ISOLATION AND CHARACTERISATION OF BACTERIOPHAGES INFECTING ENVIRONMENTAL STRAINS OF *CLOSTRIDIUM DIFFICILE.*

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Katherine Rose Hargreaves B.Sc. (Leicester)

Department of Infection, Immunity and Inflammation

University of Leicester

September 2012

Acknowledgements

Abstract

Clostridium difficile is a leading cause of nosocomial infections associated with antibiotic treatment. Despite its' pathogenic role, the bacterium can be carried asymptomatically in the GI tract and has an ubiquitous presence in the environment. The ecology of *C. difficile* outside clinical settings is not fully understood, but the evolution of pathogenic strains may occur in its zoonotic and environmental reservoirs. A major driver of bacterial genome evolution is bacteriophages. This project has investigated the prevalence and diversity of phages associated with strains from these reservoirs. A large and diverse collection of *C. difficile* isolates was established in order to isolate and characterise the phages associated with them.

27 phages were isolated. These were characterised according to their particle morphology, genome size and plaque morphology which showed they could be sub-grouped according to morphology. Host range analyses showed there is a complex network of phage-host interactions within this species.

To characterise to a genetic level, the genomes of seven phages were sequenced and annotated. Their genomes show they are related to known *C. difficile* phages. However, genes which are unusual or novel to phage genomes were also identified.

Comparative genomics identified distinct lineages within *C. difficile* phages, which correspond to their morphological sub-grouping. Also the bioinformatic analyses provide evidence of genetic exchange occurring between them, presumably during co-infection as multiple prophage carriage is common in this species. Interestingly, analysis of the *C. difficile* CRISPR system showed that in addition to the bacterial arrays, several prophages encode CRISPR arrays that target phage sequences. Evidence of their co-evolution suggests phage infection impacts both genome evolution of their host and other *C. difficile* phages.

In summary, a diverse *C. difficile* phage collection has been established which has the potential for development to allow their exploitation.

Contents.

1.   Introduction

1.1.   Introduction to *Clostridium difficile* epidemiology

1.2.   Distribution and diversity of *C. difficile* in reservoirs outside of clinical settings

1.3.   *C. difficile* infection and pathogenicity

1.4.   Introduction to bacteriophage biology

1.5.   *C. difficile* and its associated bacteriophages

1.6.   Aims and objectives of this project

2.   Materials and Methods

2.1.   Isolation and characterisation of non-clinical strains of *C. difficile*, including prophage carriage

2.2.   Isolation of phages infecting *C. difficile* in a lytic manner

2.3.   Genome sequencing and annotation of *C. difficile* phages and bioinformatic analysis

3.   Isolation of *C. difficile* from non-clinical sources

4.   Isolation of phages with lytic activity on *C. difficile*

5. Genome annotation of seven *C. difficile* phages.

6. Comparative genomes and CRISPR analysis of *C. difficile* phages

List of Figures

List of Appendices

List of Abbreviations

Media and Materials

BA	bacteriological agar
BHI	brain heart infusion
CaCl$_2$	calcium chloride
CCEY	cefoxitin-cycloserine egg yolk agar
CFU	colony forming unit
CI	Ciprofloxacin
CM	Clindamycin
DHB	defribinated horse blood
dNTPs	Deoxyribonucleotides
EDTA	ethylenediaminetetra acetic acid
FA	fastidious anaerobic
HCl	hydrogen chloride
IMS	industrial methylated spirits
MC	mitomycin c
MgCl$_2$	magnesium chloride
MgS0$_4$	magnesium sulphate
MIC	minimum inhibitory concentration
MZ	Metronidazole
NaAc	sodium acetate
NaCl	sodium chloride
NFX	Norfloxacin
OD	optical density
PCR	polymerase chain reaction
PEG	polyethylene glycerol
PFGE	Pulse field gel electrophoresis
PFU	plaque forming unit
PNACL	Protein Nucleic Acid Laboratory
RT	room temperature (21°C)
SDS	sodium dodecyl sulphate
SM	SM buffer
SSC	saline-sodium citrate
TAE	tris acetic acid EDTA buffer
TBE	tris/borate/EDTA buffer
TE	Tris:EDTA buffer
Tris HCL	tris(hydroxymethyl)aminomethane hydrochloride
UPH$_2$O	ultra pure H$_2$O
VA	Vancomycin
RT	room temperate

Bioinformatic abbreviations

aa	amino acid
ACT	Artemis Comparison Tool
bp	Basepairs
DNA	deoxyribonucleotide acid

CNI	close neighbour interchange
GC	guanine cytosine content
JTT	Jones-Taylor Thornton
Kbp	kilo basepairs
MEGA	Molecular Evolutionary Genetics Analysis
ME	minimum evolution
ML	maximum likelihood
NCBI	National Centre for Biotechnological Information
NJ	neighbour joining
ORF	open reading frame
rRNA	ribosomal ribonucleic acid
STRING	Search Tool for the Retrieval of Interacting Genes

Other

CDI	*Clostridium difficile* infection
CDRN	*Clostridium difficile* ribotyping network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
HGT	horizontal gene transfer
HPA	Health Protection Agency
HPS	Health Protection Scotland
LTM	long tailed myovirus
MBE	Mobile genetic element
MM	medium myovirus
MVLA	multilocus variable-number tandem repeat analysis
PMC	pseudomembranous colitis
PTLP	phage tail-like particles
SMV	small myovirus
SSSI	Site of Special Scientific Interest
SV	Siphovirus
TEM	transmission electron microscopy
QS	quorum sensing
VRE	vancomycin resistant *Enterococci*

Units

cm	Centimetre
ml	Millilitre
mg	Milligram
M	Molar
mm	Millimetre
mM	Millimolar
µl	Microliter
µg	Microgram
µm	Micrometre
Ng	Nanogram
Nm	Nanometre
U	Units
˚C	degrees centigrade

1. Introduction

1.1. Introduction to *Clostridium difficile* epidemiology

1.1.1. The nosocomial pathogen *C. difficile* is a significant healthcare burden.

Clostridium difficile is an endospore forming, anaerobic, low G+C, gram-positive bacillus belonging to the phylum *Firmicutes*. It was first isolated from stools of healthy newborns by Hall and O'Toole (1935) and originally named *Bacillus difficulus*. It was during the 1970's that the bacterium's pathogenic role, as a cause of antibiotic-associated diarrhoea, was first recognised (Bartlett *et al* 1978). *C. difficile* Infection (CDI) occurs following a disruption of the normal gut microbiota caused by taking broad spectrum antibiotics. CDI affects all ages, but the majority of patients are aged 65 years and over (HPA Mandatory Surveillance Data). Numerous epidemics and outbreaks have occurred across Europe, Canada and the U.S.A. in the past decade (Wiegand *et al* 2012).

CDI is a major health and economic burden, despite decreasing infection rates in the U.K. being reported: from 55,498 in 2007-2008 to 18,005 in 2011-2012 (HPA Mandatory Surveillance Data). During 2009-2011, the bacterium was involved in 6 out of every 1000 deaths in the U.K. and accounted for 1 in every 100 hospital deaths (Office for National Statistics). Whilst overall rates are falling, 26% of U.K. healthcare trusts reported an increase in CDI cases between 2010 and 2011 (HPA). CDI impacts considerably upon human morbidity and mortality, as well as places a large financial burden on healthcare trusts due to costs associated with prolonged hospital stays, additional drug prescriptions, infection control and mandatory surveillance (Wiegand *et al* 2012, Bouza 2012). It has been estimated that managing CDI costs €3000 million in Europe and $800 million in the U.S.A. (Bouza 2012). Alternative methods to treat, diagnose and prevent CDI more effectively are needed; all of which could involve the biotechnological use of bacteriophages (phages). In order to develop such applications, first understanding the fundamental biology of *C. difficile* and associated phages is necessary.

1.1.2. Epidemiology is reflected by the continued evolution of *C. difficile* strains and emergence of new clinically important ribotypes.

Genotyping of *C. difficile* to monitor the hospital strains and track epidemic outbreaks is performed using PCR ribotyping, with primers that amplify the variable intergenic region between 16S rRNA and 23S rRNA genes that allow strains to be assigned to specific ribotypes (O'Neill *et al* 1996). The prevalence of different ribotypes in clinics varies between countries and over time (Deneve *et al* 2009, Bouza *et al* 2012). Increases in incidence and severity of CDI have been partially attributed to the dissemination of emergent ribotypes, such as 027 and 078 (McDonald *et al* 2005, Kuijper *et al* 2006, Goorhuis *et al* 2008). Recently, a new ribotype, R332, has emerged in clinics in Scotland (HPS 2013), highlighting the need to understand the processes by which new, problematic, strains evolve.

Whole genome analysis of several epidemic strains and clinical ribotypes reveals *C. difficile* strains to encode multiple and diverse mobile genetic elements, including prophages, which contribute to strain evolution (Sebaihia *et al* 2006, Stabler *et al* 2006, Stabler *et al* 2009, He *et al* 2010, Cairns *et al* 2012, Stabler *et al* 2012, He *et al* 2012). One study tracked the global transmission of R027 and revealed its evolution via loss or gain of mobile elements, as well as its transfer between humans and livestock (He *et al* 2012). Zoonotic and environmental reservoirs may be where new strains may evolve, for example R078 is thought to have evolved from an atoxigenic strain outside of the clinical environment (Dingle *et al* 2011, Stabler *et al* 2012). The factors which drive genome evolution within reservoirs of *C. difficile* have not been focused on previously. This project investigates the potential impact phage infection could have in these populations.

1.2. Distribution and diversity of *C. difficile* in reservoirs outside of clinical settings

1.2.1. Distribution and diversity of *C. difficile* in zoonotic reservoirs.

C. difficile has been repeatedly isolated from livestock and animals in geographically dispersed locations (e.g. Rodriguez-Palacios *et al* 2007, 2009, Weese *et al* 2009, Indra *et al* 2009, Hensgens *et al* 2012). The bacterium can cause CDI in a range of animal species, as well as be carried asymptomatically (e.g. Borriello *et al* 1983, Keel and Songer 2006, Bojesen *et al* 2006, Simango and Mwakurudza 2008, Jhung *et al* 2008, Keel and Songer 2006, Indra *et al* 2009, Avbersek *et al* 2009).

Strains of *C. difficile* isolate from animals include ribotypes that also infect humans, including the clinically important ribotypes, R027 and R078 (Rodriguez-Palacios *et al* 2006, Stabler *et al* 2006, Keel *et al* 2007, Janezic *et al* 2012). Comparative genotype analysis has revealed that human and animal isolates fall into integrated and distinct clades, highlighting the potential transmission of strains between the two sources (Stabler *et al* 2006, Scaria *et al* 2010). It is thought that strains in zoonotic reservoirs may transfer up the food chain (Indra *et al* 2009, Gould and Limbago 2010, Rupnik 2010), although whether *C. difficile* is a zoonotic disease is not certain (Hensgens 2012). Whilst R027 has been shown to have been transmitted between humans and livestock (He *et al* 2012), it is not known whether these animal and human isolates are evolving together or separately and undergo horizontal genetic exchange (HGT) for the majority of ribotypes. No studies have examined phages (which can mediate HGT) associated with animal strains, although the whole genome sequencing of a bovine isolate has shown that it encodes a prophage (strain 2007855: Accession NC_017178.1). This project will investigate *C. difficile* strains from asymptomatic animals and determine their active temperate phage content.

1.2.2. Prevalence of *C. difficile* in environmental reservoirs and its potential transmission to humans or animals.

The bacterium is ubiquitous in the environment, as it has been isolated from water, soil and mud samples (Al Saif and Brazier 1996, Kim *et al* 1981, Borriello and Brazier 2000, Båverud *et al* 2003, del Mar Gamboa *et al* 2005, Simango and Mwakurudza 2008, Zidaric *et al* 2010, Pasquale *et al* 2011, Higazi *et al* 2011). It is unclear whether *C. difficile* is metabolically active in these environments or is present as spores, but studies suggest that it has a transient presence which may originate from contamination (Zidaric *et al* 2010, Higazi *et al* 2011). Importantly, there is a cross-over of ribotypes that are isolated from the environment, human and animal samples (Janezric *et al* 2012).

Transmission of the bacteria between environmental sources, animals and humans is likely. Evidence of direct transmission of R027 has been reported between humans and animals, occurring at least twice (He *et al* 2012). Also, spores have been isolated from retail meat (cooked and uncooked), which provides evidence of one way in which *C. difficile* enters the food chain (Songer *et al* 2007, Weese *et al* 2009, Abercron *et al* 2009, Curry *et al* 2012, Visser *et al* 2012).The bacterium has also been isolated from fish and shellfish which are sometimes consumed raw and offer another way in which *C. difficile* can transmit to humans (Metcalf *et al* 2010, Pasquale *et al* 2011). Additionally, *C. difficile* was isolated from bagged salads and raw vegetables, which presumably have been contaminated either during fertilisation with manure, or from water used to wash the salads prior to packing which harboured *C. difficle* (Al Saif 1996, Bakri *et al* 2009, Metcalf *et al* 2010). Although the bacterium is often detected at very low abundance, and sometimes not detected at all (e.g. Kalchayanand *et al* 2013), the presence of *C. difficile* in different food suggests different ways in which strains are transferred between reservoirs.

1.3. *C. difficile* infection and pathogenicity

1.3.1. Asymptomatic carriage of *C. difficile* occurs with protection supplied by the healthy gut microbiota.

Normally the natural gut microbiota acts as a defence against *C. difficile* colonisation and infection of the host's large intestine. This was demonstrated *in vitro* by Yamamotoosaki *et al* (1994) who reported an inhibition of *C. difficile* growth by co-culture of different gut bacteria. Possesion of an immature gut microbiota has been one explanation for why infants are more frequently colonised by *C. difficile* than adults (Ozaki *et al* 2005).

The majority of CDI cases follow treatment with broad spectrum antibiotics which severely disrupt the composition of the gut microbiome, thus compromising its conferred protection and permitting the outgrowth of *C. difficile* (Dethlefsen *et al* 2008, Reeves *et al* 2011). A decrease in gut bacterial diversity and over growth of *C. difficile* was reported by Chang *et al* (2008) following antibiotic treatment in patients with either single episode or recurrent CDI. The disruption caused by antibiotics is evident but why it may occur in some patients and not others is not known. Specific organisms that can prevent *C. difficile* infection have been identified. For example, *Saccharomyces boulardii* produces a protease that detoxifies two toxins produced by *C. difficile* (Castagliuolo *et al* 1999) and *Lactobacillus delbrueckii* ssp. *bulgaricus* produces unidentified soluble molecules which have inhibitory actions against the cytotoxic effect of *C. difficile* toxins and its attachment to Caco-2 epithelial cells (Banerjee *et al* 2009). Growth of *C. difficile* may also be inhibited by another natural component of the human gut microbiota, bacteriophages, although no work has examined the extent of endogenous phages killing *C. difficile* during CDI episodes.

1.3.2. Virulence factors of *C. difficile* impacting upon disease, colonisation and transmission.

CDI produces a wide spectrum of disease symptoms and severity, which results, in part, from the virulence of the infecting *C. difficile* strain (Delmée and Avesani 1990, Poxton *et al* 2001, Warny *et al* 2005, Denève *et al* 2009, Stabler *et al* 2009). For example, toxin production varies between strains and 25 different toxinotypes can be identified depending on variant toxin gene carriage (Rupnik 2008). Other major virulence determinants of *C. difficile* include its cell surface S-layer, production of flagella and ability to sporulate (Davies and Boriello 1990, Boriello 1998, Waligora *et al* 2001, McCoubrey and Poxton 2001, Poxton *et al* 2001, Sebaihia 2006, Dingle *et al* 2011, Ho *et al* 2011, Miura *et al* 2011), as well as genes involved in chemotaxis towards nutrients or away from noxious environments, antibiotic resistance and survival in harsh conditions, such as outside the host (Stabler *et al* 2009).

The expression of virulence genes is often regulated by quorum sensing (QS) (reviewed by Caetano *et al* 2010) and strains vary between which types of QS system loci they encode. For example, both encode the LuxS system but *C. difficile* strain CD630 encodes an *agr1* locus containing two genes of the *agr* QS system, and strain CD196 encodes an *agr2* locus containing all four genes of the *agr* QS system (Sebaihia *et al* 2006, Stabler *et al* 2009). Inhibition of the Lux system in *C. difficile* has been found not to influence toxin production (Carter *et al* 2005). In a related bacterial spp, *C. perfringens*, two different *agr* loci control toxin production and sporulation (Li *et al* 2011, Chen and McClane 2012), but no research in to the *agr* QS system of *C. difficile* has been performed.

Infection occurs either from through germination of ingested of *C. difficile* spores or outgrowth of the patient's own *C. difficile* population. Considering the high numbers of *C. difficile* infection within healthcare settings, it not surprising that it has been cultured from contaminated floors, curtains and surfaces within this environment (Kim *et al* 1981, MacFarland *et al* 1989, Walker *et al* 2006). The pathogen's persistence and spread between patients via contact with contaminated surfaces and staff is attributed to its heat- and alcohol-resistant spores (Lawley *et al* 2009). Although

the bacterium is anaerobic, the spores can persist within an aerobic environment for up to 5 months (Kim *et al* 1981).

Once inside the host, disease is toxin-mediated with different strains of *C. difficile* able to produce three known toxins; TcdA, TcdB and *Clostridium difficile* Binary Toxin (CDT) resulting in several variant toxinotypes (Rupnik 2008). The two toxins TcdA and TcdB are encoded in a 19 kbp Pathogenicity Locus (PaLoc) which also encodes three other genes, *tcdR*, *tcdC* and *tcdE*. TcdR encodes an alternative RNA polymerase sigma factor, positively regulating *tcdA* and *tcdB* (Mani and Dupuy 2001) and undergoes autoregulation (Mani *et al* 2002). TcdC negatively regulates toxin production by inhibition of the TcdR, halting *tcdA*/*tcdB* transcription (Matamouros *et al* 2007). TcdE is required for toxin release (Tan *et al* 2001, Govind and Dupuy 2012). The binary toxin, an actin specific ADP-ribotransferase, is encoded by two genes, *cdtA* and *cdtB* and regulated by CdtR, a global response regulator (Popoff *et al* 1988, Perelle *et al* 1997, Carter *et al* 2007). Injury to the colonic epithelium during CDI occurs due to production of bacterial toxins and pro-inflammatory response (Riegler *et al* 1995, Mahida *et al* 1996, Mahida *et al* 1998, Pothoulakis 2000, Linevsky *et al* 2008, Sun *et al* 2010). TcdA and TcdB are both cytotoxic and induce a loss of cell shape from the disruption of the actin cytoskeleton triggering apoptosis (Voth and Ballard 2005, Nam *et al* 2010). Destruction of the epithelial cell barrier produces watery, non-bloody diarrhoea due to the resulting electrolyte imbalance. Infiltration by neutrophils and vascular congestion across the damaged epithelial membrane results in formation of the characteristic lesions on the colonic wall (Kang *et al* 2009, Lee *et al* 2009). During heavy infection lesions on the surface merge together forming a psuedomembrane (Poxton *et al* 2001).

C. difficile is problematic to treat with relapse and reccurrence of the disease widely reported, as well as cases that do not respond to either antibiotic treatment, metronidazole or vancomycin (Kuijper and Wilcox 2008). Increasing antibiotic resistance has been reported in the most clinically seen ribotypes of *C. difficile,* alongside transient resistance to metronidazole (Huang

et al 2009). In *C. difficile*, antibiotic resistance genes have been identified on mobile genetic elements (MGE) such as transposons (Sebaihia *et al* 2006, Stabler *et al* 2009, He *et al* 2012). In other species, other virulence genes are encoded on MGE, including prophages (Wagnor and Waldor 2001, Frost *et al* 2005). The introduction and expression of such genes conferring a fitness advantage include toxins, adhesions and enzymes that alter bacterial antigens and or host tissues, such as hyaluronidases (Wagnor and Waldor 2001). Identifying the contribution MGEs, including prophages, make to *C. difficile* virulence and resulting CDI would provide valuable insight into the evolution and virulence of this pathogen.

1.4. Introduction to bacteriophage biology

1.4.1. Bacteriophages are abundant and diverse viruses of bacteria.

Bacterial viruses, bacteriophages (or phages), are abundant in nature, estimated to number globally approximately 10^{31} particles and outnumber their hosts by 10 to 100 per bacterial cell (Wommack *et al* 2009). Phage particles consist of a proteinous shell encapsulating a nucleic acid genome, which can be dsDNA, ssDNA, dsRNA or ssRNA. As a collective, phages contain the largest pool of genetic diversity in the biosphere as metagenomic studies have shown (Breitbart *et al* 2002, Breitbart and Rohwer 2005, Suttle 2007, Breitbart *et al* 2007). Next generation sequencing has led to a rapid expansion in the genome databases of whole genome sequences of phages. There are over 1000 phage genomes currently in the EMBL database, a figure which has doubled in the past four years (Hatfull 2008). Yet the total is still less than half the number of bacterial genome entries, despite having significantly smaller genomes that are cheaper to sequence. Excitingly, the virosphere remains a massive untapped source for biodiscovery.

Phages are obligate parasites, fully dependant on infection of a bacterial cell to replicate. Infection generally follows a lytic or lysogenic life cycle, according to the type of phage or conditions during infection (Birge 2006).

Virulent phages can only replicate through lytic infection, a rapid process resulting in lysis of the bacterial cell. The first step is phage attachment to the bacterial cell followed by injection of the phage nucleic acid, resulting in transcription of the phage genome, translation of phage proteins, phage genome replication and phage particle self-assembly. The final step is the production of specific phage proteins during the late stages of transcription and translation. Both holin and endolysin proteins are produced, the holin forms pores through the cell membranes and the endolysin breaks down the peptidoglycan layer (Young and Bläsi 1995). The resulting lysis of the cell releases phage progeny into the environment.

Temperate phages replicate in a lytic manner but have the ability to enter a lysogenic state. Lysogeny occurs when a phage genome is inserted into the bacterial host's chromosome at a specific attachment site recognisable in both genomes via a phage encoded integrase. Once in this quiescent state, the phage is a prophage and its genome is replicated alongside the bacteria's own during cell division. In this way, the phage is propagated in the population of daughter cells, all of which are infected with the prophage. This insertion is reversible and has been well studied in phage lambda which has a repressor, Cl, that prevents phage excision under normal conditions. However, if the cell undergoes stress, such as radiation from UV damage, the activity of this repressor is mediated by bacterial protein RecA (Dodd *et al* 2005), and the phage resumes the lytic pathway of replication. To force phage excision, chemical stressors can also be used such as mitomycin c, an antibiotic that cross-links DNA (Tomasz 1995) or norfloxacin, a quinolone that inhibits DNA gyrase (Drlica and Zhao 1997). The response to a signal triggering the SOS response ensures escape of the phage from a host that is unlikely to survive.

1.4.2. Bacteriophages and phage therapy.

Due to their ability to lyse and kill bacteria, phages offer an attractive therapeutic alternative to antibiotics. With concern over antibiotic resistance growing, interest in phage therapy is increasing in the U.K. and U.S.A; for example, stage II clinical trials were performed for a phage preparation successfully treating infectious otitis media caused by *P. aeruginosa* (Wright *et al* 2009). This has long been in practice in other countries, such as Georgia and Poland (O'Flaharty *et al* 2009).

Phages have potentially useful qualities, including their ability to evolve and counteract bacterial resistance (which antibiotics cannot), the replication at the site of infection and host specificity. However, of concern are the possible spread of genes via horizontal gene transfer (facilitated by phages) and the introduction of novel virulence factors encoded and/or transferred by infecting phages, into bacterial genomes. Also, some phages have narrow host ranges, infecting only single or a few strains of a species.

To address these issues, understanding the biology of phage and host is essential. For example, depending on their host ranges, phage cocktails can be produced containing more than one phage in order to infect a broad spectrum of strains (Fiorentin *et al* 2005, Loc Carrillo *et al* 2005, Wagenaar *et al* 2005, Wright *et al* 2009). Also, to avoid introduction of phage encoded toxins to the host bacteria, sequencing of phage genomes can be performed to identify putative virulence factors. If found, genetic modification to remove such genes or widen host ranges via altering attachment genes for otherwise suitable phages could be an eventual outcome of such research. Use of virulent phages may reduce the amount of HGT ongoing and ensure that lysogeny does not occur (Brüssow 2005). Another option is to use phage encoded endolysins, removing the need to consider the problem posed by phage-mediated transfer of genetic material, and first trial in human patients of a lysin is underway this year (Borrell 2012).

1.4.3. Bacteriophages' ecological niche.

At the most basic level, phages are important contributors to organic matter cycling in food chains through the release of dissolved organic matter (DOM) from their lysed hosts, as well as making up particulate organic matter (POM) (Suttle 2007). Release of cell contents by lysis makes available free bacterial DNA that can be taken up and used in genetic exchange via transformation in viable bacteria.

Phages also directly facilitate horizontal genetic transfer between bacteria via transduction. Incorrect packing of phage particles results in bacterial chromosome fragments filling the phage head, which is then introduced into a new bacterium following infection. This occurs widely in different microbial populations (e.g. Ripp and Miller 1997, Jiang and Paul 1998, Paul *et al* 2002, Kenzaka *et al* 2010, Muniesa *et al* 2011, del Casale *et al* 2011, Picozzi *et al* 2012) and includes the transfer of microbial virulence factors such as antibiotic resistance genes (Colomer-Lluch *et al* 2011).

Additionally, phages provide a source of new genetic material from their own genomes, encoding genes that can confer fitness advantages to the bacterial host when expressed. These genes can change the phenotype of the bacteria during phage infection and can enhance pathogenicity. Phage encoded virulence genes include additional toxins (van der Wilk *et al* 1999), bacterial adhesions (Wagner and Waldor 2002), surface proteins for serotype conversions (Wagner and Waldor 2002), proteins involved in biofilm formation (Rice *et al* 2009, Schuch and Fischetti 2009), sporulation modulating proteins (Silver-Mysliwiec and Bramucci 1990, Schuch and Fischetti 2009) and genes involved in metabolic processes (Mann *et al* 2005). Phages encoding virulence factors have the ability to convert previously non-pathogenic strains to those capable of causing disease (Wagner and Waldor 2002, Moran *et al* 2005). A growing body of work is reporting the complex ecological relationships that can occur between phages and their hosts (e.g. Abedon and Lejeune 2005, Shan *et al* 2008, Carrolo *et al* 2010, Zeng and Chrisholm 2012). An example of this are lysogenic phages, infecting *Bacillus anthracis*, are able to block or promote sporulation as well as

activate survival pathways for the bacteria in soil but not in mammalian tissue (Schuch and Fischetti 2009). This is an interesting example of a pathogenic bacterium behaving differently outside of a host. Whether *C. difficile* exhibits a a difference in physiology between environments is completely unknown.

1.5. *C. difficile* and its associated bacteriophages

1.5.1. Temperate phages infecting *C. difficile.*

To date no virulent phages have been identified for *C. difficile*, however, prophage carriage within *C. difficile* is widespread and several temperate phages have been identified. In clinical strains, the induction of active temperate phages, as well as the use of molecular markers to identify prophage gene carriage has shown that they are prevalent and diverse, with often carrying multiple prophages (e.g. Nagy and Foldes 1991, Fortier and Moineau 2007, Goh *et al* 2007, Shan *et al* 2012, Nale *et al* 2012). The sequences of several prophages are available in *C. difficile* genome entries deposited in NCBI, including two prophages in strain CD630 (Sebaihia *et al* 2006). There have been several temperate phages with lytic activity that have been isolated, but the degree to which each is characterised varies; from particle morphology (using transmission electron microscopy (TEM)) to molecular and proteomic analysis (e.g. Sell 1984, Mahoney *et al* 1985, Goh *et al* 2005a, Govind *et al* 2006, Goh *et al* 2007, Mayer *et al* 2008, Horgan *et al* 2010, Sekulovic *et al* 2011). Seven *C. difficile* temperate phages have been sequenced: five myoviruses and two siphoviruses. (Govind *et al* 2006, Goh *et al* 2007, Mayer *et al* 2008, Horgan *et al* 2010, Sekulovic *et al* 2011, Meesen-Pinard *et al* 2012).

Sequencing has revealed unusual genomic features in *C. difficile* phages that could influence host physiology. Firstly, a *C. difficile* riboswitch gene is present in phiCD119 genome, identified through sequence homology (Sudarsan *et al* 2008). Riboswitching is a process involving regulatory

mRNA with domains able to recognise metabolic-state cyclic di-GMP–binding RNAs, and can alter gene expression resulting in differences in bacterial physiology such as cell differentiation, motility, biofilm assembly and virulence factor production (Sudarsan *et al* 2008). Secondly, *C. difficile* phages may assist in phage defence of their host cell. PhiC2 encodes an *abiF* gene (Goh *et al* 2007), in bacteria, AbiF is predicted to be involved in phage abortive infection, a mechanism by which host bacteria can inhibit phage replication (reviewed by Labrie *et al* 2010). Also, both of the two prophages carried in *C. difficile* strain CD630, encode CRISPR arrays (Sebaihia *et al* 2006). CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) are another host mechanism of phage defence, encoded across many different bacterial phyla (Pourcel *et al* 2005, Bolotin *et al* 2005, Sorek *et al* 2008, Horvath *et al* 2009). Lastly, the holin gene encoded by phiC2 is highly similar to the tcdE gene in the PaLoc, which has led researchers to suggest the PaLoc may have a phage origin (Goh *et al* 2005).

1.5.2. Impact of phage infection on *C. difficile.*

Whilst the genomic sequencing of *C. difficile* phages shows that they encode interesting genes, no phages have been found that encode toxin genes. As a result, research examining the impact of phage infection on *C. difficile* has focused on two aspects; their application as therapeutic agents and the effect of phage infection on toxin production.

Therapeutic potential of phages for treating *C. difficile* has been investigated both *in vivo* (Ramesh *et al* 1999) and *in vitro* (Meader *et al* 2010), as well as a phage encoded endolysin (Mayer *et al* 2008). The biotechnological potential of alternative antimicrobials for treating *C. difficile* has been realised, with patents licensed for this endolysin (encoded by phiCD27, Patent Reference: International Patent Publication Nos. WO2009/068858 and WO2010/136754) and 'Diffocins' which

are bacteriocins (thought to be degenerated prophage) encoded by several *C. difficile* strains (US Patent Application No: 2011/0293,566).

Interestingly, while none of the sequenced *C. difficile* phages encode identified toxin genes, phage infection can modulate toxin production, albeit with conflicting results (Goh *et al* 2005b, Govind *et al* 2009, Meader *et al* 2010, Sekulovic *et al* 2011). Furthermore, in a therapeutic model it was shown that phiCD119 can lysogenise its host (Govind *et al* 2011). These results, and the fact that no group has isolated a virulent phage, are partly responsible for the focus on phage derived products, rather than the use of whole phage to target *C. difficile*. Yet the use of whole phages offers advantages which phage products cannot; the amplification of phage at the site of infection via replication and the ability to evolve with the host. For phage treatment to be effective, penetration of the psuedomembrane on the colonic epidermal layer formed by *C. difficile* to clear infection is needed. Although there has been no research of this nature, phage penetration has been investigated for the eradication of bacterial biofilms, a similar dense, cell- and cell debris-rich environment (Donlan 2009).

The contribution that *C. difficile* phages have on their host's evolution and pathogenicity has been explored in few studies. For example, work by Nale *et al* (2012) investigated the diversity and prevalence of prophage carriage in a large group of clinical R027 strains. Also, Lawley *et al* (2009) recently determined the entire spore proteome of *C. difficile* strain CD630, determining a "core" set of spore proteins common across Clostridial species, as well as 29 proteins unique to *C. difficile*, three of which are encoded by its two prophages (CD0939, CD0940 and CD2926). Although of unknown function, their presence suggest that the prophages contribute to the mature spore physiology. The relationship between phages and sporulation has been documented in other bacterial host systems, for example phage-enhanced sporulation of *Bacillus subtilis* (Silver-Mysliwiec and Bramucci 1990).

1.6. Aims and objectives of this project

The aims of this PhD were to investigate and characterise phages associated with non-clinical strains of *C. difficile*. To do so, a collection of infant, animal and environmental *C. difficile* isolates was establihed to use as hosts for phages as well as sources of temperate phages (section 3). To determine the diversity of this collection, ribotype analysis was performed on all isolates and toxin gene carraige, motility and antibiotic resistance was examined. Also, the prevalence, diversity and extent of lysogeny in these *C. difficile* isolates were investigated. The isolate collection represents a diverse panel of strains to use phage isolation. Additionally, their characterisation has revealed insights into the ecology of *C. difficile* outside of the hospital environment and highlights the potential for strain evolution within these reservoirs.

The *C. difficile* collection was used to isolate phages from sample extracts, enrichment cultures and induced cultures of the non-clinical isolates (section 4). The resulting phages which were isolated were then characterised according to particle morphology, genome size and host range. Based on these characterisations and origins, the phage panel represents a diverse collection that can be sub-grouped on the basis of morphology and genome size. The host range analysis showed that these phages have potential clinical applications, as they can infect several different ribotypes. The isolation work showed that several isolates release phages spontaneously which are viable, suggesting that temperate phage infection occurs frequently, *ex situ*. Also, the isolation and host range work showed that there is considerable intra-ribotype variation with respect to phage sensitivity and efficiency of plaqueing, demonstrating the tight phage-host interactions within *C. difficile*.

Whole genome sequencing and annotation was performed for six of the isolated phages and one prophage from strain T6 (section 5). Sequencing also was performed to identify potential virulence factors or genes with lysogenic conversion in order to determine their suitability in therapeutic applications and/or impact on host pathogenicity. The phages were selected to

maximise genetic diversity and the chance of identifying virulent phages. The sequencing identified highly conserved genes, such as integrases, as well as highly unusual genes encoded by the phages. These include homologs of genes in the bacterial *agr* QS system in phiCDHM1, which is the first time such genes have been found in a phage genome. Phylogenetic analyses of these genes suggest they are a novel locus which is susceptible to HGT. Also, analysis of the protein domains identified in the phage genes identified those which have been only found in bacteria and aquatic viruses. The genetic analysis in this chapter provides insight into the genome evolution of the phages as well as *C. difficile*.

In order to examine the phylogeny and genome diversity of *C. difficile* phages, comparative genomics of the six phages sequenced in this project were performed (section 6). The comparisons were to one another, as well as to known *C. difficile* phages which include those associated with clinical strains. These analyses were performed using pairwise blastn alignments, which identified regions of the phage genomes that are highly conserved and regions which are variable. Also, genome alignments at the nucleotide level were performed to overall genome architecture of *C. difficile* phages, which identified distinct genomic lineages within the *C. difficile* phages, as well as showing evidence of genetic transfer between these lineages. Lastly, analysis of the *C. difficile* CRISPR system was performed to examine whether the CRISPR system impacts on phage infection. The CRISPR system of *C. difficile* appears to be active, targeting specific genes encoded in the phage genomes. Evidence of evolution in the phages was found, as well as support for the hypothesis that phage infection is limited by the CRISPR content of *C. difficile* strains. Unusually, *C. difficile* prophages encode CRISPR arrays, and an analysis of multiple genomes showed that different prophages encode multiple and diverse arrays and can target other *C. difficile* phages.

The work in this thesis describes the tight phage-host relationship in *C. difficile*, as well as ways in which its phages impact on its biology and genome evolution.

2. Materials and Methods

2.1. Isolation and characterisation of non-clinical strains of *C. difficile*, including prophage carriage

2.1.1. Routine laboratory culture of *C. difficile* and reference strains.

C. difficile was routinely cultured at 37°C under anaerobic conditions (10% H_2, 10% CO_2 and 80% N_2) in a MiniMACS chamber (Don Whitley, U.K.). Strains were maintained in cryogenic stocks at -80°C, and cultured on Brain Heart Infusion (BHI: Oxoid, U.K) 7% Defibrinated Horse Blood (DHB) 1% Bacteriological Agar (BA: Oxoid, U.K) plates. Liquid culture was performed using either BHI or Fastidious Anaerobic broth (FA: Oxoid, U.K). Before use, BHI was pre-reduced inside the chamber for a minimum of 6 hours and FA broth was warmed to room temperature (RT). Phages stocks were maintained in BHI at 4°C, and titred using plaque assays.

C. difficile strains used include: those isolated from environmental, animal and infant sources in this project; *C. difficile* strain NCC11204; twelve environmental *C. difficile* strains kindly donated by Maja Rupnik, University of Maribor and strain CD630 kindly donated by Prof. Mark Wilcox, University of Leeds Hospital Trusts.

2.1.2. Sampling of infant, animal and environmental sources.

Nineteen stool samples were obtained from different infants attending a Leicestershire nursery, all under the age of 3 years old. Three infants were sampled over monthly intervals, one infant for 3 months, one infant for 5 months and one infant for 12 months.

Faecal samples were obtained from horses (*n*=31), pigeons (*n*=6), cows (*n*=5), cats (*n*=2), rabbits (*n*=2), a fox (*n*=1), a tortoise (*n*=1), a sheep (*n*=1) and a deer (*n*=1). Horses sampled were from a commercial livery yard in Leicestershire (*n*=30) and one from in Hampshire (*n*=1). Pigeon, cow,

horse, rabbit and fox samples were kindly provided by Julian Clokie. Cats sampled were domestic, asymptomatic animals. The deer sample was obtained from Bradgate Park, Leicester, U.K.

Four soil samples and one water sample were obtained from Bradgate Park, Leicestershire, U.K. Single estuarine sediment samples were obtained from Langtsone Harbour, Hampshire, U.K. and from Cornwall, U.K., both gratefully received from Julian Clokie. As a result of the successful isolation of *C. difficile* from the Havant estuarine sample, this location was sampled more extensively over two consecutive years. In October 2009, 21 samples were taken throughout the Havant estuarine system in Langstone and Chichester Harbours designated as a Site of Special Scientific Interest (SSSI) by Natural England (13 Unit ID 1016965. Sites were chosen to include those that were thought to be polluted and those to be less polluted (clean sites). These sites were re-sampled in October 2010 and a subset resampled in June 2011. Assistance in selecting locations and performing the sampling was gratefully received from Julian Clokie. All samples were stored at 4°C until processed.

2.1.3. Method for the isolation of *C. difficile*.

The standard method for *C. difficile* isolation provided by the Health Protection Agency (HPA) was optimised to include an enrichment step (accessed online at http://www.hpa-standardmethods.org.uk/documents/bsop/pdf/bsop10.pdf). Approximately 2 cm^3 of sample was transferred into 10 ml FA or BHI supplemented with 250 µg ml^{-1} cycloserine, 8 µg ml^{-1} cefoxitin and 0.1% sodium taurocholate to enhance spore germination (as Wilson *et al* 1982 used in solid media). Enrichment cultures were incubated for 10 days and afterwards centrifuged for 10 min at 3,398 *x* g. The supernatant was removed from the pellet for later use in phage isolation. An equal volume of Industrial Methylated Spirit (IMS) was added to the pellet and vortexed for 30 sec. The sample was incubated for 30 min at RT then 5 – 15 µl sub-cultured on Brazier's cycloserine-cefoxitin egg yolk (CCEY) agar plates (Bioconnections, U.K.) for 24-48 hours. Clonal strains were obtained by growing

single colonies through two rounds of purification on BHI 7% DHB 1% BA plates. Cryogenic stocks were produced by innoculating single colonies into 5 ml FA for overnight growth. Cultures were centrifuged at 10,000 x g for 10 min and the pellet resuspended in glycerol and stored at -80°C.

C. difficile growth was indicated from colony morphology, detection of the distinctive *C. difficile* odour of horse dung/urine, and observation of the characteristic chartreuse fluorescence under long wave UV light (365 nm). As required, confirmation using the *C. difficile* Test Kit (Oxoid, U.K.), based on latex agglutination, was performed, according to the manufacturer's guidelines.

2.1.4. DNA extraction and molecular confirmation of *C. difficile* isolates.

C. difficile DNA was extracted using one of two methods: either phenol:chloroform or Chelex resin (Biorad, U.K.). For the phenol:chloroform method, 2 ml of overnight BHI culture was centrifuged at 16,000 x g for 5 min. The pellet was washed with 0.1x sodium chloride/sodium citrate (SSC) buffer (prepared as 20x stock of 3 M NaCl, 0.3 M sodium citrate dihydrous adjusted to pH 7 using 1 M HCl diluted to 0.1x with UPH$_2$O for use). The pellet was resuspended in 0.3 ml 10mM Tris-HCl containing 2.5 mg ml^{-1} lysozyme and incubated at 37°C for 45 min. Afterwards 0.5 ml lysis buffer containing 10 mM Tris pH 8.0, 1 mM EDTA, 1% SDS and 0.2 mg ml^{-1} proteinase K was added to the sample and further incubated for 30 min. Then 0.8 ml phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v) was added to the sample and centrifuged at 16,000 x g for 5 min. The top aqueous layer was retained and to it was added an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v). The centrifugation was repeated, the top layer retained and to it was added an equal volume of chloroform:isoamyl alcohol (24:1). It was then centrifugated again. DNA was precipitated by incubation on ice for 1 hour with 2 volumes isopropanol and 0.4 volumes 1.5 M NaAc. DNA was pelleted by centrifugation at 16,000 x g for 30 min at 4°C with a wash in 70% ethanol and repeated centrifugation step. The pellet was air dried for 10 min and resuspended in either UPH$_2$O or elution buffer (10 mM Tris-HCl, pH 8.5).

Chelex DNA extraction was performed by resuspending cells from overnight culture into 150 μl solution of UPH$_2$O and 5% Chelex resin. Samples were heated to 100°C for 12 min, cooled for 5 min and centrifuged at 16,000 x g for 10 min. The supernatant was stored until use at 4°C.

PCR confirmation for *C. difficile* was performed by amplification of the 16S rRNA gene using *C. difficile* specific primers: g-cdif-F (5'-TTGAGCGATTTACTTCGGTAAAGA-3') and g-cdif-R (5'-CCATCCTGTACTGGCTCACCT-3') designed by Rinittala *et al* (2006), producing an expected 157 bp product. PCR reactions were performed in a total volume of 50 μl, containing template DNA, 4 μM forward and reverse primers, 0.25 mM dNTPs, 4 mM MgCl$_2$, 1× PCR buffer, and 0.5 U of BioTaq DNA polymerase (Bioline, U.K.). Amplification conditions were: 95°C for 5 min; 40 cycles of 95°C for 30 sec, 44°C for 30 sec and 72°C for 45 sec; with a final extension of 5 min at 72°C. The expected 157 bp products were visualised alongside 10μl of GeneRuler a 1 kbp DNA ladder (Fermentas, U.K.) on a 1% Helena Agarose gel prepared in 1x TAE buffer stained with GelRed and run at 90 volts for 1 hour in TAE buffer. Gels were imaged using SynGene software.

During optimisation of the method, identification of contaminating organisms was made on several isolates using universal bacterial 16S rRNA primers 8F (5'-AGAGTT TGATCCTGGCTCAG-3') and 1391R (5'-GACGGGCGGTGTGTRCA-3') producing a 1800 bp product as designed by Edwards *et al* (1989) and Lane *et al* (1986), respectively. PCR reactions were performed in a volume of 50 μl with: template DNA; 2 μM forward and reverse primers; 0.25 mM dNTPs; 2 mM MgCl$_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. Amplification was carried out at: 95°C for 5 min; 30 cycles of 95°C for 30 sec, 55°C for 60 sec and 72°C for 2 min; with a final extension step at 72°C for 10 min. PCR products were gel-purified using Qiagen Gel Elute kit. Sequencing was carried out at GATCbiotech, U.K., and data analysis performed using Chromasv1.45. Sequences were blasted against the NCBI non-redundant nucleotide collection using the Blastn algorithm accessed online at (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Nucleotides) to assign bacterial species.

2.1.5. Gel-based and Capillary-based PCR ribotyping of *C. difficile* isolates.

PCR ribotyping was performed on Chelex extracted DNA using primers from Bidet *et al* (2000), forward (5'-GTGCGGCTGGATCACCTCCT-3') and reverse (5'-CCCTGCACCCTTAATAACTTGACC-3'), under conditions modified by Jinyu Shan, University of Leicester. PCR products were concentrated by evaporation at 75°C on a heating block, reducing the 50 µl reaction volumes to approx 20 µl before loading for gel electrophoresis. Ribotype profiles were visualised by separation of PCR products on a 3% Hi-Res agarose (GeneLink, U.K.) gel prepared using 1× Tris-acetate-EDTA pH8 (TAE) stained with 10 µg ml^{-1} GelRed (Biotium, Hayward, Carlifornia, U.S.A.), run in TAE buffer for up to 8 hours at 90 volts with a 100 bp molecular marker, Hyperladder IV (Bioline, U.K.) and viewed using SynGene software.

