE35015)
98
Enterobacteria phage RB43 (AAX78749)
PseudoT-evens
Enterobacteria phage RB69 (AAP76086)
99
Enterobacteria phage T4 (AAD42428)
T-evens
53
W46
98
W411 (2)
n3c (ACZ73356) AL
51
KuCf-Jun12-17 (BAF91544) PF
99
ChCf-Sep22-11 (BAF91551) PF
Ungrouped

Figure 5.3. Neighbor-joining phylogenetic tree containing gene translations from wastewater phages and representative environmental T4-type phages. Wastewater clones are in bold (identified as W followed by sample number 1-4 and then clone number – see Table 2) followed by the number of clones with identical or nearly identical (>99% of amino acid identity) sequences. The GenBank accession numbers of reference sequences are shown in parentheses. AL (Antarctic lake), PF (Paddy field), BS (Black soil) and M (Marine) indicate the location of the three largest reference groups. Only bootstrap values of >50% are shown, represented by black (>90%) and grey (<90%) circles.

5.3.4. Global scale diversity

A broad sample of all g23 sequences available in GenBank was then added to this analysis, including all (>95% identity) PCR-based environmental sequences (Table 5.2) and those from the non-PCR-based GOS metagenomic sequence study (Rusch *et al.*, 2007). Sequences were aligned with MUSCLE v3.6 (Edgar, 2004) (Figure 5.2), edited with Jalview (Waterhouse *et al.*, 2009) and the neighbor-joining tree was computed with Phylip v3.67 (Felsenstein, 2005) and drawn with iTOL v2.1 (Letunic & Bork, 2011) rooted using the RM378 sequence. In the resulting global tree (Figure 5.4) the T4-type sequences formed several deep-branching clusters that could be grouped in three major groups, similar to previously described groupings (Comeau & Krisch, 2008): “Near T4” (44 sequences most closely resembling T4), “Far T4” (70 sequences including the cultured phage RM378 most divergent from T4, and hits with RM378 gp23) and “Cyano T4” (748 sequences incorporating the PCR amplified environmental sequences and cyanophage sequences). The gp23 sequences from the single wastewater site in this study were distributed in all of the three major groupings around the global phylogenetic tree. Two sequences (W110 and W112) grouped in the Near T4 cluster containing cultured representatives. Two sequences (W46 and W411) were found in the Far T4 cluster. The remaining 26 sequences were distributed across the tree in the CyanoT4 group clustering with the aquatic, terrestrial and GOS environmental sequences.

5.3.5. Comparison with the viral metagenome

A BLAST search of reads from the viral metagenome described in Chapter 3 resulted in 1 read that matched with the g23 sequences amplified in this study. No significant (e-value 0.001) similarity was found between g23 sequences and metagenomic contigs.

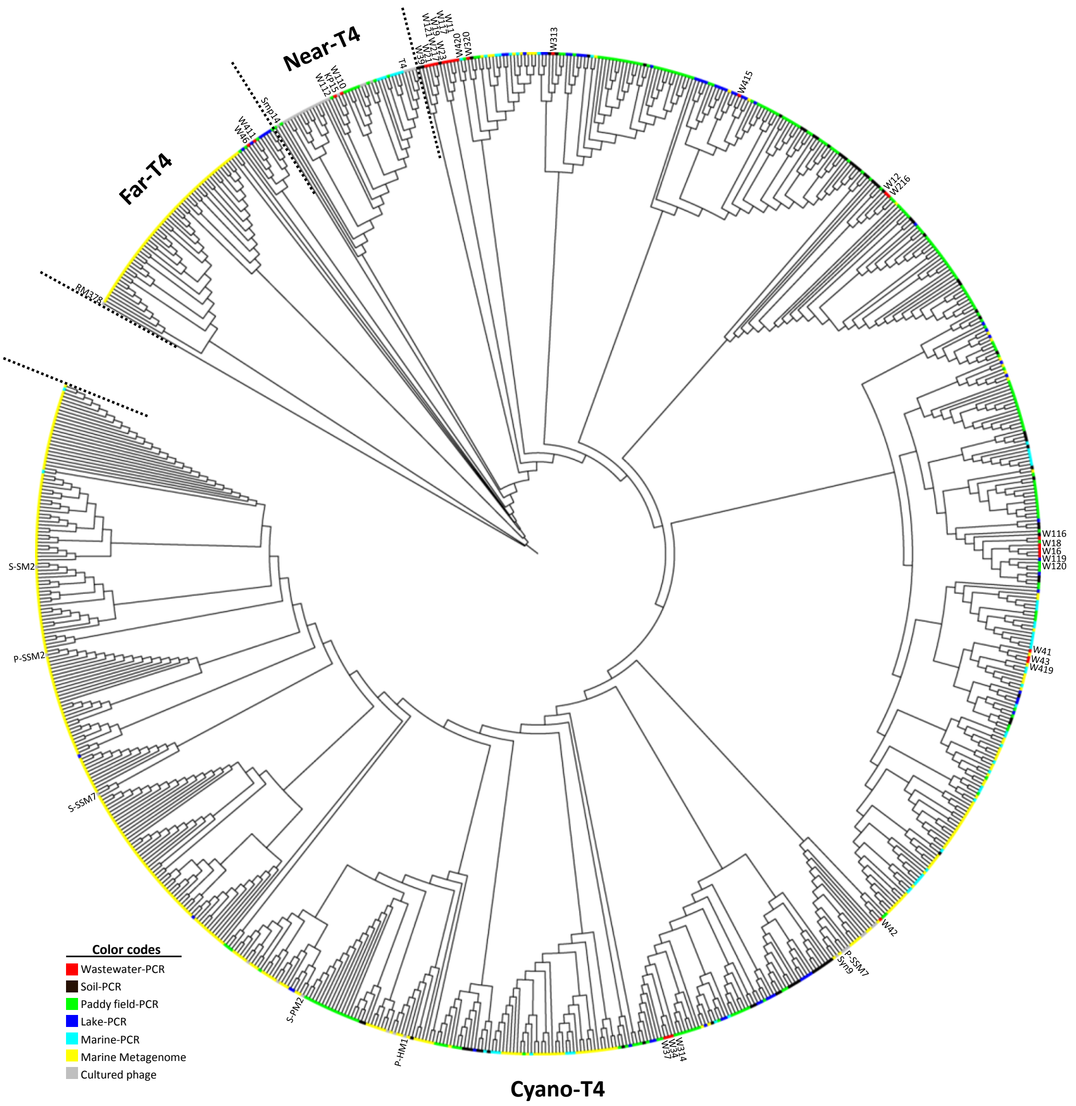


Figure 5.4. Neighbor-joining phylogenetic tree showing the genetic relationships among 862 g23 translated sequences of cultured and environmental T4-type phages. Groups ‘Near T4’, ‘Far T4’ and ‘Cyano T4’ were previously established by (Comeau *et al.*, 2007). Colour codes indicate the type of habitat from which the g23 sequences originated. The tree was rooted using *Rhodothermus* phage RM378 gp23 sequence.

5.4. Discussion

Major capsid protein (MCP, gp23) sequences have been extensively used to study genetic diversity of the T4-like genus within the *Myoviridae* family. Use of the *g23* gene has divided T4-like phages into four cultured clades (T-evens, SchizoT-evens, PseudoT-evens, and Exo-T-evens) (Desplats & Krisch, 2003; Tetart *et al.*, 2001) and at least 14 'environmental' clades that have been associated with a particular environment such as marine, freshwater and soil (Table 5.2). Comeau and Krisch (Comeau & Krisch, 2008) expanded this classification by adding metagenomic data from the Global Ocean Sampling (GOS) Expedition, a study of 37 sites from different marine environments (Rusch *et al.*, 2007) and divided T4-type phages into three main groups: Near T4, those closely related to T4; Far T4, those most divergent from T4; and Cyano T4, those related to marine cyanophages. The objective of this study was to investigate the sequence diversity among major capsid protein genes amplified from four wastewater samples collected from the dairy wastewater treatment plant on different occasions and to compare those sequences to those from cultured isolates, those amplified directly from other terrestrial and aquatic environments and to non-PCR-based marine viral metagenome sequences.

Our results suggest that the dairy plant wastewater is dominated by diverse and previously uncharacterized phages. Phylogenetic analysis revealed the existence of at least three previously uncharacterized groups of environmental T4-like capsid sequences (novel clusters W1a, W1b and W2), while the genetic diversity of other groups has been expanded (Figure 5.3). The majority of the sequences recovered by degenerate PCR were not closely related to previously cultured members of the T4-type Myophages. Previously established environmental phage groups associated with rice field floodwater and black upland soil (Paddy groups) (Fujii *et al.*, 2008; Jia *et al.*, 2007; Wang *et al.*, 2009b) and seawater (Marine groups) (Filee *et al.*, 2005) incorporated a minority of sequences in this study (12 out of the 30 representative sequences). The distribution of sequences between Paddy groups (6) and Marine groups (6) was equal, suggesting T4-type phage communities in dairy wastewater are distinct from those of both soil and seawater.

Some sequences were very similar (>90% identity at the amino acid level) to database sequences originating in different locations distant from Ireland (an Antarctic lake, sewage in Portugal, soil in China), and different wastewater, lake and soil environments (Table 5.4). A high level of similarity in samples from a single location with phages from geographically distant environments distributed around a phylogenetic tree, as in this study (Figure 5.4, Table 5.4), supports the hypothesis that phages (or phage genes) move freely between biomes, and that global phage diversity is limited by this horizontal exchange (Breitbart & Rohwer, 2005a; Kunin *et al.*, 2008). The high diversity of T4 phages we have found from a single wastewater site in Ireland (Figure 5.4) mirrors that reported from a freshwater Antarctic lake, Lake Limnopolar (Lopez-Bueno *et al.*, 2009), and suggests that this high level of local diversity is common in freshwater environments.