Capillary gel electrophoresis was used for all isolates and performed on Chelex extracted DNA following methods described in Griffiths *et al* (2010). The PCR products were analysed using capillary gel electrophoresis performed by PNACL at the University of Leicester on an ABI 3730 genetic analyser (Applied Biosystems, U.K.). Data analysis was performed on PeakScanner v1.0 (AppliedBiosystems) with peaks counted if >3% the highest peak for quality control with band resolution of 1.5 bp and unknown ribotypes sent to Prof. Mark Wilcox and Dr. Warren Fawley, University of Leeds, U.K.

2.1.6. PCR based toxin gene screening of *C. difficile* infant isolates.

PCR was carried out to detect carriage of ToxinA and ToxinB genes (Kato *et al* 1991 and 1998) as well as the two binary toxin genes: *cdtA* and *cdtB* (Stubbs *et al* 2000) in order to determine potential for toxin product. Primer pair NK2 (5'-CCCAATAGAAGATTCAATATTAAGCTT-3') and NK3 (5'-GGAAGAAAAGAACTTCTGGCTCACTCAGGT-3') which amplify the partial sequence of *tcdA,* producing an expected product of 251 bp. Primer pair NK9 (5'-CCACCAGCTGCAGCCATA-3') and NK11 (5'-

TGATGCTAATAATGAATCTAAAATGGTAAC-3') which amplify an essential repeat region within *tcdA* and produce an expected product of 1,265 bp. These primers were used in a multiplex reaction in a 50 μl volume: template DNA; 1 μM forward and reverse primers; 0.25 mM dNTPs; 1.5 mM MgCl$_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. Primers NK104 (5'-GTGTAGCAATGAAAGTCCAAGTTTACGC-3') and NK105 (5'-CACTTAGCTCTTTGATTGCTGCACCT-3') produce an expected product of 203 bp targeting the *tcdB*. PCR reactions were in 50 μl volumes with: template DNA; 2 μM forward and reverse primers; 0.25 mM dNTPs; 1.5 mM MgCl$_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. Amplification conditions for all primers were: 95°C for 5 min; 32 cycles of 95°C for 20 sec, 62°C for 120 sec and 72°C for 2 min; with final extension step at 72°C for 5 min. Primers were used to amplify *cdtA* (forward 5'-TGAACCTGGAAAAGGTGATG-3' and reverse 5'-AGGATTATTTACTGGACCATTTG-3') with a product size of 375 bp and *cdtB* (forward 5'-CTTAATGCAAGTAAATACTGAG-3' and reverse 5'-AACGGATCTCTTGCTTCAGTC-3') with a product size of 510 bp. These primers were used in a 50 μl multiplex reaction: template DNA; 1 μM forward and reverse primers; 0.25 mM dNTPs; mM MgCl$_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. PCR reaction conditions were: 95°C for 5 min; 30 cycles of 94°C for 45 sec, 52°C for 60 sec and 72°C for 2 min; with a final extension step at 72°C for 5 min. Products were visualised alongside 10 μl of Generuler 1 kbp DNA ladder. Samples were loaded on a 1% Helena Agarose gel prepared in 1x TAE buffer and stained with GelRed. It was run at 90 volts for 1 hour in TAE buffer. Gels were imaged using SynGene software.

2.1.7. Motility test of environmental *C. difficile* isolates.

Two protocols were tested to determine strain motility, a liquid assay as described by Stabler *et al* (2009) and a soft agar plate assay adapted from Be'er *et al* (2009). Overnight liquid cultures were inoculated into the top 2-5 mm of 7 ml 0.05% agar BHI in 7 ml volume sterile bijous or

stabbed into 7 ml BHI 0.3% BA 60 mm Petri dishes and incubated 24 hours. Growth was scored for levels of motile growth.

2.1.8. Antibiotic Minimum Inhibitory Concentrations (MICs) determination by Etest of *C. difficile*.

The MICs of isolates for vancomycin (VA), metronidazole (MZ), clindamycin (CM) and ciprofloxacin (CI) antibiotic were determined. 300 µl of overnight FA culture was used to produce confluent growth (a lawn) on a BHI 1% BA plate. Lawns were air dried for five minutes in a flow hood and Etest strips (Biomerieux, U.K.) were applied onto the prepared lawn. Plates were inverted and incubated overnight.

2.1.9. Prophage carriage determined using Transmission Electron Microscopy (TEM).

Isolates were grown overnight in BHI and induced with either mitomycin c (MC) at 3 µg ml^{-1} or norfloxacin (NFX) at 6 µg ml^{-1} and incubated for 24 hours. Also untreated cultures were included as controls to assess the spontaneous release of phages. Cultures were centrifuged at 3400 x g for 10 min and the supernatant filtered using 0.22 µm pore size filters (Millipore, U.K.). The lysates were purified and concentrated using PEG 6000. NaCl was added to lysates to a final concentration of 0.5 M, mixed thoroughly and incubated on ice for 1 hour. Samples were centrifuged at 3,400 x g for 40 min at 4°C. The supernatants were transferred to a sterile tube and 10% w/v PEG 6000 dissolved at RT for 1 hour, followed by overnight incubation at 4°C. Phages were precipitated by centrifugation at 3,400 x g for 40 min at 4°C, and resuspended in SM buffer. Samples were treated with 1.4 µg µl^{-1} DNAse I (New England Biolabs, U.K.) and 3 µg µl^{-1} RNase H (New England Biolabs, U.K), and incubated for 1 hour at 37°C. An equal volume of chloroform was added to each sample and sampled were centrifuged at 3,398 x g for 10 minutes at 4°C. The supernatant was stored at 4°C in the dark until use. Samples for TEM were negatively stained with 0.1% uranyl acetate on 0.25% pioloform coated

copper electron microscopical grids (Athene type 3 mm, Agar Scientific, U.K.) that had been carbon coated and high-voltage glow discharged. TEM was performed on a JEOL 1220 (JEOL, U.K.).

2.2. Isolation of phages infecting *C. difficile* in a lytic manner

2.2.1. Isolation of phages with lytic activity *C. difficile.*

To extract free phages from samples, 1-2 g of sample was rocked in an equal volume of SM buffer (5.8 g NaCl, 2 g hydrous MgSO$_4$) for a minimum of 1 hour at RT. This was centrifuged at 3,400 x g for 10 min. The supernatants were filtered through 0.22 μm pore size filters and stored in the dark at 4°C until use. Several extracts were also pooled and PEG 6000 concentrated in order to maximise chances of isolating phages from extracts. Enrichment cultures from the *C. difficile* isolation were also used to isolate phages. The supernatants were filtered through 0.22 μm pore size filters and stored in the dark at 4°C until use.

Culture lysates were screened for temperate phages, capable of propagating lytically on different *C. difficile* hosts. The culture lysates were screened first following centrifugation and filtration using 0.22 μm pore size filters, and secondly following a PEG 6000 concentration step performed after filtration.

Screening for active phages was performed by spot test assay using the double agar overlay lawn method. 100-250 μl of overnight *C. difficile* culture was mixed with 3 ml molten BHI 0.4% BA supplemented with 0.4M MgCl$_2$ and 0.1mM CaCl$_2$ (Goh *et al* 2005). This mixture was poured onto BHI 1% BA plates and set. Samples were spotted onto lawns in 10 μl volumes. Plates were incubated overnight.

In the first round of screening, 33 strains of *C. difficile* were used as a panel of indicator strains. This included strain NC11204, strain CD630, animal, infant, estuarine and Slovenian isolates.

Variations on the spot test assays were performed in an effort to increase phage isolation which included the addition of 3 μg ml^{-1} and 5 μg ml^{-1} ampicillin (Sigma, U.K.) to the BHI 1% BA plates (Łoś *et al*, 2008). Another optimisation attempt performed was addition of 0.25%, Ox Bile (Sigma-Aldrich, U.K. as previously described for Bacteroides phage assays by Araujo *et al* (2001). Also spot tests were performed with and without supplementing the soft agar layer with 0.4 mM MgCl$_2$ and 0.1 mM CaCl$_2$ salts were performed as also carried out by Goh *et al* (2005).

The second round of screening was performed by Holly Colvin during her undergraduate project. Four of the most commonly infected strains (S1, S7, H10 and T6) were used as indicator strains for screening the enriched cultures for active phages following the second sampling (2010) of the Hampshire sites (*n*=25), as well as NFX induced and control lysates of *C. difficile* strains isolated (*n*=32).

The third set of samples, obtained from Hampshire sites in 2011 (*n*=25), were processed and screened by Kirstin Norman during a summer studentship. Twelve clinical strains from Leicestershire hospital trusts and environmental isolate T6 were used as indicator strains.

To obtain clonal phage stocks, plaques underwent four rounds of purification. To do this, single plaques were picked, using sterile Pasteur pipettes, and dispensed into BHI that was incubated for a minimum of one hour at 4 °C. The resuspended phages were then centrifuged at 10,000 x g for 5 min and the supernatant was used in the next round of purification. Serial dilution using BHI broth was performed as required to obtain single plaques.

Plaque assays were used to produce high phage stocks. 10-100 μl of purified phage samples were added to 300 ml overnight liquid *C. difficile* cultures and 3 ml BHI 0.4% BA supplemented with 0.4 M MgCl$_2$ and 0.1 mM CaCl$_2$ which was mixed thoroughly before adding onto a BHI 0.1% BA plate. Plates were incubated overnight. If confluent lysis was produced, the top layer of soft agar was removed into a sterile 15 ml falcon tube. To this, an equal volume of BHI was added to enable

resuspension of phage particles out of the soft agar. Phage samples were incubated for a minimum 1 hour at 4°C. Stocks were centrifuged at 4,000 x g for ten min and the supernatant filtered through 0.22 µm filters. Stocks were examined using TEM, titred through spot tests and stored at 4°C.

2.2.2. Phage induction growth dynamics.

Two growth curves of *C. difficile* were performed using mitomycin c (MC) treated and untreated cultures, whilst monitoring the OD at 550nm, CFU ml^{-1} and PFU ml^{-1} for ten time points over 96 hours. A single colony taken from a BHI 7% DHB 1% BA plate incubated overnight was transferred into 50 ml pre-reduced BHI broth MC at a concentration of 3 µg ml^{-1} was added to one culture when OD$_{550nm}$ 0.3 was reached. Aliquots of 3 ml of culture were obtained at each time point. OD was measured from 1 ml volumes. To calculate the CFU ml^{-1}, culture was serially diluted with BHI. 10 µl volumes of each dilution were applied on to BHI 1% BA plates (in triplicate) and air dried before incubation. PFU ml^{-1} was determined by centrifugation of 1 ml culture at 10,000 x g for 5 min and the supernatant serially diluted. 10 µl volumes of each dilution were applied in triplicate onto lawns of the *C. difficile* indicator strain. Counts for CFU ml^{-1} and PFU ml^{-1} were performed after 24 hour incubation.

2.2.3. Analysis of phage genome size using Pulsed Field Gel Electrophoresis (PFGE).

Phage genome size was determined using PFGE. Phage samples were set into agarose plugs which were produced by adding 40 µl of molten 2% Seaplaque agarose (Lonza, U.K.) in x0.5 TBE to 40 µl of phage stock (>10^7 PFU ml^{-1}). Plugs were set for 30 min at 4°C and digested in 1 ml lysis buffer (100 mM EDTA, 100 mM Tris-HCl, pH 9.0, 1% SDS and 0.5 mg ml^{-1} proteinase K) at 55°C overnight. Plugs were then washed three times in TE buffer (Tris:EDTA ,pH 8.0) and assembled into a 1% PFGE agarose (Biorad, U.K.) x0.5 TBE gel. A low weight molecular marker ladder (New England BioLabs,

U.K.) was used for comparison to determined phage genome size. The gel was run at 6 volts/com for 17 hours at 14°C (Ramped pulse times were: initial 5 sec; final 13 sec) Gels were stained with 10 µgml^{-1} ethidium bromide for 30 min, then visualised using SynGene software.

2.2.4. Phage DNA extraction and restriction enzyme digest.

Extraction of phage DNA was performed using phenol:chloroform on PEG 6000 concentrated samples. Samples were treated with 1.4 µg µl^{-1} DNAse I (New England Biolabs, U.K.) and 3 µg µl^{-1} RNase H (New England Biolabs, U.K.), and incubated for 1 hour at 37°C. 20 µg µl^{-1} of Proteinase K was added to samples and incubated at 55°C for 1 hour. An equal volume of phenol was added to samples and mixed by inversion. Samples were centrifuged at 16,000 x g for 10 min and the top aqueous layer retained. This was added to an equal volume phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v) and mixed, centrifuged at 16,000 x g for 10 min and the top aqueous layer retained. The last step was repeated twice, first using phenol:chloroform:isoamyl alcohol, followed by one using chloroform:isoamyl alcohol (24:1 v/v). After the final purification, the DNA was precipitated using 2 volumes isopropanol and 0.4 volumes 7.5 M NaAcetate. Samples were incubated on ice for 1 hour or overnight at -20°C. DNA was pelleted following centrifugation at 21,000 x g for 30 min at 4°C. The pellet was washed with 70% ethanol and air dried before resuspension in either EB buffer (10 mM TrisCl, pH 8.5) or UPH$_2$O.

DNA was quantified using a NanoDrop ND-1000 spectrophotometer. Genomic DNA contamination was examined by PCR, using the primers g-cdif-F and g-cdif-R that amplify the *C. difficile* 16S rRNA gene. Approximately 100 ng ml^{-1} of DNA was used in digestion assays with pne of the following restriction enzymes: *Bam*HI; *Nde*I; *Nhe*I; *Sty*I; *Sma*I; *Hind*III; *Dyn*I; *Eco*RV or *Eco*R1 (New England Biolabs, U.K.). Enzymes were used with their recommended buffers according to the manufacturer's guidelines, and digests were performed from 1 hour to overnight, during optimisation of the methods. Digestions were mixed with 6x loading dye (New England Biolabs, U.K.)

and separated alongside Hyperladder IV by gel electrophoresis on a 1% Helena Agarose TAE gel that was run at 90 volts for 1 hour, stained with GelRed and visualised using SynGene software.

2.2.5. Determining phages host range, including environmental and clinical strains.

The host range for the first eight purified phages was performed using all the environmental strains obtained by this date, including the infant, animal, estuarine and Slovenian strains. The phages were also tested against 23 strains of R027 (representing 23 different multilocus variable-number tandem repeat analysis (MVLA) types, Fawley *et al* 2008) kindly provided by Prof. Mark Wilcox, Leeds Hospital Trusts and seven clinical strains of R078 from Leicestershire Hospital Trusts. Anisha Thanki expanded this host range analysis during an M.Sc project (01/2011– 08/2011) to include all the purified phages (*n*=23) obtained against a range of several ribotypes from the Leicestershire Hospital Trusts (isolates *n*=114, ribotypes *n*=21). A third expansion was performed using all purified phages (*n*=23) and included all strains isolated from the neonate, animal and Hampshire 2009, 2010, and 2011 sample sets (isolates *n*=61; ribotypes *n*=21).

2.3. Genome sequencing and annotation of *C. difficile* phages and bioinformatic analysis

2.3.1. Phage genome sequencing and annotation.

Genomes of two phages were submitted for sequencing using the Roche 454 sequencing with a 300x coverage, kindly performed by Prof Andrew M. Kropinski and Erika Lingohr, University of Guelph, Canada. Reads were assembled using Phred/Phrap (Ewing *et al* 1998a and Ewing *et al* 1998b). For phiCHM23, the results indicated there were two phages in the sample and their genomes could not be separated from one another. The genome of phiCDHM1 was assembled into two contigs. The genome was visualised in Artemis Genome Browser and Annotation Tool

(Rutherford *et al* 2000). Open Reading Frames (ORFs) were identified using GeneMark.hmm 2.0 (Besemer *et al* 1999). The amino acid sequences of ORFs were blasted using BLASTP through the NCBI online nr database, Pfam and Uniprot. Protein domains were also searched for using the NCBI Conserved Domain Database (Marchler-Bauer *et al* 2011). The contigs and fragmented ORFs were joined by PCR using primers designed using Primer3v0.4.0 (Rozen *et al* 2000). To close contigs: 003AR 5'-TCACAAGCCTCAATTGCATTA-3' and 004AR 5'-TGGCATTATTGTTAACAGCATCA-3' with a 456 nt product and 003BF 5'-TTTGATATGAACAATGAAAATGAACA-3' and 004BF 5'-TCCATATACTCATCGGAATTTTCA-3' producing a 689 bp product were used. PCR reactions were performed in 50 µl volumes containing: template DNA; 4 µM forward and reverse primers; 0.25 mM dNTPs; 3 mM $MgCl_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. Amplification conditions were: 95°C for 5 min; 30 cycles of 95°C for 30 sec; 48°C for 30 sec; 72°C for 60 sec; with a final extension of 5 min at 72°C. The NACHT-NTPase and histidine kinase genes of phiCDHM1 both required re-sequencing as they were fragmented into separate ORFs. Primers used targeting the putative histidine kinase gene were WHKF 5'-AGGATTTGTAATCCATAGGAACAT-3' and WHKR 5'-TTTTCGTTCGTTTTATTATTACAGTTT-3' producing a 1657 bp product. Primers targeting the NTPase gene were NTPaseF 5'-CGCAAGTTACTGAAAAACTCCA-3' and NTPaseR 5'-TTTCTCCCAATTTTTACACTGTTGA-3', which amplify a 840 bp product. PCR reactions were carried out in 50 µl volumes containing: DNA template; 4 µM forward and reverse primers; 0.25 mM dNTPs; 3 mM $MgCl_2$; 1× PCR buffer and 0.5 U of BioTaq polymerase (Bioline, London, UK). Amplification conditions were: 95°C for 5 min; 30 cycles of 95°C for 30 sec, 48°C for 60 sec and 72°C for 120 sec; with a final extension of 5 min at 72°C. All products were visualised on a 1% Helena Agarose gel prepared in 1x TAE buffer and stained with GelRed. This was run at 90 volts for 1 hr in TAE buffer and sampled were loaded alongside 10µl of GeneRuler 1 kbp DNA ladder. Sequencing was performed on gel-purified PCR products using the QiAquick Gel Extraction kit (Qiagen, U.K.) following the manufacturer's instructions. Sanger sequencing was carried out at GATCbiotech Ltd

(U.K.). Data was analysed using Chromas v1.45 and ClustalW2 (Larkin *et al* 2007, Goujon *et al* 2010); the genome was re-annotated in Artemis.

Genomes of phages phiCDHM19, phiCDHM13, phiCDHM11, phiCDHM14 were sequenced and assembled at the Centre for Genomic Research, University of Liverpool, resulting in multiple contigs which were joined using PCR with subsequent re-sequencing. A further six phage genomes were submitted but failed the quality control at the Centre. These are currently being sequenced. Due to the phage's sequence similarity, the four phages genomes were joined using scaffolds and PCR to fill gaps, resulting in fewer contigs. Open reading frame (ORF) prediction was performed using: OrfFinder (set to bacterial code 11), accessed at http://www.ncbi.nlm.nih.gov/projects/gorf/; GeneMark.hmm 2.0 (default settings: Besemer and Borodovsky 1999), accessed at http://exon.gatech.edu/heuristic_hmm2.cgi; and GLIMMER (bacterial code 11: Salzberg *et al* 1998, Delcher *et al* 1999), accessed at http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi. Artemis Genome Browser and Annotation Tool (Rutherford *et al* 2000) was used to assemble the genome annotation. The amino acid sequences of ORFs were blasted using BLASTP through the NCBI online nr database, Pfam and Uniprot. Protein domains were also searched for using the NCBI Conserved Domain Database (Marchler-Bauer *et al* 2011) and Interproscan used to identify SignalP and transmembrane helices sequences. Genome maps were generated using DNAplotter (Carver *et al* 2009).

Genomic sequence information for phiT6 was acquired from the whole genome sequence of isolate T6, as it is a prophage. The sequencing of bacterial strain T6 was performed by Trevor Lawley at the Sanger Centre, Cambridge.

2.3.2. Phylogenetic analysis of phage ORFs encoding the capsid gene and the *agr* cassette genes.

The capsid genes of the sequenced *C. difficile* phages were aligned at the amino acid level using the MUSCLE/Alignment Explorer add in of in Molecular Evolutionary Genetics Analysis (MEGA)

version 5.05 (Tamura *et al* 2007, Kumar *et al* 2008, Tamura *et al* 2011). Using MEGA, Neighbour Joining (NJ: Saitou and Nei 1987) phylogenetic analysis was performed on the alignments, with parameters set for the Poisson Model substitution model (Zuckerkandl and Pauling 1965) with uniform rates, using partial deletions and bootstrapped with 500 replicates (Felsenstein 1985). Maximum Likelihood (ML) phylogenetic analysis was also performed, with parameters set for the Jones Taylor Thornton (JTT) nucleotide substitution model (Jones *et al* 1992), with uniform rates, using all sites and Close-Neighbour-Interchange (CNI) for Tree Inference and bootstrapped with 500 replicates. The two tree topologies changed slightly with the two siphovirus sequences split and lower bootstrap values obtained in the ML tree.

Homologous sequences for each of phiCDHM1 *agrB* and histidine kinase ORFs were found using the BLASTP algorithm searching the NCBI nt/nr database. The *agrD* genes were identified by manually searching candidate genes immediately upstream or downstream of the *argB* and histidine kinase gene in deposited *C. difficile* genomes. Amino acid sequences for each gene were aligned using the MUSCLE/Alignment Explorer add in of in Molecular Evolutionary Genetics Analysis (MEGA) version 5.05 (Tamura *et al* 2007, Kumar *et al* 2008). ML phylogenetic analysis was performed, with parameters set for the JTT nucleotide substitution model, with invariant rates, using all sites and CNI for Tree Inference and bootstrapped with 500 replicates. Alternative trees were also constructed using the Poisson nucleotide substitution model, and Neighbour Joining and Minimum Evolution phylogenetic analyses. ML phylogenetic analysis was also performed on sequences aligned using CLUSTALW/Alignment Explorer in MEGA v5.01. Tree topologies remained largely the same, showing the same clustering of taxa, although branch lengths differed slightly for some.

The amino acid sequence of the histidine kinase gene was searched in the Search Tool for the Retrieval of Interacting Genes (STRING) database (Jensen *et al* 2008), accessed online at http://string-db.org/ to identify putative protein interactions of the kinase.

2.3.3. Screening of myovirus capsid gene in environmental *C. difficile* genomes.

Four sets of primers were used to screen *C. difficile* isolates for the presence of different subclasses of myovirus prophages in the bacterial genomes. DNA samples were extracted using the Chelex protocol described in 1.3.4. One set targets the phiCDHM1 group: with the following oligo sequences: K12groupF (5'-AACTTCGGGGATTTGTATGC-3') and K12groupR (5'-CAACAAATTGTATTGCATCAGC-3'), producing an expected product of 814 bp. K6group oligos: K6groupF (5'-TGGTTGGATGGATTCTAATGCT-3') and K6groupR (5'-GACCAAGCATTTGCTGTTTG-3') produce an expected product of 771 bp. CD27group oligos; CD27groupF (5'-GAGGGCAGGAATAAGAAAAGC-3') and CD27groupR (5'-GATTCCCTATCCTCAACTACGC-3') produced an expected product of 711 bp, and the K7group oligos: K7groupF (5'-GAGCGGAAGTTCAAACAAGC-3') and K7groupR (5'-AGCAAGAATCTCGCCATCTG-3') produce an expected product of 707 bp. The primers were designed using Primer3, based on alignments of capsid nucleotide sequences performed using CLUSTALW.

PCR reactions were performed for each primer set, individually, in a total volume of 25 μl containing: template DNA; 4 μM forward and reverse primers; 0.25 mM dNTPs; 3 mM MgCl$_2$; 1× PCR buffer and 0.5 U of BioTaq DNA polymerase. Amplification conditions were: 94°C for 5 min; 30 cycles of 94°C for 45 sec, 48°C for 45 sec and 72°C for 60 sec; with a final extension of 10 min at 72°C. Products were visualised on a 1% Helena Agarose gel prepared in 1x TAE buffer and stained with ethidium bromide. Samples were loaded ith 10μl of GeneRuler 1 kbp DNA ladder and the gel was run at 90 volts for 1 hour in TAE buffer. Gels were imaged using SynGene software.

2.3.4. Screening for phage specific histidine kinase gene in environmental *C. difficile* genomes.

PCR using two primer sets 003AR/004AR and WHKF/WHKR described previously, was performed to screen *C. difficile* isolates to indicate the distribution of the histidine kinase gene, either in a prophage or in the bacterial chromosome. DNA samples were extracted using Chelex

from strains isolated in this project. Set 003AR/004AR are positioned inside the histidine kinase gene on phiCDHM1, and therefore serve as "internal" primers and the second set WHKF/WHKR are positioned to flank the histidine kinase gene and serve as "external" primers, indicating the presence of the gene on the phage genome. Products were visualised on a 1% Helena Agarose gel prepared in 1x TAE buffer stained with ethidium bromide and run at 90 volts for 1 hour in TAE buffer with 10µl of GeneRuler 1 kbp DNA ladder. Gels were imaged using SynGene software.

2.3.5. Comparative genomics using ACT and Mauve.

Genomic comparisons of phiCDHM1, phiCDHM19, phiCDHM13, phiT6 were performed against each phage in addition to the phage sequences publically available: phiC2 (NC_009231.1), phiCD119 (NC_007917.1), phiCD27 (NC_011398.1), phiCD38-2 (NC_015568.1) and phiCD6356 (NC_015262.1) and the prophages encoded in strains CD630 (NC_009089.1) and CD196 (NC_1727596). Prophage sequences were re-arranged to start with the Small Subunit Terminase gene in Artemis. Comparison files were generated using Double Act v2 accessed at http://www.hpa-bioinfotools.org.uk/pise/double_act.html# visualised using the ACT: Artemis Comparison Tool (Carver *et al* 2005).

Genome alignment of 12 phages and 14 prophages was performed using the progressiveMauve algorithm. These include the phages used in the ACT comparisons, in addition to the phages phiMMP02 (NC_019421.1), phiMMP04 (NC_019422.1) and the prophages from epidemic and non-epidemic, recent and historical strains: two prophages of R012 strain CD630 (NC_009089.1), prophages from each R017 strain CF5 (NC_017173.1) and M68 (FN668375.1) and the prophage phiCD20 in R027 strains, in this paper designated with 'phi' preceeding the originating strain: CD196 (NC_013316.1), R20291 (NC_013316.1), BI-L (NC_017179.1), QCD66c26 (NZ_ABKK02000022), QCD76w55 (NZ_ABKK02000022), QCD37x79 (NZ_ABKK02000022), QCD97b34 (NZ_ABKK02000022),

CIP107932 (NZ_ABKK02000022), 207855 (NC_017178.1). Whole genome alignment was performed using Mauve v.2, using the progressiveMauve algorithm (Darling *et al* 2010).

2.3.6. CRISPR protospacer identification in *C. difficile* phage genomes.

The 26 phage and prophage sequences used previously in the Mauve alignment were searched against the CRISPRdb (Grissa *et al* 2007), accessed at http://crispr.u-psud.fr/crispr/BLAST/CRISPRsBlast.php, to identify putative protospacers with a working E-value cut off of <0.005. CRISPR arrays on phage genomes were identified by searching sequences with the CRISPRfinder (Grissa *et al* 2007) accessed at http://crispr.u-psud.fr/Server/.

A measure of relative phage relatedness was calculated using the blastn results generated from pairwise comparison of genomes using Double Act v2 accessed at http://www.hpa-bioinfotools.org.uk/pise/double_act.html#. The relative genetic relatedness value was calculated from what proportion of the genome (number of bases) were in regions of homology that had <80% identity and were at least 20 bp in length. All phages were included in this analysis with the R027 prophages represented by phiR20291 and phiCD196. Data analysis and correlation tests were performed in MicrosoftOffice Excel using student's t-test and correlation.

3. Isolation of *C. difficile* from non-clinical sources

3.1. Introduction

The prevalence and diversity of clinical strains of *C. difficile* in healthcare settings changes between countries and over time (Bouza 2012). Unsurprisingly, as a nosocomial infection, the bacterium can be isolated from the hospital environment and its persistence on surfaces is of major concern as permits the bacterium's transmission between patients (Fekety *et al* 1981, Kaatz *et al* 1988, Samore *et al* 1996). Also, it can be isolated from asymptomatic carriers, with a typically higher carriage rate detected in infants than adults (Adlerberth and Wold 2009). Colonisation is the result of contact with contaminated environments or persons and carriage is thought to be transient (Kato *et al* 1998, Chang *et al* 2012). Factors impacting *C. difficile* growth and removal from the gut are not well understood, but are linked to microbial composition of the GI tract (Rousseau *et al* 2011b), and an important component in any microbial community is bacteriophages. The impact phages have on *C. difficile* colonisation in asymptomatic carriers has not been explored.

C. difficile also causes disease in wild and domesticated animals, and, similarly, can be carried asymptomatically, with infant animals more frequently colonised (e.g. McBee 1960, Miller *et al* 2010, Norman *et al* 2009, Al Saif and Brazier 1996, Pasquale *et al* 2011). Few studies have examined the bacterium's presence in the environment, but from these it appears to be ubiquitous (Al Saif and Brazier 1996, Zidaric *et al* 2010, Higazi *et al* 2011, Pasquale *et al* 2011).However, its ecology in the environment and potential for strain evolution within these reservoirs is not known.

Reservoirs of pathogenic bacteria may be 'testing' grounds for strain evolution and populations of environmental bacteria are widely thought to serve as genetic reservoirs or "resistomes" containing antibiotic resistance genes that are transferrable to pathogenic bacteria (e.g. Davies 1994, D'costa *et al* 2006, Wright 2007, Zhang *et al* 2008, Cantón 2009, Wright 2010). In this study, the antibiotic MICs of isolates from environmental sources was performed to test

whether these populations contain resistance genes. The antibiotics selected for the resistance assays in this study were chosen because fluoroquinolones and clindamycin have been identified as risks for developing CDI (for example McCusker et al 2003; Loo et al 2005), and metronidazole and vancomycin as they are the drugs of choice in treating CDIs (Teasely et al 1983) despite disease relapse and *in vivo* treatment failure (Kuijper and Wilcox 2008).

Another key physiological trait of *C. difficile* is production of petrichious flagella (multiple flagella in multiple positions over the cell), although the role of motility in pathogenicity is unclear (Dingle *et al* 2011, Tasteyere *et al* 2001). However, flagella have been suggested to aid adaptation of the bacterium to the environment (Borriello 1998). Therefore, in this project, the motility of environmental strains has been investigated, as this also has not been examined previously.

The genetic basis for strain variation and genome evolution can be impacted by phage infection. Strains of *C. difficile* are known to be highly lysogenic and prophage carriage across *C. difficile* strains is diverse (Sell *et al* 1983, Schallehn 1985, Nagy and Foldes 1991, Mayer *et al* 2008, Goh *et al* 2005a, Fortier and Moineau 2007, Shan *et al* 2012, Nale *et al* 2012). High proportions of inducible lysogens have been reported previously (Nagy and Foldes 1991, Fortier and Moineau 2007, Shan *et al* 2012, Nale *et al* 2012). Whole genome sequencing of *C. difficile* strains has also shown that acquisition of prophages has impacted on strain evolution (Stabler *et al* 2009, He *et al* 2010). In other species, high rates of lysogeny have been demonstrated in soil bacterial communities and between 30% - 60% of cultured cells are lysogens (Williamson 2011). In the aquatic environment lysogeny is considered to be an important driving force on bacterial population evolution in marine systems (reviewed by Wommack and Colwell 2000), as well as bacterial population composition (Maurice *et al* 2011). Prophages are likely to shape environmental populations of *C. difficile* but their prevalence, activity and diversity is currently unknown.

In this project stools samples from asymptomatic infants and animals, as well as samples from soils, water and estuarine sediments were obtained and used to isolate *C. difficile*. Three of the

infants were sampled over monthly intervals. Also, repeated sampling was performed in an estuarine system in Hampshire, which included sites that include those polluted with sewage and those considered to be less polluted. These were sampled over two years, and further sampling performed in a third year of some sites.

The resulting collection of *C. difficile* isolates was assessed by ribotyping analysis, motility assays and determination of their antibiotic MICs to examine their genotypic and phenotypic diversity. Also, their prophage carriage was investigated using TEM on induced and untreated culture lysates to assess the extent and diversity of lysogeny within these strains.

3.2. Results

3.2.1. Isolation of *C. difficile* from asymptomatic human infant and animal samples.

C. difficile was isolated from 8/90 stool samples from asymptomatic human infants and various species of asymptomatic animals. Of these, infants and horses were the most intensively sampled cohorts, with >30 samples for each. The recovery rate from infants was 19% (7/36) and from horses was 3% (1/31). The positive horse stool sample was obtained from Hampshire and the negative samples were from a stable yard in Leicester. No *C. difficile* was isolated from pigeon (*n*=6), cow (*n*=6), sheep (*n*=3), cat (*n*=2), rabbit (*n*=2), fox (*n*=1), chicken (*n*=1), deer (*n*=1) or rat (*n*=1) samples.

The enrichment step appeared to be necessary for *C. difficile* recovery, as in a trial of 10 samples, the HPA method was followed but no *C. difficile* was isolated. Using enrichment culture, *C. difficile* was isolated from 8 of the the same samples. Optimisation of this enrichment protocol showed that a shorter time or use of pasteurisation, replacing the alcohol shock step, resulted in an increase of contaminating organisms. Partial 16S rRNA gene sequences were obtained from these

contaminants. They included both spore forming and non-spore forming Gram positive species: *Enterococcus faecalis* (E value = 0.0, identity 100%), *Robinsoniella peoriensis* (E value = 0.0, identity 99%), *Lactobacillus casei* (E value = 0.0, identity 99%), *Clostridium perfringens* (E value = 0.0, identity 100%), *Lactobacillus fermentum* (E value = 0.0, identity 99%) and *Lactobacillus mucosae* (E value = 0.0, identity 99%).

3.2.2. PCR ribotype analysis of human and animal *C. difficile* isolates.

The eight infant and animal isolates belong to seven ribotypes (Table 1). Two isolates are R021, and single isolates are R003, R014, R085, R010, R027 and R076. The toxin gene carriage of these isolates was determined. Two isolates (belonging to R085 and R010) are atoxigenic, as the *tcdA*, *tcdB* and binary toxin genes were not detected. Four isolates (belonging to R003, R021, R014 and R076) encode the *tcdA* and *tcdB* genes, but neither binary toxin gene. Two isolates (belonging to R021 and R027) encode all the toxin genes. The results demonstrate that the strains carried by the asymptomatic infants and horse include *C. difficile* isolates belonging to different ribotypes with different genetic toxin backgrounds.

Table 1. Ribotype and toxin gene carriage of asymptomatic human and animal *C. difficile* isolates.
Toxin genes: A= *tcdA*; B= *tcdB*; CDT= *cdtA* and *cdtB*

Source	Isolate	Ribotype	Toxin genes
Infant	SS1	003	A+B+CDT-
Infant	B1	021	A+B+CDT-
Infant	B2	021	A+B+CDT+
Infant	OA	014	A+B+CDT-
Infant	HB	027	A+B+CDT+
Infant	R8	085	A-B-CDT-
Infant	R10	010	A-B-CDT-
Horse	T6	076	A+B+CDT-

3.2.3. Isolation and ribotyping of *C. difficile* isolates from Hampshire samples from 2009 and 2010.

A preliminary sampling of an estuarine environment was performed, with a single sediment sample (2 cm deep, Thorney Island, Hampshire) taken during the sampling of animal stools in the same locality. *C. difficile* was isolated from this single sample, isolate RBI. It was decided to expand the sampling of estuarine sediments, due to the low frequency of isolation in other environmental samples (rivers and soils). The first set of sampling was performed at 19 sites around the Langstone Harbour estuarine system, Hampshire, U.K., in October 2009. This is an area designated by Natural England (13 Unit ID 1016965) within the Chichester Harbour as a Site of Special Scientific Interest (SSSIs). Sites include those considered polluted with human sewage contamination, as well as sites considered clean based on the presence of certain species of macroflora sensitive to pollution (Julian Clokie, pers. comm.). Multiple samples were obtained from some sites to include sediments from different depths, 1 cm ad 5 cm to include anoxic sediment, which resulted in 21 samples obtained (Figure 1). The sites were resampled the following year and an additional 3 samples obtained of different sediment depths.

C. difficile was isolated from 57% (12/21) and 60% (15/25) of samples in 2009 and 2010 respectively (Figure 1). Of the 19 sites, 11 had *C. difficile* isolated from in both 2009 and 2010 irrespective of sample type. However, the isolation of *C. difficile* varied by sites between location, with 8 sites having the same result in both years but 8 sites with the recovery of *C. difficile* changing between years. Also, *C. difficile* was isolated from both clean and polluted sites. For example, Thorney Island (site 9) which was clean and Sewage Pipe (site 5) which was polluted. *C. difficile* was isolated from all but two sites in either year; Langstone Bridge (site 2) and Kench boathomes (site 17). Also, sediment depth did not appear to predict *C. difficile* isolation where examined, as the bacterium was isolated from both surface and anoxic sediment samples, although recovery was not consistent for either depth between sites. However, the type of sample did appear to matter, as no *C. difficile* was isolated from the 3 water samples in either year.
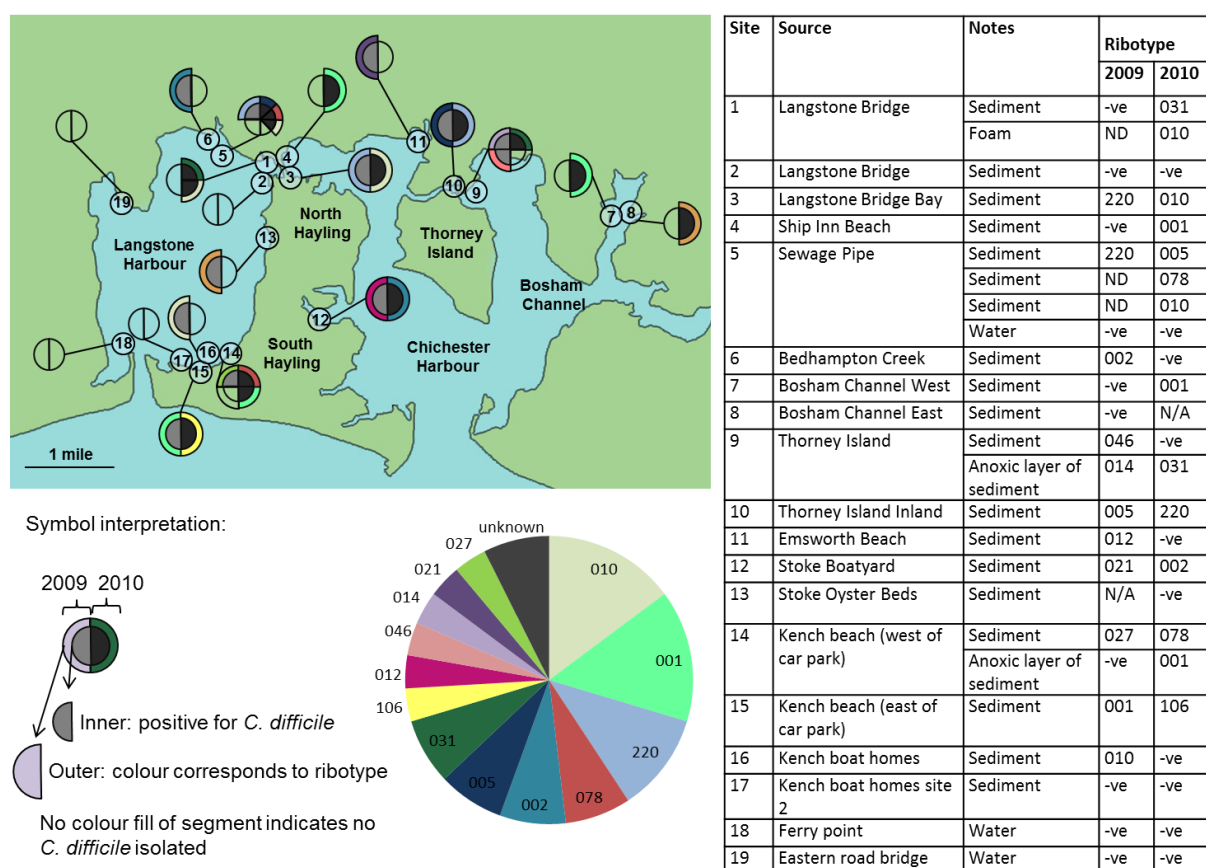
Figure 1. Presence and ribotype diversity of *C. difficile* across the Langstone and Chichester Harbours. Left: map of sampling sites around the estuarine system. Sites are numbered to correlate with the table on the right. Symbols indicate the isolation of *C. difficile*: inner circle indicates year 2009 (left side) and 2010 (right side), colour indicates *C. difficile* isolation and no fill indicates no isolation; outer circle indicates ribotype according to colour which corresponds to pie-chart below the map. Segmented circles indicate multiple samples from a single site, as detailed in the table on right. Right: table of sampling sites, sample type and ribotype. Numbers of sites correspond to map on left and are referred to within text. Abbreviation ND= not done and NA= not available to indicate unassigned ribotype.

Ribotyping of the Hampshire isolates showed the sediment population to be genetically diverse, as 25 isolates belong to 13 ribotypes and 2 isolates could not be assigned to a known ribotype. The ribotype, R010, is prevalent with 4 representative isolates which is interesting as this ribotype is associated with environmental strains but also a prevalence of R001 also with 4 isolates that is more associated with clinical strains. The other ribotypes include the highly clinically relevant R027 and R078. There is no clear correlation between year and ribotype; 5 ribotypes were isolated in 2009 and 2010; 5 ribotypes in 2009 only and 3 ribotypes in 2010 only. *C. difficile* isolates recovered in both years from the same site belong to different ribotypes in 6/7 cases. In the exception, Sewage

Pipe (site 5) the isolation of the same ribotype may be due to the fact that more samples were taken from this site. Notably, two ribotypes that are considered to be clinically important were isolated (R027 and R078) at the same site between years (Kench beach: site 14). Despite this, there was no definite correlation between site and ribotype. For example, *C. difficile* strains were isolated from Thorney Island (site 9), which is considered clean, that belong to clinically known ribotypes (R046, R014, R031, R005, and R220). Also, the environmentally associated R010 was isolated from a sample taken from Sewage Pipe (site 5).

3.2.4. Estuarine sediments contain mixed populations of *C. difficile*.

The results of the first and second sampling of estuarine sediments showed there was a diverse and mixed population of *C. difficile* within this ecosystem. To further investigate the population diversity at a smaller spatial scale, a third set of estuarine samples was obtained in June 2011, which included four sites from the Langstone Harbour area, as well as sites the harbour area surrounding Pagham, Hampshire, U.K. The Pagham sites were sampled as they were considered to be less polluted by sewage (Julian Clokie, pers. comm.). This set also included samples from different depths of sediment, 1 cm and 5 cm.

A relatively high recovery rate was found 92% (Figure 2), which could be due to the fact that 8 sites were sampled repeatedly; resulting in 19 samples obtained from these sites, or that there is possible seasonality in *C. difficile* presence in estuarine sediments. The results show that repeated sampling can result in varying isolation of *C. difficile* and that there is no correlation between surface or subsurface sampling and *C. difficile* isolation detected.

PCR ribotyping analysis of these strains showed that isolates belonging to 13 ribotypes were isolated, with 8 ribotypes also detected in 2009 and 2010 (R002, R220, R002, R010, R001, R078, R021 and R005). Five ribotypes, R100, R015, R175, R026 and R137 had not been previously isolated from either the 2009 or 2010 samples. There is also prevalence of ribotype, R220 in these samples.

The ribotyping of isolates from different samples obtained at the same sites showed that there was a mixed population of *C. difficile* present within the surface and subsurface sediments.

Sample Site		Notes	Ribotype
Pagham	1	Surface	021
	1	Sub-surface	100
	1	Surface	026
	1	Surface	137
	1	Surface	005
	2	Surface	014
	2	Sub-surface	-ve
	3	Surface	-ve
	3	Sub-surface	220
Bedhampton	1	Surface	220
	1	Sub-surface	001
	2	Surface	010
	2	Sub-surface	137
	3	Surface	031
	3	Sub-surface	-ve
	4	Surface	220
	4	Sub-surface	014
	5	Surface	002
	5	Sub-surface	078
Farlington	1	Surface	-ve
	2	Surface	220
Kench	1	Wormcast	002
	2	Surface	015
	3	Surface	002
	3	Sub-surface	010

Figure 2. Isolation and ribotype of *C. difficile* isolates from sampling sites at Pagham and Kench. Left: table with site, location, sample types and ribotype of *C. difficile* isolated. Right: pie-chart showing prevalence of *C. difficile* isolate ribotypes**.**

Table 2. Isolates and their ribotype from subset of Kench sample sites.

Site	Sample	Colonies	Ribotype
1	A	1	003
	B	-	-ve
	c	1	002
	D	10	010 (*n*=10)
2	A	-	-ve
	B	1	014
	C	1	009
	D	10	012 (*n*=7) 014 (*n*=3)

A sub-set of 4 samples from Kench was analysed to further investigate the genotype diversity at one location and in one sample (Table 2). For two of the samples at each site (3d and 4d),

ten colonies were purified from the CCEY plate and ribotyped. The isolation of *C. difficile* was variable between samples at the same site, with no *C. difficile* isolated from one sample in both sites. The remaining samples had *C. difficile* belonging to different ribotypes isolated. In the two samples where ten *C. difficile* colonies were picked, ribotype analysis showed one had *C. difficile* isolates belonging to the same ribotype, R010 and the other had seven isolates belonging to R012 and three isolates to R014. This is interesting as it shows there are genetically distinct strains of *C. ifficile* in close proximity in the estuarine sediments but also indicates that there is a dominant strain type within these micro-populations.

3.2.5. Environmental *C. difficile* isolates can switch motile phenotype.

C. difficile produces flagellum and motility has been linked to virulence *in vivo*, but cell motility may be advantageous in other environmental conditions. In order to further characterise the environmental isolates, their motile growth was assessed.

The first motility assays performed, following the method in Stabler *et al* (2009), using BHI liquid assays supplemented with 0.05% agar, resulted in growth that was difficult to interpret as the extent of motile growth for the reference strain CD630 had exceeded its normal level (Stabler *et al* 2009). A second motility assay was performed by adapting protocols from Be'er *et al* (2009), by stab inoculation of cultures into 3 ml BHI 0.4% agar 51 mm petri dishes. Overnight incubation produced growth which could be assigned as non-motile, intermediate or highly motile. Examples are shown in Figure 3.

The assays showed that motile growth was detected for 87.8% (29/33) of isolates tested. However, although assays were performed in triplicate from the same culture which resulted in a consistent phenotype, the assay repeats using fresh culture showed that the motile phenotype changed in 45% of cases. In the remaining cases, 16/33 were motile in both, and 5/33 were non-motile in either.

There was no clear correlation between ribotype and motility observed. It appears that motility is variable within specific ribotype groups, although there were too few isolates available within groups for statistical analysis. Similarly, there did not appear to be a relationship between isolate source and motility. For example, the R010 group contains isolates which were motile in both repeats (e.g. R10) and isolates which were non-motile in both repeats (e.g. K16, H5d and H1b). The motile isolate is from an infant and the non-motile are from estuarine sediments. However, other ribotypes from sediments did show motility. The only ribotype with a consistent motility observed was R021; with three isolates, two from infants and one from sediment, all possessing the same level of motility across repeats.
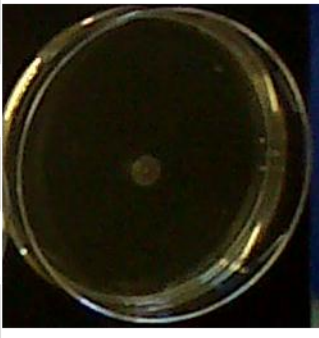
Ribotype	Isolate	Source	Motility
010	H1b	Sediment	-/-
010	H3	Sediment	+/-
010	H5d	Sediment	-/-
010	K16	Sediment	-/-
010	R10	Infant	+/+
001	H8	Sediment	++/-
001	H4	Sediment	+/+
001	H18	Sediment	++/+
001	K25	Sediment	+/+
220	H12	Sediment	+/+
220	K3	Sediment	++/-
220	K12	Sediment	+/+
002	H15	Sediment	++/++
002	K14	Sediment	++/-
031	H11	Sediment	++/++
031	H1a	Sediment	+/-
005	H5b	Sediment	+/-
005	K18	Sediment	++/+
078	H17	Sediment	++/++
078	H5c	Sediment	-/-
046	K4	Sediment	-/-
014	K6	Sediment	++/-
014	OA	Infant	N/A
021	B1	Infant	+/+
021	B2	Infant	+/+
021	K9	Sediment	+/+
027	K15	Sediment	+/++
027	HB	Infant	+/-
012	K10	Sediment	+/-
106	H19	Sediment	++/-
003	SS1	Sediment	-/-
085	R8	Infant	+/+
076	T6	Equine	-/+
Unknown	K20	Sediment	+/-
Unknown	RBI	Sediment	+/-
Unknown	H16b	Sediment	-/-

Motile growth

Non-motile growth

Figure 3. Variable motility observed for *C. difficile* isolates during repeated assays. Left: isolates are grouped according to ribotype with origin. The results of the two assays are separated by a "/". A negative symbol (-) meant no motility or low motility was observed, a positive (+) indicates that motility was observed and a double symbol ++ indicates a high level of motility. Right: top photograph shows motile growth; below shows non-motile growth. Hazy growth indicates *C. difficile* motile dispersion through the BHI 0.4% agar media.

3.2.6. Determination of antibiotic MICs for environmental *C. difficile* isolates.

Another key phenotypic trait is resistance to specific antibiotics, including those used to treat CDI, as *C. difficile* is multi-drug resistant (Sebaihia et al 2006). In other species, environmental bacteria are thought to be resivoirs of antibiotic resistant genes that can be exhcanged throughout the population. In order to determine whether the environmental *C. difficile* isolates harbour such genes, the MICs to four antibiotics, vancomycin (VZ), metronidazole (MZ), clindamycin (CM) and ciprofloxacin (CI), were determined for 27 isolates (Figure 4 and Table 3).

Figure 4. Antibiotic MICs of environmental *C. difficile* isolates show ranges in sensitivity. The MIC values of 27 environmental *C. difficile* isolates for vancomycin (VZ), metronidazole (MZ), clindamycin (CM and ciprofloxacin (CI) were plotted in an Individual Value plot and error bars calculated using MiniTab. MICs were obtained using Etest strips.

All isolates were sensitive to VZ and MZ, although one isolate H17 (R078) has an MIC of 4 μg ml^{-1} for MZ, indicating it has borderline resistance to this antibiotic. Five isolates are resistant and one borderline resistant to CM. Twelve isolates are resistant and seven borderline resistant to CI. There was a difference between the MIC values for each antibiotic (Figure 4) with the greatest range for CI.

Table 3. ETest results for environmental isolates with corresponding clinical ribotypes.

Ribotype	Isolate	MIC in μg ml^{-1}			
		Vancomycin	Metronidazole	Clindamycin	Ciprofloxacin
010	H1b	0.19	0.094	32	4
	H3	1	0.064	0.016	3
	H5d	0.38	0.19	0.016	4
	K16	1	0.19	0.016	6
001	H8	0.19	0.094	0.016	32
	H4	0.38	0.016	0.016	2
	H18	1.5	0.25	2	0.016
	K25	0.38	0.25	0.016	32
220	H12	0.75	0.64	0.6	12
	K3	0.25	0.032	0.064	3
	K12	1	0.125	4	4
002	H15	0.38	0.23	0.125	0.5
	K14	0.016	0.032	0.032	8
031	H11	0.5	0.25	0.016	32
	H1a	0.38	0.25	0.016	8
005	H5b	1	1	32	32
	K18	1.5	0.25	0.016	2
078	H17	4	0.125	0.5	6
	H5c	0.5	0.94	0.5	32
046	K4	0.25	0.25	8	32
014	K6	0.5	0.094	0.032	8
021	K9	0.19	0.19	0.032	4
027	K15	0.38	0.25	0.094	0.75
012	K10	0.19	0.064	0.047	32
106	H19	0.19	0.19	0.032	32
Unknown	K20	0.38	0.19	0.032	4
Unknown	H16b	0.5	0.023	32	4
076	T6	0.125	1.5	0.38	32

There is no clear correlation between ribotype and ribotype, except perhaps in specific groups: for example, isolates belonging to R001 include those resistant and sensitive to CI whereas all the tested R010 isolates have borderline resistance to this antibiotic. Also, there is no apparent link between isolate origin and MIC, as isolates from both clean (e.g. Thorney Island site 9) and polluted sites (e.g. Sewage Pipe site 5) were resistant to CI.

The results provide evidence that there are both antibiotic resistant and sensitive strains are present in the environment, and related strains have different sensitivities, leading to the possibility of antibiotic resistance genes being transferred via horizontal gene transfer within this population.

3.2.7. Temperate phage carriage in non-clinical strains of *C. difficile.*

In order to determine the prophage carriage within the *C. difficile* isolates obtained in this project, lysates from the culture of 35 human, animal and environmental isolates were examined for the presence of temperate phages. They include lysates from cultures subjected to stressors for prophage induction to mediate phage release, using either antibiotic mitomycin c (MC) or norfloxacin (NFX). Also lysates from untreated cultures were examined for spontaneously released phages. Samples were concentrated using PEG before negative staining and examined using TEM.

Particles were observed for 85% (30/35) of isolates, including phage particles and phage tail-like particles (PTLPs) which are predicted to be bacteriocins (Figure 5 and Table 4). Phage particles were observed for 68.5% (24/35) of isolates and could be classified into morphological groups (Figure 5). They are: medium tailed myoviruses (MMs) with tail lengths of 150 nm long and a capsid diameter of 90 nm; long tailed myoviruses (LTMs) with tail lengths of 200 nm and a capsid diameter of 90 nm; small headed myoviruses (SMVs) with tail length 100 nm and a capsid diameter of 30 nm and siphoviruses with capsids of 50-60 nm and tail lengths of 180- 300 nm.
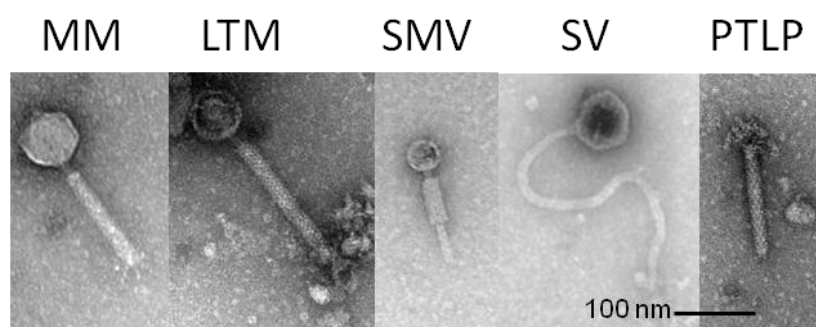
Figure 5. Morphological groups of temperate phages and phage tail-like particles. TEM images of PEG concentrated lysates from either MC or NFX induced cultures or those with no inducing agent which had released phage in a spontaneous manner. Phage particles observed could be classified into different morphological groups based on capsid and tail sizes: medium myovirus (MM), long tailed myovirus (LTM) and small myovirus (SMV). Also siphoviruses with different capsid sizes and phage tail-like (PTLPS) particles were present in several lysates. Scale bars represent 100 nm.

Table 4. Temperate phage carriage within non-clinically associated *C. difficile* isolates.
Abbreviations: N/A= not available; MMs= medium myovirus; LTMs = long tailed myoviruses; SMVs = small myoviruses; SVs = siphoviruses and PTLPs = phage tail like particles.

Ribotype	Isolate	Source	MMs	LTMs	SMVs	SVs	PTLPs
010	H1b	Sediment	-	√	√	-	√
010	H3	Sediment	N/A				
010	H5d	Sediment	-	-	√	√	-
010	K16	Sediment	√	√	-	-	√
010	R10	Infant	√	-	-	-	-
001	H8	Sediment	-	-	-	√	-
001	H4	Sediment	N/A				
001	H18	Sediment	-	-	-	-	√
001	K25	Sediment	√	-	-		-
220	H12	Sediment	√	√	-	√	-
220	K3	Sediment	√	-	-	√	-
220	K12	Sediment	√	-	-		√
002	H15	Sediment	-	-	-	√	√
002	K14	Sediment	-	-	-	√	-
031	H11	Sediment	-	-	-	-	√
031	H1a	Sediment	-	-	-	-	√
005	H5b	Sediment	-	-	-	-	√
005	K18	Sediment	√	-	-	-	√
078	H17	Sediment	N/A				
078	H5c	Sediment	-	-	-	-	√
046	K4	Sediment	√	-	-	-	-
014	K6	Sediment	-	-	-	√	√
014	OA	Infant	-	√	-	-	-
021	K9	Sediment	-	-	-	-	√
021	B1	Infant	√	-	-	√	-
021	B2	Infant	√	-	-	-	-
027	K15	Sediment	√	-	-	-	√
027	HB	Infant	√	-	-	-	√
012	K10	Sediment	√	-	-	-	√
106	H19	Sediment	N/A				
076	T6	Equine	√	-	-	-	-
085	R8	Infant	-	-	-	-	√
003	SS1	Infant	-	-	-	-	√
Unknown	H16	Sediment	N/A				
Unknown	K20	Sediment	-	-	√	√	√
Unknown	H9	Sediment	N/A				

The isolates have different temperate phage carriage (Table 4). Of the 36 isolates, 15 had a single type of phage particle observed in their lysates; 7 are MMs; 2 are LTMs and 4 are SVs. Several isolates were found to produce phage particles belonging to different morphologies, with 2 different phages observed for 6 isolates and 2 phage types for 3 isolates. This included the only isolate to produce SMVs particles. There is therefore considerable variation in the carriage and co-carriage of temperate phages in non-clinical *C. difficile* isolates.

There was no clear correlation between ribotype and temperate phage carriage. The 36 isolates examined belong to 16 ribotypes and 3 unassigned ribotype groups. Of these, 9 ribotypes are represented by 2 or more isolates. In these groups it can be seen that temperate phage carriage can vary, for example R010 has 5 representative isolates, and between them carry all 4 morphologies of phage particles. However, there may be a correlation in specific ribotypes- for example both isolates of R002 had only SVs observed and both isolates of R031 had no phage particles observed; a greater number of isolates examined would confirm this.

3.2.8. Optical density as an indicator of phage release in *C. difficile* cultures.

To determine if culture density could be used as an indicator of phage release, the OD_{550nm} of cultures for 33 animal, infant, Hampshire 2009 and Slovenian *C. difficile* isolates were measured before and after treatment with MC and NFX and compared to no treatment controls.

The results show that the OD_{550nm} decreased in all antibiotic induced cultures and 26 of the untreated cultures when compared to their starting OD_{550nm} (Figure 6). In comparison, decreased OD_{550nm} values were observed more often in inductions using NFX (55%) than MC (43%). To investigate the size of the reduction in OD_{550nm} between the different conditions, the percentage difference of the final OD_{550nm} to its starting value was calculated.
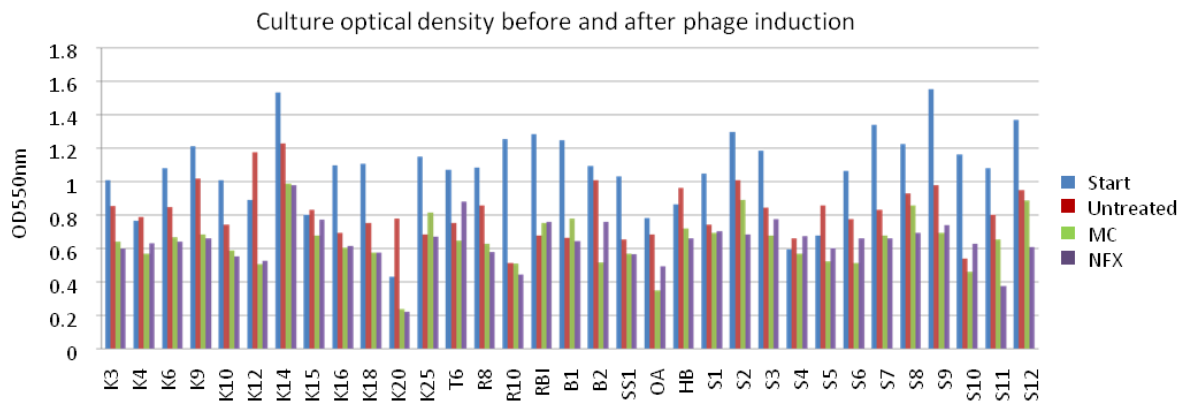
Figure 6. Culture optical density before and after phage induction. OD_{550} nm values were measured for 33 isolates following overnight growth in BHI. Cultures were split into three volumes and treated with either mitomycin C (MC) or norfloxacin (NFX), or untreated. OD_{550} nm was monitored following 24 hours culture.

On average both MC and NFX treated cultures decreased in OD_{550nm} by 39% and 38%, respectively, and control cultures decreased by an average of 30% (not including cultures which were higher than the starting OD_{550nm}). In the case of six isolates, at least one of the antibiotic treated cultures had ODs higher than that of their corresponding control. These results indicate that antibiotic treatment does decrease culture growth, but that this is also observed in some untreated cultures and OD_{550nm} may not, therefore, be a good indicator of phage release for *C. difficile*.

3.2.9. Differential release of temperate phages following spontaneous release and antibiotic induction.

A sub-set of 21 isolates was examined using TEM to observe phages released following induction by MC or NFX (Table 5). Also, the control cultures were investigated for phage particles to determine the extent of spontaneous release after prolonged culture, because of the observation that OD values decreased for the controls as well as the induced cultures.

For most of the isolates (14/20) treatment with either NFX or MC induced the release of similar phage particles. Of these, 20 lysates, representing 10 isolates, produced phage particles, and 6, representing 3 isolates, produced only phage tail-like particles. The remaining 12 cultures (of 6 isolates) were where phage was observed in only a single type of antibiotic treatment or there was an additional type of phage observed; for 4 isolates it was after induction with MC and for 2, following induction with NFX.

Table 5. Differential release of temperate phage following antibiotic induction and spontaneous release. Abbreviations: MMs= medium myoviruses; LTM= long tailed myoviruses; SMVs = small myoviruses; SVs = siphoviruses, PTLPs = phage tail like particles. Symbol (-) indicates not done.

Ribotype	Isolate	MMs			LTMs			SMVs			SVs			PTLPs		
		MC	NFX	C	MC	NFX	C	MC	NFX	C	MC	NFX	C	MC	NFX	C
031	H11			-			-			-			-	√	√	-
010	R10	√	√	-			-			-			-	√		-
010	K16	√	√	√	√	√	√							√	√	√
001	K25	√	√	√										√	√	√
001	K4	√	√											√	√	√
001	K6										√		√	√	√	√
220	K12	√	√											√	√	
220	K3	√	-	√			-			-	√	-	√			-
002	K14										√		√	√	√	√
002	H15			-			-			-		√	-	√	√	-
005	K18	√	√	√										√	√	√
005	H5b			-			-			-			-	√	√	-
021	K9													√	√	
021	B1	-	√		-			-			-			-	√	
021	B2		√	-			-			-	√	√	-		√	-
027	K15	√	√	√										√	√	√
027	HB	√	√	-			-			-			-	√	√	-
012	K10	√	√	√										√	√	√
003	SS1			-			-			-	√	√	-	√	√	-
U/K	K20			-			-	√	√	-	√	√	-	√	√	-

The result of antibiotic treatment compared to no treatment could be examined for 13 isolates: for 9 isolate both treated and untreated cultures produced the same phage-types. In the remaining 4, the results differed and phage particles were only observed in the antibiotic treated lysates. There were no cases when additional phages were observed from untreated culture lysates.

There did not appear to be any relationship between ribotype and phage release, either induced or spontaneous, although the sample sets for different ribotypes were small.

Using TEM analysis to determine phage carriage is problematic when examining samples with a low phage titre. The phages were PEG concentrated before TEM, but frequently samples resulted in poor quality grids with no phages observed (the ND samples) and where phage were observed it could be seen that they were present at low titre. While phages which were observed are unarguably present, a negative sample does not necessarily indicate that no phages were produced from that culture.

These results do show that phages are released spontaneously from cultures of *C. difficile,* grown under laboratory conditions following a prolonged culture of 48 hours and that induction by either antibiotic frequently releases temperate phages.

3.3. Discussion

3.3.1. Isolation rates of *C. difficile* from non-clinical sources.

In this project *C. difficile* was isolated from stools of asymptomatic infants and one horse, and environmental sediment samples. There was a low rate of isolation from the assorted asymptomatic animal faeces, 1.8% (1/54), a higher rate from asymptomatic infants, 22% (8/36), and a much higher rate from estuarine sediments, 71% (52/73). This study is the first to document *C. difficile* recovery from estuarine sediments.

Rates of isolation compared to previous studies in the literature are difficult to compare as they use different isolation methods. One example is the use of enrichment methods by Simango and Mwakurudza (2008) and another example is the use of different selective media by Al Saif and Brazier (1990). The use of a *C. difficile* specific 16S rRNA PCR step, performed in this project, was useful to confirm identification of isolates (example of PCR products is Appendix Figure 1). The genotyping of strains from this project shows the isolate collection is diverse with multiple ribotypes detected. It is also difficult to assess how the distribution and prevalence of these strains compares with other studies as genotyping data may be absent (e.g. Simango 2006, Miller *et al* 2010, Higazi *et al* 2011) or different genotyping methods may have been used, for example toxinotyping (Norman *et al* 2011) or in-house nonmenclature for ribotypes (e.g. Zidaric *et al* 2010). As capillary ribotyping becomes more accessible, future studies may be able to assign ribotypes more easily by comparing digital data. Examples of gel based ribotyping and capillary ribotyping are shown in the Appendix (Figures 2 and 3).

Despite this, the isolation rates of *C. difficile* obtained in this project are consistent with those reported in previous studies, despite the different isolation methods used in different research groups. This ambiguity means that direct comparison may not be accurate, but, that general trends are present in the recovery rates of *C. difficile*. Also, rates can be compared where a single study isolates from multiple sources (e.g. Al Saif and Brazier 1996, Miller *et al* 2010) or where sampling has been performed in a longitudinal manner (e.g. Norman *et al* 2009, Higazi *et al* 2011).

3.3.2. Isolation of *C. difficile* strains from asymptomatic infants.

The isolation of *C. difficile* from asymptomatic infants in this work falls within the wide range of carriage rates reported in the literature; from 15% to 90% (Hall and O'Toole 1935, Synder 1937 and 1940, Viscidi and Bartlett 1981, Stark *et al* 1982, Zedd *et al* 1984, Fallani *et al* 2006, Rousseau *et al* 2011a, Stoesser *et al* 2011). The frequency of *C. difficile* carriage in infants is attributed to their

possession of an immature gut microbiome and rapid initial colonisation of the gut by "safe" commensal organisms reduces *C. difficile* numbers (Stark *et al* 1982, Cooperstock *et al* 1982, Penders *et al* 2006, Adlerberth *et al* 2007).

In this project, one infant was followed from birth for 12 months, with *C. difficile* isolated from two samples, two months apart. These isolates were different ribotypes, R085 and R010 respectively, which supports the suggested transient presence of *C. difficile* in the infant gut, through loss and subsequent re-colonisation (e.g. Hall and O'Toole 1935, Bolton *et al* 1985, Kato *et al* 1998, Chang *et al* 2012). Ribotype analysis of infant strains in one Japanese study showed that *C. difficile* carriage is likely to be a product of contamination from the environment as well as contact with other carriers (Matsuki *et al* 2005).

Ribotyping analysis of the *C. difficile* isolates from this study showed there to be a diverse carriage of strains, including detection of clinically relevant ribotypes (for example R027), in addition to those more associated with environmental samples (for example R010). This is consistent with ribotyping analysis in the literature and there is a cross-over between ribotypes found, for example R010 and R014 (Rousseau *et al* 2011a). Also, the toxin gene assays showed that the infants can harbour strains which are positive for all three toxin genes and those which are atoxigenic, which is consistent with the literature (Stoesser et al 2011). This suggests that infants may be one reservoir of this pathogen and one hypothesis is that passage in the immature gut may serve as a proving ground for pathogenic strain evolution.

3.3.3. Isolation of *C. difficile* from asymptomatic animals.

C. difficile was isolated from 2% (1/49) of the asymptomatic animal faecal samples, which, although low, is consistent to most rates in the literature (e.g. Al Saif and Brazier 1996, Simango 2006, Indra *et al* 2009, Hopman *et al* 2011a, Rodriguez-Palacios *et al* 2011, Queen *et al* 2012). These include studies which have sampled a wide variety of farmyard animals in relatively low numbers

e.g. <100 (Simango 2006, Hopman *et al* 2011a) as well as fewer species tested but in larger numbers >100 (Al Saif and Brazier 1996, Simango and Mwakurudza 2008, Rodriguez-Palacios *et al* 2011). For example, *C. difficile* was isolated from 1.8% (17/944) of cattle faecal samples (Rodriguez-Palacios *et al* 2011). Household pets have a slightly higher reported carriage rate of *C. difficile* than livestock (Borriello *et al* 1983, Hawkins *et al* 1986, Al Saif and Brazier 1996, Weese *et al* 2010). Typically carriage rates in adult animals are low, but *C. difficile* has been isolated from the majority of species examined. The few samples obtained for several of the animals tested in this study may be why *C. difficile* was not isolated from these species. For example, no *C. difficile* was isolated from either of the feline samples in this study, but Weese *et al* (2010) isolated *C. difficile* from 21% of 14 cats samples, and Al Saif and Brazier (1996) found 2% of 100 cats were carriers. Also, similar to the situation in humans, infant animals have higher carriage rates than adults (Båverud *et al* 2003, Rodriguez-Palacios *et al* 2006, Norman *et al* 201, Costa *et al* 2011).

The largest cohort of animals sampled in this project was that of asymptomatic horses. All samples from 30 horses in a Leicestershire livery yard tested were negative, but *C. difficile* was isolated from the single horse sample obtained from Thorney Island, Hampshire. The results suggest that in this case external factors were more important than sample size for *C. difficile* carriage. This is consistent also with other studies sampling adult horses, with isolation rates ranging from 0%, despite sampling 122 and 273 horses respectively (Thean *et al* 2011, Båverud *et al* 2003), to 7.59% after sampling 540 horses (Medina-Torres *et al* 2011), as well as infant studies (e.g. Matsuki *et al* 2005).

The single equine *C. difficile* isolate in this study was R076. Where strain analysis has been performed, the diversity of strains is evident as is the change in strain carriage (e.g. Al Saif *et al* 1998, Rodriguez-Palacios *et al* 2006, Indra *et al* 2009, Costa *et al* 2011, Hopman *et al* 2011, Rodriguez-Palacios *et al* 2011). Cross-over of ribotypes between animal species is known, including calves, neonatal pigs, dogs, horses and humans (Keel *et al* 2007, Janezic *et al* 2012) although R076 was not

found in these studies. This strain was sequenced and its genome is currently being annotated in collaboration with Trevor Lawley, at the Wellcome Trust Sanger Institute.

3.3.4. Isolation of *C. difficile* from environmental sources.

C. difficile was not isolated from the four soils samples collected in this study, but has been found in soils across the world (Kim *et al* 1981, Hafiz 1974, Båverud *et al* 2003, del Mar Gamboa *et al* 2005, Simango 2006, Simango and Mwakurudza 2008, Higazi *et al* 2011). Sample size may be a factor in failing to isolate the bacterium, as well as being from a single sample location, as typically isolation rates are low (4% to 11%), but one study has a high proportion of samples with *C. difficile* isolated from 37% (Simango 2006). Its presence in soils may result from contamination from infected people (Kim *et al* 1981), or animals (Båverud *et al* 2003, Simango 2006, Simango and Mwakurudza 2008).

C. difficile was isolated from 71% (52/73) of the Hampshire sediment samples from this work. The isolation contradicts findings of one previous study which did not isolate *C. difficile* from estuarine sediments (Pasquale *et al* 2011). However, isolation rates of *C. difficile* from other aquatic samples reported are relatively higher than those from soils (Al Saif and Brazier 1996, Zidaric *et al* 2010). For example, from the Cardiff area, *C. difficile* was isolated from: 21% of soil samples; 87.5% of river water samples; 46.7% of lake water samples and 44% of Bristol Channel sea water samples (Al Saif and Brazier 1996). Why isolation rates are higher in aquatic sources is not clear but may increase the bacteria's viability or involve niche competition as environmental samples are populated by highly diverse microbial communities (Lepage et al 2012; Gordon et al 2005). These differences may result in different genotypes evolving within environmental systems (termed ecogenomics).

3.3.5. Ecology of the *C. difficile* populations in the Hampshire estuarine environment.

Repeated sampling of the Hampshire estuarine system carried out in the project has provided valuable insights into the potential transmission, persistence and ecology of *C. difficile* in this environment, as well as demonstrating that estuarine sediments are a diverse and dynamic reservoirs of this pathogenic species.

The presence of *C. difficile* within estuarine sediments builds on our current understanding of the potential origins and transmission of *C. difficile* within the marine system. The bacterium has been isolated from the marine water column (Al Saif and Brazier 1996, Pasquale *et al* 2011), as well as from the marine organisms inhabiting it, throughout the food chain including plankton, shellfish, fish and marine mammals (McBee 1960, Pasquale *et al* 2011, Metcalf *et al* 2011, Miller at al 2010). The finding of *C. difficile* in estuarine sediments, therefore has implications for recreation and fishery industries which occur in these ecosystems.

Sources of *C. difficile* in the system may include anthropomorphic inputs. The Langstone Harbour area is known to be contaminated with sewage entering the system from several outflow pipes (Townend 2007, Travis 2012). In a Californian study examining *C. difficile* carriage in sea otters, the researchers found that they were more likely to be positive for *C. difficile* at sites with sewage contaminated sites, indicating that this could be one source of *C. difficile* entering the marine system (Miller *et al* 2010). The Sewage Pipe site sampled in this study was positive for *C. difficile* in both years. Human wastewater has been shown to contain *C. difficile* (Norman *et al* 2011, Romano *et al* 2012), and PCR ribotyping analysis revealed that there is a cross-over of strains isolated throughout these studies. This includes the clinically associated ribotype R078 and the environmentally associated ribotype R010. Also, agricultural run-off may be one way of entry for *C. difficile* into the estuary system. For example, in the U.S.A. *C. difficile* has been detected in two different types of treated biosolids that are routinely used as agricultural fertilizers (Viau and Peccia 2009). Also, in the study by Miller *et al* (2010) the otter faecal samples obtained from the interface between fresh and

marine waters were highest in bacterial numbers, a finding that was suggested by the authors to be caused by both agricultural runoff and effects of urbanisation. The harbour areas sampled in this study have a known problem with agricultural runoff, as input of fertiliser has caused the area's current crisis concerning algal overgrowth within the estuarine harbours (Environmental Agency).

The distribution and diversity of *C. difficile* throughout the Hampshire estuarine did not show any clear correlation between sites or year sampled. It may be that the bacteria enter the system from a source and are dispersed throughout the estuaries by water currents or the action of macrofauna (Withers and Thorp 1978). The movement of water throughout the harbours is extensive and the sites sampled are likely to be subject to several different sediment dispersal mechanisms, such as tidal, fluvial, coastal drift and terrestrial land erosion (Figure 7). The results show that the distribution and prevalence of *C. difficile* in estuarine populations change between years. This suggests that populations are transient, which is consistent with results of studies examining *C. difficile* presence in soils and river water (Higazi *et al* 2011, Zidaric *et al* 2010).

Figure 7. Sampling sites around Langstone and Chichester Harbours. Clockwise: Bosham Shipyard, Stoke Oyster beds, Langstone Bridge andclose-up of muddy sediment with sterile sampling tube. Photographs courtesy of Julian Clokie.

As *C. difficile* is a spore former, the isolation of the bacteria may be from contaminating spores that are not active in the system. The bacterium can undergo a process of super-dormancy where spores remain viable for extended periods of time (Rodriguez-Palacios and LeJeune 2011). Alternatively the bacterium could be active, even if just for part of the year, as the increased rates of isolation from summer samples (54% and 64% versus 92%) may be due to seasonal growth of *C. difficile* growth within the sediments, or its survival in them, as the bacteria is known to grow in the laboratory at $22^{o}C$ (Karlsson *et al* 2003). Seasonality in *C. difficile* isolation rates has been observed in human wastewater samples; there was a significantly higher isolation rate during the spring months than the autumn, winter and summer months (Norman *et al* 2011). If a greater amount of *C. difficile* in human wastewater entered the Hampshire sediments during spring, this could explain why there was a higher recovery of *C. difficile* from the summer. Similarly, seasonality has also been observed in numbers of a related species, *C. perfringens* (e.g. Berthe *et al* 2008). Furthermore, adaption for survival in environmental conditions was examined using the enteric bacterium, *E. coli,* in a study that found improved fitness of a strain isolated from sediments compared to a laboratory strain, when both were grown in sediment cultures (Berthe *et al* 2008). Experiments into the growth dynamics for strains isolated from this study and clinical isolates under different physiochemical conditions present in sediments are now being performed in the laboratory to establish whether *C. difficile* in the estuarine system could be active.

3.3.6. Diverse and novel strains of *C. difficile* are present within estuarine sediments.

Ribotyping analysis has shown that the population of *C. difficile* in estuarine sediments is mixed and diverse with 21 different ribotypes detected. The ribotypes represent both clinically important strains, for example R027 and R078, as well as those rarer in clinical samples and associated with environmental sources e.g. R010 (Brazier *et al* 2001, Barbut *et al* 2007). This is consistent with other environmental and marine organism studies and there is a number of strains that cross-over between studies, such as R014 and R010 (Al Saif *et al* 1998, Zidaric *et al* 2010,

Pasquale *et al* 2010, Metcalf *et al* 2011). The results showed that different ribotypes of *C. difficile* were isolated from the same site, a y**e**ar apart which suggests that the populations' structure could turnover as a result of active growth, or that populations are transient and replaced by different strains.

Interestingly, the environmentally associated ribotype R010 was the most frequently isolated in this study. The reason for a single ribotype to be dominant in the environment may be due to multiple factors. The factors include those which involve bacterial physiology, such as the viability of strains in this environment, as well as the sensitivity of strains to the isolation methods. For example, in this project, several colonies which grew on the CCEY plates and were identified as *C. difficile,* failed to grow on blood plates or in liquid culture. A difference in strain physiology may also explain why different strains of *C. difficile* were recovered from the same samples using different methods of detection and isolation (Zidaric *et al* 2010). Factors which are not linked to strain physiology include the prevalence of strains in inputs to the system, such as sewage or agriculture run off, as well as their dispersal through the system. The observed dominance of specific strains is seen in other systems, including R014 in Slovenian river water and R010 in environmental samples around Cardiff, Wales (Zidaric *et al* 2010, Al Saif *et al* 2008).

The prevalence of these ribotypes within these samples supports their environmental associations, but as they are also isolated from human and animal samples, they are not exclusively environmental genotypes (ecotype). In this study, three isolates could not be assigned to a known ribotype, indicating that they have not previously been isolated from clinical samples. Their genomes are currently undergoing sequencing and the results will confirm their taxonomy, as well as their phylogeny to clinical isolates.

3.3.7. Micro-populations of *C .difficile* within sediments.

When investigating the diversity within single sites and samples, it was found that different ribotypes could be isolated from the same sample. The presence of multiple ribotypes from a single sample is not surprising, as it has been reported before, both from clinical samples, albeit at low frequency (Borriello and Honour 1983, Sharp and Poxton 1985, van den Berg *et al* 2005) and also in an environmental study (Zidaric *et al* 2007). Therefore, the population of *C. difficile* within environmental samples may be even more genetically diverse than is currently known.

Although there is likely to be a mixed population of *C. difficile* genotypes at one location, it may be assumed that the ribotype of isolates from the Hampshire samples is likely to represent the most common strain of *C. difficile.* It is important to consider that other ribotypes could be present. Also, there is the possibility that the bacterium may be present, despite a negative result, as the replicate samples in this study demonstrate.

3.3.8. Motility testing of environmental strains of *C. difficile.*

The results of this study show that there are non-motile and motile strains present within the sediment population. The role flagella have in *C. difficile* virulence is not fully determined. This is the first study to examine motility of environmental strains, but it seems likely that it would influence bacterial survival in different niche environments.

Additionally, the environmental *C. difficile* isolates produced variable degrees of motility, with no clear correlation to ribotype and motility observed. Few strains of *C. difficile* have been designated as non-motile such as M120 which lacks the entire region F3 (Stabler *et al* 2009). Another strain, CD-26, appeared through a motility assay to be non-motile, although flagellin protein could be extracted from the cells surface which suggested that motility in *C. difficile* may be phase-variable (Twine *et al* 2009). Previous studies that have examined motility appear to have performed single

assays, as replicate or triplicate assays are not mentioned in methods (e.g. Tasteyere *et al* 2001, Twine *et al* 2009, Stabler *et al* 2009, Dingle *et al* 2011) and the variability of the motile phenotype in *C. difficile* may be underestimated. In other species, expression of flagellar proteins is phase variable and, depending on the species, may or may not contribute directly to virulence, for example in *Salmonella spp* and *Bordetella pertussis* (van der Woude and Braumler 2004). If flagella production is phase variable, this may explain conflicting results from virulence studies (Tasteyere *et al* 2010) as well as the results obtained in this study.

3.3.9. Determination of antibiotic MICs reveals diverse resistance profiles and potential for strain evolution within *C. difficile* estuarine population.

The estuarine *C. difficile* isolates were all sensitive to vancomycin and metronidazole, although borderline resistance to this antibiotic was detected for one isolate belonging to R078. In contrast, there was considerable variation in MICs for clindamycin and ciprofloxacin, with several isolates resistant to them. The MICs results are consistent with data from clinical studies, as variation of antibiotic sensitivities to the clindamycin and fluoroquinolones is frequently observed in *C. difficile* isolates, both clinical and non-clinical (e.g. Wilcox *et al* 2000, Alonso *et al* 2001, Simango 2006, Barbut *et al* 2007, Simango and Mwakurudza 2008, Rousseau *et al* 2011a, Spigaglia *et al* 2011), but resistance to metronidazole and vancomycin is very rare (Peláez *et al* 2002) and often not detected (e.g. Barbut *et al* 2007, Huang *et al* 2009, Huang *et al* 2010, Taori *et al* 2010, Lin *et al* 2011). There was no clear correlation between ribotype and MIC, which is consistent also with results in other studies (e.g. Aspevall *et al* 2006, Peláez *et al* 2008, Kim *et al* 2012).

It is of concern that *C. difficile* may acquire resistance to certain antibiotics via HGT. For example, strains may acquire vancomycin resistance from the vancomycin-resistant *Enterococci* (VRE), another problematic nosocomial infection. Transfer of resistance genes between VRE and *C. difficile may* be possible as Jasni *et al* (2010) demonstrated *in vitro* conjugative transfer between the

species but in the opposite directiob, for example, Tn*5397* from *C. difficile* CD630 to *E. faecalis* JH2-2. *In vitro* VRE are frequently isolated from stools of CDI patients, presenting a clear opportunity for the two bacteria species to come in to close contact (Raffety *et al* 1997). Similarly, resistance to clindamycin is conferred by presence of the *ermB* gene (Solomon *et al* 2011), which can be acquired via HGT as is encoded on a mobilisable non-conjugative element Tn5398 (Farrow *et al* 2001).

In addition to conjugation and transformation, HGT can be performed via transduction and infection by bacteriophages. Phage populations encoding resistance genes have been found in environmental sources, including livestock faecal slurries (Colomer-Lluch *et al* 2011a), sewage and water samples (Muniesa *et al* 2004, Colomer-Lluch *et al* 2011b). While it is unknown whether *C. difficile* phages could transfer antibiotic resistance genes, the results from this study show that related strains are present within the same reservoir with different genetic backgrounds that could allow antibiotic resistance to be transferred throughout the population.

3.3.10. Temperate phage carriage in non-clinical *C. difficile* isolates.

The collection of *C. difficile* isolates examined in this study contains a high proportion of lysogens. The proportion of inducible prophage carriages is slightly higher than reported in studies examining clinical, infant and animal strains: 76% (Goh *et al* 2005); 54% (Fortier and Moineau, 2007); 64.3% (Nagy and Foldes 1991) and 61% (Nale *et al* 2012). This is the first study to investigate prophage carriage within environmental strains.

The phage particles observed in the analysis found all were tailed bacteriophages and belong to the *Caudovirales,* the most common morphology of phages isolated in studies (Ackermann 2007). In this study, the myovirus particles have been grouped according to their morphology and size. There were also PTLPs present in the culture lysates which are predicted to be bacteriocins produced by the bacterium (Nagy and Foldes 1991, Fortier and Moineau 2007, Shan *et al* 2012, Nale *et al* 2012).

Multiple prophage carriage was detected in the environmental isolates, which is also consistent with studies of clinical and infant isolates (Nagy and Foldes 1991, Fortier and Moineau 2007, Shan *et al* 2012, Nale *et al* 2012). However, the rate of co-carriage detected was high than reported (e.g. Nale *et al* 2012). It is clear that there is different prophage carriage within ribotype groups, which was also found by Nale *et al* (2012) but in contrast to the findings of Fortier and Moineau (2007).

The total diversity of prophages is likely to be underestimated from TEM analysis, which can only identify across morphological characteristics and provides no insight into the phage genotypes. This was demonstrated by Shan *et al* (2012) who used the holin gene to probe the diversity and phylogeny within *C. difficile* prophage. Concluding that a similar sized myoviruses are genetically related is impossible, and therefore it is difficult to estimate the level of conserved prophage carriage within ribotypes.

Based on their phage carriage and ribotype, 19 strains have been selected for whole genome sequencing by Trevor Lawley, at the Wellcome Trust Sanger Institute. Two of these genomes presently are undergoing annotation, T6 (R076) and K12 (R220). Chromosomal regions encoding the prophage within each isolate have been identified. Futher work will determine the genetic diversity of their prophages.

3.3.11. Differential release of temperate phages following induction with two antibiotics and spontaneous release.

The results of the comparison, using a subset of isolates induced with MC, NFX and untreated culture lysates, showed that in several cases, different phages could be detected. This suggests that control of these prophages could be under different genetic mechanisms. While the majority (68%) of strains tested had the same or similar phage particle observed in both MC and NFX induced lysates, two NFX induced lysates produced phage particles which were not observed in the

MC lysates and for three isolates this was reversed. Release of different phages depending on antibiotic treatment was reported in Nale *et al* (2012) who suggested the sensitivity of these strains to each antibiotic varied as the mode of actions of MC and NFX are different, resulting in a different degree of SOS response that impacted phage excision. Alternatively, I propose that for at least some of the data in this study, it may be that using TEM analysis to confirm absence of phage is erroneous due to difficulties in detecting phage particles.

In nine cases the same morphology of phages observed in the non-induced culture supernatants the antibiotic induced lysates. While temperate phages are considered generally stable when maintained in a lysogenic state, the observation of low levels of spontaneous phage release has been documented for other bacterial strains, which may be one effect of laboratory manipulation of the host (Bratbuk *et al* 1994). Whether this low level of prophage release occurs in the environment is depentant on the state of the *C. difficile*, either as vegetative cells or as endospores. Phage release is unlikely to occur if *C. difficile* is present primarily as endospores, as phages are not released during spore dormancy in other endospore producing species (Hendry and Fitz-James 1974, Osbourne and Sonenshein 1976). However, the observation of induced and spontaneous release of active prophages in these strains is further evidence for potential HGT to occur within reservoirs of *C. difficile*.

3.4. Summary

C. difficile was isolated from estuarine sediments at a relatively high frequency and at levels consistent across the literature from infants and animal samples. Whether the bacterium is growing actively in the marine ecosystem or is present from contamination by faecal waste is unclear as its distribution did not show any correlation in regards to pollution status of the sites.

Ribotype analysis of the environmental *C. difficile* isolates revealed that they are genetically diverse, and include recognisable ribotypes as well as those which could not be assigned to a known

ribotype. The prevalance of specific ribotypes in the environmental samples is consistent with other studies, and suggests that the population may be active or these are the most common contaminating strains. Importantly, the ribotype analysis revealed that there is genotypic crossover between these and clinical isolates.

The investigation of two key phenotypic traits, motility and antibiotic resistance, of the environmental isolates showed these to vary between isolates and ribotype groups. Due to the variability of these traits within related strains, there is the potential for genes conferring these phenotypes to undergo HGT in the mixed population present in the estuarine sediments.

One way in which such genes could be transferred would be via phage infection. TEM analysis showed that prophage carriage in the environmental strains is widespread and diverse, with co-carriage frequently detected. There was no clear correlation between prophage carriage and ribotype observed. Also, differential release of temperate phages from induction with different antibiotics occurred, suggesting that these prophages have different control mechanisms. Interestingly, spontaneous release of temperates phage was also observed and this could directly impact the dynamics and structure of estuarine *C. difficile* populations, if active in this environment.