Only two of the phage groups identified with good bootstrap (a statistical method used to evaluate the reliability of the phylogenetic tree) support contained sequences from different sample times (W12 and W216, in Paddy group V and W21, W23, W217, and W39, in Wastewater cluster W2). Nucleotide and translated sequence groups from the four sample libraries WL1-4 were significantly different (P <0.05) in pairwise comparisons using the ∫-LIBESHUFF community comparison test (Schloss *et al.*, 2004). The detection of different g23 groups in samples taken at the same site at different times is compatible with rapid turnover in phage–host interactions characterised as ‘Kill the winner’ dynamics (Thingstad, 2000). Rapid turnover with little crossover between samples at different time points has also been found in metagenomic sequencing of wastewater phage fractions (Skennerton *et al.*, 2011). However, other factors capable of influencing the bacterial host population include changes in pH and chemical oxygen demand (COD) over time resulting from input of waste from distinct production processes of milk, cheese, butter, and milk powder to the tank (Table 5.1).

Despite the high number of diverse g23 sequences were obtained by PCR, the clone library sequences were almost not detectable in the wastewater metagenome described in Chapter 3 (only 1 read in the wastewater metagenome dataset matched amplified g23 genes), suggesting that capsid genes (or T4-type phages) were present in the metagenome but at a very low level. Similar results were obtained from another study that used the same degenerate primers and found high diversity of major capsid protein genes in an Antarctic Lake (Lopez-Bueno *et al.*, 2009). However analysis of Antarctic Lake viral metagenomes using the MetaVir, a web server enabling automated construction of phylogenies for selected marker genes from publicly available metagenomic data sets (Roux *et al.*, 2011), revealed that the lake metagenome contained only 2 reads with similarity to gp23 proteins (data not shown). Under-sampling of g23 sequences in metagenomic datasets may be a result of bias introduced during the MDA used prior to metagenomic sequencing.

Chapter 6:

Conclusions and future directions

Viruses are the most abundant entities in the aquatic, terrestrial and animal-associated environments and the majority of them are bacteriophages. They can control microbial community composition, contribute to bacterial evolution and affect the cycling of nutrients (Fuhrman, 1999; Fuhrman & Schwalbach, 2003). It has been estimated that there are approximately 10^{31} virus particles in the world (Fuhrman, 1999), however, only about 6,300 types have been identified so far (Ackermann & Prangishvili, 2012), indicating that the current knowledge about their diversity is widely underestimated. Characterising new viruses is difficult because many viruses and their hosts cannot be cultured in the laboratory, and methods such as degenerate PCR are restricted to particular viral groups as no gene is universally present in all viruses (Amann *et al.*, 1995; Rohwer & Edwards, 2002). These limitations can be overcomed by sequencing of viral nucleic acids isolated directly from the environment (viral metagenomics) and no virus cultivation or prior knowledge about the viral types present in the samples is required (Edwards & Rohwer, 2005). Viral metagenomics provides an exhaustive view at virus diversity and it has so far revealed that a large number of uncharacterised viruses exist in nature (Breitbart *et al.*, 2002; Edwards & Rohwer, 2005). It this study metagenomic and genetic analyses were applied to evaluate the diversity of viruses present in a sputum sample collected from a patient with a cystic fibrosis, and an activated sludge sample collected from the dairy wastewater treatment plant in Ireland.

Metagenomic analysis was successful in assembling surprisingly long contigs from short sequence reads. Confirmation of some of the assembly was possible by PCR. Assembled contigs revealed much more detailed information than short read analysis alone. The largest contig (> 60 kb) had weak similarity to phage infecting *Mycobacterium smegmatis* (Figure 2.2), and non-tuberculosis mycobacteria are known to be prevalent in airways of CF patients (Levy *et al.*, 2008). Assembly of such a large contig indicate that the combination of Illumina sequencing and *de novo* assembly seems suitable for future applications of phage metagenomics in monitoring the effects of phage therapy in cystic fibrosis, or phage discovery. Sputum from a cystic fibrosis patient contained DNA typical of phages of bacteria that are traditionally involved in CF lung infections (see Chapter 2 Figure 2.1, Table 2.2), including *Staphylococcus aureus, Pseudomonas aeruginosa, Haemophilus influenzae* and *Streptococcus pneumoniae* (Bittar & Rolain, 2010). Phages of other bacteria that are part of the normal oral flora but have also been described in the context of cystic fibrosis (Bittar & Rolain, 2010) were identified, including those infecting *Prevotella* species, *Streptococcus* species and *Veillonella* species (Table 2.2). Large numbers of anaerobic bacteria within genera *Prevotella* and *Veillonella* recently identified in the sputum of CF patients suggest that they potentially contribute to infection and lung damage (Tunney *et al.*, 2008). Identification of phages infecting CF pathogens may have practical application in phage therapy. Phages have been successfully used to treat bacterial infection in a patient with cystic fibrosis (Kvachadze *et al.*, 2011).

The only eukaryotic virus detected in the CF sputum was Torque Teno virus (TTV), a small circular ssDNA virus from the *Anelloviridae* family (Figure 2.8). TTV has been associated with lower respiratory disease in nasal secretions and bronchoalveolar lavage fluid (Maggi *et al.*, 2003; Wootton *et al.*, 2011), but is also highly prevalent in healthy individuals (Vasilyev *et al.*, 2009). The presence of TTV has been also demonstrated in a viral metagenome recovered from CF lung in another continent, (Willner *et al.*, 2012), therefore it is a consistent presence in CF sputum and potential transmission of this virus by cystic fibrosis sputum is possible. The total virus diversity in CF sputum was predicted to be 89 different viral genotypes (Table 2.4). The majority of reads and contigs did not match any known sequence in the database. Sequence assignment as phages infecting bacteria in CF patients was, however, supported by analysis of the CRISPR spacer matches from CRISPR spacer database (containing collection of CRISPR spacers from 1206 genomes of bacteria and Archaea) to the CF phage reads and contigs (Table 2.9). This provides direct evidence of interactions between the bacteriophages and their hosts. Bioinformatic analysis demonstrated that phages in CF sputum contain genes encoding metallo-β-lactamases (Figure 2.4, Table 2.8), genes potentially conferring resistance to β-lactam antibiotics. The potential of phages to transfer antibiotic resistance genes to bacteria could complicate antibiotic treatment therapy in cystic fibrosis patients. Phages in CF sputum were also potential sources of bacterial virulence genes, including platelet binding factors (Table 2.5) and *Staphylococcus aureus cin* toxin (Table 2.8).

The wastewater phage metagenome showed higher diversity than that of cystic fibrosis sputum, and was estimated to contain 409 different species. The largest contig assembled was 114 kb long contig, which apparently contained complete circular sequence (Figure 3.7). Sequence analysis showed that it had weak similarities to phage SP10 infecting *Bacillus subtilis,* therefore it might represent a novel *Bacillus* phage. Wastewater metagenome had higher proportion of unassigned reads compared to CF metagenome, which may reflect the fact that dairy wastewater is less studied environment. The most abundant sequences identified in dairy wastewater metagenome were prophages of bacteria typically present in wastewater samples (Table 3.2). The directly identified viral fraction in assembled contigs was dominated by double-stranded (ds) DNA bacteriophages from *Siphoviridae*, *Myoviridae* and *Podoviridae* families, represented by phages infecting *Vibrio, Mycobacterium*, *Synechococcus*, *Pseudomonas* and *Burkholderia* species (Figure 3.3). The next most abundant group were single-stranded (ss) DNA bacteriophages from *Microviridae* family, closely related with phages infecting obligate intracellular bacteria, such as Chlamydiae (Figure 3.6). Eukaryotic viral sequences were dominated by viruses containing single-stranded DNA circular genomes, including plant pathogens from the *Geminiviridae* and *Nanoviridae* families, and animal pathogens from the *Circoviridae* family. A major component of the wastewater assembly comprised phages presumably infecting *Chlamydia*-like bacterial symbionts of freshwater amoebae (Corsaro *et al.*, 2009; Pizzetti *et al.*, 2012). The mechanism and effects of phage predation uncovered by this analysis on the life cycle of the enclosed symbiont and its eukaryotic host remain to be investigated. The potential pathogenicity for humans of such *Chlamydia*-like organisms makes such future investigation necessary.

Phage-related genes including structural proteins and genes involved in nucleic acid metabolism, such as DNA replication and synthesis were prevalent in dairy wastewater metagenome (Table 3.5 and Table 3.6). Phage metagenome was particularly enriched for phage-encoded DNA methyltransferase genes. Similar genes overrepresented another wastewater viral metagenome (Tamaki *et al.*, 2012). These genes may play role in switching between phage lytic and lysogenic life cycles (Bochow *et al.*, 2012). Antibiotic resistance genes conferring resistance to β-lactam, vancomycin and trimethoprim antibiotics were found (Table 3.7), and could reflect extensive use of β-lactams and trimethoprim to treat dairy cattle infections in Ireland (More *et al.*, 2012).

PCR amplification of the capsid genes using degenerate primers specific to T4-like phages revealed that dairy wastewater treatment plant harbour diverse and previously uncharacterized phages. The phylogenetic analysis indicated that some wastewater capsid sequences formed clades with previously established phage groups containing PCR-based environmental capsid sequences and cultured isolates (Figure 5.3) and non-PCR-based uncultured marine viral metagenome sequences (Figure 5.4). A high level of similarity in samples from dairy wastewater with phages from geographically distant environments supports the hypothesis that phages (or phage genes) move freely between biomes (Breitbart & Rohwer, 2005a; Kunin *et al.*, 2008) and that T4-like phages are widely dispersed on a global scale. The results also revealed the existence of several wastewater-specific groups of distantly related and previously unknown T4-like viruses.

Preliminary experiments with a promoter trap library confirmed that the wastewater metagenome was a potential source for gene regulatory elements for bacteria. Screening of the metagenomic library revealed that of 20 selected clones that showed strong promoter activity, 10 clones (50%) were derived from single-stranded (ss) DNA viruses from the *Geminiviridae* and *Circoviridae* family, with the majority of ORFs associated with the replication initiation protein (Figure 4.5, Table 4.3). Two inserts contained an ORF associated with phage structural proteins. Six inserts lacked any significant homology to known database sequences (Table 4.3) and three inserts showed only partial sequence homology (Table 4.2). Putative promoters were predicted for 15 inserts. Three of these inserts were characterized by the presence of phage-specific early transcription regulatory cassettes, characterized by presence of the stem-loop structure and the σ^{70}-like promoter located upstream of the ORF (Figure 4.6). These structures may play a role in phage-mediated horizontal gene transfer (Arbiol *et al.*, 2010; Cornelissen *et al.*, 2012). Future work is needed to identify other elements of phage genomes with bacterial regulatory capacity.