To conclude, a large and diverse strain collection of environmental isolates of *C. difficile* has been established and characterisation of their traits show inter-ribotype variation. This panel has been used to isolate of phages associated with these strains in the next chapter. To date no lytic phages have been reported isolating *C. difficile.* Isolation of such phages from estuarine sediments would provide further insights into the pathogens' survival this ecosystem.

4. Isolation of phages with lytic activity on *C. difficile.*

4.1. Introduction

There have been several studies that have isolated phages infecting *C. difficile,* but none of these have focused on environmentally associated phages (e.g. Sell *et al* 1983, Schallehn 1984, Mahoney *et al* 1985, Goh *et al* 2005, Horgan *et al* 2011). The prevalence of prophages encoded in *C. difficile* strains was exploited, with early work using induced phages as a typing scheme for clinical isolates but due to their narrow host ranges this was abandoned (Sell *et al* 1983, Dei 1989, Mahoney *et al* 1991). The difficulty in obtained lytically active phages to infect *C. difficile* is well documented (e.g. Sell 1984, Mahoney *et al* 1985, Goh *et al* 2005a, Mayer *et al* 2008, Horgan *et al* 2010). The isolation of free phages is highly unusual, with only one study isolating released temperate phages from patient stool samples, despite several other attempts (Goh *et al* 2005a, Horgan *et al* 2010, Meessen-Pinard *et al* 2012). No truly virulent phages have therefore been isolated which infect *C. difficile*.

By definition, all temperate phages are capable of undergoing lytic infection, otherwise they would be cryptic prophage elements in the bacterial chromosome with no exit strategy (via lytic replication) available to them. The switch between a lysogenic and lytic lifecycle, as well as how lysogeny is established is well understood for several phages, such as phage lambda (Echols 1972). The ability of a temperate phage to infect only in a lytic manner is subject to multiple factors influencing the state of replication as the switch to lysogeny involves rapid repression of late stage transcription. Also required for lysogeny to occur is the presence of an attachment site (which may or may not be specific) and an integrase (Birge 2006).

Truly virulent phages do not encode integrases and are unable to insert into their bacterial hosts DNA. As the process of lysogeny introduces novel phage-encoded genes into the bacterial host, virulent phages are considered most useful in therapeutic purposes, as well as to maximise phage killing of the target host cells (Brüssow *et al* 2005). More interestingly is why there have been

no virulent phages isolated to date for *C. difficile*. Currently, there are 790 entries of phages belonging to the *Caudovirales* (the most commonly isolated order of bacteriophages) in the NCBI genome database. Of these 790, only 41% encode a predicted integrase showing that the majority of isolated phages are virulent. Also, virulent phages infecting related species have been isolated, for example, phiCTP1 which infects *Clostridium tyrobutyricum* (Mayer *et al* 2010).

The activity on the isolated *C. difficile* phages is variable, and host ranges are typically narrow (e.g. Mahoney *et al* 1985, Goh *et al* 2007) but there are exceptions (e.g. Sekulovic *et al* 2011). Specialist and generalist phages can be seen in other species and their patterns of infections (interactions) can reveal distinct evolutionary relationships through their co-evolution with their host bacteria and resulting phage population structure (Flores *et al* 2011). As previously discussed, the epidemiology of *C. difficile* varies between countries and over time (McDonald *et al* 2005, Kuijper *et al* 2006, Goorhuis *et al* 2008, Deneve *et al* 2009), and the ability of phages to infect different strains have a direct impact on the contribution phages make to *C. difficile* biology.

In order to isolate phages that are associated with environmental strains of *C. difficile*, as well as maximise the chance of isolating virulent phages, several screenings of culture lysates, environmental phage extracts and enrichment cultures were performed. The environmental *C. difficile* isolates were used as sources of phages, as lysates from their cultures were used to identify active phages, as well as indicators of active phages in the screening panels. To confirm the ability of *C. difficile* strains to release phages from cultures after induction and without induction, growth dynamics of strain K12 were investigated. Isolated phages were characterised according to their morphology determined using TEM, their genome size and plaque morphology. Host range analyses were also performed, which included environmental and clinical *C. difficile* isolates to assess the potential usefulness of these phages for therapeutic applications, as well as gain insights into phage-host relationships in *C. difficile*.

4.2. Results

4.2.1. Isolation of 15 phages from enriched phage culture lysates.

In an effort to isolate virulent phages which infect *C. difficile*, phage extracts and enrichment cultures were tested for the presence of viable phages on different host isolates. Throughout this Ph.D., three rounds of phage isolation were performed, due to repeated sampling of the Hampshire area. The first round of phage isolations were performed using phage extracts from the infant, animal and 2009 Hampshire samples (*n*=118). Phage extracts from these samples were tested on a panel of 33 *C. difficile* isolates (comprising of animal, neonate, 2009 Hampshire and Slovenian isolates). None of the samples produced single plaques that could be propagated further than the first round of purification. Repeated assays were performed using subsets of pooled and PEG concentrated samples, but these also failed to isolate purified phages. Enriched phage culture lysates from the same samples were tested on 5 isolates as indicator hosts (S1, S2, S11, RBI and NC11204). Lysis was observed in several cases, but no individual plaques were obtained following serial dilution of the samples or from resuspended agar plugs from zones of lysis.

A second round of phage isolation was performed using the 2010 Hampshire sediment samples in which 25 phage enrichment culture lysates were tested on 4 isolates: S1, 10, T6 and S7. The initial screening assay and the first round of purification was performed by Holly Colvin as part of her undergraduate laboratory project (09/2010-01/2011). Second and third rounds of plaque purification were carried out by myself and this combined effort, resulted in isolation of 10 phages: phiCDHM2, phiCDHM4, phiCDHM7, phiCDHM8, phiCDHM9, phiCDHM10, phiCDHM11, phiCDHM12, phiCDHM13 and phiCDHM14.

In a third round of phage isolation performed in this project, 25 enriched phage culture lysates were tested on 13 clinical strains belonging to different ribotypes, with testing carried out by Krystin Norman during a summer studentship in the laboratory. The clinical isolates were AIU (R027), AUG (R078), ARJ (R107), APL (R015), ATJ (R014), ATH (R002), APA (R005), ASH (R013), AKL (R023), AII

(R026), ANO (R081) and AIH (R087), and as well as the non-clinical isolate T6 (R076). From this isolation work, five phages were isolated: phiCDHS1, phiCDHM5, phiCDHM6, phiCDHM15 and phiCDHM16.

4.2.2. Isolation of 12 temperate phages with lytic infective cycles on genetically different *C. difficile* strains.

As no one has successfully isolated a virulent phage that infects *C. difficile*, this project also included the isolation and characterisation of phages known to be temperate but that had virulent activity on certain hosts. To identify such phages, culture lysates of 33 isolates that were induced either with mitomycin C (MC) or norfloxacin (NFX), as well as untreated cultures, were used to test for active temperate phages on a panel of 33 isolates (Table 6).

Before the large screening was carried out, three preliminary method optimisations were performed to assess: 1) use of $MgCl_2$ and $CaCl_2$ to supplement the media, 2) use of ampicillin to supplement the media and 3) use of bile salts to supplement the media. Firstly, a comparison of supplemented and non-supplemented soft agar was performed. The optimisation was performed on six isolates (S1, S5, S11, HB, RBI and NC11204) using MC induced lysates from 19 isolates. The results showed that supplementing the agar was important, as 5 of the samples produced a positive result (with plaques observed); without supplements, plaques could not be seen for 4 of these 5 positive samples. Further optimisation was performed by the addition of ampicillin at two different concentrations, 3 µg ml^{-1} or 5 µg ml^{-1}, to the soft agar for two isolates S1 and NC11204.The resulting growth of the bacterial lawns was inhibited so this method not used. Another optimised method was performed, with the addition of ox bile salts at 0.25% to the soft agar. However, this was unsuccessful as it also inhibited the growth of the lawn. All subsequent screenings were therefore performed using only soft agar supplemented with the salts.

The results showed that lysates of 8 *C. difficile* isolates produced lysis on a single host isolate and 5 lysates produced lysis on more than one host isolate. The isolates of either lysate or host represented several different ribotypes, and included pairings where the lysates from the one ribotype could infect the same, as well as different ribotypes. Lysates of four isolates could produce lysis or plaques on more than one *C. difficile* strain with interesting patterns of activity which suggested they produced related phages. For example, lysates of S8 (R002) cleared three of the same isolates as those from S1, 16 and S11 (all R010). Also of note was the fact that isolates belonging to the same ribotype were lysed by the same samples, showing there is intra-ribotype variability with regards to potential phage sensitivity.

In some cases the type of lysate appeared to influence the results; for example, lysates from isolate R10 (R010) produced clearing on four *C. difficile* isolates: S1, S11, K3 and K4. However, while both MC and NFX lysates were active on isolates S1 and K3, only the MC lysate was on S11 and K4. Interestingly, the uninduced lysates also showed activity in 6 isolates suggesting that the phages in these lysates were spontaneously released.

Serial dilution of lysates was performed to obtain single plaques for the purification of the active phages. Several clearings or plaques could not be propagated further than the initial screen. The induction and propagation work resulted in the isolation of 8 phages phiCDHM1, phiCDHM20, phiCDHM21, phiCDHM22, phiCDHM23, phiCDHS2, phiCDHS3 and phiCDHM24.

Table 6. Temperate phages with lytic activity on environmental hosts.
Columns: isolate lysates M= mitomycin C treated, N= norfloxacin treated, C= untreated control
Rows: isolate as host

Ribotype		R010												R220						R005		
		S1			R10			K16			S11			K12			K3			K18		
	Isolate	M	N	C	M	N	C	M	N	C	M	N	C	M	N	C	M	N	C	M	N	C
R010	S1	-	-	-	√	√	-	√	√	-	√	√	-	-	-	-	-	-	-	-	-	-
	R10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S11	-	-	-	√	-	-	√	√	√	-	-	-	-	-	-	-	-	-	-	-	-
R220	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K3	-	-	-	√	√	-	-	-	-	√	√	-	-	-	-	-	-	-	-	-	-
R005	K18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R001	K25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R002	S8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R027	HB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R014	K6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	OA	-	-	-	-	-	-	√	-	√	-	-	-	-	-	-	-	-	-	-	-	-
R003	SS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R085	R8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R021	K9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R012	K10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	√
R076	T6	-	-	-	-	-	-	-	-	-	-	-	-	√	√	√	-	-	-	-	-	-
R046	K4	-	-	-	√	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R031	S7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	K20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	RBI	√	√	-	-	-	-	-	√	-	-	√	-	-	-	-	-	-	-	-	-	-
R023	S2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R103	S3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R035	S6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R100	S9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6. *continued*

Ribotype	Isolate	R001 K25 M	N	C	R002 S8 M	N	C	K14 M	N	C	R027 HB M	N	C	K15 M	N	C	R014 K6 M	N	C	OA M	N	C
R010	S1	-	-	-	√	√	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	R10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S11	-	-	-	√	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R220	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K3	-	-	-	√	√	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R005	K18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R001	K25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R002	S8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R027	HB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R014	K6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	OA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R003	SS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R085	R8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R021	K9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R012	K10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R076	T6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R046	K4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R031	S7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	K20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	RBI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R023	S2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R103	S3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R035	S6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R100	S9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6. *continued*

Ribotype	Isolate	R003 SS1 M	N	C	R085 R8 M	N	C	R021 K9 M	N	C	B1 M	N	C	B2 M	N	C	R012 K10 M	N	C	R076 T6 M	N	C
R010	S1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	R10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R220	K12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R005	K18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R001	K25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R002	S8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R027	HB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R014	K6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	OA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R003	SS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R085	R8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R021	K9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R012	10	-	-	√	-	-	-	-	-	-	-	-	√	√	√	√	-	-	-	-	-	-
R076	T6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R046	K4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R031	S7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	√	-	-	-	-	-
U/K	K20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	RBI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R023	S2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R103	S3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R035	S6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R100	S9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6. *continued*

Ribotype	Isolate	R046			R031			UK			UK			R023			R103			UK		
		K4			S7			K20			RBI			S2			S3			S4		
		M	N	C	M	N	C	M	N	C	M	N	C	M	N	C	M	N	C	M	N	C
R010	S1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	R10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R220	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R005	K18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R001	K25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R002	S8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R027	HB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R014	K6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	OA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R003	SS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R085	R8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R021	K9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R012	K10	-	-	-	-	-	-	-	-	-	-	-	√	-	-	-	-	-	-	-	-	-
R076	T6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R046	K4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R031	S7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	K20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	RBI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R023	S2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R103	S3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R035	S6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R100	S9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6. *continued*

Ribotype	Isolate	UK S5 M	N	C	R035 S6 M	N	C	R100 S9 M	N	C	UK S10 M	N	C	UK S12 M	N	C
R010	S1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	R10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R220	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R005	K18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R001	K25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R002	S8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R027	HB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	K15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R014	K6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	OA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R003	SS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R085	R8	-	-	-	-	-	-	√	√	√	-	-	-	-	-	-
R021	K9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	B2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R012	K10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R076	T6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R046	K4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R031	S7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	K20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U/K	RBI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R023	S2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R103	S3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R035	S6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R100	S9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	S12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

From the 2010 sampling, 16 isolates of *C. difficile* were used to identify active temperate

phages with lytic activity from cultures treated with NFX and without. The lysates were tested on a

set of four isolates: S1, T6, S7 and K10 (by Holly Colvin). Subsequent second and third rounds of

plaque purification were carried out by myself and four phages phiCDHM17, phiCDHM19,

phiCDHM18 and phiCDHM3 were isolated.

4.2.3. Growth dynamics of induced and uninduced cultures show the cost of phage production.

To investigate the dynamics of spontaneous release in *C. difficile*, the OD, PFU ml^{-1} and CFU

ml^{-1} were measured following culture, with and without MC treatment for K12 (R220). Isolate K12

was chosen for this experiment as it releases the temperate phage phiCDHM1 which propagate

lytically on strain T6 (R076), allowing phage titration assays to determine PFU.

Both cultures follow a standard bacterial growth curve with a lag phase, exponential growth

phase and a stationary phase, as observed from OD$_{550nm}$ values and CFU ml^{-1} counts (Figure 8). The

PFU ml^{-1} counts show that active phages are released in both antibiotic treated and untreated

cultures. However, the dynamics of the two cultures appear different in terms of the cell number

and cell density reached over the course of the growth curve. Also, drops in the OD$_{550nm}$ values can

be seen for both induced and uninduced cultures, but this drop occurs at an earlier time point in the

former. These values may reflect the cost of phage induction on the culture as reflected in the higher

PFU ml^{-1} values observed in the antibiotic treated culture. The PFU ml^{-1} counts show that release of

phage following antibiotic induction is rapid, with a peak of 1.1×10^4 two hours after MC induction.

Phage production appears to fall and rise during the course of the growth curve and further

experiments could confirm if this is consistently observed.
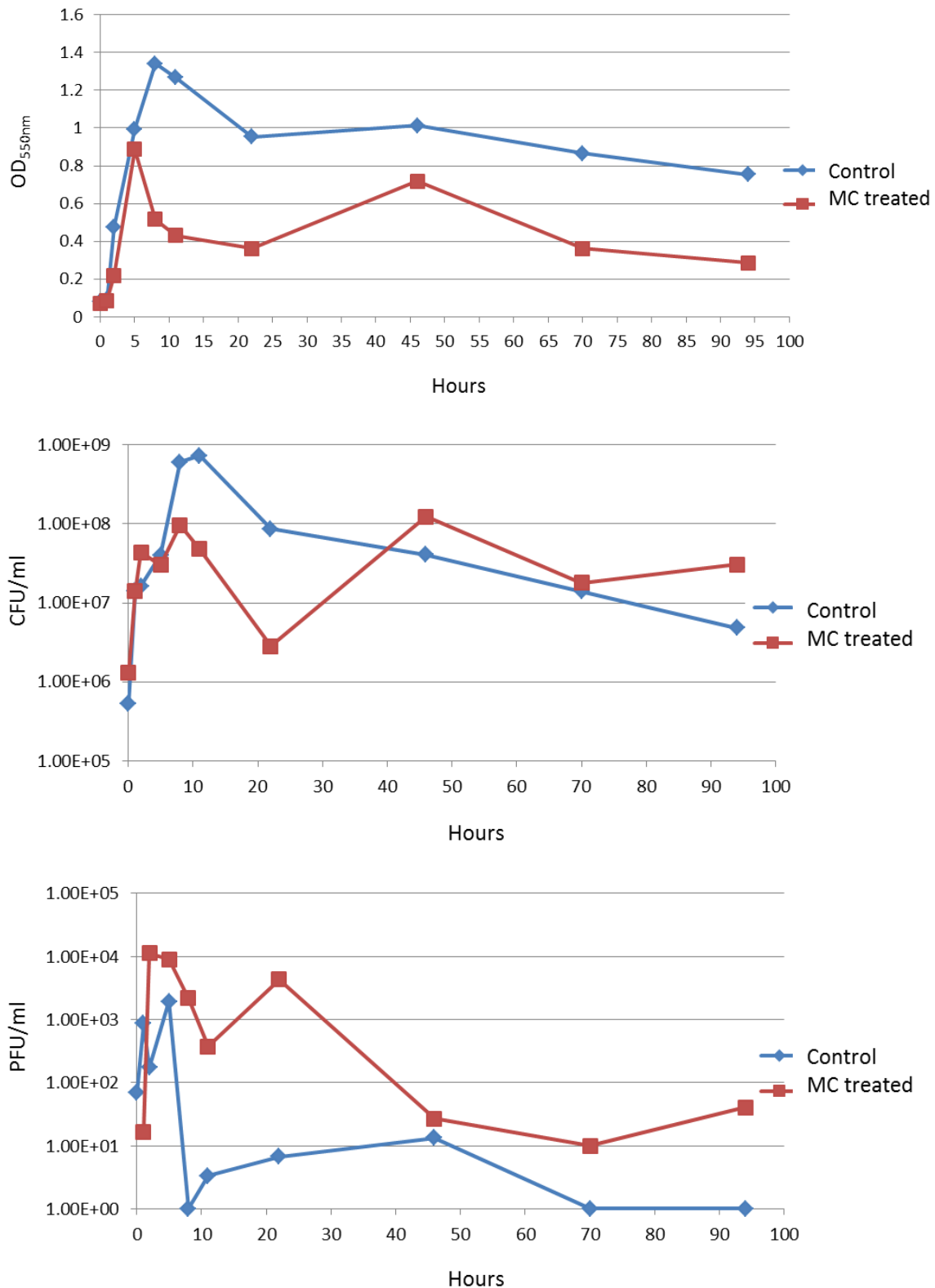
Figure 8. Bacterial growth curves and viable phage release from mitomycin c treated and untreated *C. difficile* cultures of strain K12. Top: OD_{550}nm growth curve over 94 hours. Middle: CFU ml^{-1} and bottom graph shows PFU ml^{-1} of phage released, visualised on host strain T6 (R076). Trend lines: red = culture induced with mitomycin c and blue = no antibiotic added.

4.2.4. TEM analysis of phage particle morphology reveals distinct morphotypes.

In order to begin their characterisation, the particle morphologies were determined for the 27 phages isolated. TEM analysis was performed on negatively stained samples to observe their phage morphologies. All phages were found to belong to the tailed bacteriophages family, the *Caudovirales* and phages belonging to either the *Myoviridae* or *Siphoviridae* were detected (Figure 9). These two phage families differ in regards to their particle morphology as myoviruses have contractile tail and siphoviruses have non-contractile tails. These families can be sub-typed into morphotypes as described by Ackermann (2009). All 21 myoviruses belong to morphotype A1 and all 3 siphoviruses belong to B1 morphotype. However, the phage particles can vary in size from one another and can be split into sub-types accordingly. The myoviruses fall into three main groups: those with capsid diameters of 60-70 nm and tail lengths of ~130 nm which are termed medium myoviruses (MMs); those with capsid diameters of ~60 nm and tail lengths of ~220 nm which are termed long tailed myoviruses (LTMs) and those with capsid diameters of ~50 nm and tail lengths of ~105 nm which are termed small myoviruses (SMVs). These classifications are referenced throughout the remainder of the thesis. The siphoviruses (SVs) also vary in size, as 2 have capsid diameters of ~55 nm and a tail length of ~280 nm, and the third is slightly larger with a capsid diameter of ~60 nm and tail length of ~300 nm.

There is no clear relationship between morphology and phage origin, as in each group there are known temperate phages released from lysogens, as well as phages from the enrichment cultures. There may be a trend of morphology and host indicator isolate used in the isolation, as for example no SV was isolated on one of the most common hosts of the MMs. However, this may be a bias from the use of subsets of indicator strains rather than a biological relationship.

.

MMs

1	phiCDHM1
2	phiCDHM3
3	phiCHM7
4	phiCDHM8
5	phiCDHM17
6	phiCDHM18
7	phiCDHM19
8	phiCDHM20
9	phiCDHM21
10	phiCDHM22
11	phiCDHM23
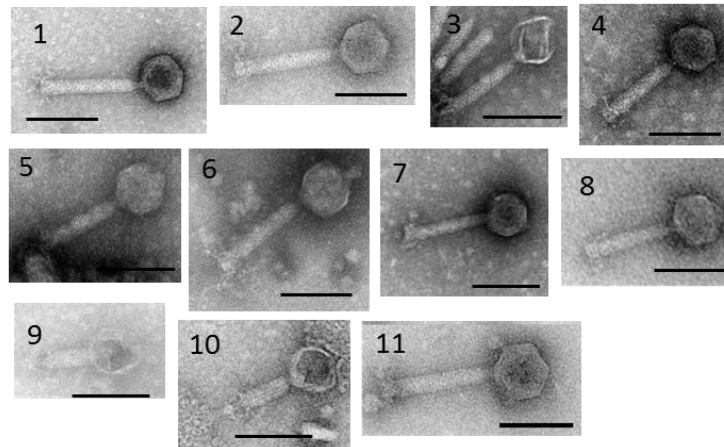
LTMs

12	phiCDHM2
13	phiCDHM4
14	phiCDHM5
15	phiCDHM6
16	phiCDHM15
17	phiCDHM16
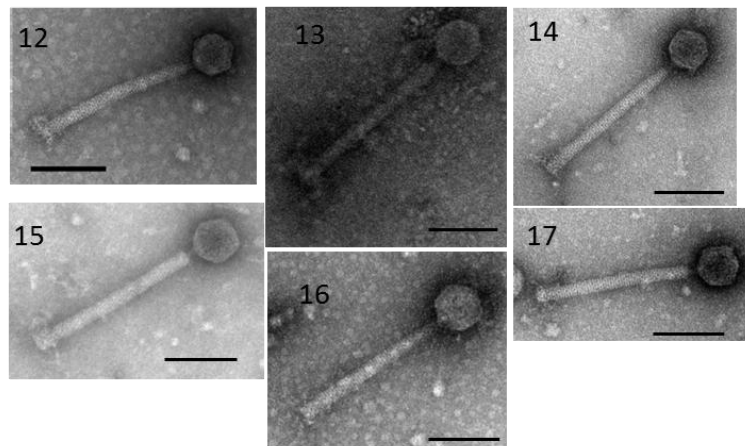
SMVs

18	phiCDHM9
19	phiCDHM10
20	phiCDHM11
21	phiCDHM12
22	phiCDHM13
23	phiCDHM14
24	phiCDHM24
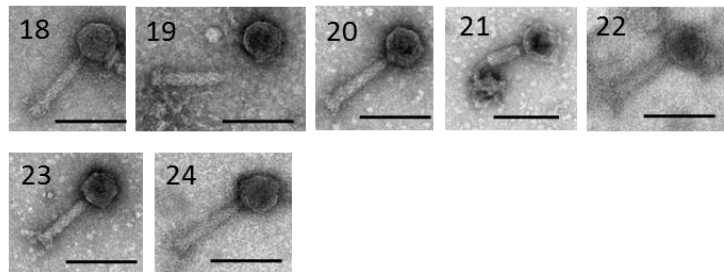
SVs

25	phiCDHS1
26	phiCDHS2
27	phiCDHS3
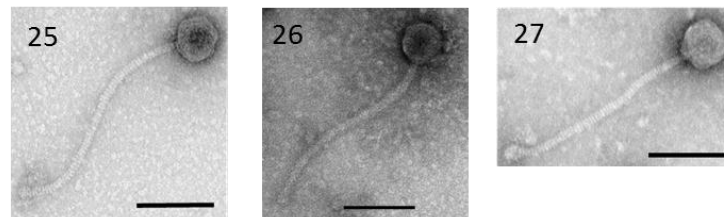
Figure 9. Morphology of purified phage particles with lytic activity against *C. difficile* host strains. 27 phages belong to the families *Myoviridae* (*n*=24) and *Siphoviridae* (*n*=3). The observed myoviruses fall into three subtypes according their size and particle morpohology: 11 medium sized myoviruses (MMs): 6 long tailed myoviruses (LTMs): 7 small myoviruses (SMVs) and 3 siphoviruses (SVs). The SVs include two distinct particle sizes, with phiCDHS1 having larger capsid and long tail length than phiCDHS2 and phiCDHS3, which both have similar sized particles

4.2.5. Pulsed Field Gel Electrophoresis analysis indicates relationship between phage morphology and genome size.

To assess genome diversity, the genome sizes of the phages were obtained using PFGE. The results for 21 of the 27 phages show that their sizes range from 30-55 kbp (Figure 10). The sizes of the genomes appear to correlate with their particle morphology as the phages with capsid diameters of 60-70 nm all have similar sized genomes ranging from 50-55 kbp. These phages are all representatives of either the MMs or LTMs. The MMs had the greatest amount of variation in genome size observed. Phages which are SMVs and have smaller capsid diameters (of~ 50 nm), all have smaller genomes ranging from 30-35 kbp. The SV phiCDHS1 has a capsid size ~60 nm and a genome size of ~ 42 kbp. The genome bands detected in the PFGE analysis varied in intensity and in resolution, as obtaining high enough phage titres for use in the analysis was difficult. The phage titre has a direct result on visualisation of a genome band and, although >10^7 PFU ml^{-1} stocks were used for plug production; several samples consistently did not produce bands indicating that this was not a high enough titre. These phages were phiCDHM17, phiCDHM22, phiCDHM21, phiCDHS2, phiCDHS3 and phiCDHM24.
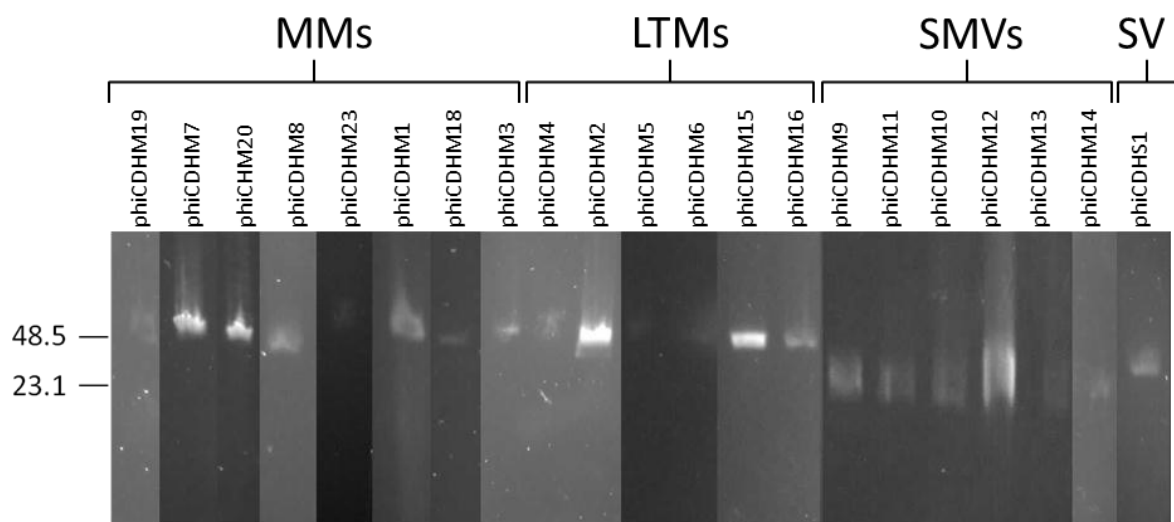
Figure 10. Phage genome analysis using PFGE. Composite image of phage genomes from several PFGE gels. Phage genomes ranged from 30-54kbp. There were three main sizes of genome bands: 50-54kbp (including the MMs and LTMs), 30-35 kbp (the SMVs) and the single SV ~ 40kbp.

4.2.6. Plaque morphology analysis of purified phages reveals diversity of infection dynamics.

Plaque morphology can be used to aid characterisation of phages and it was seen that plaque type (turbid or clear) and size varied between phages. The plaque types could be grouped into three and phages produced either small clear plaques, medium sized clear plaques or large turbid plaques (Figure 11).

There was no clear relationship between phage morphology and plaque morphology. Of the 7 phages producing small clear plaques, 3 are myoviruses (belonging to MMs, LTMs and SMVs) and 3 are siphoviruses. Medium sized clear plaques are produced by 11 phages and include those belonging to all 3 myovirus subtypes. Large, turbid plaques were produced by 5 phages (MMs and SMVs).

While turbid plaque production indicates the phage is lysogenic, there was no clear relationship between plaque type and the phage origin as both induced and enriched phages produced all three types in similar proportions (30.7% and 25% respectively).
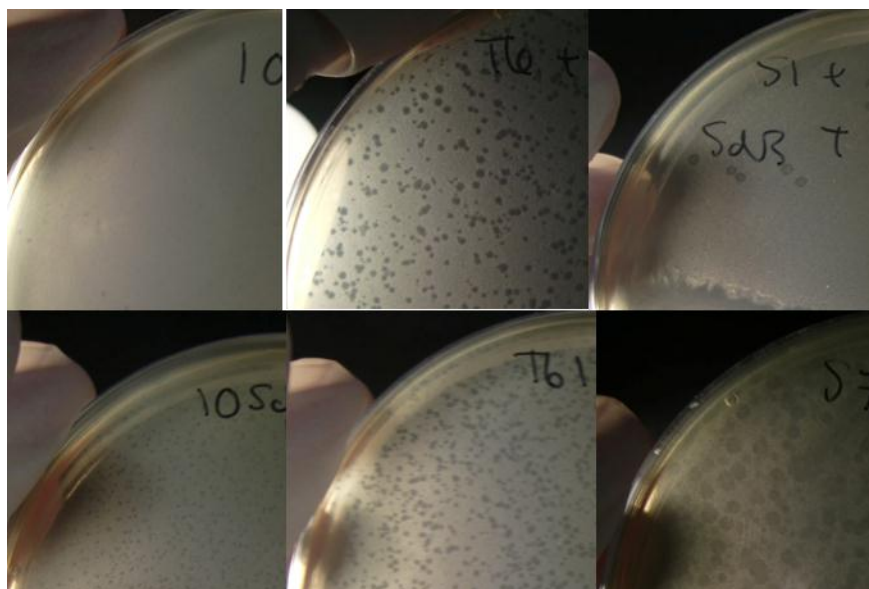
Figure 11. Examples of the three types of plaque morphology observed for purified *C. difficile* phages. Left (top and bottom): small clear plaques produced by phages phiCDHS3 and phiCDHM4. Middle (top and bottom): medium clear plaques produced by phiCDHM1 and phiCDHM2. Right (top and bottom): large turbid plaques produced by phiCDHM8 and phiCDHM12.

4.2.7. Phage restriction enzyme profile analysis suggests modification of phage DNA.

In order to compare the potential diversity of the phage genomes, their restriction enzyme (RE) profiles were investigated. Digests of genomic DNA from phiCDMH1, phiCDHM20 and phiCDHM23 with *Eco*RI, *Bam*HI and *Hind*III were performed, following the manufacturer's guidelines as well as including optimisation steps. However, restriction enzyme digestion profiles could not be obtained (data not shown). The panel of restriction enzymes was then expanded to include *Sty*I, *Eco*RN, *Nde*I, *Dpn*I, *Nhe*I and *Sma*I to digest phiCDMH1 DNA (Figure 12). These digestions also failed to produce satisfactory band profiles, or to even digest the genomic DNA, and conditions were repeated to increase digestion times yet this also did not produce RE profiles. The three enzymes that could digest the phage DNA were *Bam*HI, *Nde*I and *Nhe*I: *Bam*HI produced a two band profile; *Nde*I could cleave at multiple sites although the bands produced were poorly resolved and *Nhe*I digested the DNA extensively, so bands could be observed.
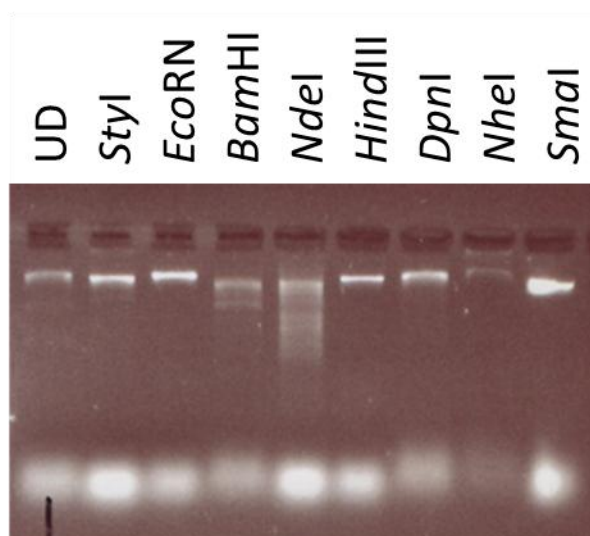
Figure 12. DNA digests of phiCDMH1 DNA with eight restriction endonuclease enzymes. Left to right: phiCDMH1 gDNA undigested and treated with 8 restriction enzymes: *Sty*I; *Eco*RN; *Bam*HI; *Nde*I; *Hind*III; *Dpn*I; *Nhe*I and *Sma*I. Phage gDNA is cut by *Bam*HI, *Nde*I and *Nhe*I but did not produce multiple bands and generate multiple bands. DNA was visualised under longwave U.V. after electrophoresis in a 1% Helena Agarose gel stained using GelRed.

4.2.8. Phage host range analysis reveals specialist and generalist phages, phage relatedness and host intra-ribotype variation in phage sensitivity.

The host ranges of the first 8 phages to be purified in this project in 2009 (phiCDHM1, phiCDHM20, phiCDHM21, phiCDHM22, phiCDHM23, phiCDHM24, phiCDHS2 and phiCDHS3) were established on 33 environmental and non-clinically associated *C. difficile* isolates (Table 7). They were also tested against 23 clinical isolates belonging to R027 which represented 23 MLVA (Multiple Loci VNTR Analysis) types, as well as 7 clinical isolates belonging to R078, but none of the phages could infect any of the clinical isolates in this analysis.

The host ranges of the 8 phages varied. Some appeared to be specialist, as they could only infect 1 or 2 isolates, for example phiCDHM1. Others were more generalist, for example phiCDHM21 and phiCDHM23 which infected 8 and 11 isolates respectively. It also appears that for some, their host range may have increased following their passage in an indicator host from the initial lysate used to isolate them. Alternatively, the results could reflect a variable efficiency in phage infection, as the titres of the purified phage stocks used in the host range analysis were higher than in the initial lysates.

Also the patterns of infections from the host ranges indicate that certain phages may be related as they infect the same hosts. For example, the MMs, phiCDHM21 and phiCDHM22, infect 3 of the same hosts. Also, the other MMs, phiCDHM20 and phiCDHM23, infect 8 out of 11 of the same hosts. In contrast, phiCDMH1 shares only one host with phiCDHM23, and none with the other phages, and phiCDHM13 does not infect any of the same hosts as the other phages.

Of the 33 isolates, 14 (42%) could be infected by at least a single phage. Of these 14, 2 were infected by a single phage, 10 were infected by 2 phages and 2 were infected by 4 phages. In total, the 8 phages could infect isolates belonging to 10 ribotypes. They could not infect isolates belonging to 4 ribotypes (R085, R027, R005 and R220) and 8 isolates that have not been assigned a known

ribotype. Interestingly, the phages could not necessarily infect all isolates of the same ribotype and this intra-ribotype variation included both the ability to infect as well as the type of resulting infection (confluent lysis or turbid). For example, phiCDHM21 produced clear plaques on one isolate belonging to R010 and turbid plaques on 2 other R010 isolates.

Table 7. First set of eight phages host range on environmental strains from 2009.
Symbol ◉ and ○ indicate turbid clearing and confluent lysis. Symbol * indicates where the host isolate is the same of the tested isolate. Abbreviation U/K = unknown ribotype.

		Phage						
Ribotype	Isolate	phiCDHM1	phiCDHM20	phiCDHM21	phiCDHM22	phiCDHM23	phiCDHM24	phiCDHS3
R010	S1		○*	◉*	◉	○*		
	R10			○*		○*		
	S11		○*	◉*	○	○*		
	16							
R031	S7						◉	
R003	SS1		○			○		
R085	R8							
R001	K25							
R012	K10					○		○
R021	B1		○			○		
	B2		○			○		
	K9							
R075	T6	○						
R027	HB							
	K15							
R046	K4			○	○			
R014	K6							
	OA		○			○		
R005	K18							
R220	K3							
	K12							
R002	S8	○				○		
	K14							
U/K	K20							
U/K	RB1		○			○		
U/K	S4		○			○		
U/K	S5							
U/K	S6							
U/K	S9							
U/K	S10							
U/K	S12							
U/K	S2							

4.2.9. Expanded host range analysis reveals phage cross-infectivity of both spatially and temporally separated hosts.

An expanded host range was performed that included 23 phages and 61 *C. difficile* isolates with additional isolates from the Hampshire 2010 and 2011 sampling, as well as the phages isolated from these samples (Table 8).

The host range analysis produced results that were consistent with the previous analysis but determined the phage-host interactions for this species to include more diversity as this was on a larger scale. In total, the phages could infect 49 isolates out of 61 tested (80%) and the infected isolates belong to 21 known and 6 unknown ribotypes. Collectively, the phages could infect at least one strain in 17/21 of the known ribotype groups and 4/6 unknown ribotypes. Results were consistent with the previous host range analysis, as there were specialist and generalist phages could be identified. For example, the phage phiCDHM22 could infect a single isolate, whereas phiCDHS1 infected 29.

Also consistent with the previous analysis, was the observation of extensive intra-ribotype variation in phage sensitivity. A clear example of this is for R010, the ribotype group with the most isolates in the analysis. The 7 isolates belonging to R010 could be differentially infected by 11/23 of the phages tested. The results may be subject to bias as the indicator hosts for several of these phages were of the same ribotype, but they include those which were not; for example phiCDHS1 can infect all isolates belonging to R010 except 1, but was isolated on a host belonging to R027. Again, consistent with the previous host range analysis was the variation in infection type by the same phage between different isolates, with some phages producing confluent lysis on some hosts but turbid clearing on others. Similarly, single hosts could be produce both infections, with confluent lysis from some phages, and turbid from others. This implies that the interaction between phage and host determines the outcome of infection in *C. difficile* and is very phage/strain specific.

Table 8. Expanded host range of phages against environmental strains of *C. difficile*. Phages are in columns and isolates in rows, grouped according to ribotype. Symbol: ○ indicates clear lysis and ◉ indicates turbid zones.

Ribotype	Isolate	phiCDHM21	phiCDHM22	phiCDHM20	phiCDHM8	phiCDHM19	phiCDHM17	phiCDHM7	phiCDHS3	phiCDHM4	phiCDHM13	phiCDHM14	phiCDHM11	phiCDHM9	phiCDHM10	phiCDHM12	phiCDHM1	phiCDHM2	phiCDHM3	phiCDHM15	phiCDHM16	phiCDHM5	phiCDHM6	phiCDHS1
R010	H1b																◉							○
	H3						◉					◉												○
	H5d																							○
	K16																							○
	R																							○
	I												○					○		◉	○	○	○	
	R10					◉	○																	◉
R002	H15																◉						○	○
	K14																◉							
	O																							◉
	P																◉							◉
	S																							◉
R031	H11			○			○			○	○	○	◉	◉	○			○						
	H1a			○							○	○	◉					○	○					
	M			○					○	○	◉	◉	○	○	○									◉
R012	K10									○	○	○	○				○							
R005	H5b																							○
	K18						○																	○
	E										○		○											
R106	H19																							○
R220	H12																			◉	◉	◉	◉	
	K3										○		○	○		○				◉	◉	◉	◉	
	K12												○	○						○	○	○	○	
	N																							◉
	K		○				○						○				○	○	○					
	Y												○	○				○	○	○	○	○	○	○	
	U																							
	L				○	○																		○
R076	T6										○		○		○		○	○	○	○	○	○	○	
R014	0A										○						◉	○		○	○	○	○	◉
	K6																			○	○	○	○	
	V										○		○	○				○	○					○
R046	K4						○																	○
R001	H8					◉																		◉
	H4																							○
	H18																							◉
	K25																							○
	F																							○

Table 7. *continued.*

Ribotype	Isolate	phiCDHM21	phiCDHM22	phiCDHM20	phiCDHM8	phiCDHM19	phiCDHM17	phiCDHM7	phiCDHS3	phiCDHM4	phiCDHM13	phiCDHM14	phiCDHM11	phiCDHM9	phiCDHM10	phiCDHM12	phiCDHM1	phiCDHM2	phiCDHM3	phiCDHM15	phiCDHM16	phiCDHM5	phiCDHM6	phiCDHS1
R027	K15					○	○															○		○
R027	HB						○																	○
R085	R8																							
R021	B2								◉		◉		◉											
R021	B1															◉								
R021	K9									◉	○	○												
R021	A																							
R003	SS1	◉	◉		◉			◉								◉								
R078	H17																							
R078	H5c																							
R078	J																							
R137	G																							
R137	D																							
R026	C																	○						
R100	B																							
R015	Q																				◉			
U/K	K20																							◉
U/K	RBI																◉							◉
U/K	H																○		◉	◉				
U/K	H16b																							○
U/K	X																							
U/K	H14																							

The expanded host range also provided an opportunity to investigate whether any evolutionary relationship could be detected between the phages and hosts from the sediment reservoir, as the analysis included those isolated from the same environment but from different years. Also, as the sites sampled remained the same, the relationship between location and phage-host interaction could be assessed. Phages isolated from the 2009 samples could not be tested on past strains, only contemporary (2009), future (2010) and far-future (2011) isolates. Those from

2010 could be tested on past (2009), contemporary (2010) and future strains (2011) and those from 2011 samples could only be tested on far-past (2009), past (2010) and contemporary isolates (2011).

There was no clear relationship between host-phage interaction and year, as where it could be determined, most phages could infect at least one isolates from the same year, past year/s and future year/s. There was no obvious relationship between the site of isolation and phage-host interaction as phages could infect isolates of the same ribotype that were isolated from samples taken at different sites around the estuarine system.

4.2.10. Phage host range analysis of diverse clinical isolates demonstrates strain cross-infectivity of phages.

Following the second round of sampling, *C. difficile* isolation and phage isolation, a host range analysis was performed to establish whether the phages could infect a broader range of clinical isolates (as previously only R027 and R078 were tested). The host range included 20 of the phages and 21 clinical isolates. This was performed by Anisha Thanki. In total, the phages infected 88/114 (77%) isolates tested which belonged to 19/21 (90%) ribotypes tested (R002, R003, R005, R009, R010, R013, R014/020, R015, R023, R026, R027, R031, R035, R050, R078, R081, R087, R100, R106 and R107) and could not infect R017 and R103. There were different sample sizes of the representative isolates tested for each ribotype, for example R017 and R081 both had one isolate, whereas group R078 contained 13 isolates. Again, variation in sensitivity to the same phage between isolates of the same ribotype was observed. For example, phiCDHM3 can infect all strains tested of R027, but only 3/13 strains of R078. The host ranges between phages also varied for example phiCDHM3 infected strains belonging to R027, R078, R087, R023, R015, R014/020, R013, R005 and R002 but phiCDHM11 could infect only R005.

4.2.11. Host range analysis shows a variable efficiency of plaquing and potential impact on natural populations is significant.

During the expanded host range analysis it was observed that some phages produced different numbers of plaques on certain isolates compared to others. To further investigate this observation, the host ranges and efficiency of plaquing (EOP) was examined for subset of 5 phages on 20 isolates. The 5 phages all belong to the SMVs and were isolated on the same host: phiCDHM9, phiCDHM10, phiCDHM11, phiCDHM13 and phiCDHM14. The isolates include those at least a single phage can infect, as well as other isolates that belong to the same ribotype as the infected and consist of 20 isolates belonging to 7 ribotypes (Table 9). The EOP can be compared between hosts as the same phage stock with a titre of at least 1×10^{6} was used for all infections (this titre is high enough to produce clear lysis on at least one host for all phages).