References

Abedon, S. T. (2008). Phages, ecology, evolution. In *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*, pp. 1-27. Edited by S. T. Abedon. New York, US: Cambridge University Press.

Abedon, S. T., Kuhl, S. J., Blasdel, B. G. & Kutter, E. M. (2011). Phage treatment of human infections. *Curr Pharm Biotechnol* **1**, 66-85.

Ackermann, H. W. & Krisch, H. M. (1997). A catalogue of T4-type bacteriophages. *Arch Virol* **142**, 2329-2345.

Ackermann, H. W. (2011). The first phage electron micrographs. *Bacteriophage* **1**, 225-227.

Ackermann, H. W. & Prangishvili, D. (2012). Prokaryote viruses studied by electron microscopy. *Arch Virol* **157**, 1843-1849.

Alemayehu, D., Casey, P. G., McAuliffe, O., Guinane, C. M., Martin, J. G., Shanahan, F., Coffey, A., Ross, R. P. & Hill, C. (2012a). Bacteriophages phiMR299-2 and phiNH-4 can eliminate Pseudomonas aeruginosa in the murine lung and on cystic fibrosis lung airway cells. *MBio* **3**, e00029-00012.

Alemayehu, D., Casey, P. G., McAuliffe, O., Guinane, C. M., Martin, J. G., Shanahan, F., Coffey, A., Ross, R. P. & Hill, C. (2012b). Bacteriophages phiMR299-2 and phiNH-4 can eliminate Pseudomonas aeruginosa in the murine lung and on cystic fibrosis lung airway cells. *MBio* **3**.

Alhamlan, F. S., Ederer, M. M., Brown, C. J., Coats, E. R. & Crawford, R. L. (2013). Metagenomics-based analysis of viral communities in dairy lagoon wastewater. *J Microbiol Methods* **92**, 183-188.

Alisky, J., Iczkowski, K., Rapoport, A. & Troitsky, N. (1998). Bacteriophages show promise as antimicrobial agents. *J Infect* **36**, 5-15.

Allander, T., Andreasson, K., Gupta, S., Bjerkner, A., Bogdanovic, G., Persson, M. A., Dalianis, T., Ramqvist, T. & Andersson, B. (2007). Identification of a third human polyomavirus. *J Virol* **81**, 4130-4136.

Allen, H. K., Looft, T., Bayles, D. O., Humphrey, S., Levine, U. Y., Alt, D. & Stanton, T. B. (2011). Antibiotics in feed induce prophages in swine fecal microbiomes. *MBio* **2**, 1-9.

Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell* **42**, 599-610.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**, 143-169.

Anandham, R., Weon, H. Y., Kim, S. J., Kim, Y. S., Kim, B. Y. & Kwon, S. W. (2010). Acinetobacter brisouii sp. nov., isolated from a wetland in Korea. *J Microbiol* **48**, 36-39.

Anderson, R. E., Brazelton, W. J. & Baross, J. A. (2011). Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol Ecol* **77**, 120-133.

Angly, F., Rodriguez-Brito, B., Bangor, D. & other authors (2005a). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41.

Angly, F., Rodriguez-Brito, B., Bangor, D. & other authors (2005b). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41.

Angly, F. E., Felts, B., Breitbart, M. & other authors (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**, e368.

Angly, F. E., Willner, D., Prieto-Davo, A. & other authors (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**, e1000593.

Arbiol, C., Comeau, A. M., Kutateladze, M., Adamia, R. & Krisch, H. M. (2010). Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes. *Genome Biol Evol* **2**, 140-152.

Baker, K. S., Leggett, R. M., Bexfield, N. H. & other authors (2013). Metagenomic study of the viruses of African straw-coloured fruit bats: Detection of a chiropteran poxvirus and isolation of a novel adenovirus. *Virology* **x**, x.

Barr, J. J., Auro, R., Furlan, M. & other authors (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci U S A* **x**, x.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712.

Bateman, A., Birney, E., Cerruti, L. & other authors (2002). The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280.

Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K. & Wommack, K. E. (2007). Metagenomic characterization of Chesapeake Bay virioplankton. *appl environ microbiol* **73**, 7629-7641.

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

Besemer, J. & Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* **27**, 3911-3920.

Bhaya, D., Davison, M. & Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45**, 273-297.

Bibby, K. & Peccia, J. (2013). Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environ Sci Technol* **47**, 1945-1951.

Bittar, F., Richet, H., Dubus, J. C., Reynaud-Gaubert, M., Stremler, N., Sarles, J., Raoult, D. & Rolain, J. M. (2008). Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. *PLoS ONE* **3**, e2908.

Bittar, F. & Rolain, J. M. (2010). Detection and accurate identification of new or emerging bacteria in cystic fibrosis patients. *Clin Microbiol Infect* **16**, 809-820.

Blahova, J., Hupkova, M., Babalova, M., Krcmery, V. & Schafer, V. (1993). Transduction of resistance to Imipenem, Aztreonam and Ceftazidime in nosocomial strains of Pseudomonas aeruginosa by wild-type phages. *Acta Virol* **37**, 429-436.

Blinkova, O., Rosario, K., Li, L., Kapoor, A., Slikas, B., Bernardin, F., Breitbart, M. & Delwart, E. (2009). Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. *J Clin Microbiol* **47**, 3507-3513.

Bochow, S., Elliman, J. & Owens, L. (2012). Bacteriophage adenine methyltransferase: a life cycle regulator? Modelled using Vibrio harveyi myovirus like. *J Appl Microbiol* **113**, 1001-1013.

Boujelben, I., Yarza, P., Almansa, C., Villamor, J., Maalej, S., Anton, J. & Santos, F. (2012). Virioplankton community structure in Tunisian solar salterns. *appl environ microbiol* **78**, 7429-7437.

Boyd, E. F., Heilpern, A. J. & Waldor, M. K. (2000). Molecular analyses of a putative CTXphi precursor and evidence for independent acquisition of distinct CTX(phi)s by toxigenic Vibrio cholerae. *J Bacteriol* **182**, 5530-5538.

Boyd, E. F. & Brussow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol* **10**, 521-529.

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

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**, 6220-6223.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2004a). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**, 565-574.

Breitbart, M., Miyake, J. H. & Rohwer, F. (2004b). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* **236**, 249-256.

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

Breitbart, M. & Rohwer, F. (2005b). Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39**, 729-736.

Breitbart, M., Rohwer, F. & Abedon, S. T. (2005). Phage ecology and bacterial pathogenesis. In *Phages: Their Role in Bacterial Pathogenesis and Biotechnology*, pp. 66-91. Edited by M. K. Waldor, D. I. Friedman & S. L. Adhya. Washington DC: ASM Press.

Breitbart, M., Haynes, M., Kelley, S. & other authors (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**, 367-373.

Britz, T. J., van Schdkwyk, C. & Hung, Y. T. (2005). *Treatment of Dairy Processing Wastewaters*. New York: CRC Press.

Browning, D. F. & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57-65.

Brum, J. R. & Steward, G. F. (2010). Morphological characterization of viruses in the stratified water column of alkaline, hypersaline Mono Lake. *Microb Ecol* **60**, 636-643.

Brussaard, C. P., Marie, D. & Bratbak, G. (2000). Flow cytometric detection of viruses. *J Virol Methods* **85**, 175-182.

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

Brussow, H., Canchaya, C. & Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**, 560-602, table of contents.

Butina, T. V., Belykh, O. I., Maksimenko, S. Y. & Belikov, S. I. (2010). Phylogenetic diversity of T4-like bacteriophages in Lake Baikal, East Siberia. *FEMS Microbiol Lett* **309**, 122-129.

Cahyani, V. R., Murase, J., Asakawa, S. & Kimura, M. (2009a). Change in T4-type bacteriophage communities during the composting process of rice straw: Estimation from the major capsid gene (g23) sequences. *Soil Science and Plant Nutrition* **55**, 468–477.

Cahyani, V. R., Murase, J., Ishibashi, E., Asakawa, S. & Kimura, M. (2009b). T4-type bacteriophage communities estimated from the major capsid genes (*g23*) in manganese nodules in Japanese paddy fields. *Soil Science and Plant Nutrition* **55**, 264–270.

Calendar, R. & Inman, R. (2005). Phage biology. In *Phages: Their role in bacterial pathogenesis and biotechnology*, pp. 18-36. Edited by Waldor M.K., Friedman D.I. & A. S.L. Washington: ASM Press.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brussow, H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**, 417-424.

Cann, A. J., Fandrich, S. E. & Heaphy, S. (2005). Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**, 151-156.

Cantalupo, P. G., Calgua, B., Zhao, G. & other authors (2011). Raw sewage harbors diverse viral populations. *MBio* **2**.

Carlier, J. P., K'Ouas, G., Bonne, I., Lozniewski, A. & Mory, F. (2004). Oribacterium sinus gen. nov., sp. nov., within the family 'Lachnospiraceae' (phylum Firmicutes). *Int J Syst Evol Microbiol* **54**, 1611-1615.

Carmody, L. A., Gill, J. J., Summer, E. J., Sajjan, U. S., Gonzalez, C. F., Young, R. F. & LiPuma, J. J. (2010). Efficacy of bacteriophage therapy in a model of Burkholderia cenocepacia pulmonary infection. *J Infect Dis* **201**, 264-271.

Carvalho, C. M., Alves, E., Costa, L. & other authors (2010). Functional cationic nanomagnet-porphyrin hybrids for the photoinactivation of microorganisms. *ACS Nano* **4**, 7133-7140.

Casas, V., Miyake, J., Balsley, H. & other authors (2006). Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. *FEMS Microbiol Lett* **261**, 141-149.

Casas, V. & Rohwer, F. (2007). Phage metagenomics. *Methods Enzymol* **421**, 259-268.

Cassman, N., Prieto-Davo, A., Walsh, K. & other authors (2012). Oxygen minimum zones harbour novel viral communities with low diversity. *Environ Microbiol* **14**, 3043-3065.