Table 9. Example of variable efficiency of plaquing for five related phages on host strains of *C. difficile* belonging to seven ribotypes

Red = confluent lysis. Orange = individual plaques. Yellow = turbid zones.

Ribotype	Isolate	Phage (SMVs)				
		phiCDHM13	phiCDHM14	phiCDHM11	phiCDHM9	phiCDHM10
R031	H11	Red	Red	Yellow	Yellow	Red
R031	H1a	Red	Red	Yellow		
R031	M	Yellow	Yellow	Orange	Orange	Orange
R012	K10	Red	Red	Red		
R005	H5b					
R005	K18					
R005	E	Orange		Orange		
R106	H19					
R220	H12					
R220	K3	Orange		Red	Orange	
R220	K12			Orange	Orange	
R220	N					
R220	K			Orange		
R220	Y			Red	Red	
R220	U					
R220	L					
R076	T6	Red		Red		Red
R014	0A	Orange				
R014	K6					
R014	V	Red		Red	Orange	

Clear zones of lysis which indicated a high EOP were observed on at least one isolate for every phage (n= 1-5); intermediate lysis (assigned from the production of plaques in the spot test, 10^3 to 10^4 PFU ml^{-1}) was observed for 4 phages. Turbid clearing which indicates lysogeny was also observed for all five phages. The results show that the efficiency of infection changes for each phage between different strains, even of the same ribotype, and that despite their morphological similarity, the host ranges of these 5 SMVs differs and some switch from lytic to lysogenic lifecycles for the same host.

4.3. Discussion.

4.3.1. Isolation of phages associated with environmental isolates of *C. difficile.*

In this work, 27 phages which can infect in a lytic manner of *C. difficile* strains have been isolated. These include 12 known temperate phages isolates from culture lysates and 15 from enriched samples. They infect a number of different host indicator strains. The phages isolated in this project represent the largest collection of *C. difficile* phages reported to date, and the only study using *C. difficile* strains from environmental sources.

The majority of known *C. difficile* phages are temperate, and have been isolated following large scale screenings of induced lysates against indicator panels (Sell *et al* 1983; Mahoney *et al* 1985; Goh *et al* 2005; Horgan *et al* 2011). Despite the varied sources of isolates and strains used, relatively low proportions of lysates containing phages are capable of lytic infection on indicator strains have been reported: 3.9% (10/254) (Sell *et al* 1983); 2.1% (2/94) (Mahoney *et al* 1985); 7% (4/56) (Goh *et al* 2005); 4.6% (2/43) (Horgan *et al* 2011); and 20% (1/5) (Schallehn 1984). Some studies have failed to find temperate phages able to propagate on host strains (Nagy and Foldes 1991; Nale *et al* 2012, Shan *et al* 2012). Nagy and Foldes (1991) did not obtain any plaques from their 28 induced strains, despite observing phage particles in 18/28 lysates. The difficulty in

obtaining lytically active phages is evident within the literature, with descriptions of filtrates that lysed strains but did produce individual plaques (Sell *et al* 1983). Also, one study used enrichment culture for isolation work, but no phages were isolated from them (Horgan *et al* 2010). In comparison those in the literature, the rates of isolation in this study are slightly higher, with 24-26% from induced isolates and 0 – 40% from enriched cultures. Whether this relates to the fact these are environmental strains or to just to the strains that were used in the panels is not known.

No phages were isolated directly from the environmental sample extracts. There has been only one study to isolate free phages (which were found to be lysogenic) reported for *C. difficile* (Meessen-Pinard *et al* 2012), despite several attempts to isolate them (e.g. Goh *et al* 2005, Horgan *et al* 2010, Meessen-Pinard *et al* 2012). The fact that no phages were isolated directly from extracts in this study may be explained by: the total absence of *C. difficile* phages in these samples (as not all samples yielded *C. difficile* isolates); that some did contain *C. difficile* phages, but these were not infective against strains used in the screening panel or that infective phages were present but in low numbers and were not detected. Phages isolated from enrichment cultures may be virulent phages that replicated to a detectable level on the naturally present co-cultured *C. difficile* strains. Alternatively, they may be prophage released from these strains or another bacterial species grown in the cultures. The cross infectivity of temperate phages from another species, *Clostridium sordellii,* to *C. difficile* has previously been described (Schallehn 1984). In this study, multiple phages were isolated from enrichment cultures and they may be phages that are temperate or virulent.

The difficulty in isolating and working with *C. difficile* phages is well known (e.g. Horgan *et al* 2010). In this study, often lysis on strains was observed but single plaques could not be produced. It may be that the samples produced bacteriocins which were responsible for the observed lysis (Sell *et al* 1983), or that they contain an inducing agent and cell lysis is the result of the release of the isolates own prophages. For example, in soil bacterial communities phages can be induced by quorum sensing (Ghosh *et al* 2009). Alternatively, phages in the samples could cause lysis of *C.*

difficile without completing a full infection and replication cycle by two mechanisms: a phenomenon termed "lysis from without" or abortive infection (Abedon 2011).

4.3.2. Release of active prophage following antibiotic treated cultures versus spontaneously released phages.

In this work it was found that lysates from untreated as well as induced cultures could release phages. This is consistent to the previous TEM analysis performed in section 3.2.7. where active phages were detected in culture lysates, but the results in this chapter demontstrate lysates contain viable phages. Most of the published *C. difficile* phages have been isolated following MC induction (Sell *et al* 1983, Schallehn 1985, Mayer *et al* 2008, Goh *et al* 2005, Fortier and Moineau 2007). Mahoney *et al* (1985) isolated phages following their spontaneous release. Previously, the differential release of prophages using the two antibiotics MC and NFX was reported by Nale *et al* (2012) based on TEM and PFGE genome analysis. This implies that the phages released from these isolates may be repressed under slightly different control mechanisms. The phages in this work that were induced using NFX are the first propagating phages for *C. difficile* to be induced using this antibiotic.

The release of phages from strain K12 was observed during growth of both induced and untreated. It could be observed that there was a cost of phage induction in the treated culture compared the uninduced, but at the end of the growth curve, their growth had stabilised. The impact of a phages' switch from lysogenic to lytic lifecycle on a bacterial population is complex, as alongside lysis and reduction of the lysogen population there also may be an increase in virulence of the remaining bacterial population. This increase could be due to acquisition of novel/beneficial traits through horizontal gene transfer mediated by the released phages (as suggested by Allen *et al* (2011)), including antibiotic resistance which was demonstrated after exposure of the pig microbiome to antibiotics. Also of importance is the mechanistic effect of cell lysis following

prophage induction, shown to facilitate *Streptococcus pneumoniae* biofilm production, a known *in vivo* virulence factor (Carrolo *et al* 2010). Therefore, consideration of the implications of prophage release *in situ* of *C. difficile* both through spontaneous or antibiotic induction is necessary since such could occur within treatment of clinical patients, farmyard livestock and the environment.

4.3.3. Characterisation of the phages reveals their diversity and allows sub-classification.

The phages isolated in this study were characterised according to their morphology and genome sizes, as well as their origin and host indicator strains (Figure 13 and Table 10). TEM analysis showed that all the phages belong to the *Caudovirales*. They can be grouped into four main subtypes based on particle morphology: the medium myoviruses (MMs); long tailed myoviruses (LTMs); small myoviruses (SMVs) and siphoviruses (SVs). *C. difficile* phages in the literature can be included into these classifications, for example, the MMs phiC2 and phiCD119 are MMs, phiCD27 and phiMMP02 are LTMs, phiMMP04 is a SMV and phiCD6356 and phiCD28-3 are SVs (Goh *et al* 2005, Govind *et al* 2006, Mayer *et al* 2008, Meessen-Pinard *et al* 2012, Mahoney *et al* 1985, Goh *et al* 2005, Horgan *et al* 2010, Sekulovic *et al* 2011). In the literature there has been a report of a putative podovirus (Mahoney *et al* 1985) but no podoviruses were isolated in this study.

The PFGE analysis showed that the genome sizes correlate to phage particle morphology. The LTMs and MMs have genome sizes of 50-55 kbp, with most variation in the MM group. The published MMs and LTMs have varied genome sizes from 44-54 kbp (Goh *et al* 2005, Govind *et al* 2006, Goh *et al* 2007, Mayer *et al* 2008, Meesen-Pinard *et al* 2012). The siphovirus genome size obtained is 40 kbp, which compares to that of the other siphoviruses genome sizes reported: 36 kbp (Goh *et al* 2005), 41 kbp (Sekulovic *et al* 2011) and 38 kbp and 50 kbp (Horgan *et al* 2010). The SMV genome sizes of 30-35 kbp are considerably smaller. The published phage phiMMP02 is also this size and was isolated from patient stools (Meessen-Pinard *et al* 2012). The abundance of SMVs in the enrichments may indicate this morphology is more stable as free phage particles.
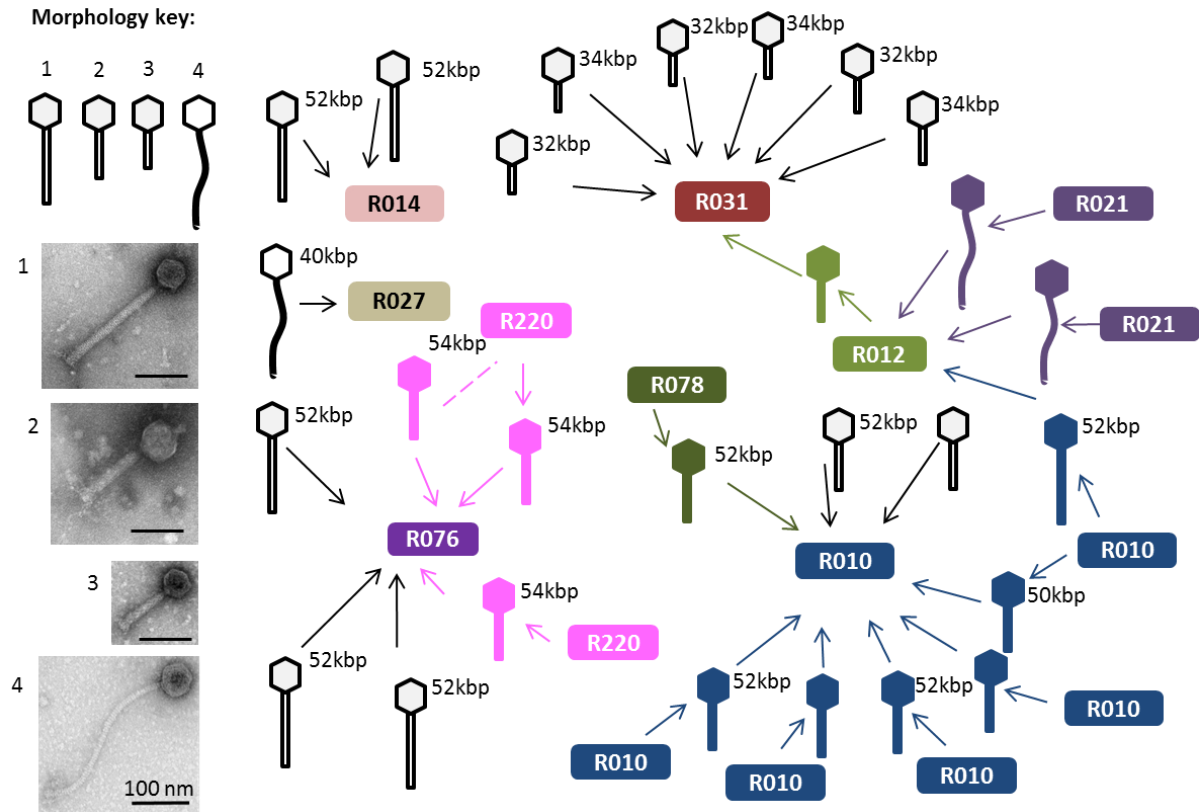
Figure 13. Isolated phages: their diversity and network of infecting strains. Virion particles are shown according to morphology. Infection is indicated by an arrow from phage capsid to bacterial cell. If a phage has been induced from antibiotic treatment arrow from its host to the phage (and this line is dashed if phage was isolated from an uninduced lysate). Colours of phages and cells correlate to ribotype. Each isolate is a separate cell. Where phages have been isolated from enrichment broths and their host unknown the phages are colourless. Genome sizes where known are provided by capsids.

Table 10. Phages isolated in this study with assigned characteristics indicating their diversity. Phage numbers correspond to those in Figure 9.

	Phage	Origin	Particle morphology	Plaque morphology	Genome size(kbp)	Ribotype host/origin	Indicator host prophage/s
1	phiCDHM1	L	M	C	54	076/220	M
2	phiCDHM3	L	M	C	54	076/220	M
3	phiCDHM7	L	M	SC	52	076/220	M and S
4	phiCDHM8	L	M	T	50	010/010	M and S
6	phiCDHM18	L	M	C	52	076/220	M
7	phiCDHM19	E	M	T	52	010/-	M and S
8	phiCDHM20	L	M	SC	52	010/010	M and S
9	phiCDHM21	L	M	T	-	010/010	M and S
10	phiCDHM22	L	M	T	-	010/002	M and S
11	phiCDHM23	L	M	C	-	010/010	M and S
12	phiCDHM2	E	LTM	C	52	076/-	M
13	phiCDHM4	L	LTM	SC	54	012/010	M
14	phiCDHM5	E	LTM	C	52	014/-	-
15	phiCDHM6	E	LTM	C	52	014/-	-
16	phiCDHM15	E	LTM	C	52	076/-	M
17	phiCDHM16	E	LTM	C	52	076/-	M
18	phiCDHM9	E	SM	C	34	031/-	PTLPS
19	phiCDHM10	E	SM	SC	32	013/-	PTLPS
20	phiCDHM11	E	SM	C	34	031/-	PTLPS
21	phiCDHM12	E	SM	T	32	031/-	PTLPS
22	phiCDHM13	E	SM	C	32	031/-	PTLPS
23	phiCDHM14	E	SM	C	34	031/ -	PTLPS
24	phiCDHM24	L	SM	C	-	031/012	PTLPs
25	phiCDHS1	E	S	SC	40	027/-	-
26	phiCDHS2	L	S	SC	-	012/021	M
27	phiCDHS3	L	S	SC	-	021/021	M

The plaque morphology of the phages did not appear to reflect to particle morphology or origin. Interestingly there was no overall trend for of plaque size for the phage types, for example if the smaller genome phages, the SMVs, produced larger plaques as might have a shorter replication cycles. Production of turbid lysis indicates the phage is lysogenic and these included those both from culture lysates as well as enrichment cultures. The results are consistent with observations of phage-host interactions in other species. Plaque formation is influenced by the specific phage and host relationship, and as such, burst size, phage attachment, latent period and the diffusion of phage

particle through the media (Abedon and Yin 2009). Only when conditions are standardized can the contribution of singular factors can be assessed independently (Gallet *et al* 2011). Their findings indicate that both absorption rate and morphology of phage affect plaque sizes, and that latent periods have a concave correlation, and therefore, phage traits directly influence plaque morphology. The infections in this system are not standardised and are unlikely to represent the optimum for any given phage and host. Instead, the host-phage interactions observed appear to be situated along a shifting scale ranging from their optimum phage production to a host-phage interaction which produces few phages, due to the variation in plaque morphologies observed between phage sub-types and host strains. A report by Hächler and Wüst (1984) described different plaque sizes of a single phage on different *C. difficile* strains during phage typing, highlighting the close host-phage relationship on plaque morphology in this system.

Unfortunately the DNA digestion assays were not successful and failed to produce genetic profiles for comparison. Other *C. difficile* phage genomes have been cleaved using the restriction enzyme *Hind*III (Goh *et al* 2005, Govind *et al* 2006, Fortier and Moineau 2007, Sekulovic *et al* 2011). Phage DNA can be heavily modified through methylation as a way to evade the host's own endonucleases, although enzymes can be found that can cleave common modifications (Huang *et al* 1982). Also, phages can incorporate modified bases in their DNA which can protect against specific nucleases but not all do (Warren 1980).The failure in this study to obtain useful RE profiles for comparison may be due to insufficient optimisation of the protocol or DNA purification or possible novel DNA methylation during DNA replication in the tested phages.

4.3.4. Host range analyses of the phages provide insight into the fundamental biology of *C. difficile*.

Results of the host range analyses performed in this project demonstrate that the phages infect multiple and diverse strains of *C. difficile*, including clinical and environmental strains. The analyses also revealed several observations that have biological significance. Firstly, that phages could be identified that have narrow or wider host ranges, and that they could be designated as specialist or generalist phages. The observation of generalist phages is unusual in *C. difficile* (e.g. Meyer *et al* 2008). For potential therapeutic purposes it would be assumed that generalist phages may be more useful in order to effectively kill a greater number of diverse *C. difficile* strains. However, the narrower host range phages often infected isolates, or ribotypes, that other phages did not and highlight the fact that a combination of phages would be most useful in any future therapeutic approach for *C. difficile*. In other species, phage cocktails have been successfully developed which show promise for effective treatment including *Staphylococcus aureus (*Kelly *et al* 2011) and *Klebsiella pneumoniae* (Gu *et al* 2012). All therapeutic studies performed for *C. difficile* have used only single phages (Ramesh *et al* 1999, Meader *et al* 2010, Govind *et al* 2011) and investigation whether a phage cocktail could be more effective is needed.

Interestingly, there was extensive intra-ribotype variation in the host ranges that could be observed where there were more than one isolate belonging to the same ribotype. This highlights the genetic diversity present within ribotypes and that differential phage predation could shape *C. difficile* strain evolution within even ribotype groups. This observation is consistent with previous host ranges for other *C. difficile* phages (Sekulovic *et al* 2011, Mayer *et al* 2008, Horgan *et al* 2010). Whole genome comparisons of *C. difficile* strains have shown that presence of core genes corresponds to ribotype (He *et al* 2010), however genes involved in phage resistance may be exchanged horizontally and be confined to individual ribotypes, or present in all representative isolates. Additionally, isolates of the same ribotype could also encode different CRISPR arrays in their

CRISPR/cas system, which is a form a acquired bacterial resistance that incorporates spacers from past, unsuccessful, phage infections (Pourcel *et al* 2005, Bolotin *et al* 2005, Sorek *et al* 2008, Horvath *et al* 2009). *C. difficile* encodes several CRISPR/cas loci and CRISPR content can vary between isolates of the same ribotype, as seen using CRISPRcompar tool (Grissa *et al* 2008). Lastly, prophage can confer protection for their hosts, by inhibiting secondary infection of phages and in *C. difficile*, the prophage content of strains may considerably impact phage sensitivity, as different prophages are encoded in isolates of the same ribotype. Whatever resistance mechanisms *C. difficile* uses to inhibit phages, they appear to either be subject to horizontal gene transfer or evolve rapidly.

The host range analyses demonstrate that a large panel of phages are required to maximise the number of *C. difficile* isolates infected as well as their diversity. Several ribotypes present in the clinical which were not in the environmental screens (and vice versa). As the phages were all isolated from environmental or non-clinical sources, it was interesting that the phages had mixed abilities to infect the isolates only from the clinical patients in this study, for example R107 was not recovered from the environmental samples yet four phages could infect isolates of it, but in the case of R103 no phages could infect it. The increased rate of phage infection observed in the clinical screen compared to the environmental may be due to the fewer ribotype groups tested (21 versus 27). There may be bias in the results from the different sample sizes of each ribotype in the screens but the ability of later isolated phages to infect a larger proportion of clinical isolates may reflect the altered strategy in isolating phages on clinical ribotypes. The host range analysis does suggest that these phages could serve as useful tools with the potential translation applications such as phage therapy, diagnostics or a source of useful genes for working with different *C. difficile* strains.

4.3.5. Variation in the efficiency of plaquing suggests there are tight phage-host interactions in *C. difficile*.

In the subset of phages examined it could be seen that phages could infect in a lytic manner on one host, but in a lysogenic manner on another, related, isolate. The results of this observed lysogenic/lytic switch are consistent to other reports for another *C. difficile* phage, phiCD6356 (Horgan *et al* (2010). This suggests that the switch between lytic and lysogenic lifestyles for many *C. difficile* phages is closely balanced, again emphasis the tight relationship between strains and phages.

Also there were observed differences between the efficiency of plaquing (EOP) on strains infected by the same phage. Multiple phage defence mechanisms can alter the efficiency in which phage can infect a cell (Hyman and Abedon 2010). These are normally associated with full immunity such as restriction/modification response, phage-genome uptake blocking, CRISPR/cas system and abortive infection. The ability of such defences reducing EOP without fully inhibiting phage infection across the bacterial population *in situ* is considered underappreciated (Hyman and Abedon 2010).

The change in EOP reflects the host-phage interaction rather than the viability of phages in the stock. However, difficulties in producing high titre stocks of these phages from plaque assays and liquid culture were experienced. *C. difficile* phages are notoriously difficult to manipulate in the laboratory and often require optimisation of techniques to isolate and propagate them (e.g. Mahoney *et al* 1985, Goh *et al* 2005, Horgan *et al* 2010). Also experienced are the difficulties in working with *C. difficile* strains to permit phage work, for example frequently there are inconsistencies in lawn growth (Dei 1989).

The plaque morphologies assigned in this chapter are those for each phage when infecting their indicator hosts. The hosts chosen to act as propagating hosts were which showed clear plaques where possible, in order to promote high titre phage production, something which has been difficult

to achieve with many of these phages, the highest titres often at 10^8-10^9 PFU ml^{-1}. In other species, producing high titre stocks of temperate phages is known to be problematic (e.g. Serwer *et al* 2007). Screening for optimum hosts may help to increase these titres, as well as infecting liquid cultures at lower OD (Horgan *et al* 2010). Further optimisation of phage infection protocols is required if *C. difficile* phages are to be potential therapeutic products and this work is currently being undertaken in the laboratory with these phages.

4.4. Summary

27 phages associated with environmental *C. difficile* strains were successfully isolated. They include known temperate phages which were isolated following their induction from a host lysogen and lytic replication on an indicator strain. They also include phages that were isolated from enrichment cultures and could be temperate or virulent. Different approaches used to isolate *C. difficile* phages were performed with the aim to maximise the number, (by using the most permissive hosts detected in previous isolation rounds) as well as the diversity of phages isolated (by the specific targeting of ribotypes that no previous isolated phages could infect). This has resulted in the largest reported collection of *C. difficile* phages and the first that is associated with environmental strains. Morphological characterisation of the phages showed that they are similar to known *C. difficile* phages, but on the basis of genome size, DNA digest and origin they are a diverse panel.

Host range analyses showed that the phages can infect both non-clinical and clinical strains. They also revealed that there are generalist and specialist phages within the panel and that some phages may be closely related as have a similar host range pattern. Also, it was seen that there is extensive variation between isolates of ribotype in regards to phage sensitivity, suggesting that the genetic basis of phage resistance may be horizontally transferred throughout the *C. difficile*

population or evolve rapidly. This has implications for the potential impact of phages in this species and the extent phages can shape the genome evolution of this pathogen.

Their host ranges individually are typically narrow but collectively the phages can infect multiple strains and ribotypes demonstrating their potential usefulness for a translational purpose either therapeutically or in diagnostic applications. Further characterisation will determine if these phages are virulent or temperate phages (for those not induced from strains of *C. difficile*) as well as to ensure their genomes do not encode putative virulence factors. Genomic sequencing will also reveal their genetic content and provide the sequence data to enable phylogenetic analyses and comparative genomics to be performed in order to assess their diversity.

5. Genome annotation of seven *C. difficile* phages.

5.1. Introduction

To date there are several *C. difficile* phage and prophage sequences available in public databases and all infect clinical, human or livestock strains of *C. difficile* (Goh *et al* 2005, Sebaihia *et al* 2006, Stabler *et al* 2006, Govind *et al* 2007, Mayer *et al* 2008, Stabler *et al* 2009, He *et al* 2010, He *et al* 2010, Horgan *et al* 2010, Sekulovic *et al* 2011, Meessen-Pinard *et al* 2012). Despite the prevalence of *C. difficile* in environmental sources, this project is the first study to isolate phages associated with these strains.

The first complete genome sequence of a *C. difficile* phage was obtained for the myovirus phiCD116 (Govind *et al* 2006), in addition to two the prophage sequences encoded in *C. difficile* strain CD630 (Sebaihia *et al* 2006). Four more myoviruses have been sequenced since; phiC2 (Goh *et al* 2007), phiCD27 (Mayer *et al*, 2008), phiMMP02 and phiMMP04 (Meessen-Pinard *et al* 2012), and two siphoviruses: phiCD6356 (Horgan *et al* 2010) and phiCD38-2 (Sekulovic *et al* 2011). All the phage genome sizes fall within 30-55 kbp.

Sequencing has shown that the phage genomes encode genes identified with typical phage essential functions, such as endolysin, DNA methylases, Tape measure protein and terminase genes. Proteomic analysis of their structural proteins has identified putative capsid and tail genes also (Goh *et al* 2006). There are no obvious virulence factors identified in the genomes of the 'lytic' phages, but there is a putative VirE, an unknown virulence associated gene, in one of the prophage sequences in NCBI, CD_1450 in phiCD196. They do however encode some unusual genes, such as the AbiF gene in phiC2 which is a predicted phage abortive infection protein. Also, the sequence similarity of the holin gene in phiC2 to the *tcdE* gene which is encoded on the PaLoc led researchers to suggest that it may have a phage origin or been transferred via phages (Goh *et al* 2007).

Whether these environmental *C. difficile* phages differ from those associated with clinical, human or livestock strains on a genetic basis is therefore completely unknown. Also, all the published phage genomes encode integrase genes and no truly virulent phage has yet been described. In order to ascertain whether there are genomic differences between environmental and clinical associated phages, as well as to identify key genes such as integrase and toxin genes or genes characteristic of environmental *C. difficile* phages, whole phage genome sequencing and annotation was performed on six of the phages and the prophage from *C. difficile* isolate T6, isolated in this project.

Phages to be sequenced were selected according to their particle morphology, genome size, host range and origin in an effort to maximise the range of genomic diversity sampled and increase the chances of identifying virulent phages. Those selected include three temperate MMs, which were induced from different isolates and propagated through lytic replication on different hosts, and three SMVs that were isolated on the same host from enrichment cultures and may be virulent phages (Table 11). Additionally, the genome sequence of a prophage, phiT6, was obtained from the whole genome sequencing of the isolate T6 (R076). PhiT6 appears to be an active prophage, as the lysate from induced cultures of T6 produced phage particles belonging to the MMs.

Table 11. Summary of *C. difficile* phages sequenced in this study, their morphology and origin.

Phage	Morphology	Origin	Indicator Strain
phiCDHM1	MM	Lysogen (R220)	R076
phiCDHM19	MM	Lysogen (R012)	R010
phiCDHM23	MM	Lysogen (R010)	R010
phiCDHM13	SMV	Enrichment	R031
phiCDHM14	SMV	Enrichment	R031
phiCDHM11	SMV	Enrichment	R031
phiT6	MM	Prophage (R076)	N/A

In this chapter the full genome annotations in table and map format are presented for the phages phiCDHM1, phiCDHM19, phiCDHM13 and phiT6. Unusual or important features of the genomes are investigated further through phylogenetic analysis and a comparison of their genes to one another and to previously published *C. difficile* phages is conducted in the discussion

5.2. Results

5.2.1. Genomic annotation of the medium myovirus phiCDHM1.

PhiCDHM1 is a temperate MM with a dsDNA genome of 54,122 bp and encodes 84 predicted ORFs (Table 12). Predicted coding sequences account for 88.2% of the genome and the majority (75/84) of genes are orientated on the sense strand when starting the genome with the small subunit terminase gene. Of the 84 ORFs, 44 have been assigned a predicted function based on either sequence homology and/or identification of a protein domain; 38 have been assigned as hypothetical proteins without identified functions but share significant identity to genes in the NCBI database, and two have no significant hits to sequences in the NCBI database.

The genome is closely related to that of the other *C. difficile* myoviruses as seen from the entries in the top blastp hits. However, the genome displays a high degree of mosaicism as affinities vary on an individual gene basis, for example there are 8 ORFs with homology to genes in phiC2, 4 to phiCD27 and 3 to phiCD119. Despite this mosaicism, it can be seen that the genome is most closely related to the prophages encoded in *C. difficile* strain CD630 as 24 ORFs are homologous to their sequences with identities ranging from 61-100%.

In the genome there are several highly conserved ORFs (27/84) as indicated by having multiple entries as joint top blastp hits. For example, the endolysin gene has joint blastp hits to sequences from eight *C. difficile* phages/strains. The most conserved genes are generally located in the structural and lysis regions and the most divergent genes are located in the DNA replication

region according to the gene identities to their top blastp hits and occurrence of multiple joint entries.

Notably, there is a region encoding 3 unique ORFs, whose closest blastp hit are to bacterial genes in strains of *C. difficile*. The set of 3 ORFs encode predicted homologs of the accessory gene regulator (*agr*) quorum sensing (QS) proteins. These genes are not encoded in any other sequenced phage genome to date. Another ORF, located in the DNA replication region, is also homologous to a bacterial gene in *Clostridium kluyveri* strain DSM 555 and is predicted to encode an NTPase, which also has not been found in any other phage genome.

Table 12. Genome annotation for phiCDHM1. Symbols indicate sense (+) or antisense (-) strand. For top blastp hit where no bacterial species is indicated this is *C. difficile* strains, with their strain name.

ORF	Start	End	Aa size	Strand	Function	Domain	Top blastp hit	Identity %	E-value
1	1	675	224	+	terminase small subunit	COG5484	ATCC 43255	99	5.00E-124
2	653	1906	417	+	terminase large subunit	Terminase_3 (CL0023)	630	99	0
3	1912	3351	479	+	portal protein	Phage_prot_Gp6	630	99	0
4	3341	4834	497	+	minor head protein	Phage_Mu_F	630	97	0
5	4791	5006	71	+	hypothetical protein		630	99	3.00E-31
6	5113	5307	64	+	hypothetical protein		phiCD119	90	1.00E-13
7	5352	5915	187	+	minor structural protein	Phage_GP20	630	99	6.00E-101
8	5927	6826	299	+	major capsid protein		phiC2/QCD-63q42/ ATCC 43255	99	5.00E-174
9	7060	7470	136	+	hypothetical protein		630/CD196/R20291	97	3.00E-69
10	7464	7811	115	+	hypothetical protein		630	100	9.00E-60
11	7811	8236	141	+	tail protein	DUF646 (CL0348)	phiC2 / QCD-63q42	99	2.00E-77
12	8229	8666	145	+	hypothetical protein		630	95	7.00E-74
13	8659	8835	58	+	hypothetical protein		QCD-66c26/CIP 107932/QCD-76w55/ QCD-97b34/QCD-37x79/QCD-32g58	95	2.00E-21
14	8836	10146	436	+	sheath tail protein	Phage_sheath_1	ATCC 43255	96	0
15	10171	10641	156	+	core tail protein	DUF2001(CL0504)	phiC2	99	1.00E-85
16	10695	10952	85	+	hypothetical protein		ATCC 43255	84	2.00E-36
17	10957	11397	146	+	hypothetical protein	XkdN	phiC2	98	8.00E-79
18	11718	11843	41	+	hypothetical protein		ATCC 43255	100	1.00E-15
19	12152	12394	80	+	hypothetical protein		QCD-76w55	98	3.00E-35
20	12512	13399	295	+	anti-repressor protein	Bro-N, N terminal domain	QCD-66c26/CIP 107932/QCD-76w55/QCD-97b34/QCD-37x79/CD196/R20291	81	2.00E-131

21	13678	14448	256	+	hypothetical protein		QCD-66c26/QCD-76w55/QCD-97b34/QCD-37x79/CD196/R20291/QCD-32g58	50	2.00E-57
22	14518	16860	780	+	tape measure protein	tape_meas_nterm	QCD-63q42	85	0
23	16876	17565	229	+	cell wall hydrolase /LysM	LysM (CL0187)	ATCC 43255	95	7.00E-122
24	17558	19447	629	+	cell wall hydrolase	Phage_GPD and NLPC_P60 (CL0125)	QCD-66c26]/CIP 107932/QCD-76w55/QCD-97b34/QCD-37x79	86	0
25	19461	19718	85	+	hypothetical protein	DUF2577	630 /phiC2/ATCC 43255	91	1.00E-38
26	19726	20145	139	+	hypothetical protein	DUF2634	QCD-63q42	98	3.00E-72
27	20158	21192	344	+	BaseplateJ-like tail protein	Baseplate_J	QCD-63q42	94	0
28	21185	21802	205	+	tail protein	DUF2313	QCD-66c26/CIP 07932/QCD-76w55/QCD-97b34/CD196/R20291/QCD-32g58	93	9E-110a
29	21814	22617	267	+	putative tail fiber		630	97	2E-141a
30	22635	24329	564	+	hypothetical protein	Gly_Rich	630	83	0
31	24354	24647	97	+	hypothetical protein		630	92	1.00E-44
32	24647	24829	60	+	hypothetical protein		ATCC 43255/QCD-76w55/QCD-97b34	95	9.00E-24
33	24891	25322	143	+	Holin	Phage_holin_4	QCD-63q42/ATCC 43255/QCD-76w55/QCD-37x79	96	3.00E-73
34	25322	26134	270	+	Endolysin	Amidase_3 (CL0035)	QCD-66c26/CIP 107932/QCD-76w55/QCD-97b34/QCD-37x79/CD196/R20291	98	1.00E-151
35	26176	26289	37	-	*agrD*		QCD-23m63/ NAP08/NAP07	58	0.002
36	26315	26908	197	-	*agrB*	AgrB	NAP08/NAP07	56	3.00E-47
37	26901	28262	454	-	histidine kinase	HATPase_C	NAP08/NAP07	61	1.00E-144
38	28748	28936	62	-	sigma factor/hypothetical protein		phiC2/phiCD27/QCD-63q42/ATCC 43255	99	1.00E-25
39	29844	30233	129	+	putative regulatory protein	Penicillinase_R (CL0123)	ATCC 43255	82	4.00E-54
40	30273	30659	128	+	putative regulatory protein	Penicillinase_R CL0123	630	99	2.00E-66
41	31019	32167	382	-	phage integrase	Phage_integrase (CL0382)	630	100	0
42	32645	34213	522	-	hypothetical protein/regulatory protein	T5orf172 (CL0418)	630	97	0
43	34944	35309	121	-	XRE family transcriptional regulator/repressor	HTH_3 (CL0123)	QCD-63q42 /ATCC 43255	100	5.00E-61
44	35454	35690	78	+	XRE family transcriptional regulator, ORF44	HTH_3(CL0123)	QCD-63q42/ATCC 43255	100	1.00E-37
45	35683	36474	263	+	Orf6N like protein DNA binding/anti-repressor	ORF6N	QCD-63q42/ATCC 43255	100	1.00E-151
46	36533	36670	45	+	hypothetical protein		ATCC 43255	89	1.00E-14
47	36656	36967	103	-	hypothetical protein		ATCC 43255	96	1.00E-47
48	37071	37259	62	+	hypothetical protein		ATCC 43255	86	2.00E-20
49	37262	37912	216	+	anti-repressor	Bro-N N-terminal domain	phiC2	85	8.00E-101
50	37959	38144	61	+	hypothetical protein		630	96	1.00E-25
51	38183	38851	222	-	hypothetical protein		630	97	2.00E-113

52	38822	39019	65	+	repressor	HTH_3 (CL0123)	phiC2	99	1.00E-27
53	39036	39365	109	+	DNA binding protein/excisionase	HTH_17 (CL0123)	QCD-63q42/ATCC 42355	100	6.00E-54
54	39467	39667	66	+	hypothetical protein		630	89	5.00E-14
55	39664	39909	81	+	no significant hits		-	-	-
56	39992	40501	169	+	Gp-157 like protein	Sipho_Gp157	phiC2	99	9.00E-86
57	40511	41099	195	+	recombination protein	ERF	630	99	4.00E-107
58	41110	42009	299	+	DnaD/replication initiation/membrane attachment	HTH_23 (CL0123) DnaB_2	phiC2	94	7.00E-155
59	42070	42243	57	+	hypothetical protein		phiC2	100	4.00E-24
60	42259	42540	93	+	ssDNA binding protein	SSB (CL0021)	630	61	9.00E-38
61	42615	42869	84	+	hypothetical protein		phiCD119/CD630	100	4.00E-42
62	42943	43200	85	+	hypothetical protein		phiCD119	91	4.00E-38
63	43197	43562	121	+	hypothetical protein		phiCD119	91	4.00E-56
64	43566	43946	126	+	hypothetical protein		630	96	4.00E-63
65	43930	44076	48	+	hypothetical protein		630	98	9.00E-18
66	44063	44209	48	+	no significant hits		-	-	-
67	44206	44379	57	+	hypothetical protein		phiCD119/CD630	100	1.00E-24
68	44379	44504	41	+	hypothetical protein		QCD-63q42/ATCC 43255	79	3.00E-08
69	44497	44832	111	+	hypothetical protein		630	100	2.00E-56
70	44932	45207	91	+	hypothetical protein		630	100	2.00E-44
71	45303	46376	357	+	C-5 cytosine-specific DNA methylase	DNA_methylase (CL0063)	630	99	0
72	46453	46674	73	+	hypothetical protein		QCD-63q42/ATCC 43255	69	9.00E-19
73	46671	47675	334	+	hypothetical protein		phiC2	92	1.00E-168
74	47650	47976	108	+	hypothetical protein		QCD-63q42	37	6.00E-09
75	47996	48163	55	+	hypothetical protein		QCD-63q42/ATCC 43255	81	2.00E-14
76	48157	48513	118	+	putative transcriptional regulator, XRE family, bacteriophage lambda repressor C1 like protein		QCD-76w55	37	0.002
77	48607	48945	112	+	peptide chain release factor/hypothetical protein	RF-1 (CL0337)	phiCD27	94	6.00E-53
78	48961	49074	37	+	hypothetical protein		phiCD27	87	2.00E-06
79	49071	49478	135	+	endodeoxyribonuclease	RusA	QCD-63q42/ATCC 43255	97	1.00E-68
80	49493	50218	241	+	anti-repressor	AntA	630	99	3.00E-135
81	50312	50800	162	+	possible sigma factor /hypothetical		phiCD119/CD630	100	5.00E-87
82	51182	51307	41	+	hypothetical protein		phiCD27	100	3.00E-15
83	51277	52038	253	+	regulatory protein	Bro-N	phiCD27	100	2.00E-128
84	52136	54211	691	+	signalling protein/NTPase,	NACHT (CL0023)	*Clostridium kluyveri* DSM 555	47	2.00E-147

The genome follows a modular arrangement according to predicted gene functions associated with different stages of the phage lifecycle: packaging; structural; attachment and lysis; lysogenic conversion; lysogeny control and DNA replication (Figure 14). It has an average GC% of 28.4 but there are regions throughout the genome with lower and higher than average GC% content. The GC skew indicates that the origin of replication is near to the integrase gene.
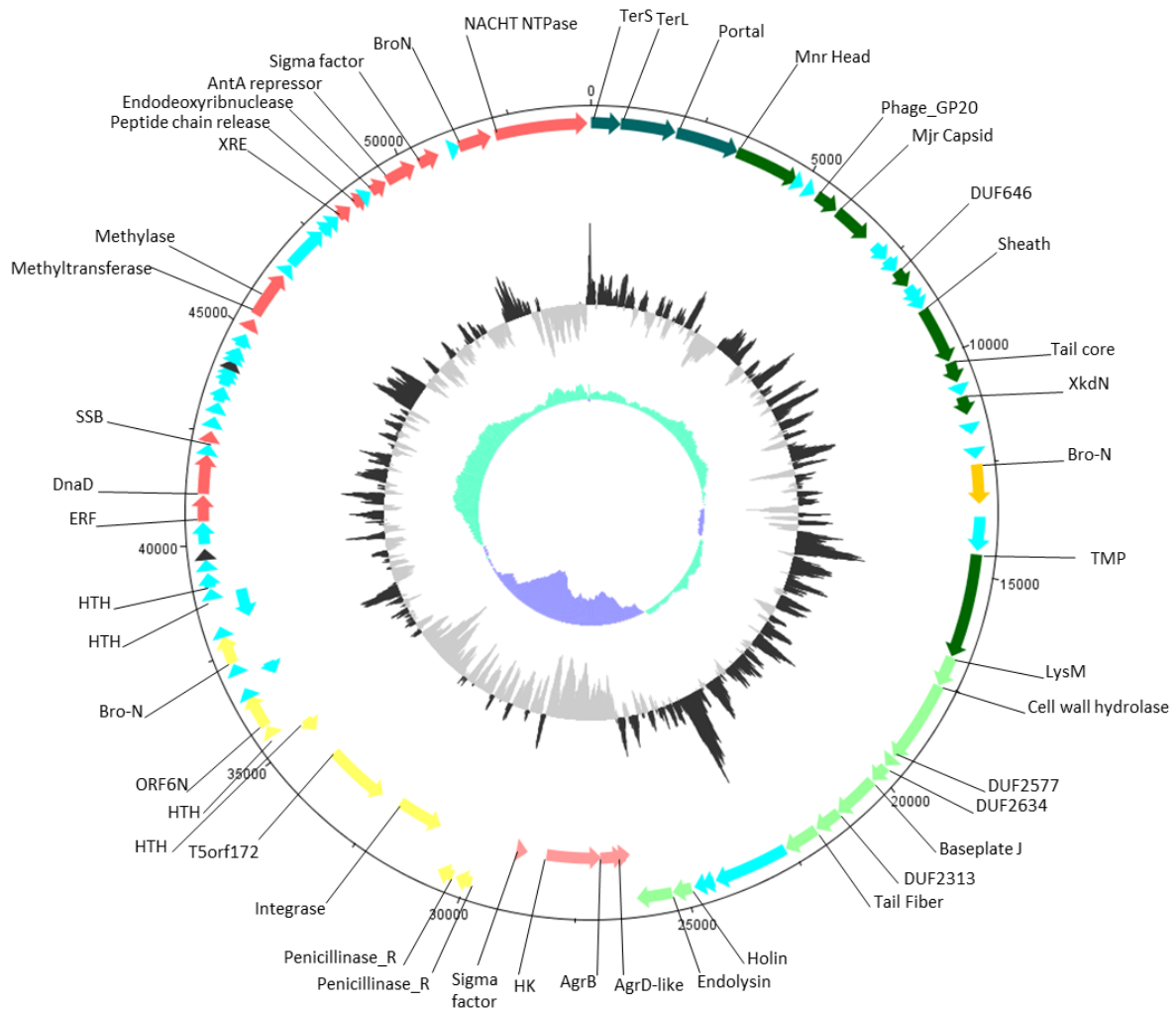
Figure 14. Genome map of phiCDHM1. Circular double stranded genome map with GC skew, GC content and modular regions of the genome shown drawn in DNAplotter (Carver *et al* 2009). Track 1 = GC skew, light green is above average and purple is below average. Track 2 = GC%, dark grey is above average and light grey is below average. Track 3 (antisense) and Track 4 (sense) encode ORFs. Different colours of genes indicate functional modules: head packaging (blue); structural (green); lysis and attachment (light green); lysogenic conversion (pink); lysogeny control (yellow) and DNA replication (red). Aqua indicates hypothetical proteins, orange is a regulatory protein in the structural region and charcoal are ORFs with no significant hits.

A linear 'heat' map of the genome has been generated with ORFs coloured to indicate their individual GC% (Figure 15). The structural genes have mostly >26 GC% content, whereas ORFs in the lysogeny conversion and control regions have a lower GC%. The average GC% in *C. difficile* strains ranges from 28.4 to 29.1, although this is effected by the amount of mobile elements such as prophage and transposons encoded in the genomes (Sebaihia *et al* 2006, Stabler *et al* 2009, He *et al* 2010, Monot *et al* 2011, Brouwer *et al* 2012). Comparison of the average phage and host gene GC% shows which genes may be considered typically phage genes, and which may have been acquired from the host genome, such as the QS homologs. These three genes have significantly lower GC% on average than genes in the structural, lysis and attachment and DNA replication regions (p=0.009321011, 0.010523 and 0.02602586 respectively). The heat map provides insight into the evolutionary patterns for different genes and regions of the phiCDHM1 genome, highlights which genes may have been acquired or evolved more recently, as well as shows in which modules gene acquisition is most rapid.
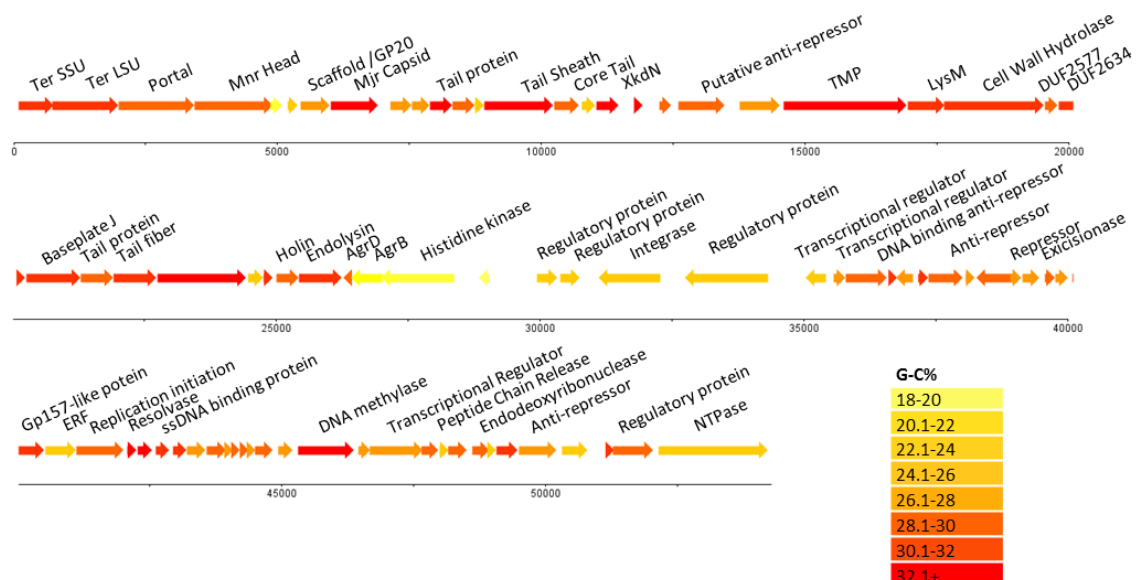
Figure 15. Linear genome map of phiCDHM1 with GC%. The genome follows a modular organisation according to assigned function of genes, head packaging, structural genes, lysis and attachment genes on the positive strand with relatively high GC%, followed by lysogenic conversion genes and lysogeny control on the antisense and sense strands of the genome with lower GC%, the QS homolog genes have the lowest GC%, and then DNA replication genes which are more mixed in their GC%.

5.2.2. Genomic annotation of the medium myovirus phiCDHM19.

PhiCDHM19 is a temperate medium myovirus with a dsDNA genome that was induced from isolate 10 (R012) and propagated in a lytic manner on an isolate belonging to R010. Assembly of the sequencing data resulted in several contigs that were joined into 2 contigs (the final result), which have a combined length of 53,956 bp. 88 ORFs were identified and predicted coding sequences cover 86.8% of the genome (Table 13). The majority of ORFs (80/88) are encoded on the sense DNA strand. The ORFs on the antisense strand are mainly clustered between ORFs 43-50 and encode the integrase gene, a transcriptional regulator gene, a putative repressor gene and four hypothetical protein genes. In total 35 of the 88 ORFs have been assigned putative functions, 49 are hypothetical proteins and four have no significant matches to database entries.

The genome of phiCDHM19 is mosaic and gene affinities vary to different *C. difficile* myoviruses and strains, however it is most related to phiCD119 as 37/88 ORFs have genes in this phage as their top blastp. These ORFs are mainly located in the structural and replication regions of the genome, and have high sequence identity, between 86-100%. Interestingly, fewer genes in phiCDHM19 are as highly conserved when compared to those encoded in the genome of phiCDHM1, with 18/88 ORFs having multiple joint entries in their top blastp hits.

Notably, of the 30 ORFs in the structural region, 25 have phiCD119 as their top blastp hit, but of the remaining five ORFs, two are not present in any of the previously sequenced *C. difficile* phage or prophage genomes. These are Orf10, a predicted Rho termination which has 56% identity to *Streptococcus suis* R61 and Orf7, a hypothetical protein with 36% identity to *C. kluyveri* DSM 555. Additionally, in the rest of the genome there are two more ORFs that do not share identity to any *C. difficile* strain. These, Orf51 and Orf58, are located in the DNA replication and both encode hypothetical proteins which have low identity to genes from *Enterococcus faecalis* Fly1 and *Tepidanaerobacter* sp. Re1 (38% and 40% respectively).

Table 13. Genome annotation for phiCDHM19.
Symbols indicate sense (+) or antisense (-) strand. For top blastp hit where no bacterial species is indicated this is *C. difficile* strains, with their strain name.

ORF	Start	End	AA size	Strand	Function	Domain	Top blastp hit	Identity %	E-value
1	1	456	152	+	small subunit Terminase		QCD-63q42	91	1.00E-79
2	443	2842	800	+	large subunit Terminase	Terminase_6 (CL0023)	phiCD119	99	0
3	2857	4212	452	+	putative portal protein	Phage_prot_Gp6	phiCD119	99	0
4	4225	5265	347	+	putative head protein	Phage_Mu_F	phiCD119	99	0
5	5333	5962	210	+	hypothetical protein		phiCD119	100	4.00E-150
6	5984	6178	65	+	hypothetical protein		phiCD119	100	1.00E-34
7	6220	6450	77	+	hypothetical protein		*Clostridium kluyveri* DSM 555	36	7.00E-06
8	6571	7167	199	+	putative scaffold protein/minor capsid	Phage_GP20	phiCD119	86	2.00E-111
9	7191	8129	313	+	hypothetical protein		phiCD119	100	0
10	8142	8363	74	+	hypothetical Rho termination factor	Rho_N	*Streptococcus suis* R61	56	5.00E-07
11	8392	8742	117	+	head connector	Phage_connect_1	phiCD119	100	1.00E-55
12	8744	9094	117	+	hypothetical protein		phiCD119	100	2.00E-80
13	9095	9505	137	+	hypothetical protein	DUF646 (CL0348)	phiCD119	100	2.00E-81
14	9516	9968	151	+	hypothetical/head connector		phiCD119	99	3.00E-104
15	9969	11039	357	+	phage sheath protein	Phage_sheath_1	phiCD119	99	0
16	11054	11485	144	+	core tail protein	DUF2001 (CL0504)	phiCD119	100	4.00E-100
17	11517	11987	157	+	hypothetical protein	XkdN	phiCD119	100	7.00E-107
18	12014	12160	49	+	hypothetical protein		QCD-76w55/QCD-97b34	70	1.00E-12
19	12164	12429	88	+	tape measure protein		phiCD119	100	4.00E-50
gap	12430	12768	-	-	-	-	-	-	-
20	12769	14997	730	+	tape measure protein		phiCD119	99	0
21	15068	15820	251	+	hypothetical protein		phiCD119	99	1.00E-139
22	16565	16756	64	+	hypothetical protein		QCD-63q42	60	3.00E-15
23	16780	17409	210	+	LysM	LysM (CL0187)	QCD-63q42/QCD-76w55/QCD-97b34/QCD-37x79/	73	5.00E-114
24	17415	17696	94	+	hypothetical protein		phiCD119	100	5.00E-65
25	17717	17959	81	+	hypothetical protein		phiCD119	100	6.00E-52
26	18218	19738	507	+	cell wall hydrolase	NLPC_P60 (CL0125)	phiCD119	99	0
27	19754	20080	109	+	hypothetical protein		phiCD119	99	5.00E-71
28	20080	20508	143	+	tail protein	DUF2634	phiCD119	100	1.00E-83
29	20501	21553	351	+	baseplate-J protein	Baseplate_J	phiCD119	98	0
30	21553	22161	203	+	tail protein	DUF2313	phiCD119	100	3.00E-147
31	22169	23179	337	+	putative tail fiber protein		phiC2	75	0
32	23196	24884	563	+	hypothetical protein	Gly_rich	ATCC 43255	81	0
33	24898	25191	98	+	hypothetical protein		ATCC 43255	93	5.00E-58
34	25191	25373	61	+	hypothetical protein		phiC2	95	1.00E-32

35	25407	26171	255	+	hypothetical protein		phiCD27/QCD-63q42/ATCC 43255	44	8.00E-42
36	26211	26666	152	+	no sig hits		-	-	-
37	26666	26965	100	+	hypothetical protein		QCD-63q42/QCD-37x79	87	4.00E-59
38	26967	27104	46	+	hypothetical protein /XkdK	Phage_XkdX	phiCD27/QCD-63q42/ATCC 43255	96	1.00E-22
39	27166	27597	144	+	Holin	Phage_holin_4	QCD-63q42/ATCC 43255/QCD-76w55/QCD-97b34/QCD-37x79	96	7.00E-96
40	27597	28409	271	+	Endolysin	Amidase_3 (CL0035)	phiC2	100	0
41	28490	28852	121	+	hypothetical protein		phiC2/phiCD27	100	4.00E-81
42	29018	29536	173	+	hypothetical protein		phiC2/ phiCD27/QCD-63q42/ATCC 43255	100	2.00E-115
43	29600	31186	529	-	Hydrolase	Nudix_Hydrolase_39	phiC2/phiCD27	100	0
44	31162	31683	174	-	hypothetical protein	DUF955 (CL0126)	phiC2/ phiCD27/QCD-63q42/ATCC 43255	100	2.00E-126
45	32385	32573	63	-	hypothetical protein		phiC2/ phiCD27/QCD-63q42/ATCC 43255	100	5.00E-36
46	33503	33886	128	+	transcriptional regulator	Pencillinase_R (CL0123)	630	91	3.00E-70
47	34243	35361	373	-	Integrase	Phage_integrase (CL0382)	phiCD119	96	0
48	35420	35821	134	-	hypothetical protein	UPF0150	phiCD119	10%	2.00E-94
49	36335	36472	46	+	no sig hits		-	-	-
50	36469	36864	132	-	RepR putative repressor	HTH_19 (CL0123)	phiCD119	100	1.00E-88
51	37017	37250	78	+	hypothetical protein		*Enterococcus faecalis* Fly1	38	1.00E-08
52	37255	37380	42	+	no sig hits		-	-	-
53	37449	37760	104	-	hypothetical protein		ATCC 43255	94	4.00E-63
54	37848	38051	68	+	hypothetical protein		ATCC 43255	88	2.00E-28
55	38054	38809	252	+	anti-repressor protein	Bro-N/ANT	phiC2	65	1.00E-82
56	38893	39087	65	+	hypothetical protein		phiC2	92	5.00E-35
57	39088	39285	66	+	no sig hits		-	-	-
58	39272	39472	67	-	hypothetical protein		*Tepidanaerobacter* sp. Re1	45	3.00E-08
59	39525	39722	66	+	putative repressor	HTH_3 (CL0123)	630	94	7.00E-36
60	39739	40068	110	+	putative repressor/excisionase	HTH_17 (CL0123)	phiCD119	100	2.00E-72
61	40069	40173	35	+	hypothetical protein		ATCC 43255	85	3.00E-13
62	40170	40364	65	+	hypothetical protein		630	95	5.00E-22
63	40367	40684	106	+	hypothetical protein		ATCC 43255	99	3.00E-64
64	40768	41277	170	+	Gp157-like protein	Sipho_Gp157	phiC2	99	3.00E-112
65	41287	41895	203	+	essential recombination function protein	ERF	phiCD119	98	4.00E-141
66	41896	42786	297	+	DnaD/replication protein	DnaB_2	phiCD119	95	0
67	42845	43018	58	+	hypothetical protein		630	100	2.00E-33
68	43034	43450	139	+	putative single strand DNA binding protein	SSB (CL0021)	phiCD119/phiC2	94	2.00E-91
69	43478	44041	188	+	hypothetical protein		phiCD27	100	8.00E-133
70	44034	44420	129	+	DNA-binding helix-turn-helix based	NUMOD1	phiCD27	100	4.00E-84

71	44420	44554	45	+	hypothetical protein		phiCD119/phiC2	100	6.00E-23
72	44556	44873	106	+	hypothetical protein		phiCD119	100	4.00E-67
73	44981	46054	358	+	C-5 cytosine-specific DNA methylase	DNA_methylase (CL0063)	630	99	0
74	46131	46352	74	+	hypothetical protein		QCD-63q42/ATCC 43255	68	8.00E-26
75	46349	47353	335	+	37 kDa nucleoid associated protein	NA37	phiC2	91	0
76	47328	47654	109	+	hypothetical protein		QCD-63q42/ATCC 43255	37	6.00E-12
77	47674	47841	56	+	hypothetical protein		630	78	8.00E-21
78	47838	48191	118	+	putative transcriptional regulator, XRE family		QCD-76w55	36	1.00E-04
79	48285	48623	113	+	peptide chain release factor RF1	RF-1 (CL0337)	phiCD27	94	2.00E-69
80	48749	49156	136	+	endodeoxyribonuclease	RusA	QCD-63q42/ATCC 43255	96	2.00E-89
81	49171	49896	242	+	AntA/AntB antirepressor	AntA	630	99	1.00E-174
82	49990	50478	163	+	hypothetical protein		phiCD119/630	100	1.00E-112
83	51205	51399	65	+	hypothetical protein		phiCD119/630/phiC2	98	6.00E-37
84	51399	52010	204	+	hypothetical protein		CD119/630/QCD-63q42/ATCC 43255	100	1.00E-143
85	52087	52275	63	+	hypothetical protein		630/phiC2/QCD-63q42/ATCC 43255	91	1.00E-65
86	52255	52572	106	-	hypothetical protein		630/QCD-63q42/ATCC 43255	100	9.00E-38
87	52678	53532	285	+	hypothetical protein		phiCD119	99	0
88	53670	54230	187	+	hypothetical protein		ATCC 43255	93	4.00E-113

The genome organisation of the phage is very similar to that of phiCDHM1 as shares the same arrangement of genes into functional modules (Figure 16). The average GC% of the genome is 28.42. There is a clear split between the genome in terms of GC% as the structural and attachment and lysis modules contain sequence with mostly higher GC%, but the remaining genome is mostly lower than average GC%. However a GC skew can still be observed and its location indicates that the origin of replication is near to the integrase gene, similar to phiCDHM1.
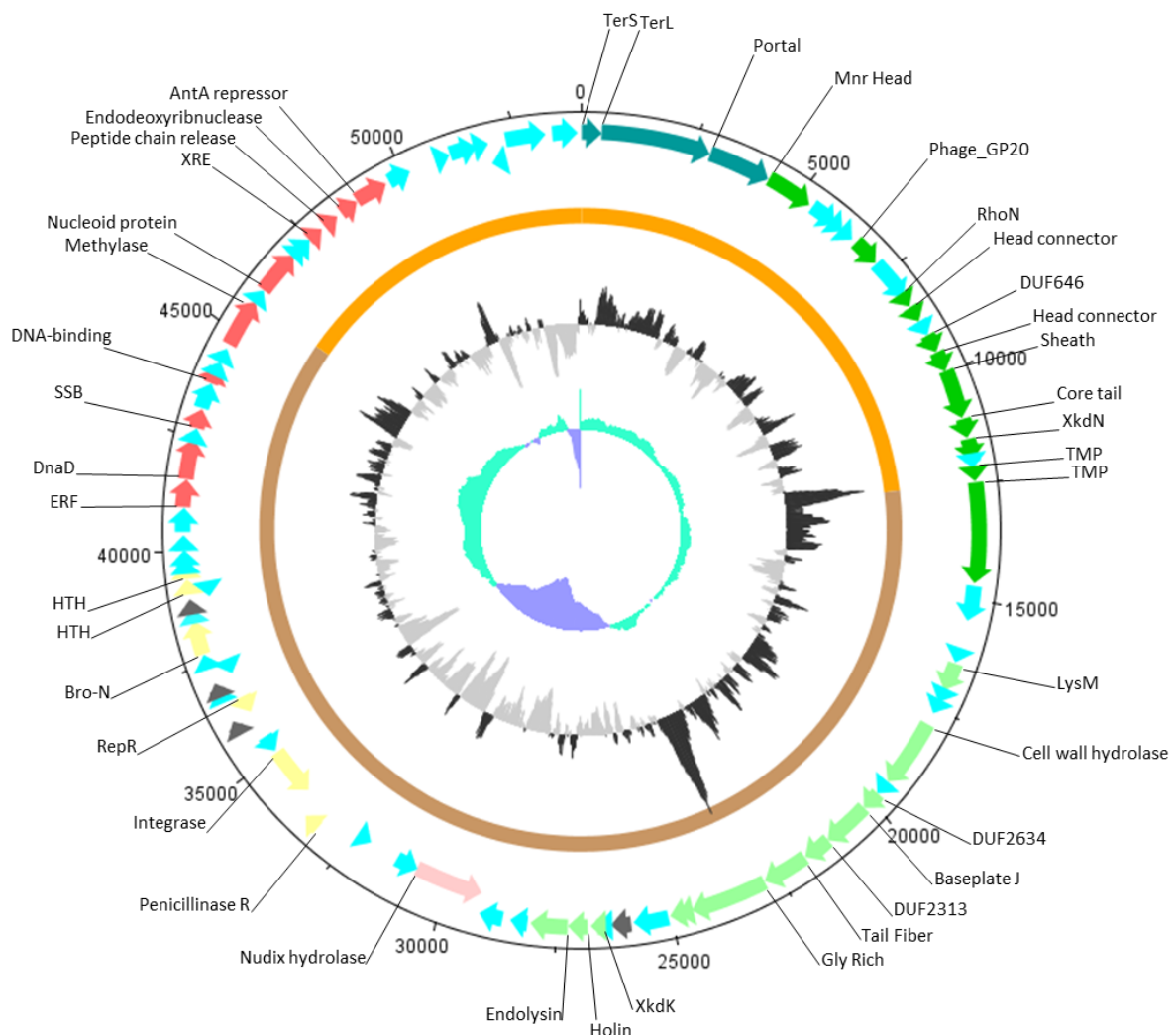
Figure 16. Genome map of phiCDHM19. Circular double stranded genome map with GC skew, GC% and modular regions of the genome shown drawn in DNAplotter (Carver *et al* 2009). Track 1 = GC skew, light green is above average and purple is below average. Track 2 = GC%, dark grey is above average and light grey is below average. Track 3 displays the contigs, orange is contig 1 and brown is contig 2. Track 4 (antisense) and Track 5 (sense) encode ORFs. Different colours of genes indicate functional modules: head packaging (blue); structural (green); lysis and attachment (light green); lysogenic conversion (pink); lysogeny control (yellow) and DNA replication (red). Aqua are ORFs with no predicted function assigned and charcoal are ORFs with no significant hits.

5.2.3. Genomic annotation of the medium myovirus phiCDHM23.

After the genome of phiCDHM23 was sequenced and the results analysed, it was clear that the sample contained two phages. The TEM characterisation of the phage had shown a single morphotype belonging to the MMs in the sample, and the sequencing revealed that this was actually two morphologically similar but genetically distinct phages (as seen from different LysM genes). However due to their sequence similarity and the mosaic nature of phage genomes it was not possible to assemble the reads further into larger contigs.

5.2.4. Genomic annotation of the small myoviruses: phiCDHM13, phiCDHM11 and phiCDHM14.

The phages phiCDHM13, phiCDHM11 and phiCDHM14 are SMVs with dsDNA genomes (all approximately 30-35 kbp) which were isolated from enriched environmental samples on the same host strain belonging to R031. Whole genome sequencing resulted in the assembly of several contigs for each phage. Because of a high level of sequence similarity between the three phage genomes it was possible to use larger contigs as scaffolds for others and then PCR to fill in gaps in order to close the assembled contigs to 2, 3 and 6 contigs, for phiCDHM13, phiCDHM11 and phiCDHM14 respectively. The annotations of the resulting contigs showed that the three SMVs are closely related to one another in their gene content and homologies but that they are genetically distinct from the other *C. difficile* myoviruses.

As the most complete, the genome annotation for phiCDHM13 is presented here to represent all three phages due to their genetically similarity. The genome of phiCDHM13 has been assembled into 2 contigs of 10,040 bp and 23,917 bp that total 33,959 bp. The combined contigs have 54 predicted ORFs which cover 88% of the contig sequences (Table 14). The majority of the ORFs (49/54) are encoded on the sense strand. In total 25 ORFs have been assigned a putative function, 26 encode hypothetical proteins and 3 have no significant hits. The gaps between contigs

are in the tape measure protein (TMP) gene and following Orf87 which encodes a hypothetical protein.

The genome of phiCDHM11 has been assembled into three contigs, 17,840 bp, 8,338 bp and 5,733 bp, which total 31,914 bp and together encode 54 putative ORFs. The first contig ends at Orf23, the second at Orf36 and the third after Orf53.

The genome of phiCDHM14 was assembled into 6 contigs which comprise of two larger contigs (10,588 bp and 20,567 bp) and 4 smaller contigs (632, 631, 540 and 501 bp). The two larger contigs were annotated and 48 putative ORFs were identified. The gaps between the two are located in the TMP protein gene (Orf13) and at Orf49.

Table 14. Genome annotation for phiCDHM13.
Symbols indicate sense (+) or antisense (-) strand. For top blastp hit where no bacterial species is indicated this is *C. difficile* strains, with their strain name.

ORF	Start	End	AA size	Strand	Function	Domain	Top blastp hit	Identity %	E value
1	1	1650	550	+	phage terminase	Terminase_1	*Clostridium cellulovorans* 743B	79	6E-87
2	1666	2820	385	+	portal protein	Phage_portal	*Clostridium hiranonis* DSM13275	64	4.00E-179
3	2826	3563	246	+	protease |(phage serine protease)	CLP_protease	*Bacillus cereus* AH187	53	1.00E-85
4	3556	4824	423	+	capsid protein	Phage_capsid (CL0373),	*Clostridium perfringens* E str. JGS1987	42	8E-104
5	4834	5055	74	+	head-tail connector protein	Phage_connect_1	*Clostridium botulinum* str. Loch Maree	47	7e-10
6	5126	5464	113	+	head-tail joining protein	Phage_H_T_join	*Clostridium hiranonis* DSM 13275	55	5.00E-31
7	5412	5870	153	+	DUF646	DUF646 (CL0348)	*Clostridium hiranonis* DSM 13275	54	9.00E-44
8	5876	6298	141	+	hypothetical protein		*Clostridium hiranonis* DSM13275	44	7.00E-26
9	6315	7382	356	+	sheath protein	Phage_sheath_1	*Clostridium hiranonis* DSM13275	46	6.00E-108
10	7396	7845	150	+	core protein /XkdM	DUF2001 (CL0504)	*Clostridium hiranonis* DSM13275	62	3.00E-56
11	7892	8275	128	+	XkdN	XkdN	*Brevibacillus brevis* NBRC 100599	39	2.00E-19
12	8293	8466	58	+	hypothetical protein		*Clostridium hiranonis* DSM 13275	36	2.00E-04
13	8459	10210	584	+	TMP		*Clostridium hiranonis* DSM13275	53	0
14	10213	10851	213	+	LysM	LysM (CL0187)	QCD-63q42/QCD-76w55/QCD-97b34/QCD-	62	3.00E-100
15	11091	12611	507	+	cell wall hydrolase	NLPC_P60 (CL0125)	QCD-63q42	82	0

16	12627	12953	109	+	hypothetical protein		phiCD119	85	4.00E-63
17	12953	13381	143	+	XkdS	DUF2634	630/ATCC423555	77	2.00E-78
18	13356	14426	357	+	XkdT protein	Baseplate_J	phiCD119	94	0
19	14423	15040	206	+	tail protein	DUF2313	phiCD119	85	3E-125
20	15041	15820	260	+	phage tail fiber		QCD-63q42/QCD-76w55/QCD-37x79	80	7E-137
21	15836	17116	427	+	GLY RICH hypothetical protein	Gly_rich	630	75	0
22	17116	17298	61	+	hypothetical protein		ATCC43255/QCD-76w55/QCD-97b34	95	3.00E-32
23	17335	17565	77	+	membrane protein/ Haemolysin XhlA	XhlA	630	99	1.00E-46
24	17582	17839	86	+	Holin	none	630	96	4.00E-53
25	17839	18651	271	+	Endolysin	Amidase_3 (CL0035)	ATCC 43255	94	0
26	19037	19486	150	+	no significant hits.		-	-	-
27	19498	20133	212	+	hypothetical protein	AP2 (CL0081)	*Eubacterium limosum* KIST612	37	4.00E-28
28	20210	20524	105	+	hypothetical protein		QCD-63q42/QCD-37x79/phiCD38-2	93	1.00E-61
29	20508	20681	58	+	hypothetical protein		QCD-63q42/QCD-37x79/phiCD38-2	78	4.00E-21
30	20992	21273	94	-	hypothetical protein		QCD-63q42/QCD-37x79	82	1.00E-51
31	21417	21749	111	-	hypothetical protein		QCD-63q42/QCD-37x79	99	1.00E-69
32	21750	22559	270	-	parA (ATPase)	CbiA (CL0023)	QCD-63q42/QCD-37x79	100	0
33	22690	22812	41	-	hypothetical protein		phiCD6356	80	4e-04
34	22898	23005	36	-	hypothetical protein		QCD-63q42/ATCC 43255	69	4e-09
35	23122	23859	204	-	Exonuclease	RNase_T (CL0219)	*Clostridium hiranonis* DSM 13275	55	3.00E-58
36	24291	24902	525	+	putative DNase	ResIII (CL0023)	QCD-63q42/QCD-37x79/phiCD38-2	89	2E-121
37	24871	26445	115	+	putative DNase		QCD-63q42/ QCD-37x79/phiCD38-2	89	4E-123
38	26460	26804	36	+	hypothetical protein		QCD-63q42/QCD-37x79/phiCD38-2	54	0
39	27133	27240	125	+	hypothetical protein		QCD-63q42/QCD-37x79/phiCD38-2	63	8.00E-05
40	27280	27654	61	+	transcriptional regulator, XRE family protein		QCD-63q42/QCD-37x79/phiCD38-2	66	3.00E-48
41	27830	28012	145	-	hypothetical protein	HTH_19 (CL0123)	QCD-63q42/QCD-37x79/phiCD38-2	52	1.00E-11
42	28078	28512	54	+	hypothetical protein		QCD-63q42/QCD-37x79/phiCD38-2	51	3.00E-09
43	28610	28772	115	+	hypothetical protein		phiCD6356	52	1.00E-09
44	28776	29120	132	+	hypothetical protein		phiCD6356	34	1.00E-04
45	29123	29518	49	+	single-strand DNA binding protein	SSB (CL0021)	630/QCD-66c26/CIP 107932/ATCC 43255/QCD-76w55/QCD-97b34/QCD-37x79/QCD-23m63/CD196/R20291/NAP08/NAP07/QCD-32g58/	61	5.00E-52
46	29576	29722	90	+	hypothetical protein		phiCD6356	84	2.00E-09
47	29832	30101	87	+	hypothetical protein		phiCD6356	97	4.00E-42

119

48	30120	30380	57	+	hypothetical protein		phiCD6356	92	7.00E-25
49	30629	30800	156	+	hypothetical protein /sigma70 factor		phiCD6356	92	1E-23
50	31196	31663	202	+	hypothetical /sigma F/70 - like factor		phiCD6356	42	3.00E-21
51	31806	32411	72	+	Integrase	Phage_integr ase (CL0382)	*Clostridium cellulovorans* 743B	74	4E-103
52	32608	32823	98	+	hypothetical protein		630	99	
53	32823	33116	160	+	putative HNH endonuclease	HNH (CL0263)	630	100	2.00E-65
54	33117	33596	122	+	hypothetical protein		*Clostridium beijerinckii* NCIMB 8052	58	3.00E-51

There is a clear division across the phiCDHM13 genome with respect shared with known *C. difficile* phage genes. For example, 36/54 ORFs have a top blastp hit to a *C. difficile* phage or strain sequence which are mainly located in or downstream of the lysis and attachment module. Several of these ORFs (*n*=17) share identity with the siphoviruses phiCD38-2 and phiCD6536, with these genes in their top blastp hits and range in identity between 34% and 92%. The most conserved ORFs with multiple joint hits to entries of *C. difficile* phage and strains are typically located in the lysis and attachment region of the genome. However, in the structural region of the genome, the ORFs have top blastp hits to sequences of other Clostridial species including *C. botulinum* and *C. hiranonis* strains, as well as other gram positive spore formers: *Bacillus cereus, Brevibacillus brevis* and *Eubacterium limosum* KIST612. The identities range between 36% and 79%. In addition, 5 ORFs, which are located in the lysogeny control and DNA replication regions, also have top blastp hits to sequences in other related bacterial species including *Eubacterium limosum* KIST612, *Clostridium hiranonis* DSM 13275 and *Clostridium cellulovorans* 743B. This apparent division between functional regions could indicate the divergence of the *C. difficile* phage evolution or co-evolution of different phage types to enable infection of the same bacterial species.

The three genomes of the SMVs follow a similar organisational arrangement to that of the MMs, except the positions of the lysogeny control and the DNA replication modules are rearranged. The GC% of phiCDHM13 is 29.95 which is slightly higher than in phiCDHM1 and phiCDHM19 (Figure 17). Similar to phiCDHM19 though, phiCDHM13 has higher than average GC% sequences throughout the structural regions and lysis regions of the genome but typically more sequence with lower than average GC% in the remainder of the genome. Like the MMs, the position of the GC skew indicates that the origin of replication is near to the integrase gene.
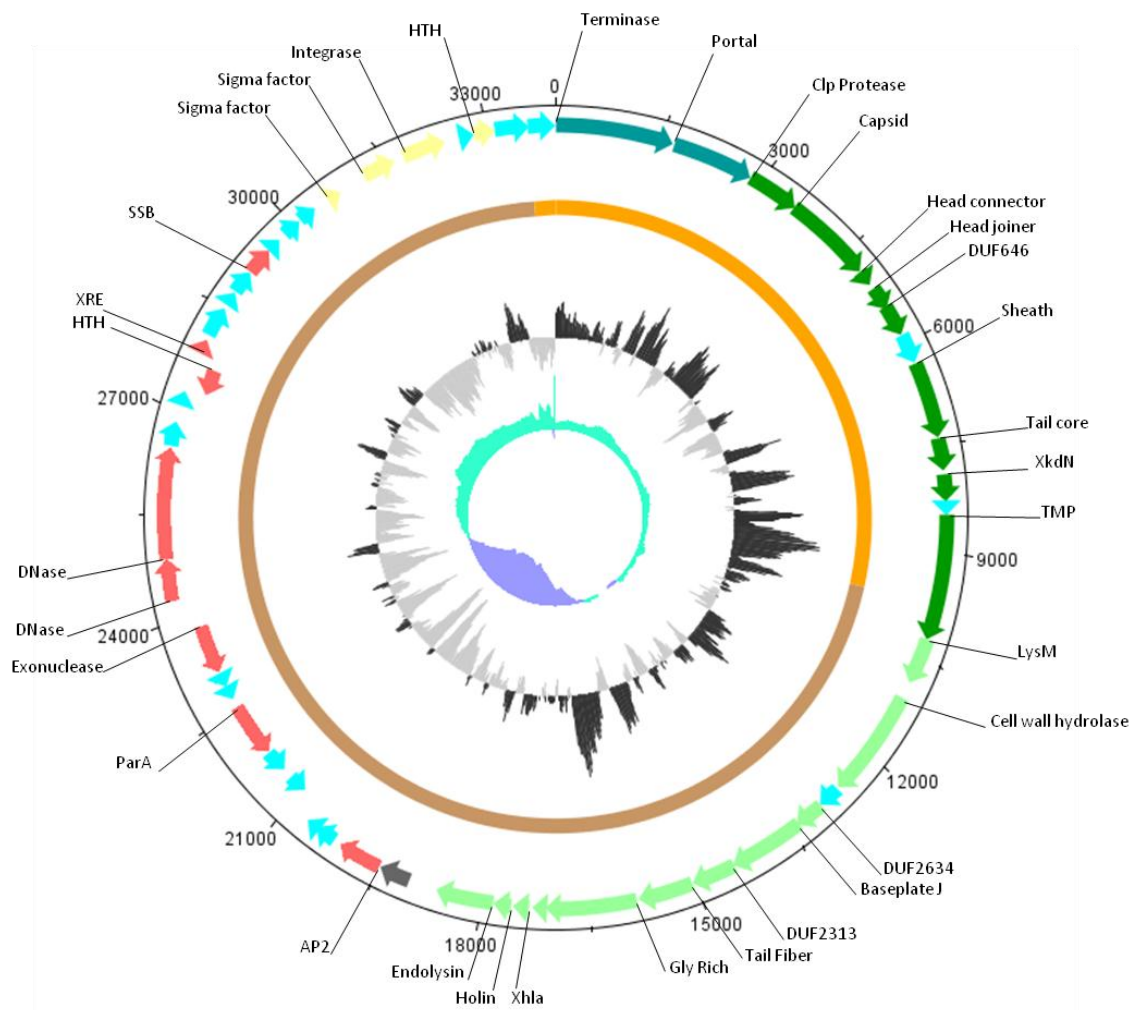
Figure 17. Genome map of phiCDHM13. Circular double stranded genome map with GC skew, GC content and modular regions of the genome shown drawn in DNAplotter (Carver *et al* 2009). Track 1 = GC skew, light green is above average and purple is below average. Track 2 = GC%, dark grey is above average and light grey is below average. Track 3 displays the contigs, orange is contig 1 and brown is contig 2. Track 4 (antisense) and Track 5 (sense) encode ORFs. Different colours of genes indicate functional modules: packaging (blue); structural (green); attachment and lysis (light green); lysogenic conversion (pink); lysogeny control (yellow) and DNA replication (red). Aqua are ORFs with no predicted function assigned and charcoal are ORFs with no significant hits.

Although most of the gene content is conserved between the genomes of phiCDHM13, phiCDHM11 and phiCDHM14, the genomes are not identical. Furthermore, there may be variation in the parts of the genomes that have not been fully assembled and where contigs end the resolution of ORFs is difficult. For example, Orf36, Orf37 and Orf38 in phiCDHM13 may be fragments of one gene encoding a putative deoxyribonuclease. This gene in phiCDHM14 is split into two ORFs and in phiCDHM11 is truncated as is on the end of a contig. However, genome evolutionary events such as the loss or acquisition of whole genes as well as point substitutions, indels and frameshift mutations were identified in each of the phage genomes.

Loss or gain of genes in the SMV genomes has occurred at least four times. For example, an ORF predicted to encode a hypothetical protein is in the genomes of phiCDHM14 and phiCDHM11, but is not present in phiCDHM13 (and would be located between Orf25 and Orf26). A second ORF, also encoding a hypothetical protein, is present in the genome of phiCDHM14 but absent in both phiCDHM13 and phiCDHM11 (and would be located between Orf46 and Orf47). Lastly, two adjacent ORFs, Orf46 and Orf47 (which are also hypothetical proteins), are encoded by phiCDHM13 and phiCDHM14 but are absent in the genome of phiCDHM11. Presumably these genes are non-essential, but their differential carriage between the SMVs may indicate that they have complimentary functions. However, as no predicted function is assigned to any of the ORFs in question their true importance is unknown.

Point substitutions that result in a change to the translated amino acid sequence are present in three ORFs which are conserved across all three phage genomes. Orf39 encodes a hypothetical protein that contains a point substitution in the phiCDHM11 gene, (a guanine to a thymine) which results in an amino acid residue change from aspartic acid to glutamic acid. This mutation may not affect protein function as both amino acid residues have acidic R groups. Orf41 encodes a hypothetical protein and in phiCDHM14 the gene has a nucleotide substitution (a cytosine to a thymine) which results in a change from a valine to an alanine amino acid residue. This substitution

also may not affect protein function as both amino acid residues have small, uncharged R groups. However, Orf16 encodes a hypothetical protein and contains a point nucleotide substitution in the phiCDHM14 version of the gene, (a cytosine to a thymine) and results in a change from a threonine to an isoleucine in the translated amino acid sequence. This substitution could affect the function of the protein, as threonine has a polar R group but isoleucine has a non-polar R group. Depending on the position of the residue this could change the protein's solubility or molecular interactions.

An indel region was identified in Orf11, a gene that is present in all three phage genomes (Appendix Figure 4). This indel region is 33 bp long in the phiCDHM11 version of the gene and causes truncation of the translated amino acid sequence which could destroy gene function. However, Orf11 encodes a structural protein, XkdN, and as this phage is able to replicate either the indel is a sequencing error or the mutation is not lethal.

Frameshift mutations were also identified, initially through observed differences in the values of identity to top blastp for individual ORFs. For example, Orf5 has an identity of 47% in phiCDHM13 to the top blastp hit and 43% in each phiCDHM14 and phiCDHM11. Following the alignment of the nucleotide sequences of each phages' ORF in CLUSTLW, it can be seen that a frameshift mutation due to a point nucleotide deletion results in the truncation of the gene by 66 nucleotides (Appendix Figure 5). Another example is Orf30, which contains a point insertion resulting in a frameshift mutation and truncation of the gene in phiCDHM14. This may be due to a sequencing error as it occurs in a polyT region of the gene (Appendix Figure 6).

The observed similarities and differences between the three SMVs genomes provide insight into the evolutionary origins as well as their subsequent divergence from one another. As the mutations are mainly in hypothetical proteins it is not possible to know whether they influence gene function, or whether these genes are indeed functional or essential. However, these genetic differences serve to illustrate that *C. difficile* phage genomes can undergo several types of genetic mutation and potentially account for the observed differences in their host ranges.

5.2.5. Genome annotation of the environmental *C. difficile* prophage phiT6.

The genomic sequence of the prophage phiT6 was identified following the whole genome sequencing of the environmental *C. difficile* strain T6 (R076). TEM analysis found that the induced culture lysates of T6 produced MMs (section 3.2.7.) which would predict the inducible MM to have a genome of 50 -54 kbp. The genome size of phiT6 is consistent with this prediction, as is 52,530 bp. 83 ORFs were predicted, with the majority (73) encoded on the sense strand, and predicted coding sequences account for 85.5% of the genome (Table 15.). Of the 83 ORFs, 36 have been assigned predicted functions, 39 encode conserved hypothetical proteins and 4 have no significant blastp hits.

There is a distinct separation of the genome with regards to the predicted ORF similarity to their top blastp hit. For example, 58 ORFs have identities of >90%, 20 of which are 100%. Most of these ORFs are located in the structural, lysis and lysogeny control regions of the genome. In contrast, a cluster of ORFs in the DNA replication region, Orf77 to Orf83, have low similarity (68% - 31%) to their top blastp hits and two ORFs in the cluster have no significant blastp hits (ORFs 81 and 82). Additionally, several of the ORFs in the DNA replication have most similarity to genes encoded by other bacterial species. For example, Orf75, a predicted C-5 cytosine-specific DNA methylase, has a top blastp hit to a gene in *Clostridium lentocellum* DSM 5427. Others include: Orf59 (*Candidatus Arthromitus sp.* SFB-mouse-Japan*); Orf78 (*Clostridium hiranonis* DSM 13275); Orf57 (*C. perfringens* WAL-14572); Orf77 (*C. botulinum C str* Eklund); Orf79 (*Collinsella intestinalis* DSM 13280) and Orf80 (*Clostridium cf. saccharolyticum* K10). The directionality of the genome in regards to gene similarities is interesting as it suggests there has been extensive divergence in the DNA replication region, or a large recombination event that has produced a chimeric like phage genome.

Table 15. Genome annotation of phiT6.

Symbols indicate sense (+) or antisense (-) strand. For top blastp hit where no bacterial species is indicated this is *C. difficile* strains, with their strain name.