Catalan, V., Garcia, F., Moreno, C., Vila, M. J. & Apraiz, D. (1997). Detection of Legionella pneumophila in wastewater by nested polymerase chain reaction. *Res Microbiol* **148**, 71-78.

Chen, C. (1996). Distribution of a newly described species, Kingella oralis, in the human oral cavity. *Oral Microbiol Immunol* **11**, 425-427.

Chen, F., Suttle, C. A. & Short, S. M. (1996). Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *appl environ microbiol* **62**, 2869-2874.

Chen, F., Wang, K., Huang, S., Cai, H., Zhao, M., Jiao, N. & Wommack, K. E. (2009). Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environ Microbiol* **11**, 2884-2892.

Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* **40**, D641-645.

Chen, S., Bagdasarian, M., Kaufman, M. G. & Walker, E. D. (2007). Characterization of strong promoters from an environmental Flavobacterium hibernum strain by using a green fluorescent protein-based reporter system. *appl environ microbiol* **73**, 1089-1100.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500.

Cheung, J. C. & Deber, C. M. (2008). Misfolding of the Cystic Fibrosis Transmembrane Conductance Regulator and Disease. *Biochemistry* **47**, 1465-1473.

Chiou, C. S., Li, H. Y., Tung, S. K., Chen, C. Y., Teng, C. H., Shu, J. C., Tseng, J. T., Hsu, C. Y. & Chen, C. C. (2010). Identification of prophage gene z2389 in Escherichia coli EDL933 encoding a DNA cytosine methyltransferase for full protection of NotI sites. *Int J Med Microbiol* **300**, 296-303.

Choi, J., Kotay, S. M. & Goel, R. (2011). Bacteriophage-based biocontrol of biological sludge bulking in wastewater. *Bioeng Bugs* **2**, 214-217.

Christie, G. E. & Dokland, T. (2012). Pirates of the caudovirales. *Virology* **434**, 210-221.

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

Coetzee, B., Freeborough, M. J., Maree, H. J., Celton, J. M., Rees, D. J. & Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology* **400**, 157-163.

Collier, J. (2009). Epigenetic regulation of the bacterial cell cycle. *Curr Opin Microbiol* **12**, 722-729.

Colomer-Lluch, M., Jofre, J. & Muniesa, M. (2011). Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS ONE* **6**, e17549.

Comeau, A. M., Bertrand, C., Letarov, A., Tetart, F. & Krisch, H. M. (2007). Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**, 384-396.

Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H. & Prangishvili, D. (2008). Exploring the prokaryotic virosphere. *Res Microbiol* **159**, 306-313.

Comeau, A. M. & Krisch, H. M. (2008). The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol Biol Evol* **25**, 1321-1332.

Compeau, P. E., Pevzner, P. A. & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**, 987-991.

Copeland, A., Sikorski, J., Lapidus, A. & other authors (2009). Complete genome sequence of Atopobium parvulum type strain (IPP 1246). *Stand Genomic Sci* **1**, 166-173.

Cornelissen, A., Hardies, S. C., Shaburova, O. V., Krylov, V. N., Mattheus, W., Kropinski, A. M. & Lavigne, R. (2012). Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse PhiKZ-related phages. *J Virol* **86**, 1844-1852.

Corsaro, D. & Greub, G. (2006). Pathogenic potential of novel Chlamydiae and diagnostic approaches to infections due to these obligate intracellular bacteria. *Clin Microbiol Rev* **19**, 283-297.

Corsaro, D., Feroldi, V., Saucedo, G., Ribas, F., Loret, J. F. & Greub, G. (2009). Novel Chlamydiales strains isolated from a water treatment plant. *Environ Microbiol* **11**, 188-200.

Coulon, C., Eterpi, M., Greub, G., Collignon, A., McDonnell, G. & Thomas, V. (2012). Amoebal host range, host-free survival and disinfection susceptibility of environmental Chlamydiae as compared to Chlamydia trachomatis. *FEMS Immunol Med Microbiol* **63**, 364-373.

Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature* **192**, 1227-1232.

Culley, A. I., Lang, A. S. & Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795-1798.

Dafale, N., Agrawal, L., Kapley, A., Meshram, S., Purohit, H. & Wate, S. (2010). Selection of indicator bacteria based on screening of 16S rDNA metagenomic library from a two-stage anoxic-oxic bioreactor system degrading azo dyes. *Bioresour Technol* **101**, 476-484.

Danovaro, R., Dell'Anno, A., Trucco, A., Serresi, M. & Vanucci, S. (2001). Determination of virus abundance in marine sediments. *appl environ microbiol* **67**, 1384-1387.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fabiano, M. & Corselli, C. (2005). Viruses, prokaryotes and DNA in the sediments of a deep-hypersaline anoxic basin (DHAB) of the Mediterranean Sea. *Environ Microbiol* **7**, 586-592.

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P. & Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* **13**, 299.

Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-1099.

Debarbieux, L., Leduc, D., Maura, D., Morello, E., Criscuolo, A., Grossi, O., Balloy, V. & Touqui, L. (2010). Bacteriophages can treat and prevent Pseudomonas aeruginosa lung infections. *J Infect Dis* **201**, 1096-1104.

Decostere, A., Haesebrouck, F. & Devriese, L. A. (1997). Shieh medium supplemented with tobramycin for selective isolation of Flavobacterium columnare (Flexibacter columnaris) from diseased fish. *J Clin Microbiol* **35**, 322-324.

Del Casale, A., Flanagan, P. V., Larkin, M. J., Allen, C. C. & Kulakov, L. A. (2011a). Extent and variation of phage-borne bacterial 16S rRNA gene sequences in wastewater environments. *appl environ microbiol* **77**, 5529-5532.

Del Casale, A., Flanagan, P. V., Larkin, M. J., Allen, C. C. & Kulakov, L. A. (2011b). Analysis of transduction in wastewater bacterial populations by targeting the phage-derived 16S rRNA gene sequences. *FEMS Microbiol Ecol* **76**, 100-108.

Delwart, E. & Li, L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Res* **164**, 114-121.

Delwart, E. L. (2007). Viral metagenomics. *Rev Med Virol* **17**, 115-131.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S. & other authors (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**, 340-343.

Desplats, C. & Krisch, H. M. (2003). The diversity and evolution of the T4-type bacteriophages. *Res Microbiol* **154**, 259-267.

Deveau, H., Garneau, J. E. & Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* **64**, 475-493.

Djikeng, A., Kuzmickas, R., Anderson, N. G. & Spiro, D. J. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS ONE* **4**, e7264.

Dong, X., Stothard, P., Forsythe, I. J. & Wishart, D. S. (2004). PlasMapper: a web server for drawing and auto-annotating plasmid maps. *Nucleic Acids Res* **32**, W660-664.

Downes, J., Olsvik, B., Hiom, S. J., Spratt, D. A., Cheeseman, S. L., Olsen, I., Weightman, A. J. & Wade, W. G. (2000). Bulleidia extructa gen. nov., sp. nov., isolated from the oral cavity. *Int J Syst Evol Microbiol* **50 Pt 3**, 979-983.

Drozdz, M., Piekarowicz, A., Bujnicki, J. M. & Radlinska, M. (2011). Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res* **40**, 2119-2130.

Duckworth, D. H. (1976). "Who discovered bacteriophage?". *Bacteriol Rev* **40**, 793-802.

Dugdale, B., Beetham, P. R., Becker, D. K., Harding, R. M. & Dale, J. L. (1998). Promoter activity associated with the intergenic regions of banana bunchy top virus DNA-1 to -6 in transgenic tobacco and banana cells. *J Gen Virol* **79 (Pt 10)**, 2301-2311.

Dungan, R. S., Klein, M. & Leytem, A. B. (2012). Quantification of Bacterial Indicators and Zoonotic Pathogens in Dairy Wastewater Ponds. *Appl Environ Microbiol* **78**.

Dunn, A. K. & Handelsman, J. (1999). A vector for promoter trapping in Bacillus cereus. *Gene* **226**, 297-305.

Ebright, R. H. (2000). RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* **304**, 687-698.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.

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

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B. & Banfield, J. F. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *appl environ microbiol* **78**, 6309-6320.

Emond, E. & Moineau, S. (2007). Bacteriophages and Food Fermentations. In *Bacteriophage: Genetics and Molecular Biology*. Edited by S. Mc Grath & D. van Sinderen. Norfolk, UK: Caister Academic Press.

Fabbi, M., Pastoris, M. C., Scanziani, E., Magnino, S. & Di Matteo, L. (1998). Epidemiological and environmental investigations of Legionella pneumophila infection in cattle and case report of fatal pneumonia in a calf. *J Clin Microbiol* **36**, 1942-1947.

Fancello, L., Desnues, C., Raoult, D. & Rolain, J. M. (2011). Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota. *J Antimicrob Chemother* **66**.

Fancello, L., Raoult, D. & Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology* **434**, 162-174.

Fancello, L., Trape, S., Robert, C., Boyer, M., Popgeorgiev, N., Raoult, D. & Desnues, C. (2013). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *ISME J* **7**, 359-369.

Farabaugh, P. J. (1996). Programmed translational frameshifting. *Annu Rev Genet* **30**, 507-528.

Farrell, P., Joffe, S., Foley, L., Canny, G. J., Mayne, P. & Rosenberg, M. (2007). Diagnosis of cystic fibrosis in the Republic of Ireland: epidemiology and costs. *Ir Med J* **100**, 557-560.

Fassler, J. S. & Gussin, G. N. (1996). Promoters and basal transcription machinery in eubacteria and eukaryotes: concepts, definitions, and analogies. *Methods Enzymol* **273**, 3-29.

Felsenstein, J. (2005).PHYLIP (Phylogeny Inference Package) version 3.6, pp. Distributed by the author: Department of Genome Sciences, University of Washington, Seattle.

Fierer, N., Breitbart, M., Nulton, J. & other authors (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.

Filee, J., Tetart, F., Suttle, C. A. & Krisch, H. M. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci U S A* **102**, 12471-12476.

Finkbeiner, S. R., Allred, A. F., Tarr, P. I., Klein, E. J., Kirkwood, C. D. & Wang, D. (2008). Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog* **4**, e1000011.