ORF	Start	End	AA size	Strand	Function	Domain	Top blastp hit	Identity %	E value
1	1	669	223	+	terminase small subunit	HTH_28 (CL0123)	phiCD27	94	1.00E-142
2	663	1916	418	+	terminase large subunit	Terminase_3 (CL0023)	phiC2	99	0
3	1922	3361	480	+	portal protein	Phage_prot_Gp6	phiC2	99	0
4	3351	4844	498	+	phage minor head protein	Phage_Mu_F	630/002-P50-2011	97	0
5	4822	5052	77	+	hypothetical protein		630	100	4.00E-45
6	5123	5317	65	+	hypothetical protein		phiCD119	87	9.00E-18
7	5362	5925	188	+	phage scaffold protein	Phage_GP20	630	98	1.00E-128
8	5937	6827	297	+	major capsid protein		phiC2/QCD-63q42/ATCC 43255/050-P50-2011/002-P50-2011	99	0
9	6846	7079	78	+	hypothetical protein		QCD-66c26/ CIP 107932/QCD-76w55/ QCD-97b34/QCD-37x79/QCD-32g58/ BI1	99	7.00E-45
10	7080	7490	137	+	hypothetical protein		630/ CD196/R20291	99	3.00E-91
11	7484	7831	116	+	hypothetical protein		630/002-P50-2011	100	1.00E-76
12	7831	8121	97	+	hypothetical protein (gp10)		QCD-63q42	99	4.00E-64
13	8210	9037	276	+	hypothetical protein		CD196/ R20291	99	0
14	9316	9393	26	+	hypothetical protein (HK97 gp10)		002-P50-2011	100	5.00E-07
15	9386	9823	146	+	hypothetical protein		050-P50-2011	95	6.00E-94
16	9807	9992	62	+	hypothetical protein		ATCC 43255	95	1.00E-30
17	9994	11304	437	+	sheath protein	Phage_sheath_1	630	99	0
18	11321	11791	157	+	core protein/ XkdM	DUF2001 (CL0504)	630/ATCC 43255/ 050-P50-2011/ 020-P50-2011	100	5.00E-111
19	11863	12303	147	+	XkdN	XkdN	630/QCD-66c26/CIP 107932/ATCC 43255/QCD-76w55/ QCD-97b34/QCD-37x79/ CD196/R20291/QCD-32g58/ BI1/050-P50-2011/002-P50-2011	100	4.00E-102
20	12378	12497	40	+	hypothetical protein		630	100	5.00E-16
21	13146	13250	35	+	hypothetical protein		QCD-66c26/CIP 107932/ATCC 43255/QCD-76w55/ QCD-97b34/QCD-37x79/ QCD-32g58/	100	7.00E-15
22	14130	14303	58	+	hypothetical protein		050-P50-2011	98	2.00E-32
23	14427	15254	276	+	anti-repressor	Bro-N	050-P50-2011	92	6.00E-176
24	15317	15493	59	+	hypothetical protein		QCD-63q42	100	5.00E-32
25	15906	16109	68	-	DNA-binding protein/repressor	HTH_26 (CL0123)	QCD-63q42	59	3.00E-18
26	16353	17156	268	+	putative anti-repressor	ORF6N	*Clostridium botulinum* BKT015925	73	7.00E-56
27	17183	17287	35	+	hypothetical protein		QCD-66c26/CIP 107932/ATCC 43255/QCD-76w55/ QCD-97b34/QCD-37x79/ QCD-32g58/	97	2.00E-12

28	17668	17994	109	+	hypothetical protein		NAP08/NAP07	46	1.00E-10
29	18061	18171	37	-	no significant hits		-	-	-
30	18594	19118	175	+	phage protein /stress protein	zf-ribbon_3 (CL0167)	ATCC 43255	96	1.00E-104
31	19183	21576	798	+	TMP		ATCC 43255	93	0
32	21592	22281	230	+	cell wall hydrolase	LysM (CL0187)	ATCC 43255	96	3.00E-158
33	22274	24166	631	+	cell wall hydrolase	NLPC_P60 (CL0125)	QCD-63q42	97	0
34	24180	24440	87	+	hypothetical protein	DUF2577	QCD-63q42	100	9.00E-54
35	24445	24864	140	+	hypothetical protein	DUF2634	ATCC 43255	96	3.00E-92
36	24877	25914	346	+	Baseplate-J protein	Baseplate_J	630	99	0
37	25907	26524	206	+	tail protein	DUF2313	630	98	1.00E-145
38	26536	27528	331	+	tail fiber	DUF3751	002-P50-201	99	0
39	27545	29257	571	+	hypothetical protein	Gly_rich	002-P50-201	99	0
40	29278	29625	116	+	hypothetical protein		phiC2/050-P50-2011/002-P50-2011	100	4.00E-75
41	29637	29807	57	+	hypothetical protein		050-P50-201/002-P50-201	100	2.00E-31
42	29838	30404	189	+	hypothetical protein		QCD-66c26/CIP 107932/QCD-76w55/ QCD-97b34/QCD-37x79/CD196 /R20291/ QCD-32g58/ BI1	98	1.00E-131
43	30435	30725	97	+	hypothetical protein		QCD-63q42	97	7.00E-61
44	30729	30986	86	+	Holin		phiC2/ QCD-66c26/CIP 107932/ ATCC 43255/QCD-76w55/ QCD-97b34/QCD-37x79/CD196/R20291 /QCD-32g58/BI1	94	2.00E-49
45	31247	31573	109	+	phage AbiD protein	Abi_2	BI1	100	1.00E-69
46	31613	31741	43	+	phage AbiD protein	Abi_2	050-P50-201	95	2.00E-20
47	31757	32110	118	+	phage AbiD protein	Abi_2	050-P50-201	100	8.00E-78
48	32189	33001	271	+	N-acetylmuramoyl-L-alanine amidase (cell wall hydrolase	Amidase_3 (CL0035)	QCD-66c26/CIP 107932/QCD-76w55/ QCD-97b34/QCD-37x79/CD196 /R20291/ QCD-32g58/ BI1	100	0
49	33138	34268	377	-	Integrase	Phage_integrase (CL0382)	050-P50-2011	99	0
50	34508	35014	169	-	hypothetical protein	DUF955	QCD-66c26/CIP 107932/ATCC 43255/QCD-76w55/ QCD-97b34/QCD-37x79/CD196 /R20291/ QCD-32g58/ BI1/050-P50-2011/002-P50-2011	100	4.00E-114
51	35277	36866	530	-	hypothetical protein	SIR2_2 (CL0085),	ATCC 43255/050-P50-2011/002-P50-2011	99	0
52	36890	37453	188	-	putative phage repressor	HTH_19/HTH_19 (CL0123)	ATCC 43255/002-P50-2011	100	6.00E-135
53	37606	37827	74	+	XRE family transcriptional regulator	HTH_3 (CL0123)	ATCC 43255/002-P50-2011	99	3.00E-44
54	37803	38645	281	+	anti-repressor	Phage_pRha/ANT (CL0123)	002-P50-2011	98	6.00E-158
55	38824	39003	60	-	regulatory protein	Arc (CL0057),	ATCC 43255/002-P50-2011	100	4.10E-05

56	39139	39318	60	+	repressor protein	RHH_1 (CL0057),	002-P50-2011	100	3.00E-33
57	39311	40102	264	+	anti-repressor	ORF6N	*Clostridium perfringens* WAL-14572	45	1.00E-12
58	40185	40397	71	-	hypothetical protein		QCD-63q42/TCC 42355	99	0
59	40535	40753	73	-	XRE family transcriptional regulator	HTH_3 (CL0123)	*Candidatus arthromitus* sp. SFB-mouse-Japan	52	2.00E-11
60	40845	41060	72	+	XRE family transcriptional regulator	HTH_3 (CL0123)	CD119	100	9.00E-45
61	41102	41896	265	+	anti-repressor	Phage_pRha/ANT	CD119	84	4.00E-163
62	41911	42162	84	+	hypothetical protein		050-P50-2011	99	3.00E-51
63	42172	42546	125	-	hypothetical protein		QCD-66c26/ATCC 42355/QCD-76w55/QCD-97b34/QCD-37x79/ QCD-32g58/ BI1/050-P50-2011	100	3.00E-85
64	42622	42819	66	+	regulatory protein	HTH_26 (CL0123)	QCD-66c26/QCD-76w55/ QCD-97b34/QCD-37x79/ QCD-32g58/ BI1/	95	1.00E-34
65	42850	43242	131	+	DNA binding protein		ATCC 42355 050-P50-2011	94	4.00E-79
66	43223	43372	50	+	hypothetical protein		002-P50-2011	88	2.00E-18
67	43539	43781	81	+	hypothetical protein		QCD-63q42	89	2.00E-43
68	43860	44141	94	+	no sig hits		-	-	-
69	44128	44766	213	+	essential recombination protein	ERF	630	68	1.00E-95
70	44767	45234	156	+	putative phage replication protein	HTH_36	QCD-63q42/ATCC 42355	65	5.00E-47
71	45213	45671	153	+	phage replication protein	DnaB_2	QCD-63q42/ATCC 42355	86	2.00E-71
72	45691	46110	140	+	single-strand binding protein	SSB (CL0021)	630/QCD-66c26/CIP 107932/ATCC 42355/QCD-76w55/QCD-97b34/QCD-37x79/ QCD-23M63/Cd196/R20291/NAP08/NAP07/QCD-32g58/ BI1/050-P50-2011/002-P50-2011/70-100-2010	74	2.00E-73
73	46140	46502	121	+	hypothetical protein		phiCD27	66	2.00E-44
74	46502	46636	45	+	hypothetical protein		phiC2/PHICD27	95	6.00E-21
75	46736	47587	284	+	C-5 cytosine-specific DNA methylase	DNA_methylase (CL0063)	*Clostridium lentocellum* DSM 5427	47	1.00E-71
76	47904	48101	66	+	hypothetical protein		QCD-76w55BI1	65	4.00E-17
77	48128	48544	139	+	hypothetical protein	DUF1064 (CL0236)	*Clostridium botulinum* C str Eklund	68	4.00E-55
78	48563	49063	167	+	sigma factor		*Clostridium hiranonis* DSM 13275	32	4.00E-15
79	49170	49283	38	+	hypothetical protein		*Collinsella intestinalis* DSM 13280	52	5.00E-15
80	49485	49691	69	+	hypothetical protein		*Clostridium* cf. *saccharolyticum* K10	57	5.00E-05
81	50180	51367	396	+	no sig hits		-	-	-
82	51315	51698	128	+	no sig hits		-	-	-
83	51703	52473	257	+	restriction endonuclease	Mrr_cat (CL0236)	*Staphylococcus aureus* subsp. *aureus* 21305	31	3.00E-14

The organisation of the genome has a modular arrangement similar to that of the MMs, except for a cluster of ORFs in the structural region which encode genes predicted to be involved in regulatory and DNA binding functions (Figure 18). The genome has an average GC% of 28.87 and has more sections in the genome that contain lower than average GC% compared to phiCDHM1 and phiCDHM19. The GC skew however is in the same position and indicates that the origin of replication is near to the integrase gene.
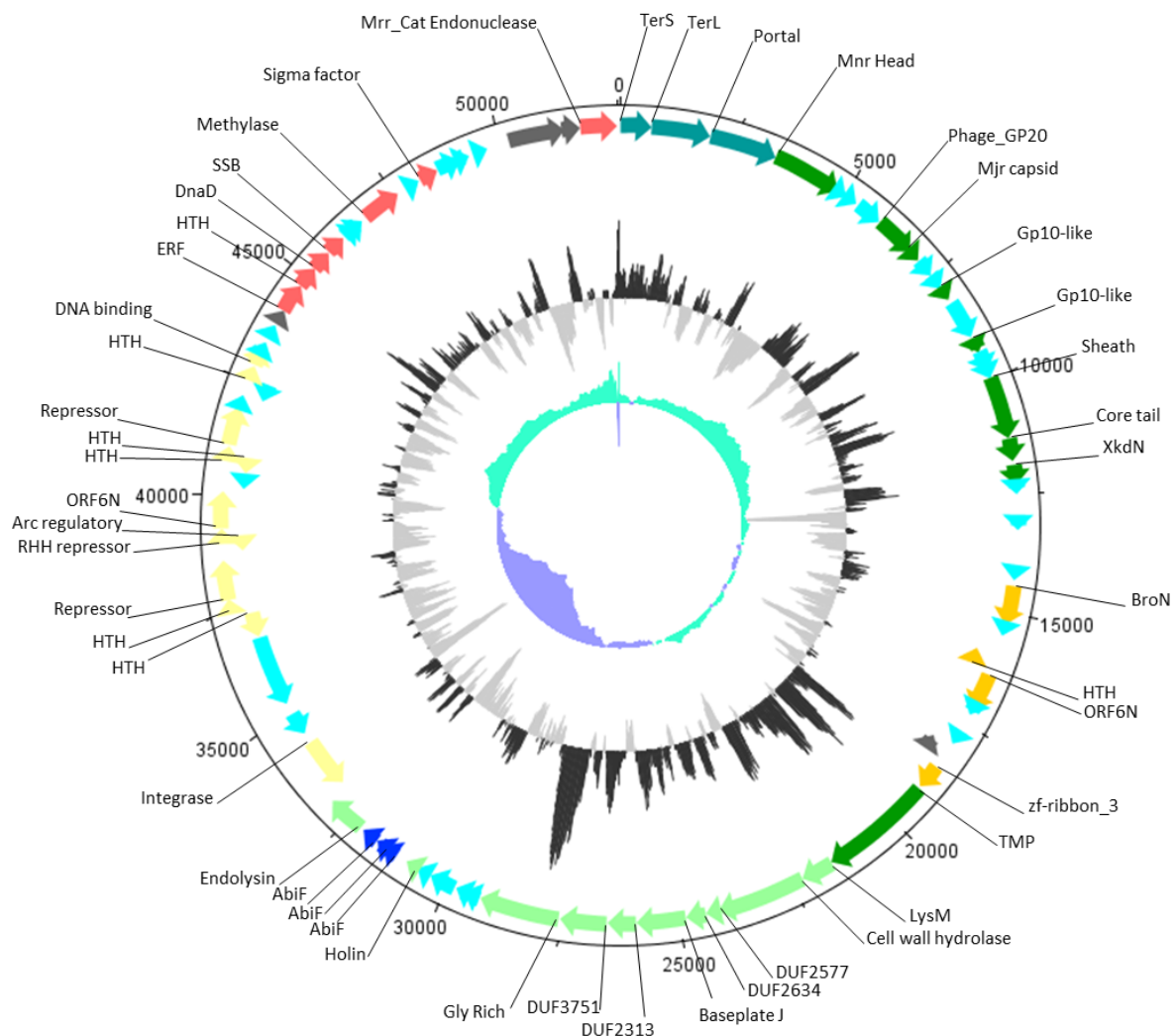

Figure 18. Genome map of phiT6. Circular double stranded genome map with GC skew, GC content and modular regions of the genome shown drawn in DNAplotter (Carver *et al* 2009). Track 1 = GC skew, light green is above average and purple is below average. Track 2 = GC%, dark grey is above average and light grey is below average. Track 3 (antisense) and Track 4 (sense) encode ORFs. Different colours of genes indicate functional modules: head packaging (blue); structural (green); lysis and attachment (light green); lysogenic conversion (pink); lysogeny control (yellow) and DNA replication (red). Aqua are ORFs with no predicted function assigned and charcoal are ORFs with no significant hits.

5.2.6. Phylogenetic analysis of *C. difficile* phage capsid genes identifies distinct myovirus types.

In order to investigate how related these phages are to one another, and to previously published *C. difficile* phages, the capsid gene was selected for use in a phylogenetic analysis (Figure 19). Capsid gene sequences from the phages phiCDHM1, phiCDHM19, phiCDHM13 and prophage phiT6, as well as the publically available sequences of phages phiC2, phiCD119, phiMMP04, phiCD38-2 and phiCD6356 and prophages from strains CD196, CD630 and R20291 were used in the analysis. Also candidate capsid gene sequences were used from each LTM, phiMMP02 and phiCD27, to identify if these could be included and serve as markers of phage diversity for this morphotype as no capsid genes have been annotated in their genomes. An alignment was performed using MUSCLE and analysed using Maximum Likelihood phylogenetic analysis in MEGAv5.01 (Tamura *et al* 2011). Importantly, the capsid gene contains enough sequence variation at the amino acid level to enable meaningful phylogenetic comparison between the phages, as 4% of sites are conserved in all the examined phages. This value can be higher between specific phages, for example 67% sites are conserved between phiCDHM1 and phiT6.

The results of the ML analysis showed that the capsid gene sequences cluster into six clades and these correspond to their particle morphology (Figure 19). All the myovirus sequences group into four clusters (1-4): those of the MMs and prophages are split into two clusters, 1 and 2; the two sequences LTMs cluster together, 3; and the SMV sequences in cluster 4. All clusters are supported by high bootstrap values. The capsid sequences of the SVs are divergent and do not cluster together.
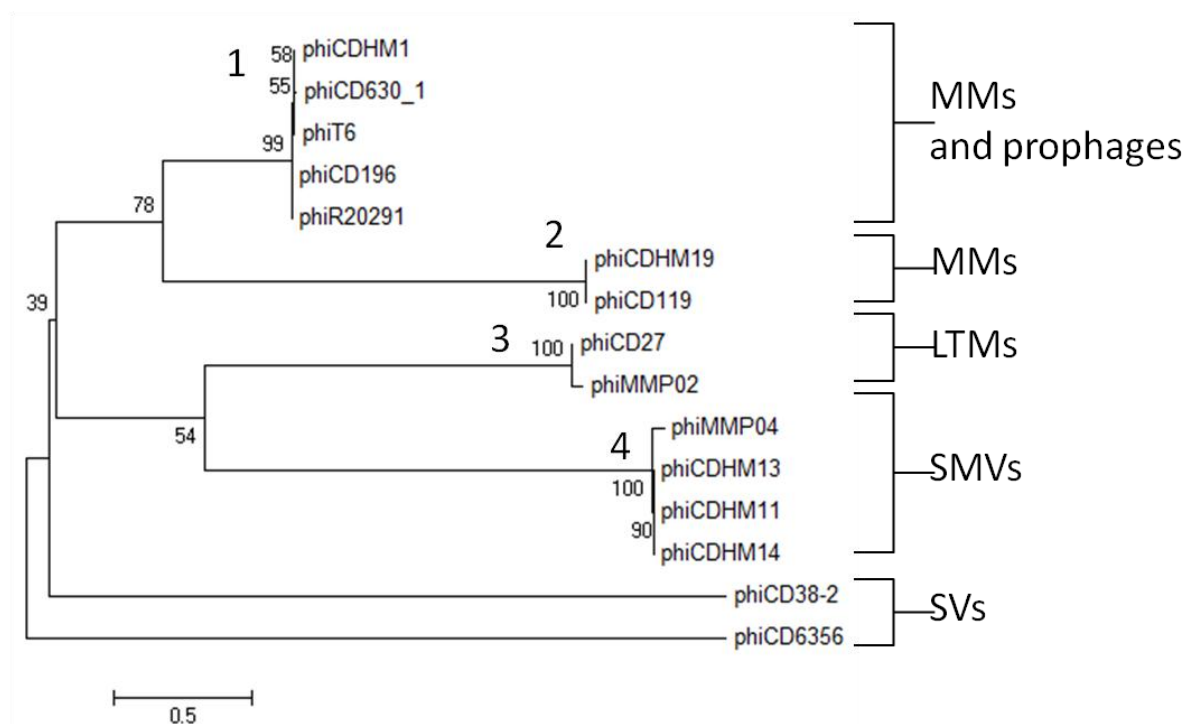
Figure 19. Phylogeny of capsid genes in the *C. difficile* phages and strains. The capsid gene tree is constructed from ML phylogenetic analyses in MEGA5v1 with evolutionary distances computed using the JTT model of substitution and rates were set at gamma invariant. Branch lengths are to scale for the number of amino acid substitutions per site. Analysis was performed using all sites and NNI as the heuristic model. Taxa cluster into six distinct groups as reflected by branch length and bootstrap scores.

5.2.7. Detection of myovirus capsid gene types in *C. difficile* isolates shows prophage prevalence, diversity and co-carriage.

In order to investigate the extent and diversity of prophage carriage in environmental *C. difficile* isolates, the capsid gene was selected for use as a molecular marker for temperate myoviruses. Four sets of primers were designed to target each of the myovirus clusters identified in the phylogenetic analysis. Using these primer sets, 60 *C. difficile* isolates that were isolated in section 3 were included in a PCR based screen.

The results of this screen show that 73% (44/60) isolates were positive for at least one type of capsid gene (Table 16). Prophage carriage of the different types is widespread, with each capsid type detected in multiple isolates: type 1 in 36 isolates, group 2 in 18 isolates, type 3 in four isolates

and type 4 in eight isolates. Also, multiple prophage carriage was detected; while 29 isolates had a single type of capsid detected, 11 had 2 types and 4 had 3 types. In isolates encoding multiple prophages, the most common co-carriage was that of types 1 and 2 (*n*=7). Overall the results are consistent with those of the TEM analysis performed in section 3.2.7., with high rates of lysogeny, dominance of myovirus prophage carriage and common occurance of multiple prophage carriage detected in both. However, the PCR based screen reveals the co-carriage of both types of MMs that are genetically distinct is prevalent, which could not be seen in the TEM analysis since the two types are not visually distinguishable.

Also similar to the TEM results was that there was no clear relationship observed between myovirus capsid carriage and ribotype. In the screen 53 isolates belong to 19 ribotypes and 7 isolates belong to an unknown ribotype. In cases where ribotype groups have more than one isolate, it was evident that for some ribotypes all isolates had the same capsid type detected, but for other ribotype groups isolates harbour phages with different capsid sequences. One example is R010, where at least one of the eight isolates of R010 had each of the different capsid groups detected. However, two ribotype groups, R031 and R137, did not have any positive isolates. For R031 this is consistent with the TEM results, where no phages were observed and together these results suggest these isolates may not encode prophage.

Table 16. Myovirus capsid gene carriage and diversity in *C. difficile* isolates.
Symbol (+) indicates presence of a PCR product, symbol (-) indicates no amplification.

Ribotype	Isolate	Capsid group				Ribotype	Isolate	Capsid group			
		1	2	3	4			1	2	3	4
220	K12	+	-	-	-	005	H5B	+	-	-	-
	K3	-	+	-	-		K18	+	+	-	-
	N	+	+	-	-		E	+	-	-	+
	L	+	-	-	-	002	O	-	+	-	+
	K	-	-	-	-		H15	NA	NA	NA	-
	H12	+	-	-	-		K14	+	-	-	+
	Y	+	-	-	-		P	+	-	-	-
	U	+	+	-	-		S	+	-	-	-
010	K16	+	+	+	-	078	H17	-	-	-	+
	H3	+	+	-	-		H5C	-	-	-	-
	H5d	-	+	-	-		J	NA	-	NA	-
	R10	+	+	-	+	021	A	+	-	-	-
	R	+	-	-	-		K9	+	-	-	-
	H1B	-	+	+	+		B1	-	-	-	-
	I	+	-	-	+		B2	-	-	-	-
027	HB	+	-	-	-	003	SS1	+	-	-	-
	K15	+	-	-	-	076	T6	+	-	-	-
014	K6	-	+	-	-	085	R8	-	-	-	-
	OA	-	+	-	-	106	H19	-	-	-	-
	V	+	-	-	-	137	G	-	-	-	-
012	K10	+	-	+	-		D	-	-	-	-
001	F	+	+	-	-	026	C	-	+	-	-
	H18	+	-	-	-	015	Q	-	-	-	-
	?	-	-	-	-	uk	H14	-	-	-	-
	H8	+	-	-	-	uk	X	+	-	-	-
	H4	+	+	-	-	uk	RBI	+	-	-	-
031	H1A	-	-	-	-	uk	H16B	+	-	-	-
	H11	-	-	-	-	uk	K20	-	+	-	-
	M	-	-	-	-	uk	H9	+	-	-	-
046	K4	+	-	-	-	uk	H	+	+	-	+

5.2.8. Phylogenetic analysis of the phage encoded *agr* QS genes reveals three types of *agr* loci in *C. difficile*.

The identification of three *agr* QS genes in phiCDHM1 is the first report of such genes to be encoded in a phage genome. QS is a method of bacterial communication between cells that is dependent on cell density and the release, and recognition of signalling peptides to elicit a response. The *agr* QS system is found throughout Gram-positive bacteria, but the gene content and organisation of the loci varies between species. The first described and best characterised *agr* locus is encoded in *S. aureus* (Novick 2003). The locus encodes four genes: AgrD (*agrD*) is a pre-peptide that is cleaved post-translationally into the autoinducing peptide (AIP); AgrB (*agrB*) is an enzyme that cleaves AgrD; AgrC (*agrC)* is a membrane located histidine kinase (HK) that recognises the AIP; AgrA (*agrA*) is the response regulator that is phosphorylated by AgrC.

PhiCDHM1 encodes homologs of three of the four genes: *agrD*, *agrB* and a HK. The genes in the phage encoded *agr* locus differ in gene content and orientation content to the *agr* locus of *S. aureus* but are the same as those encoded in *C. difficile* strain NAP07 (Figure 20). PhiCDHM1 therefore has genes that potentially can produce the signal (AgrD and AgrB) as well as detect the resulting AIP (the HK). The phage does not encode a response regulator gene homolog and may therefore use a host encoded one. Excitingly, the presence of these genes in a phage genome suggests a highly novel mechanism by which a phage could influence their host cell physiology. Due to the potential significance of this discovery, the phylogeny of each phage encoded homologs was examined to determine the evolutionary origin of the genes.
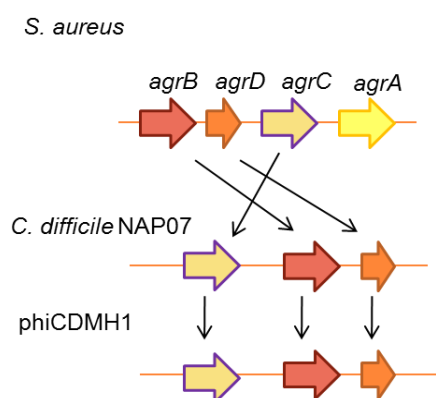

Figure 20. Variable gene content and organisation of *agr* loci. Illustration showing the gene content and organisation of the *agr* loci in the two bacterial species, *S. aureus* and *C. difficile* strain NAP07, and the phage encoded *agr* loci.

In order to confirm the predicted annotation of this gene as *agrD*, the amino acid sequence was checked for the P-X-X-P motif which is characteristic of AgrD, as is the site where AgrB binds (Novick and Geisinger 2008). The motif is present between amino acid residues 35 and 38. It also contains a signal sequence, residues 12-24 and a transmembrane domain, 12-32 as predicted by InterProScan. Additionally, the proximity of *agrD* to *agrB* suggests that this gene is likely to be the autoinducing prepeptide. The HK contains a HATPase_c protein domain which is also present in AgrA, as well as many other histidine kinases. As the sequence is homologous to several histidine kinase entries in NCBI which have not been annotated as AgrA, but its proximity to the *agrD* and *agrB* strongly suggests this HK is functioning as AgrA.

Results of the individual genes blastp searches showed that there were different numbers of homologous entries for each gene. These results are explained by the carriage of different *agr* loci in different *C. difficile* strains based on gene content and organisation. For example, strain CD630 encodes *agrB* and *agrD* in *agr1* and strain CD196 encodes *agr1* and an additional set, *agr2,* which encodes *agrB*, *agrD*, a HK and a response regulator (Sebaihia *et al* 2006, Stabler *et al* 2009). The phage *agr* genes do not belong to either *agr1* or *agr2,* so represent a third loci, *agr3* (Figure 21). In order to further investigate the evolution and origin of these genes in a phage genome, phylogenetic analysis was performed for each gene using the phage encoded and *C. difficile* homologs, as well as *S. aureus agr* genes as an outgroup (see Appendix Table 1 for accession numbers and strains).
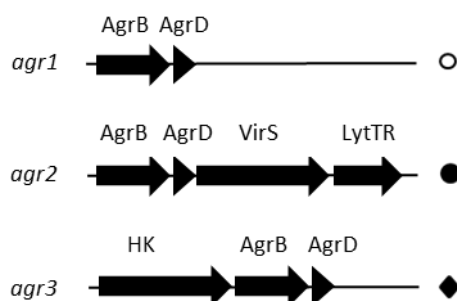

Figure 21. Illustration of the three types of *agr* locus encoded across *C. difficile* strains and phiCDHM1. Type *agr1* encode only two genes *agrB* and *agrD*. Type *agr2*, encodes *agrB*, *agrD,* a HK with homology to VirS and a response regulator LytTR. Type *agr3* encodes *agrB, agrD* and a HK gene (which phiCDHM1 belongs).
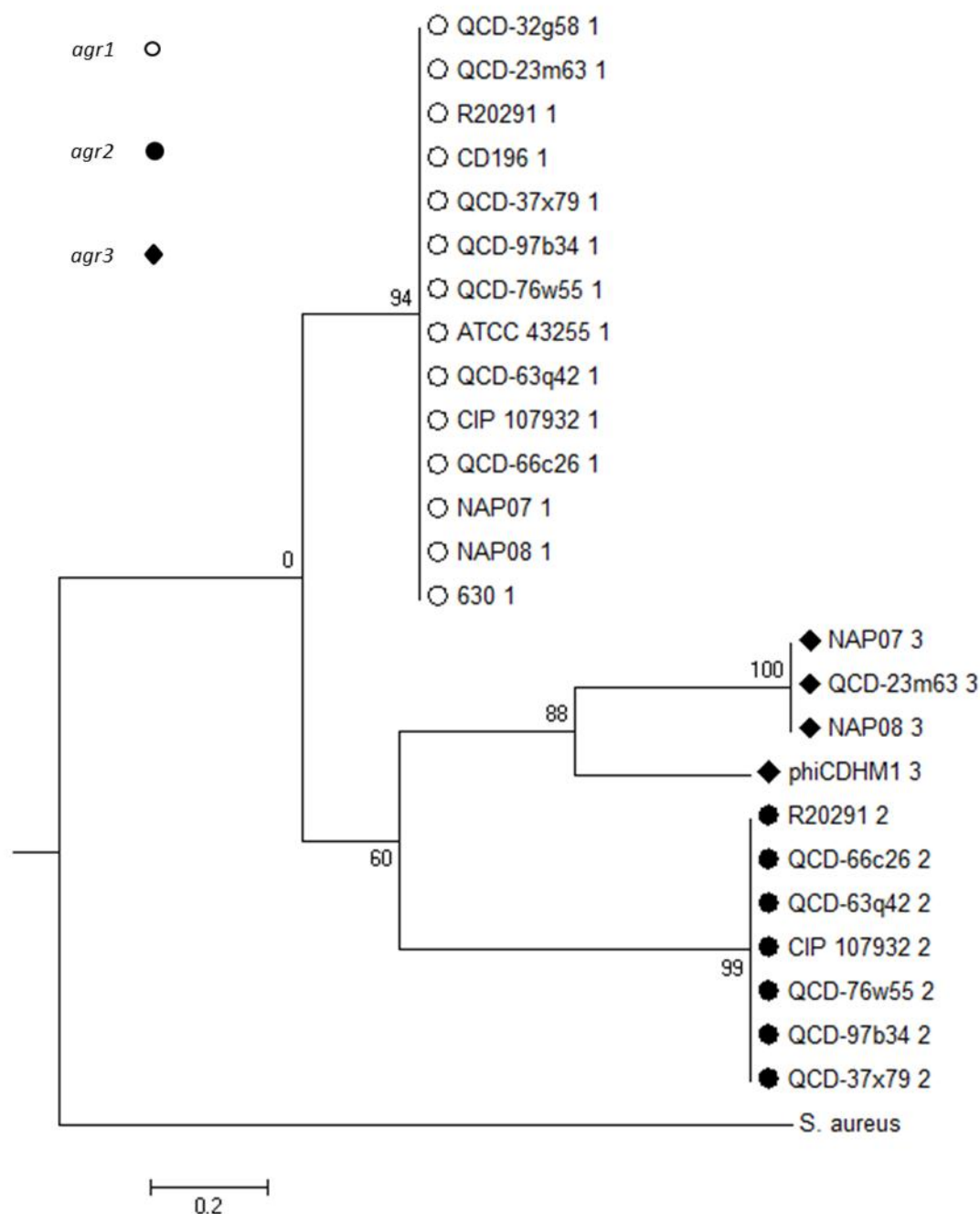
Figure 22. ML phylogenetic analysis of the *agrD* gene from phiCDHM1 and *C. difficile* strains. Amino acid sequences for the *agrD* gene carried in phiCDHM1 and in all *C. difficile* strains deposited in the NCBI to date identified (Oct/2011). ML analysis was performed using MEGAv5.01 with alignment carried out in with parameters set for the JTT nucleotide substitution model, with invariant rates, using all sites and CNI for Tree Inference and bootstrapped with 500 replicates. Taxa are abbreviated to strain names and letter indicates where taxa have two entries (A or B versions of the gene). Symbols indicate the clusters of the gene, as several strains encode two versions of the gene.

Figure 23. ML phylogenetic analysis of the *agrB* genes from phiCDHM1 and *C. difficile* strains. Amino acid sequences for the *agrB* gene carried in phiCDHM1 and in all *C. difficile* strains deposited in the NCBI to date identified (Oct/2011). ML analysis was performed using MEGAv5.01 with alignment carried out in with parameters set for the JTT nucleotide substitution model, with invariant rates, using all sites and CNI for Tree Inference and bootstrapped with 500 replicates. Taxa are abbreviated to strain names and letter indicates where taxa have two entries (A or B versions of the gene). Symbols indicate the clusters of the gene, as several strains encode two versions of the gene.
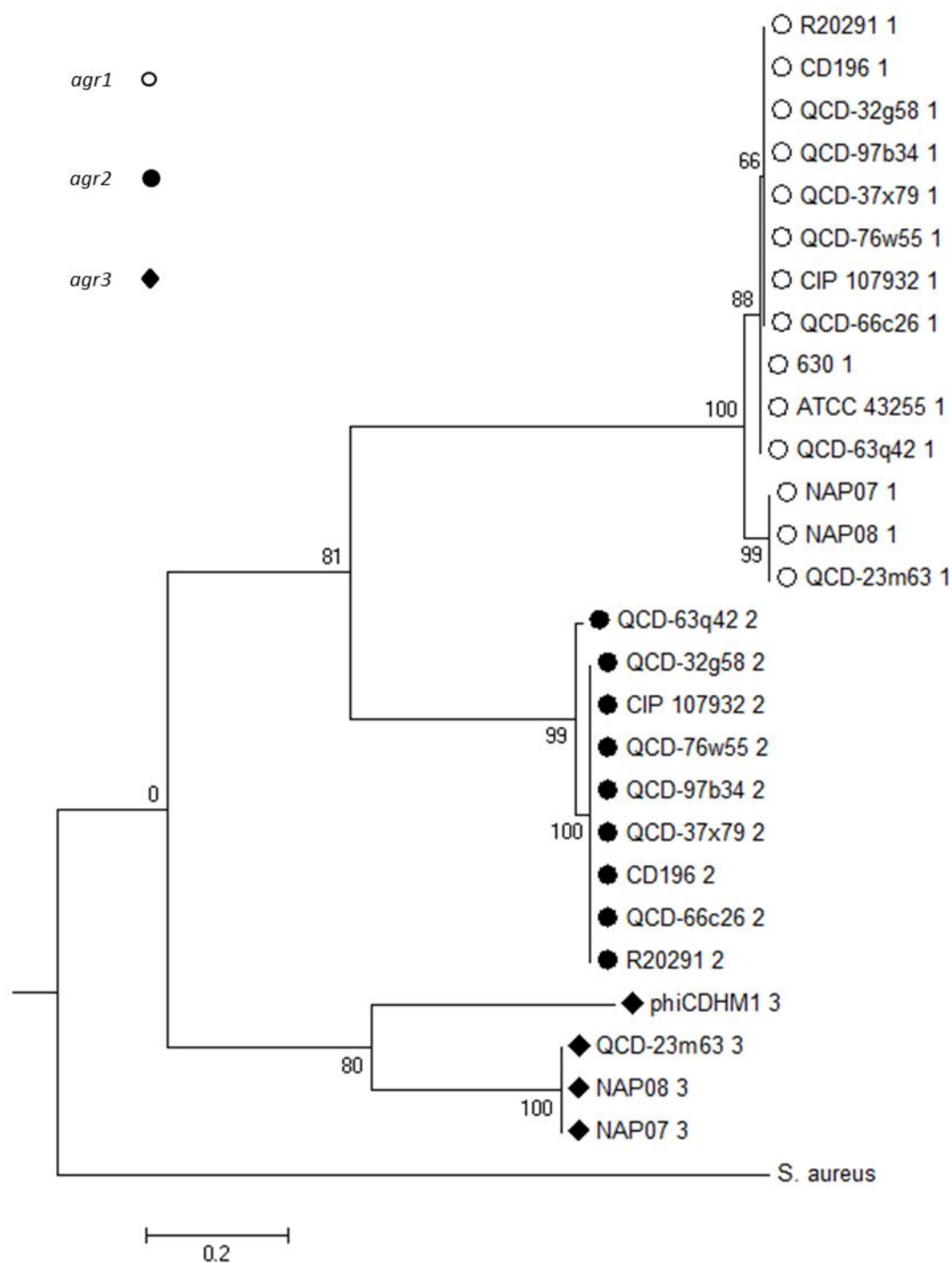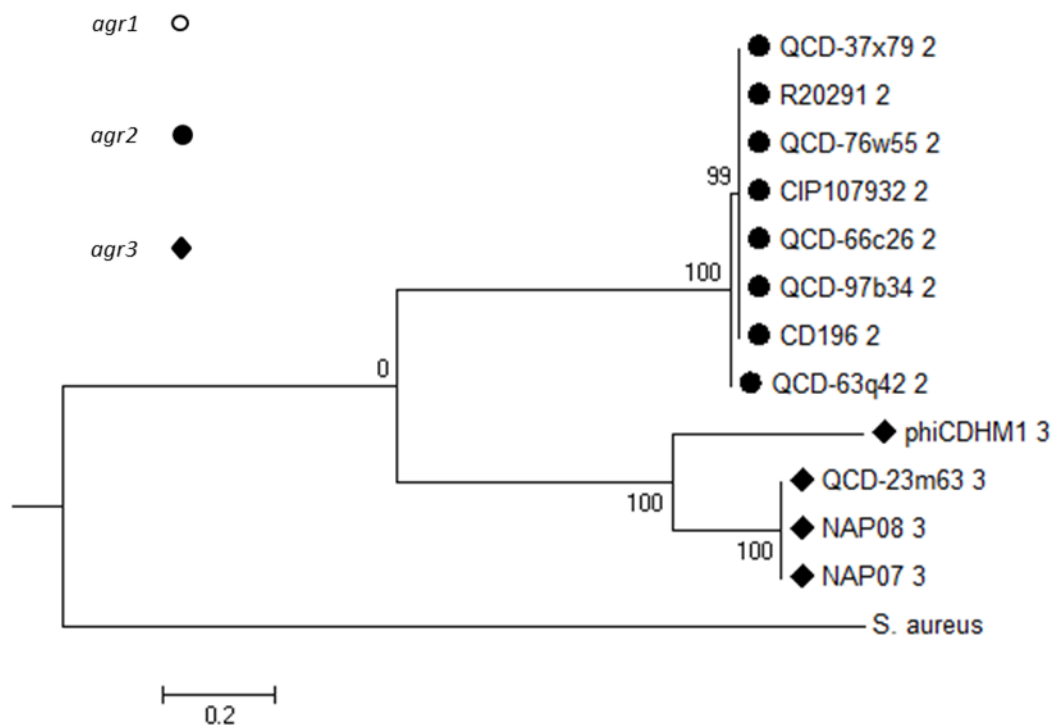
Figure 24. ML phylogenetic analysis of related HK genes from phiCDHM1 and *C. difficile* strains. Amino acid sequences for HK genes with shared sequence identity to the gene encoded in phiCDHM1 and in all *C. difficile* strains deposited in the NCBI identified using blastp (Oct/2011). ML analysis was performed using MEGAv5.01 with alignment carried out in with parameters set for the JTT nucleotide substitution model, with invariant rates, using all sites and CNI for Tree Inference and bootstrapped with 500 replicates. Symbols indicate which *agr* loci the HK genes belong to.

The resulting trees for each gene, *agrD* (Figure 22), *agrB* (Figure 23) and the HK (Figure 24) are consistent with different types of *agr* loci encoded by *C. difficile* strains. In the *agrD* and *agrB* trees taxa cluster into three clades which correspond to the different *agr* loci, *agr1, agr2* and *agr3*. As there is no HK in the *agr1* locus, the taxa in this tree are only from *agr2* and *agr3*, and two clades are present which correspond to these two loci.

In all three trees, the phage genes cluster with the bacterial homologs from strain NAP08, NAP07 and QCD-23m63 which are both their top blastp hits and in *agr3* type loci in these bacterial strains, as checked from manually searching their genomes. Interestingly these genes are not encoded by prophage in these strains, but are near to transposase and/or integrase genes.

137

While the phage genes are closely related to the bacterial homologs, as seen from tree tropology and branch lengths, the phage encoded genes are considerably divergent from the bacterial *agr3* genes. (88 for *agrD*, 75 for *agrB* and 99 for the HK) that these sequences are closely related but have undergone some evolutionary divergence. The amino acid sequence of each phage gene was aligned to the top blastp hit sequence entries from *C. difficile* strain NAP07 (the top blastp hit was joint with identical sequences in *C. difficile* strains NAP07, NAP08 and QCD-23m63). The similarities of the *agrD*, *agrB* and HK sequences to their closest homologs are 56%, 58% and 61% respectively (Figure 25). This suggests that while the phage and bacterial genes in agr3 are closely related, and that this locus can undergo horizontal gene transfer between strains via phage or transposons, the phage encoded genes on phiCDHM1 are genetically distinct.
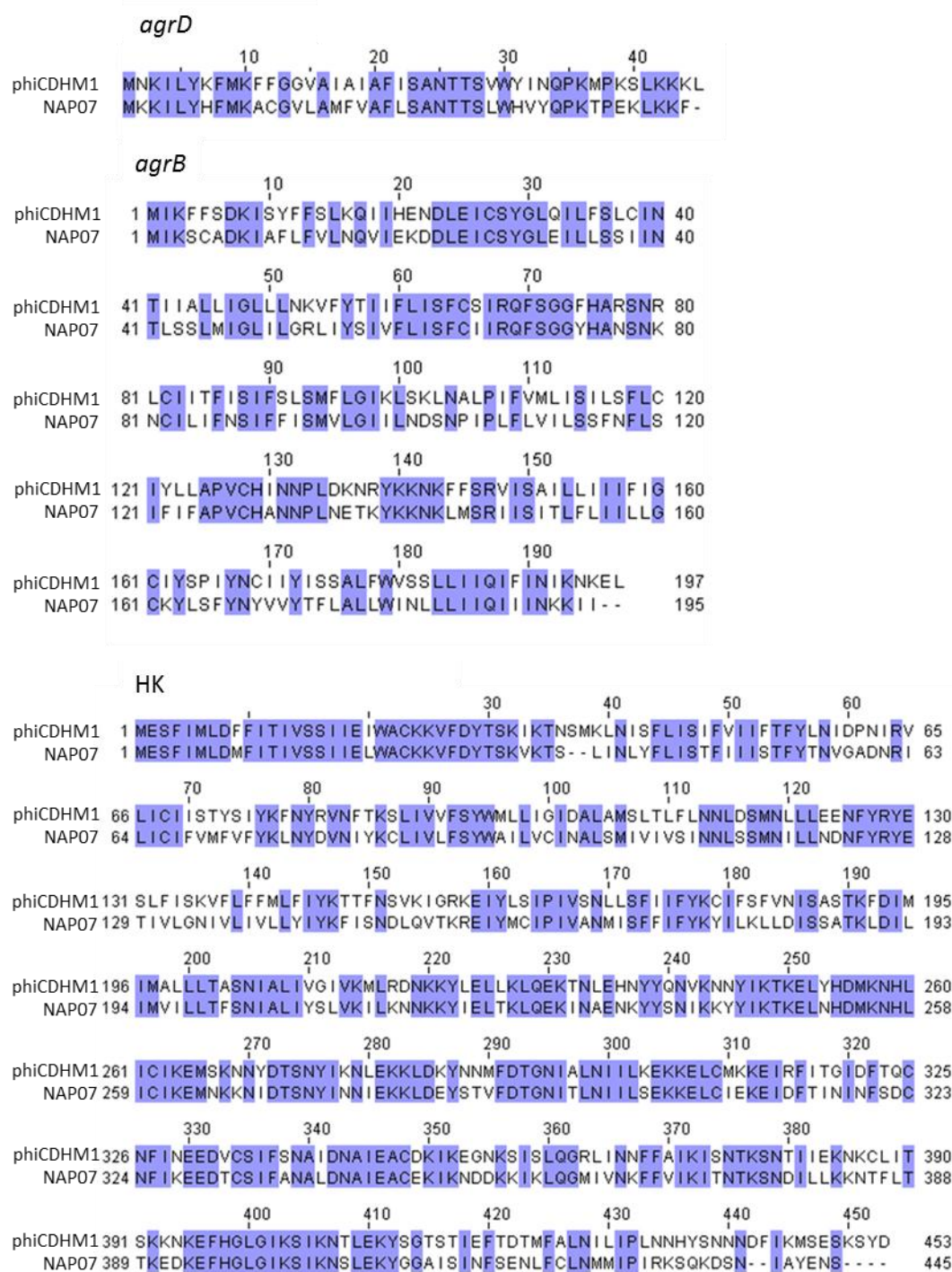
Figure 25. Alignments of *agrD*, *agrB* and HK genes of phiCDHM1 and *C. difficile* strain NAP07. ClustalW alignment of each gene at the amino acid level. Top: *agrD*, sequences share 58% identity and are 45 and 46 aa long respectively. Middle: *agrB*, sequences share 56% identity and are 197 and 195 aa long respectively. Bottom: HK gene sequences share 61% identity and are 453 and 445 aa long respectively. Purple highlights identical residues

5.2.9. Prediction of protein interactions of the HK indentifies potential host targets of the phage QS system.

As there is no response regulator encoded in the phage (or bacterial) *agr3* locus, a search of the STRING database was conducted to identify what the agr HK could interact. The STRING database links interaction data based on experimental and bioinformatic evidence with (Jensen *et al* 2008). The database includes several different *C. difficile* strains so the top blastp hits of the phage homologs were queried first. All predicted interactions were based on bioinformatic evidence, which consisted of 'Co-occurrence Across Genomes' and 'Neighbourhood in the Genome'.

The queried HK from *C. difficile* CD196 in *agr1* was predicted to interact with AgrB protein, a membrane transporter protein, a ligase and, with high confidence, CdtR (Appendix Figure 7). CdtR is a global response regulator that contains a LytTR protein domain and regulates production of the binary toxin (Carter *et al* 2007). To further investigate what response may be elicited by CdtR, it was analysed in the STRING database (Appendix Figure 8). Predicted interactions include with *cdtA*, one of the binary toxin genes, as well as a transcriptional regulator protein, two hypothetical proteins and three kinases. However, neither *C. difficile* strain K12 (the lysogen encoding phiCDHM1) or T6 (the lytic host for phiCDHM1) encode the binary toxin genes, as determined through PCR using primers described in Section 2.1.6. Therefore, the next closest homolog from a *cdtA*- strain (CD630) was analysed, which also resulted in a predicted interaction with CdtR. A further search of this CdtR resulted in predicted interactions with kinases, a sortase B protein, a transcriptional regulator, a DltD transferase protein, two hypothetical proteins and two phage proteins, all with medium confidence scores (data not shown). The results of these predictions suggest that the HK interacts with CdtR, and that the QS response involves both host and phage encoded genes, including the toxin genes.

5.2.10. Detection of phage specific HK gene in environmental *C. difficile* genomes indicates their transfer via HGT.

This is the first report of *agr* QS genes in a phage genome and to determine how widespread the carriage of these genes is in the environmental, human and animal isolates obtained in this project used in a PCR based screen for the HK gene. Two primer sets were used that are specific for the phage encode gene, with one internal set and another external set, so results identify if the HK gene is present and if it is encoded on a related phage genome.

The results of the PCR screens found three isolates were positive for the internal primer set product including two estuarine isolates H5c (R078) and H10 (046), and as expected K12 (Appendix Figure 9). However, only isolate K12 was positive for the product from the external primer set (Appendix Figure 10). This suggests that the phage HK in H5c and H10 may be encoded by a genetically divergent prophage, or on the bacterial chromosome. The results also show that this gene, while uncommon, is present in diverse *C. difficile* strains belonging to different ribotypes. The gene's presence on a prophage (phiCDHM1) in one strain (K12) but on a divergent prophage/element in others suggests that it may have been transferred via HGT in the *C. difficile* population. The genomes of both H5c and H10 are currently being sequenced which will reveal the location of this gene.

5.3. Discussion

5.3.1. Comparison of the genome content of *C. difficile* phages.

The results of sequencing and annotation have revealed that the genomes of phiCDHM1, phiCDHM19 and phiT6 are closely related, as encode several highly conserved genes and share high levels of identity to genes in other *C. difficile* myoviruses and prophage that are associated with clinical and human strains. The SMV genomes however, while genetically divergent from the MMs

and prophages, are very similar to one another. This is interesting, as phages are considered the most genetically diverse entities on Earth, in part due to their abundance as well as the fluid nature of their genomes (Breitbart *et al* 2002). High genomic diversity present across phage genomes is evident from both large scale studies of isolated phages (e.g. Pedulla 2003, Hatfull *et al* 2008, Hatfull *et al* 2012, Smith *et al* 2012) as well as from metagenomic analysis of viral fractions sampled across multiple different environments (e.g. Breitbart *et al* 2002).