Fothergill, J. L., Mowat, E., Ledson, M. J., Walshaw, M. J. & Winstanley, C. (2010). Fluctuations in phenotypes and genotypes within populations of Pseudomonas aeruginosa in the cystic fibrosis lung during pulmonary exacerbations. *J Med Microbiol* **59**, 472-481.

Foulongne, V., Sauvage, V., Hebert, C. & other authors (2012). Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS ONE* **7**, e38499.

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

Fricke, W. F., Seedorf, H., Henne, A., Kruer, M., Liesegang, H., Hedderich, R., Gottschalk, G. & Thauer, R. K. (2006). The genome sequence of Methanosphaera stadtmanae reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. *J Bacteriol* **188**, 642-658.

Friedman, M. G., Dvoskin, B. & Kahane, S. (2003). Infections with the chlamydia-like microorganism Simkania negevensis, a possible emerging pathogen. *Microbes Infect* **5**, 1013-1021.

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

Fuhrman, J. A., Griffith, J. & Schwalbach, M. S. (2002). Prokaryotic and viral diversity patterns in marine plankton. *Ecological Research* **17**, 183–194.

Fuhrman, J. A. & Schwalbach, M. (2003). Viral influence on aquatic bacterial communities. *Biol Bull* **204**, 192-195.

Fujihara, S., Murase, J., Tun, C., Matsuyama, T., Ikenaga, M., Asakawa, S. & Kimura, M. (2010). Low diversity of T4-type bacteriophages in applied rice straw, plant residues and rice roots in Japanese rice soils: Estimation from major capsid gene (g23) composition *Soil Science and Plant Nutrition* **56**, 800-812.

Fujii, T., Nakayama, N., Nishida, M., Sekiya, H., Kato, N., Asakawa, S. & Kimura, M. (2008). Novel capsid genes (g23) of T4-type bacteriophages in a Japanese paddy field. *Soil Biology & Biochemistry* **40**, 1049–1058.

Fuller, C. W., Middendorf, L. R., Benner, S. A. & other authors (2009). The challenges of sequencing by synthesis. *Nat Biotechnol* **27**, 1013-1023.

Garcia Martin, H., Ivanova, N., Kunin, V. & other authors (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**, 1263-1269.

Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., Zhang, Y. & Shi, Z. (2011). Genetic diversity of novel circular ssDNA viruses in bats in China. *J Gen Virol* **92**, 2646-2653.

Ghai, R., Rodriguez-Valera, F., McMahon, K. D., Toyama, D., Rinke, R., Cristina Souza de Oliveira, T., Wagner Garcia, J., Pellon de Miranda, F. & Henrique-Silva, F. (2011). Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS ONE* **6**, e23785.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**, 759-769.

Glinkowska, M., Los, J. M., Szambowska, A. & other authors (2010). Influence of the Escherichia coli oxyR gene function on lambda prophage maintenance. *Arch Microbiol* **192**, 673-683.

Glonti, T., Chanishvili, N. & Taylor, P. W. (2010). Bacteriophage-derived enzyme that depolymerizes the alginic acid capsule associated with cystic fibrosis isolates of Pseudomonas aeruginosa. *J Appl Microbiol* **108**, 695-702.

Goyal, S. M., Zerda, K. S. & Gerba, C. P. (1980). Concentration of coliphages from large volumes of water and wastewater. *appl environ microbiol* **39**, 85-91.

Grant, J. R. & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res* **36**, W181-184.

Greer, G. G. (2005). Bacteriophage control of foodborne bacteriat. *J Food Prot* **68**, 1102-1111.

Greub, G. (2009). Parachlamydia acanthamoebae, an emerging agent of pneumonia. *Clin Microbiol Infect* **15**, 18-28.

Grissa, I., Vergnaud, G. & Pourcel, C. (2007a). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**, W52-57.

Grissa, I., Vergnaud, G. & Pourcel, C. (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

Gronow, S., Welnitz, S., Lapidus, A. & other authors (2010). Complete genome sequence of Veillonella parvula type strain (Te3). *Stand Genomic Sci* **2**, 57-65.

Guss, A. M., Roeselers, G., Newton, I. L., Young, C. R., Klepac-Ceraj, V., Lory, S. & Cavanaugh, C. M. (2011). Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J* **5**, 20-29.

Hammerschlag, M. R., Harding, L., Macone, A., Smith, A. L. & Goldmann, D. A. (1980). Bacteriology of sputum in cystic fibrosis: evaluation of dithiothreitol as a mucolytic agent. *J Clin Microbiol* **11**, 552-557.

Han, S. S., Lee, J. Y., Kim, W. H., Shin, H. J. & Kim, G. J. (2008). Screening of promoters from metagenomic DNA and their use for the construction of expression vectors. *J Microbiol Biotechnol* **18**, 1634-1640.

Han, X. Y., Hong, T. & Falsen, E. (2006). Neisseria bacilliformis sp. nov. isolated from human infections. *J Clin Microbiol* **44**, 474-479.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245-249.

Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S. & Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virol J* **9**, 9.

Hara, S., Terauchi, K. & Koike, I. (1991). Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. *appl environ microbiol* **57**, 2731-2734.

Haraszthy, V. I., Zambon, J. J., Sreenivasan, P. K., Zambon, M. M., Gerber, D., Rego, R. & Parker, C. (2007). Identification of oral bacterial species associated with halitosis. *J Am Dent Assoc* **138**, 1113-1120.

Harper, D. R. & Enright, M. C. (2011). Bacteriophages for the treatment of Pseudomonas aeruginosa infections. *J Appl Microbiol* **111**, 1-7.

Harper, D. R. & Morales, S. (2012). Bacteriophage therapy: practicability and clinical need meet in the multidrug-resistance era. *Future Microbiol* **7**, 797-799.

Harrington, C., Del Casale, A., Kennedy, J. & other authors (2012). Evidence of bacteriophage-mediated horizontal transfer of bacterial 16s rRNA genes in the viral metagenome of the marine sponge Hymeniacidon perlevis. *Microbiology* **1**, 1.

Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D. & other authors (2006). Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet* **2**, e92.

Hatfull, G. F. (2008). Bacteriophage genomics. *Curr Opin Microbiol* **11**, 447-453.

Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G. & other authors (2010). Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol* **397**, 119-143.

Hatfull, G. F. & Hendrix, R. W. (2011). Bacteriophages and their genomes. *Curr Opin Virol* **1**, 298-303.

Hautefort, I., Proenca, M. J. & Hinton, J. C. (2003). Single-copy green fluorescent protein gene fusions allow accurate measurement of Salmonella gene expression in vitro and during infection of mammalian cells. *appl environ microbiol* **69**, 7480-7491.

Hawley, D. K. & McClure, W. R. (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res* **11**, 2237-2255.

Haynes, M. & Rohwer, F. (2011). The Human Virome. In *Metagenomics of the Human Body*, pp. 63-77. Edited by K. E. Nelson: Springer.

Helton, R. R., Liu, L. & Wommack, K. E. (2006). Assessment of factors influencing direct enumeration of viruses within estuarine sediments. *appl environ microbiol* **72**, 4767-4774.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E. & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* **96**, 2192-2197.

Henne, A., Daniel, R., Schmitz, R. A. & Gottschalk, G. (1999). Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *appl environ microbiol* **65**, 3901-3907.

Hershey, A. D. & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56.

Hilliard, T. N., Sukhani, S., Francis, J., Madden, N., Rosenthal, M., Balfour-Lynn, I., Bush, A. & Davies, J. C. (2007). Bronchoscopy following diagnosis with cystic fibrosis. *Arch Dis Child* **92**, 898-899.

Himpsl, S. D., Pearson, M. M., Arewang, C. J., Nusca, T. D., Sherman, D. H. & Mobley, H. L. (2010). Proteobactin and a yersiniabactin-related siderophore mediate iron acquisition in Proteus mirabilis. *Mol Microbiol* **78**, 138-157.

Hinton, D. M. (2010). Transcriptional control in the prereplicative phase of T4 development. *Virol J* **7**, 289.

Horvath, P. & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170.

Huang, C. Y., Patel, B. K., Mah, R. A. & Baresi, L. (1998). Caldicellulosiruptor owensensis sp. nov., an anaerobic, extremely thermophilic, xylanolytic bacterium. *Int J Syst Bacteriol* **48 Pt 1**, 91-97.

Huber, T., Faulkner, G. & Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319.

Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res* **21**, 1552-1560.

Jacob, D., Lewin, A., Meister, B. & Appel, B. (2002). Plant-specific promoter sequences carry elements that are recognised by the eubacterial transcription machinery. *Transgenic Res* **11**, 291-303.

Jenkins, F. J., Howett, M. K. & Rapp, F. (1983). Simian virus 40 promoters direct expression of the tetracycline gene in plasmid pACYC184. *J Virol* **45**, 478-481.

Jia, Z., Ishihara, R., Nakajima, Y., Asakawa, S. & Kimura, M. (2007). Molecular characterization of T4-type bacteriophages in a rice field. *Environ Microbiol* **9**, 1091-1096.

Jiang, S. C. & Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol* **64**, 2780-2787.

Jorgensen, K. S. & Pauli, A. S. (1995). Polyphosphate accumulation among denitrifying bacteria in activated sludge. *Anaerobe* **1**, 161-168.

Ju, J., Kim, D. H., Bi, L. & other authors (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **103**, 19635-19640.

Kaliniene, L., Klausa, V. & Truncaite, L. (2010). Low-temperature T4-like coliphages vB_EcoM-VR5, vB_EcoM-VR7 and vB_EcoM-VR20. *Arch Virol* **155**, 871-880.

Kazmierczak, M. J., Wiedmann, M. & Boor, K. J. (2005). Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev* **69**, 527-543.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664.

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W., Kim, M. S., Sung, Y., Jeon, C. O., Oh, H. M. & Bae, J. W. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *appl environ microbiol* **74**, 5975-5985.

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W. & Bae, J. W. (2010). Phenotypic characterization and genomic analysis of the Shigella sonnei bacteriophage SP18. *J Microbiol* **48**, 213-222.

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**, 7663-7668.

Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *appl environ microbiol* **77**, 8062-8070.

King, A. M., Elliot Lefkowitz, Michael J. Adams & Carstens, E. B. (2011). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier.

Kropinski, A. M., Mazzocco, A., Waddell, T. E., Lingohr, E. & Johnson, R. P. (2009). Enumeration of Bacteriophages by Double Agar Overlay Plaque Assay. In *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*, pp. 69-76. Edited by M. R. Clokie & A. M. Kropinski. Hatfield, UK: Humana Press.

Kruger, D. H. & Bickle, T. A. (1983). Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiol Rev* **47**, 345-360.

Kunin, V., He, S., Warnecke, F. & other authors (2008). A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* **18**, 293-297.

Kvachadze, L., Balarjishvili, N., Meskhi, T. & other authors (2011). Evaluation of lytic activity of staphylococcal bacteriophage Sb-1 against freshly isolated clinical pathogens. *Microb Biotechnol* **4**, 643-650.

Kwan, T., Liu, J., DuBow, M., Gros, P. & Pelletier, J. (2005). The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. *Proc Natl Acad Sci U S A* **102**, 5174-5179.

Kwan, T., Liu, J., Dubow, M., Gros, P. & Pelletier, J. (2006). Comparative genomic analysis of 18 Pseudomonas aeruginosa bacteriophages. *J Bacteriol* **188**, 1184-1187.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J. M. & Raoult, D. (2003). A giant virus in amoebae. *Science* **299**, 2033.

Lane, D. J. (1991). 16S/23S rRNA sequencing. In *Nucleic Acid Techniques in Bacterial Systematics*, pp. 115–148. Edited by E. Stackebrandt & M. Goodfellow. Chichester, New York: Wiley.

Larkin, M. A., Blackshields, G., Brown, N. P. & other authors (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Laserson, J., Jojic, V. & Koller, D. (2011). Genovo: de novo assembly for metagenomes. *J Comput Biol* **18**, 429-443.

Lasken, R. S. & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**, 19.

Lee, S. H., Kim, J. M., Lee, H. J. & Jeon, C. O. (2011). Screening of promoters from rhizosphere metagenomic DNA using a promoter-trap vector and flow cytometric cell sorting. *J Basic Microbiol* **51**, 52-60.

Leiman, P. G., Kanamaru, S., Mesyanzhinov, V. V., Arisaka, F. & Rossmann, M. G. (2003). Structure and morphogenesis of bacteriophage T4. *Cell Mol Life Sci* **60**, 2356-2370.

Letunic, I. & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, 475-478.

Levy, I., Grisaru-Soen, G., Lerner-Geva, L. & other authors (2008). Multicenter cross-sectional study of nontuberculous mycobacterial infections among cystic fibrosis patients, Israel. *Emerg Infect Dis* **14**, 378-384.

Lewin, A., Mayer, M., Chusainow, J., Jacob, D. & Appel, B. (2005). Viral promoters can initiate expression of toxin genes introduced into Escherichia coli. *BMC Biotechnol* **5**, 19.

Li, D., Aaskov, J. & Lott, W. B. (2011a). Identification of a cryptic prokaryotic promoter within the cDNA encoding the 5' end of dengue virus RNA genome. *PLoS ONE* **6**, e18197.

Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., Kunz, T. H. & Delwart, E. (2010a). Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J Virol* **84**, 6955-6965.

Li, L., Pesavento, P. A., Shan, T., Leutenegger, C. M., Wang, C. & Delwart, E. (2011b). Viruses in diarrhoeic dogs include novel kobuviruses and sapoviruses. *J Gen Virol* **92**, 2534-2541.

Li, L., Shan, T., Soji, O. B., Alam, M. M., Kunz, T. H., Zaidi, S. Z. & Delwart, E. (2011c). Possible cross-species transmission of circoviruses and cycloviruses among farm animals. *J Gen Virol* **94**, 768-772.

Li, R., Zhu, H., Ruan, J. & other authors (2010b). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272.

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F. & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci U S A* **101**, 11013-11018.

Little, J. W. (2005). Lysogeny, prophage induction, and lysogenic conversion. In *Phages: Their Role in Bacterial Pathogenesis and Biotechnology*, pp. 37–54. Edited by M. K. Waldor, D. I. Friedman & S. L. Adhya. Washington: ASM Press.

Liu, B., Frostegard, A. & Shapleigh, J. P. (2013). Draft genome sequences of five strains in the genus thauera. *Genome Announc* **1**, e00052-00012.

Liu, H., Fu, Y., Li, B. & other authors (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol Biol* **11**, 276.

Loperena, L., Ferrari, M. D., Diaz, A. L., Ingold, G., Perez, L. V., Carvallo, F., Travers, D., Menes, R. J. & Lareo, C. (2009). Isolation and selection of native microorganisms for the aerobic treatment of simulated dairy wastewaters. *Bioresour Technol* **100**, 1762-1766.

Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A. & Alcami, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**, 858-861.

Los, M. & Wegrzyn, G. (2012). Pseudolysogeny. *Adv Virus Res*, 339-349.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* **7**, e30087.

Lyczak, J. B., Cannon, C. L. & Pier, G. B. (2002). Lung infections associated with cystic fibrosis. *Clin Microbiol Rev* **15**, 194-222.

Lysholm, F., Wetterbom, A., Lindau, C. & other authors (2012). Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE* **7**, e30875.

Maggi, F., Pifferi, M., Fornai, C. & other authors (2003). TT virus in the nasal secretions of children with acute respiratory diseases: relations to viremia and disease severity. *J Virol* **77**, 2418-2425.

Mankertz, A. & Hillenbrand, B. (2002). Analysis of transcription of Porcine circovirus type 1. *J Gen Virol* **83**, 2743-2751.

Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**, 741.

Mann, N. H., Clokie, M. R., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A. & Krisch, H. M. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus strains. *J Bacteriol* **187**, 3188-3200.

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

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402.

Margulies, M., Egholm, M., Altman, W. E. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Marhaver, K. L., Edwards, R. A. & Rohwer, F. (2008). Viral communities associated with healthy and bleaching corals. *Environ Microbiol* **10**, 2277-2286.

Marie, D., Brussaard, C. P. D., Thyrhaug, R., Bratbak, G. & Vaulot, D. (1999). Enumeration of marine viruses in culture and natural samples by flow cytometry. *appl environ microbiol* **65**, 45-52.

Markowitz, V. M., Ivanova, N. N., Szeto, E. & other authors (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**, D534-538.

Masembe, C., Michuki, G., Onyango, M. & other authors (2012). Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus. *Virol J* **9**, 218.

Matsuzaki, S., Inoue, T., Kuroda, M., Kimura, S. & Tanaka, S. (1998). Cloning and sequencing of major capsid protein (mcp) gene of a vibriophage, KVP20, possibly related to T-even coliphages. *Gene* **222**, 25-30.

Maynard, N. D., Macklin, D. N., Kirkegaard, K. & Covert, M. W. (2012). Competing pathways control host resistance to virus via tRNA modification and programmed ribosomal frameshifting. *Mol Syst Biol* **8**.

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F. & Paul, J. H. (2008). Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS ONE* **3**, e3263.

McGarvey, J. A., Miller, W. G., Zhang, R., Ma, Y. & Mitloehner, F. (2007). Bacterial population dynamics in dairy waste during aerobic and anaerobic treatment and subsequent storage. *appl environ microbiol* **73**, 193-202.

Mekalanos, J. J., Rubin, E. J. & Waldor, M. K. (1997). Cholera: molecular basis for emergence and pathogenesis. *FEMS Immunol Med Microbiol* **18**, 241-248.

Meyer, F., Paarmann, D., D'Souza, M. & other authors (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386.

Millard, A. D. (2009). Isolation of Cyanophages from Aquatic Environments. In *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*, pp. 33-42. Edited by M. R. Clokie & A. M. Kropinski. Hatfield, UK: Humana Press.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. & Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**, 1616-1625.

Mitsialis, S. A., Young, J. F., Palese, P. & Guntaka, R. V. (1981). An avian tumor virus promoter directs expression of plasmid genes in Escherichia coli. *Gene* **16**, 217-225.

Mokili, J. L., Rohwer, F. & Dutilh, B. E. (2011). Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**, 63-77.

Mokili, J. L., Dutilh, B. E., Lim, Y. W. & other authors (2013). Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS ONE* **8**, e58404.

Molineux, I. J. & Panja, D. (2013). Popping the cork: mechanisms of phage genome ejection. *Nat Rev Microbiol* **11**, 194-204.

More, S. J., Clegg, T. A. & O'Grady, L. (2012). Insights into udder health and intramammary antibiotic usage on Irish dairy farms during 2003-2010. *Ir Vet J* **65**, 7.

Moser, M. J., DiFrancesco, R. A., Gowda, K., Klingele, A. J., Sugar, D. R., Stocki, S., Mead, D. A. & Schoenfeld, T. W. (2012). Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. *PLoS ONE* **7**, e38371.

Muniesa, M., Garcia, A., Miro, E., Mirelis, B., Prats, G., Jofre, J. & Navarro, F. (2004). Bacteriophages and diffusion of beta-lactamase genes. *Emerg Infect Dis* **10**, 1134-1137.

Nakamura, S., Yang, C. S., Sakon, N. & other authors (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* **4**, e4219.

Nakayama, N., Asakawa, S. & Kimura, M. (2009a). Comparison of g23 gene sequence diversity between Novosphingobium and Sphingomonas phages and phage communities in the floodwater of a Japanese paddy field. *Soil Biol Biochem* **41**, 179-185.

Nakayama, N., Tsuge, T., Asakawa, S. & Kimura, M. (2009b). Morphology, host range and phylogenetic diversity of Sphingomonas phages in the floodwater of a Japanese paddy field. *Soil Science and Plant Nutrition* **55**, 53-64.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* **40**, e155.

Narihiro, T., Terada, T., Kikuchi, K. & other authors (2009). Comparative analysis of bacterial and archaeal communities in methanogenic sludge granules from upflow anaerobic sludge blanket reactors treating various food-processing, high-strength organic wastewaters. *Microbes Environ* **24**, 88-96.

Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibanez, J. T. & Hanley-Bowdoin, L. (2011). Functional analysis of a novel motif conserved across geminivirus Rep proteins. *J Virol* **85**, 1182-1192.

Nayar, G. P., Hamel, A. L., Lin, L., Sachvie, C., Grudeski, E. & Spearman, G. (1999). Evidence for circovirus in cattle with respiratory disease and from aborted bovine fetuses. *Can Vet J* **40**, 277-278.

Ng, T. F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L. & Breitbart, M. (2009). Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. *J Virol* **83**, 2500-2509.

Ng, T. F., Willner, D. L., Lim, Y. W. & other authors (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579.

Ng, T. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J Virol* **86**, 12161-12175.

Niu, B., Zhu, Z., Fu, L., Wu, S. & Li, W. (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27**, 1704-1705.

Noble, R. T. & Fuhrman, J. A. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology* **14**, 113-118.

Overbeek, R., Begley, T., Butler, R. M. & other authors (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**, 5691-5702.

Paget, M. S. & Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biol* **4**, 203.

Palacios, G., Druce, J., Du, L. & other authors (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med* **358**, 991-998.

Park, E. J., Kim, K. H., Abell, G. C., Kim, M. S., Roh, S. W. & Bae, J. W. (2011). Metagenomic analysis of the viral communities in fermented foods. *appl environ microbiol* **77**, 1284-1291.

Park, S. Y. & Kim, G. J. (2010). Screening of functional promoter from metagenomic DNA for practical use in expression systems. *Methods Mol Biol* **668**, 141-152.

Parracho, H. M., Burrowes, B. H., Enright, M. C., McConville, M. L. & Harper, D. R. (2012). The role of regulated clinical trials in the development of bacteriophage therapeutics. *J Mol Genet Med* **6**, 279-286.

Parsley, L. C., Consuegra, E. J., Kakirde, K. S., Land, A. M., Harper, W. F., Jr. & Liles, M. R. (2010a). Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *appl environ microbiol* **76**.

Parsley, L. C., Consuegra, E. J., Thomas, S. J. & other authors (2010b). Census of the viral metagenome within an activated sludge microbial assemblage. *Appl Environ Microbiol* **76**, 2673-2677.

Paul, J. H., Sullivan, M. B., Segall, A. M. & Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**, 463-476.

Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* **27**, 94-101.

Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428.

Pereira, M. S., Barreto, V. P. & Siqueira-Junior, J. P. (1997). Phage-mediated transfer of tetracycline resistance in Staphylococcus aureus isolated from cattle in Brazil. *Microbios* **92**, 147-155.

Petrov, V. M., Nolan, J. M., Bertrand, C., Levy, D., Desplats, C., Krisch, H. M. & Karam, J. D. (2006). Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol* **361**, 46-68.

Petrov, V. M., Ratnayaka, S., Nolan, J. M., Miller, E. S. & Karam, J. D. (2010a). Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol J* **7**, 292.

Petrov, V. M., Ratnayaka, S., Nolan, J. M., Miller, E. S. & Karam, J. D. (2010b). Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol J* **7**, 292.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The fecal viral flora of wild rodents. *PLoS Pathog* **7**, e1002218.

Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., Tartaro, K. R., Plant, R. N., Egholm, M., Rothberg, J. M. & Leamon, J. H. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216.

Pinho-Nascimento, C. A., Leite, J. P., Niel, C. & Diniz-Mendes, L. (2011). Torque teno virus in fecal samples of patients with gastroenteritis: prevalence, genogroups distribution, and viral load. *J Med Virol* **83**, 1107-1111.

Pirnay, J. P., De Vos, D., Verbeken, G. & other authors (2010). The phage therapy paradigm: pret-a-porter or sur-mesure? *Pharm Res* **28**.

Pizzetti, I., Fazi, S., Fuchs, B. M. & Amann, R. (2012). High abundance of novel environmental chlamydiae in a Tyrrhenian coastal lake (Lago di Paola, Italy). *Environ Microbiol Rep* **4**, 446-452.

Polson, S. W., Wilhelm, S. W. & Wommack, K. E. (2011). Unraveling the viral tapestry (from inside the capsid out). *ISME J* **5**, 2010/2006/2018.

Pride, D. T. & Schoenfeld, T. (2008). Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics* **9**, 420.

Pride, D. T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R. A., 3rd, Loomer, P., Armitage, G. C. & Relman, D. A. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J* **6**, 915-926.

Rahme, L. G., Stevens, E. J., Wolfort, S. F., Shao, J., Tompkins, R. G. & Ausubel, F. M. (1995). Common virulence factors for bacterial pathogenicity in plants and animals. *Science* **268**, 1899-1902.

Ratjen, F. & Doring, G. (2003). Cystic fibrosis. *Lancet* **361**, 681-689.

Resch, G., Kulik, E. M., Dietrich, F. S. & Meyer, J. (2004). Complete genomic nucleotide sequence of the temperate bacteriophage Aa Phi 23 of Actinobacillus actinomycetemcomitans. *J Bacteriol* **186**, 5523-5528.

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F. & Gordon, J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277.

Ripp, S. & Miller, R. V. (1997). The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment. *Microbiology* **143**, 2065-2070.

Rohwer, F. & Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**, 4529-4535.

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

Rolain, J. M., Fancello, L., Desnues, C. & Raoult, D. (2011). Bacteriophages as vehicles of the resistome in cystic fibrosis. *J Antimicrob Chemother* **66**, 2444-2447.

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Res* **11**, 3-11.

Rosario, K., Duffy, S. & Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J Gen Virol* **90**, 2418-2424.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* **11**, 2806-2820.

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

Rosario, K., Duffy, S. & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**, 1851-1871.

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., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012a). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* **7**, e33641.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012b). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* **7**, 505-507.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386.

Rusch, D. B., Halpern, A. L., Sutton, G. & other authors (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**, e77.

Salinero, K. K., Keller, K., Feil, W. S., Feil, H., Trong, S., Di Bartolo, G. & Lapidus, A. (2009). Metabolic analysis of the soil microbe Dechloromonas aromatica str. RCB: indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation. *BMC Genomics* **10**, 351.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular cloning: a laboratory manual*: Cold Spring Harbor Laboratory Press.

Sander, M. & Schmieger, H. (2001). Method for host-independent detection of generalized transducing bacteriophages in natural habitats. *appl environ microbiol* **67**, 1490-1493.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. & Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467.

Sayers, E. W., Barrett, T., Benson, D. A. & other authors (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15.

Scanlan, P. D. & Buckling, A. (2012). Co-evolution with lytic phage selects for the mucoid phenotype of Pseudomonas fluorescens SBW25. *ISME J* **6**, 1148-1158.

Schloss, P. D., Larget, B. R. & Handelsman, J. (2004). Integration of Microbial Ecology and Statistics: a Test To Compare Gene Libraries. *Applied and Environmental Microbiology* **70**, 5485-5492.

Schloss, P. D., Westcott, S. L., Ryabin, T. & other authors (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *appl environ microbiol* **75**, 7537-7541.

Schmieder, R. & Edwards, R. (2011a). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864.

Schmieder, R. & Edwards, R. (2011b). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288.

Schmitz, J. E., Schuch, R. & Fischetti, V. A. (2010). Identifying active phage lysins through functional viral metagenomics. *appl environ microbiol* **76**, 7181-7187.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M. & Mead, D. (2008). Assembly of viral metagenomes from yellowstone hot springs. *Appl Environ Microbiol* **74**, 4164-4174.

Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489-491.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**, e75.

Shapiro, O. H., Kushmaro, A. & Brenner, A. (2010). Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J* **4**, 327-336.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145.

Shirasawa-Seo, N., Sano, Y., Nakamura, S. & other authors (2005). The promoter of Milk vetch dwarf virus component 8 confers effective gene expression in both dicot and monocot plants. *Plant Cell Rep* **24**, 155-163.

Short, C. M. & Suttle, C. A. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *appl environ microbiol* **71**, 480-486.

Sigdel, T. K., Easton, J. A. & Crowder, M. W. (2006). Transcriptional response of Escherichia coli to TPEN. *J Bacteriol* **188**, 6709-6713.

Sinoquet, C., Demey, S. & Braun, F. (2008). Large-scale computational and statistical analyses of high transcription potentialities in 32 prokaryotic genomes. *Nucleic Acids Res* **36**, 3332-3340.

Skennerton, C. T., Angly, F. E., Breitbart, M., Bragg, L., He, S., McMahon, K. D., Hugenholtz, P. & Tyson, G. W. (2011). Phage encoded H-NS: a potential achilles heel in the bacterial defence system. *PLoS ONE* **6**.

Soding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-248.

Sorek, R., Kunin, V. & Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181-186.

Sorek, R., Lawrence, C. M. & Wiedenheft, B. (2013). CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annu Rev Biochem* **82**, 237-266.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I. & Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. *Nat Protoc* **1**, 1965-1970.

Stalder, T., Barraud, O., Casellas, M., Dagot, C. & Ploy, M. C. (2012). Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol* **3**, 119.

Stavrinides, J. & Guttman, D. S. (2004). Nucleotide sequence and evolution of the five-plasmid complement of the phytopathogen Pseudomonas syringae pv. maculicola ES4326. *J Bacteriol* **186**, 5101-5115.

Stavrinides, J., Kirzinger, M. W., Beasley, F. C. & Guttman, D. S. (2012). E622, a miniature, virulence-associated mobile element. *J Bacteriol* **194**, 509-517.

Stein, L. Y., Arp, D. J., Berube, P. M. & other authors (2007). Whole-genome analysis of the ammonia-oxidizing bacterium, Nitrosomonas eutropha C91: implications for niche adaptation. *Environ Microbiol* **9**, 2993-3007.

Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* **22**, 1985-1994.

Stevens, R. H., Sela, M. N., McArthur, W. P., Nowotny, A. & Hammond, B. F. (1980). Biological and chemical characterization of endotoxin from Capnocytophaga sputigena. *Infect Immun* **27**, 246-254.

Steward, G. F., Montiel, J. L. & Azam, F. (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* **45**, 1697–1706.

Steward, G. F. & Preston, C. M. (2011). Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. *Virol J* **8**, 287.

Stewart, E. J. (2012). Growing unculturable bacteria. *d* **194**, 4151-4160.

Stewart, R. M., Wiehlmann, L., Ashelford, K. E. & other authors (2011). Genetic characterization indicates that a specific subpopulation of Pseudomonas aeruginosa is associated with keratitis infections. *J Clin Microbiol* **49**, 993-1003.

Sulakvelidze, A., Alavidze, Z. & Morris, J. G., Jr. (2001). Bacteriophage therapy. *Antimicrob Agents Chemother* **45**, 649-659.

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

Sun, S., Chen, J., Li, W. & other authors (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**, D546-551.

Tamaki, H., Zhang, R., Angly, F. E., Nakamura, S., Hong, P. Y., Yasunaga, T., Kamagata, Y. & Liu, W. T. (2012). Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**, 441-452.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739.

Tanzer, F. L., Shephard, E. G., Palmer, K. E., Burger, M., Williamson, A. L. & Rybicki, E. P. (2011). The porcine circovirus type 1 capsid gene promoter improves antigen expression and immunogenicity in a HIV-1 plasmid vaccine. *Virol J* **8**, 51.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D. & other authors (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Tetart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H. W. & Krisch, H. M. (2001). Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J Bacteriol* **183**, 358-366.

Thingstad, T. F. & Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* **13**, 19-27.

Thingstad, T. F. (2000). Elements of a Theory for the Mechanisms Controlling Abundance, Diversity, and Biogeochemical Role of Lytic Bacterial Viruses in Aquatic Systems. *Limnology and Oceanography* **45**, 1320-1328

Thingstad, T. F., Bratbak, G. & Heldal, M. (2008). Aquatic phage ecology. In *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Edited by S. T. Abedon. Cambridge, UK: Cambridge University Press.

Thomas, T., Gilbert, J. & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2**, 1-12.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680.

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

Tice, H., Mayilraj, S., Sims, D. & other authors (2010). Complete genome sequence of Nakamurella multipartita type strain (Y-104). *Stand Genomic Sci* **2**, 168-175.

Tocchi, C., Federici, E., Fidati, L., Manzi, R., Vincigurerra, V. & Petruccioli, M. (2012). Aerobic treatment of dairy wastewater in an industrial three-reactor plant: effect of aeration regime on performances and on protozoan and bacterial communities. *Water Res* **46**, 3334-3344.

Tucker, K. P., Parsons, R., Symonds, E. M. & Breitbart, M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**, 822-830.

Tunney, M. M., Field, T. R., Moriarty, T. F. & other authors (2008). Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. *Am J Respir Crit Care Med* **177**, 995-1001.

Urbonavicius, J., Qian, Q., Durand, J. M., Hagervall, T. G. & Bjork, G. R. (2001). Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J* **20**, 4863-4873.

van der Gast, C. J., Walker, A. W., Stressmann, F. A., Rogers, G. B., Scott, P., Daniels, T. W., Carroll, M. P., Parkhill, J. & Bruce, K. D. (2011). Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J* **5**, 780-791.

van Wamel, W. J., Rooijakkers, S. H., Ruyken, M., van Kessel, K. P. & van Strijp, J. A. (2006). The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of Staphylococcus aureus are located on beta-hemolysin-converting bacteriophages. *J Bacteriol* **188**, 1310-1315.

Vasilyev, E. V., Trofimov, D. Y., Tonevitsky, A. G., Ilinsky, V. V., Korostin, D. O. & Rebrikov, D. V. (2009). Torque Teno Virus (TTV) distribution in healthy Russian population. *Virol J* **6**, 134.

Veesler, D. & Cambillau, C. (2011). A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiol Mol Biol Rev* **75**, 423-433.

Vega Thurber, R. L., Barott, K. L., Hall, D. & other authors (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proc Natl Acad Sci U S A* **105**, 18413-18418.

Vickerman, M. M., Iobst, S., Jesionowski, A. M. & Gill, S. R. (2007). Genome-wide transcriptional changes in Streptococcus gordonii in response to competence signaling peptide. *J Bacteriol* **189**, 7799-7807.

Wagner, P. L. & Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infect Immun* **70**, 3985-3993.

Wang, G., Hayashi, M., Saito, M., Tsuchiya, K., Asakawa, S. & Kimura, M. (2009a). Survey of major capsid genes (g23) of T4-type bacteriophages in Japanese paddy field soils. *Soil Biology & Biochemistry* **41**, 13-20.

Wang, G., Jin, J., Asakawa, S. & Kimura, M. (2009b). Survey of major capsid genes (g23) of T4-type bacteriophages in rice fields in Northeast China. *Soil Biology & Biochemistry* **41**, 423-427.

Wang, G., Murase, J., Taki, K., Ohashi, Y., Yoshikawa, N., Asakawa, S. & Kimura, M. (2009c). Changes in major capsid genes (g23) of T4-type bacteriophages with soil depth in two Japanese rice fields *Biol Fertil Soils* **45**, 521-529.

Wang, G., Yu, Z., Liu, J., Jin, J., Liu, X. & Kimura, M. (2011). Molecular analysis of the major capsid genes (*g23*) of T4-type bacteriophages in an upland black soil in Northeast China. *Biol Fertil Soils* **47**, 273-282.

Wang, I. N., Smith, D. L. & Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu Rev Microbiol* **54**, 799-825.

Wang, Z., Gerstein, M. & Snyder, M. (2009d). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.

Warren, R. A. (1980). Modified bases in bacteriophage DNAs. *Annu Rev Microbiol* **34**, 137-158.

Warren, R. L., Freeman, J. D., Levesque, R. C., Smailus, D. E., Flibotte, S. & Holt, R. A. (2008). Transcription of foreign DNA in Escherichia coli. *Genome Res* **18**, 1798-1805.

Wat, D. & Doull, I. (2003). Respiratory virus infections in cystic fibrosis. *Paediatr Respir Rev* **4**, 172-177.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**, 127-181.

Weinberger, A. D., Sun, C. L., Plucinski, M. M. & other authors (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput Biol* **8**, e1002475.

Wen, K., Ortmann, A. & Suttle, C. (2004). Accurate estimation of viral abundance by epifluorescent microscopy. *Appl Environ Microbiol* **70**, 3862-3867.

Weng, F. C., Su, C. H., Hsu, M. T., Wang, T. Y., Tsai, H. K. & Wang, D. (2010). Reanalyze unassigned reads in Sanger based metagenomic data using conserved gene adjacency. *BMC Bioinformatics* **11**, 565.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578-6583.

Whon, T. W., Kim, M. S., Roh, S. W., Shin, N. R., Lee, H. W. & Bae, J. W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J Virol* **86**, 8221-8231.

Wilhelm, S. W. & Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *BioScience* **49**, 781–788.

Willi, K., Sandmeier, H., Kulik, E. M. & Meyer, J. (1997). Transduction of antibiotic resistance markers among Actinobacillus actinomycetemcomitans strains by temperate bacteriophages Aa phi 23. *Cell Mol Life Sci* **53**, 904-910.

Williamson, K. E., Radosevich, M. & Wommack, K. E. (2005). Abundance and diversity of viruses in six Delaware soils. *appl environ microbiol* **71**, 3119-3125.

Williamson, K. E., Radosevich, M., Smith, D. W. & Wommack, K. E. (2007). Incidence of lysogeny within temperate and extreme soil environments. *Environ Microbiol* **9**, 2563-2574.

Williamson, S. J., Rusch, D. B., Yooseph, S. & other authors (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**, e1456.

Williamson, S. J., Allen, L. Z., Lorenzi, H. A. & other authors (2012). Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* **7**, e42047.

Willner, D., Furlan, M., Haynes, M. & other authors (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **4**, e7370.

Willner, D., Furlan, M., Schmieder, R. & other authors (2010). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4547-4553.

Willner, D., Furlan, M., Schmieder, R. & other authors (2011). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4547-4553.

Willner, D., Haynes, M. R., Furlan, M. & other authors (2012). Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *Am J Respir Cell Mol Biol* **46**, 127-131.

Willner, D. & Hugenholtz, P. (2013). From deep sequencing to viral tagging: Recent advances in viral metagenomics. *Bioessays* **35**, 436-442.

Wilson, G. G. & Murray, N. E. (1991). Restriction and modification systems. *Annu Rev Genet* **25**, 585-627.

Winstanley, C., Langille, M. G., Fothergill, J. L. & other authors (2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of Pseudomonas aeruginosa. *Genome Res* **19**, 12-23.

Withey, S., Cartmell, E., Avery, L. M. & Stephenson, T. (2005). Bacteriophages--potential for application in wastewater treatment processes. *Sci Total Environ* **339**, 1-18.

Wommack, K. E. & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**, 69-114.

Wommack, K. E., Bhavsar, J. & Ravel, J. (2008). Metagenomics: read length matters. *appl environ microbiol* **74**, 1453-1463.

Wootton, S. C., Kim, D. S., Kondoh, Y. & other authors (2011). Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **183**, 1698-1702.

Wu, H. J., Wang, A. H. & Jennings, M. P. (2008). Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol* **12**, 93-101.

Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. (2011a). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**, s444.

Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. (2011b). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**, 444.

Yilmaz, S., Allgaier, M. & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* **7**, 943-944.

Yin, Y. & Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**, 24.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013). Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS ONE* **8**, e57271.

Yu, K. & Zhang, T. (2012). Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS ONE* **7**, e38183.

Zargar, M. A., Chaturvedi, D. & Chakravort, M. (2001). Identification of a strong promoter of bacteriophage MB78 that interacts with a host coded factor and regulates the expression of a structural protein. *Virus Genes* **22**, 35-45.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829.

Zhang, T., Breitbart, M., Lee, W. H. & other authors (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**, e3.

Zhong, Y., Chen, F., Wilhelm, S. W., Poorvin, L. & Hodson, R. E. (2002). Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl Environ Microbiol* **68**, 1576-1584.

Zhu, W., Lomsadze, A. & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**, e132.

Zuber, S., Ngom-Bru, C., Barretto, C., Bruttin, A., Brussow, H. & Denou, E. (2007). Genome analysis of phage JS98 defines a fourth major subgroup of T4-like phages in Escherichia coli. *J Bacteriol* **189**, 8206-8214.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415.