Despite the numbers of phages and their genomic diversity, closely related phages are distributed on a global scale (Breitbart *et al* 2006) and phage genome architecture is surprisingly conserved with distinct regions positioned along the genome even in divergent phages infecting different bacterial hosts (Hendrix 2003, Casjens 2005). The relatedness between the genomes of phiCDHM1, phiCDHM19 and phiT6 to the known *C. difficile* phages is therefore consistent with the observation of genetically similar phages present in geographically distinct locations. These three phages could be included into the phiCD119-like group of phages based on genome size and genetic features with slight modification to the inclusion criteria as phiT6 lacks one gene in their characteristic DNA replication cassette (Lavigne *et al* 2009). The isolation of three SMVs is more unusual as phages of their size are relatively rare in the literature (Lavigne *et al* 2009, Comeau *et al* 2012). Their genome sizes ranging in 33-35 kbp is smaller than most of the designated 'dwarf' myoviruses infecting *Proteobacteria* but their genome architecture is highly similar (Comeau *et al* 2012). Therefore the SMVs isolated in this study are proposed as a novel class of myoviruses infecting *C. difficile.* Recently another *C. difficile* SMV phiMMP04 was isolated which, as shown in the whole genome alignment has a highly similar genome to that of the SMVs and places it into this new group (Meessen-Pinard *et al* 2012).

Despite a high degree of relatedness between these phage's genome, there is a considerable amount of genetic variation between their genomes, and appear highly mosaic on the basis of gene identity from blastp searches. Mosaicism arises due to the on-going reassortment of phage genomes

which is largely facilitated by homologous recombination (Hendrix 2003). Homologous recombination acts on related sequence in phage genomes, (or host), and results in the exchange of genetic material between them. The length of material that is exchanged can vary, so resulting synteny between genomes can be extensive (in whole functional modules) or short, as in partial gene sequences (Hendrix 2003, Casjens and Thuman-Commike 2011). In the *C. difficile* phage genomes there can be seen that homology to a particular phage can be extensive through several modular regions or restricted to single genes.

Whether the mechanism behind recombination events is specific or random is uncertain, but opportunities for temperate phage to acquire novel sequences during infection are rife (Hendrix 2003). Each of the phages sequenced in this project encode novel genes as well as homologs of bacterial genes. In bacteria as a whole, the proportion of genomes that encode prophage sequences is high, on average 71% (Canchaya *et al* 2003), and their genomes provide an abundant and diverse pool of genes available for co-infecting phages to acquire (Brüssow *et al* 2004). This average is similar to the lysogeny rate in *C. difficile* as determined through TEM analysis and capsid gene carriage in this project which suggests that there is an abundant reservoir of potentially new and useful prophage genes in the *C. difficile* for infecting phage to access.

How phages acquire new genetic material can be due to erroneous DNA packaging but also mediated through the action of specific enzymes, either phage or host encoded, that are involved in DNA repair such as RecA and holiday junction resolvases (Lawrence *et al* 2002). Two of the *C. difficile* phages isolated in this project, phiCDHM1 and phiCDHM19, encode RusA orthologs indicating that these phages have the genetic potential to facilitate their own recombination events and acquisition of new genetic material.

The genome architecture of the *C. difficile* phages sequenced in this project all conform to a similar modular arrangement, by gene function, although the SMVs and MMs differ slightly in the

order of these regions. The genetic content of each module can vary between the phages and each functional module is discussed in detail below.

5.3.2. Head packaging genes.

All three MMs encode two terminase genes, TerL and TerS, and a portal protein gene which is similar to the published *C. difficile* myoviruses phiC2, phiCD119 and phiCD27, as all encode two Terminase genes, followed by a portal protein gene, with the exception of phiCD119 which has a hypothetical protein located between the TerL and Portal gene (Govind *et al* 2006, Goh *et al* 2007, Mayer *et al* 2008). However, the TerL and TerS genes in the three phages belong to different Terminase families. The variation in single genes between otherwise closely related phages has also been observed in *Streptococcus* prophage (Romaro *et al* 2009) and varying gene affinities between Stx-converting *E. coli* phages (Smith *et al* 2012).

The three small myovirus genomes are more unusual as they encode only one identified terminase gene and a portal gene and lack a TerS. Most phages encode two Terminase genes, with only one known phage, *Bacillus subtilis* φp29 encoding a TerL only (Casjens 2011). The packaging module in most dsDNA phages comprises of two terminase genes, a large (TerL) and small (TerS) subunit and a portal gene (Casjens 2011). The packaging module proteins are responsible for the translocation of newly replicated DNA into the procapsid. They form a complex which functions as a nanomotor, using energy produced from the ATPase activity. The portal protein oligomerises to form a hollow ring bound to the procapsid with a multimeric ring of TerL proteins and outermost ring of TerS. The TerS proteins binds and recognises the phage DNA, and the ATPase activity of the TerL drives the translocation of the DNA, as well as the cleavage of the DNA at an endonuclease site for ensuring one genome is packaged. Presumably, the SMVs must encode a gene with a homologous function of TerS.

5.3.3. Morphogenesis genes.

All of the phages encode predicted structural proteins and a number of these are highly conserved between phage genomes but the number of ORFs in this module varies between them. For example, the three MM genomes encode two genes predicted to encode a minor head protein and scaffold protein respectively, but inserted between these there are two ORFs in the genomes of phiCDHM1 and phiT6, and three ORFs in phiCDHM19. The number of ORFs separating the head and scaffold proteins is also varied between the published phages (Govind *et al* 2006, Goh *et al* 2007, Mayer *et al* 2008).

The structural module becomes more diverse following the predicted capsid gene. For example, this region encodes seven ORFs before the Phage Sheath domain ORF in phiCDHM1, five ORFs in phiCDHM19 and nine ORFs in phiT6. Additionally, of the five ORFs in phiCDHM19, two contain protein domains (Rho_N and Phage_connect_1) that are not encoded by either phiCDHM1 or phiT6. There is a highly variable region located between the XkdN and tape measure proteins, for example, in phiCDHM19 the tape measure protein is one ORF downstream of the XkdN, but phiCDHM1 encodes four ORFs upstream of the TMP gene, including an ORF with a Bro-N domain. In phiT6 this region is even more extensive, with ten ORFs encoded upstream of the TMP gene, which also include more regulatory genes such as ORFs with Bro_N, ORF6N and zf-ribbion_3 domains. This region of the genome seems especially variable across the *C. difficile* MMs (Goh *et al* 2006, Govind *et al* 2006). This region of the genome is where the CRISPR arrays are located in prophages phiCD630_1 and phiCDD630_2 and is discussed in more detail in section 6.2.4., as these regulatory proteins may be involved in the processing of the phage encoded CRISPR arrays.

In the SMV genomes there are fewer ORFs identified in the structural module than in the MMs, with eleven ORFs and although have complimentary functions encodes different structural genes, such as the Phage_capsid, and Phage_H_T_join which are not present in the MM genomes. Conserved genes include the ORFs containing DUF646, Phage_sheath_1, DUF2001 and XkdN protein

domains. An ORF with a Phage_connect_1 domain is conserved with phiCDHM19 only. Similar to the MMs, the Phage_sheath_1, DUF2001 and XkdN ORFs are adjacent on the genomes. However, the SMVs also encode a predicted serine protease (containing a CLP_protease domain) upstream of the capsid gene which is not present in the MM genomes, although the siphovirus phiCD3656 shares this ORF also in its structural region (Hogan *et al* 2011) and another SMV phiMMP04 (Meessen-Pinard *et al* 2012). A number of proteins with this domain (IPR001907) lack peptidase activity, and therefore may not function as an enzyme in the phages.

5.3.4. Lysis and attachment genes.

This module includes genes whose products are thought to interact with the host cell surface, either for the purpose of the phage gaining entry to the cell or release of phage progeny. Surprisingly there is a large amount of variation between the phages within this module in terms of additional genes with unknown functions although in all the genes involved in phage entry are located together before the exit genes, the predicted holin and endolysin genes.

The module starts with the LysM and ends with the Amidase_3 ORFs. This module encodes 12 ORFs in phiCDHM1, 17 ORFs in phiCDHM19, 17 ORFs in phiT6, and 13 ORFs in the SMV genomes all encoded on the sense strand. The module includes a number of expected proteins which are characteristic of phage-host cell wall interactions such as a predicted cell wall hydrolase enzymes involved in degrading the cell wall and ORFs predicted to encode tail structural proteins involved in receptor recognition such as the baseplate-J gene. Most of these genes are highly conserved, for example the LysM is present in many bacteria and phages encoded enzymes that bind and degrade peptidoglycan (Buist *et al* 2008) and the predicted cell wall hydrolase gene with a NLPC_P60 domain that also targets peptidoglycan and degrades the bacterial cell wall (Anantharaman and Aravind 2003).

Immediately downstream of these genes are ORFs predicted to encode structural tail proteins, with some containing different DUF domains which are differentially encoded by the phages. For example, in all the phage genomes an ORF encoding a predicted tail fiber protein are located after the DUF2313 protein except for in phiT6 which encodes an ORF with a DUF3751 domain instead.

The region downstream of the predicted tail proteins becomes increasingly divergent between the phage genomes. For example: phiCDHM1 encodes three hypothetical proteins tail fiber and holin genes; phiCDHM19 encodes five ORFs between the Gly_Rich and holin genes, including a Phage_XkdX domain; phiCDHMT6 encodes four ORFs and the SMVs encode two ORFs, including one with a XhlA domain. The XhlA (Pf10779) protein family has been attributed with lytic activity and is a membrane protein. It is encoded by the other myoviruses phiC2, phiCD119 and phiCD27. Interestingly the other three phages in this study do not encode ORFs with this domain, despite its essential role for host cell lysis in the *B. subtilis* prophage PBSX (Krogh *et al* 1998).

The exit genes, the holin and endolysin, are highly conserved in all the *C. difficile* phage genomes. This is most likely due to their essential function as at the end of a lytic replication cycle, phages need to exit the the host cell (Wang *et al* 2000). Two phage encoded proteins are responsible for this timed lysis: an enzyme to degrade the peptidoglycan layer of the cell, and a holin which permeabilises the bacterial cell membrane giving the phage endolysin access to peptidoglycan and causing lysis of the cell (São-José *et al* 2007).

Holins are small proteins <150 aa and contain differing numbers of transmembrane domains (TMDs), which determine their class (Wang *et al* 2000). In all the phages a holin and endolysin gene were identified. PhiCDHM1 and phiCDHM19 holin genes contain a Phage_holin_4 domain whereas phiT6 and phiCDHM13 annotations have been made from blastp matches based on sequence homology with no domain detected.

The Phage_holin_4 domain is also found in the *tcdE* gene, encoded on the PaLoc of *C. difficile* involved in toxin release (Tan *et al* 2001). The homology between *tcdE* and the phiC2 holin has been suggested as supporting the hypothesis PaLoc has a phage origin in *C. difficile* (Goh *et al* 2007). Analysis of the *C. difficile* prophage holin sequences across different clinical strains has been found to discriminate between myoviruses and siphoviruses, and also different ribotypes (Shan *et al* 2012). Interestingly ,the holin gene in phiCDHM1 is so divergent from the other myovirus holin genes the, primers targeting in Shan *et al* (2012) do not amplify the gene (data not shown) and when primers where designed specifically targeting this holin gene they were not suitable for use as non-specifically amplified bacterial *tcdE* sequences. To investigate the phylogeny of the phage holin genes and TcdE, they were aligned using MUSCLE and ML phylogenetic analysis performed (Figure 26).

The three phage holins from phiCDHM1, phiCDHM19 and phiMMP04 cluster in a separate clade with the TcdE gene supported by a bootstrap value of 100 and provides further support for the hypothesis that *tcdE*, and the PaLoc, may have a phage origin. In another species, *Shigella dysenteriae* .1, it appears that a prophage has been the source of its virulence island (McDonough and Butterton 1999). Strain .1 encodes Shiga toxin genes, which are usually encoded in a complete prophage in other *S. dysenteriae* strains. In strain .1 the prophage has become degenerate losing most of the phage genes with only the toxin and lysis genes remaining (McDonough and Butterton 1999). If similar event had occurred in *C. difficile* this would explain the presence of the holin-like TcdE in the PaLoc (Goh *et al* 2007).
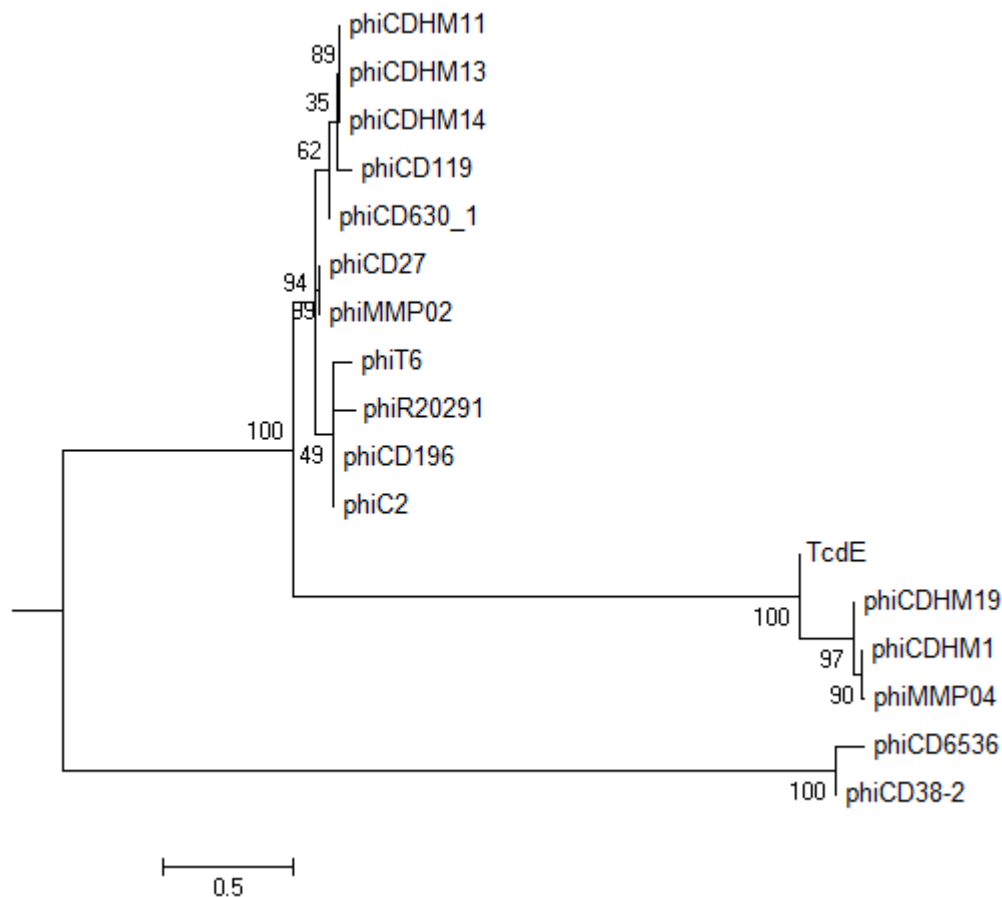
Figure 26. ML phylogenetic analysis of the holin gene of *C. difficile* prophage and the TcdE gene. Alignment was performed in MEGAv5.01 using MUSCLE, of sequences of holin genes from phages sequenced in this project, phiCDHM1, phiT6, phiCDHM19 and phiCDHM13, as well as published phages holin genes (see Appendix Table 2 for accession numbers). The ML analysis was performed using the JTT model with gamma invariant rates, NNI and all site parameters set and bootstrap values were calculated to 500 iterations. The taxa cluster into three clades which are supported by high bootstrap values: the first contains phiT6 and other other prophage, MM and LTM sequences; the second clade contains phiCDHM1, phiCDHM19, phiMMP04 and *tcdE* and the third clade contains both SV sequences.

All the phages encode an endolysin gene with an Amidase_3 protein domain, and appear highly conserved between their genomes. Again this is likely due to the essential function of this protein as the endolysin is a hydrolase which usually has one catalytic domain that cleaves

peptidoglycan and amidases specifically cut the **N**-acetylmuramic acid and l-alanine residue interpeptide bond (Borysowski *et al* 2006).

Typically, the holin and endolysin genes are located together on phage genomes (São-José *et al* 2007). While in the genomes of phiCDHM1, phiCDHM19 and phiCDHM13 the two genes are adjacent, in phiT6 there are three ORFs, (most likely a single gene requiring re-sequencing), which are inserted between these. These are predicted to encode AbiF genes, as have Abi_2 like domains. Most of the known *C. difficile* phages have holin and amidase adjacent to one another, except for phiC2, which like phiT6 also encodes a predicted AbiF protein that is inserted between them (Goh *et al* 2007). The AbiF proteins are predicted to be involved with abortive phage infection, and are typically encoded by their bacterial hosts. The presence of these genes on phage genomes demonstrates an unusual mechanism *C. difficile* phages may use to inhibit secondary phage infection, which could be useful considering the high proportion of lysogeny in this species.

5.3.5. Lysogeny control genes.

The lysogeny control genes are chiefly an integrase and its associated repressor. An integrase gene was identified in all the phage genomes, which are all orientated on the antisense strand following the lysis module which is consistent with the other *C. difficile* MMs (Govind *et al* 2006, Goh *et al* 2007). The integrase inserts the phage DNA genome into the host chromosome via site specific recombination between the phage attachment site attP and bacterial host site, attB. The integrases contain a Phage_Integrase domain, belonging to the tyrosine family of integrases, where cleavage is catalysed by a tyrosine residue at the C-terminus end of the domain (Groth and Calos 2004). These integrases normally require additional host or phage encoded co-factors responsible for binding site recognition, which could be the roles of surrounding hypothetical protein genes, the carriage and sequence of these surrounding genes vary between the phages (Groth and Calos 2004).

Maintenance of the lysogenic state is dependent on a repressor protein which binds to DNA in the promoter regions inhibiting transcription and its anti-repressor. Once induced, an excisionase excises the prophage genome. There are a number of different candidate genes for these roles across the different phage genomes. For example, phiCDHM1 encodes 3 putative repressors, Orf50 (HTH_3), Orf51 (HTH_3) and orf59 (HTH_3) and 2 anti-repressors Orf52 (ORF6N) and Orf56 (Bro-N) and a regulatory protein containing a T5orf173 domain in Orf47. However, only Orf52 is on the antisense strand, upstream to the phage integrase and makes this the most repressor. A putative excisionase (Orf61) was identified with a HTH_17 domain. In phiCDHM19, 2 repressors also have been identified, in Orf50 (HTH_19) and Orf59 (HTH_3), with Orf50 on the antisense strand and one anti-repressor, Orf55 (Bro-N/ANT) as well as a putative excisionase (Orf60) with HTH_17 domain. In phiT6, 5 repressors are encoded and the genome encodes 3 anti-repressors, 2 of which have a Phage_pRha and ANT domain (Orf54 and Orf61 respectively) and the other has an ORF6N domain (Orf57). PhiT6 also encodes proteins that are not encoded in the other phage genomes, including ORFs with a SIR2_2, a HTH_26 and an Arc protein domain but does not itself encode a predicted excisionase gene.

While the MM genomes are typical in the location of their integrase gene and lysogeny region compared to the published phages, those of the SMVs are more unusual. The integrase gene, Orf51, is located downstream of the replication module and orientated the sense strand with no obvious repressor protein nearby. This is arrangement is similar to the siphovirus phiCD38-2 and in the other SMV (Sekulovic *et al* 2011, Meessen-Pinard *et al* 2012).

5.3.6. DNA replication and transcription genes.

None of the phage genomes sequenced encode a predicted DNA or RNA polymerase so they presumably rely on host encoded replication machinery to facilitate their DNA genome replication and gene transcription. The phages do encode several genes that could be involved in the process of

replication, that have various predicted activities such as DNA binding, methylation and recombination. Several of these genes are conserved between the phage genomes but others are encoded by just one or two and the modules differ in gene content and length. For example, the DNA replication module in phiCDHM1, phiCDHM19 and phiT6 contain 28 ORFs, 22 ORFs and 15 ORFs respectively. All phages encode a predicted ERF protein, which is an essential recombination factor with single strand annealing function conserved in phages (Iyer *et al* 2002a). Also, all the phages encode an ORF with a DnaB_2 protein domain, which is involved in the initiation of DNA replication, and an ORF with a SSB protein domain, which is involved in binding single strand DNA.

Specific conserved genes in this region are one requisite genome feature for their inclusion to the phiCD119-like taxonomic group of *C. difficile* phages (Lavigne *et al* 2009). However, while both phiCDHM1 and phiCDHM19 encode the required genes to place them within this group, phiT6 does not encode the hypothetical protein within this cassette. This highlights the divergence of its genome in the DNA replication module and either means phiT6 should be classes separately from these phages or require revision of the phiCD119-like group of phages is needed considering the relatedness of phiT6 genome in other modules. The divergence is further exemplified by the absence of other ORFs in phiT6 which are conserved between phiCDHM1 and phiCDHM19. These include an ORF with a RF-1 domain, involved in release of peptide chains; an endodeoxyribonuclease with a RusA protein domain and AntA anti-repressor protein.

Despite encoding several conserved genes, all three phages encode genes that are not present in the other's genome. For example, phiCDHM1 encodes two ORFs, one with a Bro-N domain and the other a predicted NTPase with NACHT domain. PhiCDHM19 encodes two ORFs, one with a NA37 (37-kD nucleoid-associated bacterial protein) domain and the second with a NUMOD1 (DNA-binding) domain. PhiT6 encodes a putative DNA binding protein containing a HTH_36 domain, which is inserted between the ERF and DnaB_2 ORFs, another ORF with a DUF1064 (unknown protein) domain and an ORF with a Mrr_cat (restriction endonuclease) domain.

The position and arrangement of the DNA replication module in the SMVs is more similar to that of the SVs than the MMs and LTMs and is much smaller. However, the SMVs do encode the highly conserved predicted SSB protein, Orf45 which is encoded in all the known *C .difficile* phage genomes. They also encode another DNA binding protein, with a HTH_19 protein domain and a XRE family transcriptional regulator. The genomes lack a predicted DNA methylase, ERF and DnaB_2 DNA replication protein which are encoded in the MMs. Interestingly, three of the genes that have either been lost or acquired by one or two of the SMVs are located in the DNA replication module. Although these are hypothetical proteins with no known function assigned, the genetic differences in this module could account for the observed variation in host range and efficiency of plaquing in section 4.2.12. In a study of ten phages infecting *P. aeruginosa,* genetic variations of 3-17% accounted for observed differences in 5-58% host infections (Ceyssens *et al* 2011).The gene diversity in the genomes of the SMVs could highlight the impact phage genetics has on host infection in *C. difficile*.

5.3.7. Putative lysogeny conversion genes.

Bacteriophage contribution to their host physiology can directly enhance virulence; with some phages encoding toxin genes and infection converts their hosts into pathogenic strains (Wagner and Waldor 2002). The *C. difficile* phages do not encode recognisable toxin genes, however all the sequenced phage genomes do encode a region encoding putative lysogenic conversion factors. Even phages lacking obvious virulence factors, such as toxin genes, are likely to exert an influence on their host through a wide repertoire of encoded genes as reported in many phages (Pedulla *et al* 2003). Regions of additional genes are commonly encoded in temperate bacteriophages infecting environmental/non-pathogenic bacteria in lysogenic conversion regions, but whether these genes confer selective advantages is largely unknown (Canchaya *et al* 2003). The regions encode genes with no predicted association to phage structural, lysogeny control proteins, or replicative proteins. The locations of these regions are conserved between the phages, although

their content is not. These genes are encoded on the antisense strand, downstream of the lysis genes. Modules encoding putative lysogen conversion factors were identified in all the phages sequenced in this project.

The MMs genomes lysogeny conversion modules are located following the endolysin gene and upstream of the integrase. Each phage encodes different genes in their respective modules. For example, phiCDHM19 encodes an ORF containing a Nudix_hydrolase domain in this region. This protein family is associated with mediating stress response by modulating signalling molecules or metabolites through enzymatic activity of its hydrolase. PhiCDHM1 encodes three genes, *agrD, agrB* and a HK, in this region that are homologs of genes in the bacterial *agr* QS system. These genes are discussed in detail in the next section.

The SMVs have a different arrangement of their genome, but a putative lysogeny conversion module was identified downstream of the lysis proteins. This encodes six ORFs, including an ORF with a CbiA domain and another ORF with a RNase_T domain. This it the first report of a phage encoded protein with a CbiA domain. The CbiA protein family consists of cobyrinic acid a,c-diamide synthases and are involved in the biosynthesis of cobalamin (vitamin B12) (Pollich and Klug 1995). Cobalamin is a coenzyme which is essential for numerous metabolic processes and is produced by complex biosynthetic pathways including multiple enzymatic steps which are conserved among eubacteria (Raux *et al* 2000). *C. difficile* CD630 encodes a CbiA protein alongside other genes involved in cobalamin biosynthesis (Sebaihia *et al* 2006).

The cobalamin synthesis pathway genes have been subject to horizontal gene transfer between the *Firmicutes* and another bacterial phylum the *Thermotogales* (Swithers *et al* 2012), but this is the first report of this protein domain in a phage encoded gene. Other phages are known to encode genes that are involved in redirecting the host metabolic processes (Thompson *et al* 2011). It may be that the SMVs use this gene to influence *C. difficile* metabolism during infection in a similar manner. The second ORF is a predicted exonuclease as the RNase_T domain is associated with high

specificity binding and cleavage of RNA associated with production of stable RNA as well as binding and cleaving ssDNA and dsDNA, but with less specificity (Zuo and Deutscher 1999). This gene could be involved in transcription or as a restriction enzyme. Genes with a RNase_T domain are not present in the other *C. difficile* phage genomes. Also, within this module, the SMVs encode an ORf with an AP2 domain which is characteristic of endonucleases and DNA binding proteins encoded by other phages, for example the *E. coli* Stx2-converting phage 1717 (NC_011357.1) in OrfStx2-1717_gp29 (YP_002274239.1) (Zhang *et al* 2010).

5.3.8. Predicted phage encoded genes involved in secondary phage immunity and stress response.

The phages encode several genes with predicted functions that inhibit secondary phage infection. These include the sipho gp157 gene encoded by phiCDHM1 and phiCDHM19, which is thought to prevent superinfection, as well as the predicted AbiF proteins encoded by phiT6, whose predicted function is prevent phage infection although the mechanism AbiF does this is not known. The AbiF gene is unusual as although also encoded by phiC2, this system is typical of the bacterial host. Also, phiT6 encodes other genes that may be involved in phage inhibition, for example, the predicted type II restriction endonuclease, with a Mrr_cat domain. This may act as a restriction modification system involved in phage defence, by degrading invading phage DNA, as a R/M system with a type II RE is encoded by *S. auerus* phage42 (Dempsey *et al* 2005). Normally, the enzyme is adjacent to the phage integrase gene, but as this is not the case in phiT6 and the gene may have a alternative role. One such function could be as a regulatory protein, as the Mrr protein in *E. coli* is the effector of the pressure induced SOS response (Aertsen and Michiels 2005). Also, a recent paper has described the antagonistic effect of co-expression of Mrr and Res restriction endonuclease, proposed as a defence mechanism against foreign methylases (Mebrhatu, *et al* 2011). This has relevance here as the ResIII protein encoded by the SMVs may be a component in a phage encoded phage defence mechanism that has been appropriated from a host system.

Two other genes encoded by phiT6 have protein domains associated with stress response. Orf51 contains Sir2-like domain which is similar to the domain found in sirtuins. These are proteins which can be involved in aging and stress response regulation (Rajendran *et al* 2011). Orf30 contains a zf-ribbon_3 domain, which through the NCBI blastp shows homology to proteins annotated as stress response phage proteins. The presence of these genes in a phage genome suggests their role in combating the cellular response to secondary phage infection and thus avoiding a lethal infection or in enabling the phage to detect times when the prophage should excise from the cell and switch to the lytic cycle.

5.3.9. Distribution of conserved protein domains in *C. difficile* phages reveals insights to their potential evolution and niche adaption.

The *C. difficile* phages encode a number of genes with protein domains that are highly conserved in their genomes, as well as several that are novel or highly unusual. The distribution of both conserved and unusual protein domains in the *C. difficile* phage genomes highlights their mosaic nature, and also provides some clues to their origins and evolution (Table 17).

For example, there is some that suggests horizontal gene transfer has occurred in genes with predicted essential function. For example, all the phages encode a gene with a predicted large subunit terminase, but they contain domain motifs of different protein families. All three of the terminase domains are in many phage genomes and could have evolved in the *C. difficile* phage genomes horizontally or vertically. Similarly, the portal protein is conserved in all the genomes, but belongs to different protein families in the SMVs and the MMs. Other genes are highly conserved in terms of their functional role, such as the Phage_Integrase and SSB protein which is encoded in all the sequenced phage genomes. In contrast a number of genes with predicted regulation roles, such those with the HTH protein domains, appear to be highly genome specific and, for example, NA37 is only present in phiCDHM19 and is uncommon in other viral genomes.

Table 17. Protein domains in phage genomes, conservation and substitution.
Symbol √ indicates number of genes containing the domain encoded by the phage.

Structural domain	Protein Ref.	Function	No. of viral entries in NCBI genedb	phiCDHM1	phiCDHM19	phiT6	phiCDHM13
COG5484		terminase small subunit	11	√	-	-	-
HTH_28	PF13518	regulation, DNA binding, terminase small subunit	8	-	-	√	-
Terminase_1	PF03354	terminase Large Sub-unit	156	-	-	-	√
Terminase_3	PF04466	terminase Large Sub-unit	110	√	-	√	-
Terminase_6	PF03237	terminase Large Sub-unit	211	-	√	-	-
Phage_prot_Gp6	PF05133	Portal	130	√	√	√	-
Phage_portal	PF04860	Portal	76	-	-	-	√
Phage_Mu_F	PF04233	head morphogenesis	124	√	√	√	-
Phage_GP20	PF06810	structural protein	26	√	√	√	-
Phage_capsid	PF05065	capsid protein	198	-	-	-	√
Phage_H_T_join	PF05521	phage head-tail adaptor	69	-	-	-	√
Phage_connect_1	PF05135	phage head-tail adaptor	32	-	√	-	√
CLP_protease	PF00574	unknown function (protease or inactive)	67	-	-	-	√
Phage_sheath_1	PF04984	tail sheath protein	59	√	√	-	√
DUF2001(CL0504	PF09393	structural beta barrel protein	6	√	√	√	√
XkdN	PF08890	unknown function	6	√	√	√	√
tape_meas_nter	TIGR02675	tape measure protein	99	√	-	√	-
LysM	PF01476	peptidoglycan binding/lysis function	66	√	√	√	√
NLPC_P60	PF00877	membrane protein, lipoprotein motif	79	√	√	√	√
Baseplate_J	PF04865	baseplate protein	84	√	√	√	√
Gly_rich	PF12810	function unknown	4	-	√	√	√
Phage_XkdX	PF09693	unknown function	45	-	√	-	-
Phage_holin_4	PF05105	Holin	29	√	√	-	-
Amidase_3	PF01520	cell wall hydrolase	61	√	√	√	√
DUF1064	PF06356	unknown	41	-	√	-	-
DUF2313	PF10076	Tail	14	√	√	√	√
DUF646	PF04883	Tail	92	√	√	-	√
DUF2577	PF10844	unknown	5	√	-	√	-
DUF2634	PF10934	unknown	8	√	√	√	√
DUF955	PF06114	unknown, may be catalytic	57	-	√	√	-
NA37	PF04245	unknown cellular regulation	4	-	√	-	-
SIR2_2	PF13289	unknown cellular regulation	1	-	-	√	-
Arc	PF03869	repressor	19	-	-	√	-
RHH_1	PF01402	Repressor	12	-	-	√	-
AntA	PF08346	anti-repressor	22	√	√	-	-
Pencillinase_R	PF03965	repressor, recognition of DNA	728	√	√	-	-
BroN	PF02498	regulation, DNA binding	310	√	√	√	-
zf-ribbon_3	PF13248	regulation, DNA binding	11	-	-	√	-
HTH_36	PF13730	regulation, DNA binding	36	√	-	√	-
HTH_3	PF01381	regulation, DNA binding	8	√	√	√√√	-
HTH_17	PF12728	regulation, DNA binding	65	√	√	-	-
HTH_19	PF12844	regulation, DNA binding	3	-	√	√	√
HTH_26	PF13443	regulation, DNA binding	12	-	-	√√	-
T5orf172	PF10544	transcriptional regulator	63	√	-	-	-
ORF6N	PF10543	transcriptional regulator	16	√	-	√√	-
sipho gp157	PF05565	unknown function – super-infection immunity	36	√	√	-	-
Agrb	PF04647	*agrD* processor	0	√	-	-	-
Abi_2	PF07751	abortive infection, DNA binding	1	-	-	√	-
CbiA	PF01656	nucleotide binding	0	-	-	-	√
NACHT	PF05729	apoptosis associated	1	√	-	-	-
XhlA	PF10779	cell-surface associated haemolysin	20	-	-	-	√

Table 17. *contd*.

Structural domain	Protein Ref.	Function	No. of viral entries in NCBI genedb	phiCDHM1	phiCDHM19	phiT6	phiCDHM13
				phage			
HATPase_c	PF02518	ATPase domain of histidine kinase	96	√	-	-	-
HSP90	PF00183	Heat shock protein, cell viability	155				
Nudix_Hydrolase_39	cd04699	regulation through cell cycle or stress active on signally molecules or metabolites	39	-	√	-	-
DNA_methylase	PF00145	DNA methyltransferase	208	√	√	√	-
phage integrase	PF00589	Integrase	41	√	√	√	√
AP2	PF00847	transcription factor	57	-	-	√	√
ERF	PF04404	single strand annealing protein	75	√	√	√	-
SSB	PF00436	single strand binding protein	130	√	√	√	√
dnaB_2	PF07261	replication initiation and membrane attachment	24	√	√	√	-
RusA	PF05866	holiday junction resolvase	75	√	√	-	-
NUMOD1	PF07453	DNA binding	39	-	√	-	-
UPF0150	PF03681	RNA binding	9	-	√	-	-
RF-1	PF00472	peptide chain release factor	4	√	√	-	-
HNH	PF01844	endonuclease	661	-	-	-	√
RNase_T	PF00929	Exonuclease	5	-	-	-	√
ResIII	PF04851	type III restriction enzyme, res subunit	22	-	-	-	√
Mrr_cat	PF04471	type II restriction enzyme	6	-	-	√	-
Rho_N	PF07498	termination factor disengages newly transcribed RNA from its DNA	35	-	√	-	-

Further evidence that points to the evolution *C. difficile* phage genomes via horizontal gene transfer can be found in the genomes of the SMVs. They encode genes with protein domains that are commonly identified in the viral entries, but are not present in the genomes of the *C. difficile* MMs. For example, phiCDHM13 encodes 4 ORFs that have a Phage_Capsid, HNH, Terminase_1 and Phage_H_T_join, which are present in genome of the SV phiCD6356 and SMV phiMMP04. There conserved presence suggests there has been convergent evolution between the SMV and siphovirus genomes involving horizontal gene transfer. These protein domains are also in genes encoded by phages infecting other Clostridial species including *C. tyrobutyricum*, phiCTP1 (NC_014457.1) and *C. perfringens* phages phi3626 (NC_003524.1), phiSM101 (NC_008265.1) and phiS63 (NC_017978) which suggests they may have evolved across hosts.

Additionally, the protein domain data also provides evidence of convergent evolution and horizontal gene transfer from the hosts as other some uncommon protein domains are present in viral genomes that infect hosts which inhabit similar environmental niches (both the gut and marine ecosystems). For example, the Mrr_cat domain encoded in phiT6 is not in any of the published *C. difficile* phages genomes, but is encoded by a *C. botulinum* type C neurotoxin converting phage (Accession: NC_007581.1, Sakaguchi *et al* 2005), two *Propionibacterium* phage, infecting *P*. *acnes,* an anaerobic bacteria which can inhabit the gut (Lood and Collin 2011), *Thermus* phage IN93 (NC_004462.2, Matsushita and Yanase 2008), a *Synechococcus* phage P60 (NC_003390.1, Chen and Lu 2002) and *Cafeteria roenbergensis* virus BV-PW1, a virus infecting the bacterial grazing marine flagellate *C. roenbergensis* (NC_014637.1, Fischer *et al* 2010)*.* Also phiT6 encodes an Orf with a RNAse_T domain which is also present in five viral genomes that infect different species of algae: *Bathycoccus* sp. RCC1105 virus BpV1 (NC_014765.1); *Ostreococcus lucimarinus* virus OlV1 (NC_014766.1); *Micromonas* sp. RCC1109 virus MpV1 (NC_014767.1); *Ostreococcus tauri* virus 1 (NC_013288.1) and *O. tauri* virus 2 (NC_014789.1). The NACHT protein domain in an ORF encoded by phiCDHM1 has not been reported in any other phage genome, with the only other viral entry in the NCBI database, in the *Paramecium bursaria Chlorella* virus 1 infecting the marine protozoa *Paramecium bursaria (*NC_000852.5). The rarity of these protein domains suggest that they are likely to have been acquired from the host genome.

The incidence of these unusual genetic traits shared between these viruses may reflect particular selective pressures in in the terrestrial or marine environment and be adaptive factors to enhance survival under conditions. Their varied conservation also serves as further evidence to suggest horizontal gene transfer has occurred between virus and host genomes during cross-infection by similar phages. Whether they can be classed as markers for environmental origins for these phages is not fully determined, but is the most marked difference between their genomes and the clinically associated phages infection *C. difficile*.

The phages sequenced in this project also encode genes with protein domains that are not present in other phage genomes. For example, the CbiA domain in the SMV genomes is not present in any other viral entries in the NCBI database, which suggests if phages influence the vitamin B12 synthesis pathway in *C. difficile*; this is a highly novel mechanism for phages. Another is the AgrB domain in an ORF encoded in phiCDHM1 is not present in any other viral genome. It is a homolog of the bacterial gene AgrB, which is involved in QS, and the presence of this gene in the phiCDHM1 genome also presents a highly novel mechanism for phage to influence the physiology of the host.

5.3.10. Using the capsid gene as a molecular marker to detect specific myovirus types encoded as prophages in environmental *C. difficile* isolates.

One of the most conserved genes in the phage genomes is the capsid gene. Phylogenetic analysis of this gene showed that the phages fall into distinct clades which largely correspond to phage morphology. However, the MMs are split into two clades despite their morphology. To probe the myovirus diversity on a genetic level primer sets were designed that target each group and used to detect their presence in the *C. difficile* isolates used in this project.

The results of the screen were largely consistent with the TEM analysis. For example, using a genetic screen to determine lysogeny may be more sensitive as even though the screen is restricted to detecting only myoviruses, the proportion of lysogens encoding related myoviruses was 72%, and using TEM was 62% (section 3.2.7.). However, there is no distinguishing between active and degraded prophage using the genetic marker which could account for the higher detection rate. Slightly higher rates of phage sequences specific phiC2 was identified using DNA hybridization probes to screen clinical isolates with 85% of 37 positive (Goh *et al* 2007).

Also, the PCR based screen showed multiple prophage carriage was common and lysogens were dominated by prophage belonging to the MM morphotype which is consistent to the TEM results. The PCR based screen did reveal that there is often co-carriage of genetically distinct but

morphologically identical MMs that the TEM analysis could not detect. The carriage of multiple phages is consistent with reports in the literature (Sebaihia *et al* 2006, Goh *et al* 2007, Shan *et al* 2012, Nale *et al* 2012) and the markers used in this study highlight the co-carriage of related myoviruses.

Where direct comparison between the two methods (PCR and TEM) using the same isolates could be performed, generally the PCR assay was more sensitive than the TEM examination as isolates were positive for the capsid gene but did not have phage observed in their lysates. However, in three cases the TEM analysis was more sensitive, as myovirus particles were observed in the lysates but no PCR product amplified. This suggests there remains more genetic diversity within *C. difficile* myoviruses than currently known. In some instances both methods indicated the isolates are free from prophage, for example in all isolates examined belonging to R031. In a species with such a high lysogeny rate, and with all isolates sequenced containing prophages, isolates without prophage are interesting due to their rarity. A number of these isolates are currently undergoing sequencing the results will determine whether these isolates are indeed free from prophages.

5.3.11. Diversity of phage *agr* genes places them in a third locus type, *agr3*, which is susceptible to HGT in *C. difficile*.

The presence of homologs of the bacterial agr QS system in a phage genome is new. The phage encodes the genes which can produce and sense the signal, but lacks the gene for the response regulator. This suggests that the phage can detect itself and regulate its own lysogeny population. QS is a mechanism by which bacteria can communicate based on the surrounding cell population density and it may be that the phage encodes a signal to act as a release for trigger or suppression of phage replication in the surrounding culture. Alternatively the phage encoded genes could produce a response on the bacterial host and act as fitness factors that could influence host physiology. *C. difficile* strains are known to encode two types of *agr* loci, *agr1* and *agr2* (Sebaihia *et*

al 2006, Stabler *et al* 2009) based on the gene content and organisation of the loci. The phage *agr* locus is different from *agr1* and *agr2* and I propose a third type, *agr3*. The phylogenetic analysis performed on these genes and their bacterial homologs showed that their sequences correspond to the type of locus and that the *agr3* locus is present in other *C. difficile* strains. The proximity of *agr3* in the *C. difficile* strain genomes to phage-like integrase genes and transposases suggests that this locus can be transferred via HGT; its presence on a temperate phage genomes strongly supports this hypothesis and is consistent to findings by a recent study that suggested the *agr* locus may have been horizontally transferred between some species of the *Firmicutes* (Wuster and Babu 2008).

Despite the being in the same locus type, the phage genes are genetically distinct from the other bacterial *agr3* genes. Their diversity could reflect s that these *agr* loci may have different functions. In *C. difficile* only the LuxS system has been experimentally investigated. It has been found to be active (Carter *et al* 2005) and upregulates transcription of both toxin genes *tcdA* and *tcdB* (Lee and Song 2005). Despite the association of the *agr* system with regulation of virulence factors in other gram-positive bacterial species (Antunes *et al* 2010) little work has focused on the *C. difficile* encoded *agr* system. One study examining gene expression through the growth stages found increasing expression of *agrD* during late stage exponential growth (Saujet *et al* 2011). In the *agr2* locus, the HK is annotated as VirS due to its' sequence homology to that of the VirS encoded by *C. perfringens* involved in regulation of virulence factors (Ohtani *et al* 2003, Cheung *et al* 2010).

While the phage encoded genes group closely with their bacterial homologs, they are noticeably different at the aa sequence level (~40 %). This suggests that the phage genes have evolved separately within phage genomes. Whilst this is the first time that the *agr* QS cassette has been identified in a phage genome, there are examples where phages and QS systems interact.

While the phage encoded genes group closely with their bacterial homologs, they are noticeably different at the sequence level (~40 % different at the aa level, Figure 2). This suggests that the phage genes have evolved separately within phage genomes. While this is the first time that

the *agr* QS cassette has been identified in a phage genome, there are examples where phages and QS systems interact.

For example, in one study, native soil bacterial populations were shown to release phages when they were exposed to several species variants of the signalling molecule, N-acyl homoserine lactone, from the LuxS QS system (Ghosh *et al* 2009). Whether phages can actively 'listen in' to this signal is unknown, but several sequenced phages in the NCBI database encode genes that could be homologs of the LuxR, the response regulator as they contain either LuxR_C_like or HTH_LUXR protein domains. These are characteristic of transcriptional regulators, including LuxR and they are found both in known temperate and plasmid-like phages (e.g. (Toh *et al* 2006, Boyer *et al* 2008, Lan *et al* 2009) as well as in virulent phages (Mayer *et al* 2010). In contrast to listening in, one phage, phiPLPE which infects *Iodobacter* sp CdM7 isolated from river water*,* may instead block out the luxS QS signal, as it encodes a putative acylhydrolase (LeBlanc *et al* 2009). This ezyme could degrade the N-acyl homoserine lactone signal molecules and reduce the QS elicited response.

An example of why a phage may want to block the signal of the QS system is seen in another species, *E. coli*, where the phage receptor molecules for phage lambda are downregulated in via the luxS system and so inhibiting this would presumably allow a successful infection for the phage (Høyland-Kroghsbo *et al* 2013). There are fewer examples linking phages and the *agr* QS system, but, interestingly, three phages encode genes with a LytTR protein domain (and so may be homologous to the response regulator, AgrA) and may have the capacity to 'listen in' to this system. They are all pseudomonas phages, phage Lu11 (Adriaenssens *et al* 2012), phage vB_PaeS_PMG1 (NC_016765.1) and phage D3 (Kropinski 2000), two of which encode predicted integrases. Clearly phages could benefit from interacting with their bacterial QS systems through listening in and blocking the signals and the phage phiCDHM1 is the first example of a phage encoding the necessary genes to be capable of doing the 'talking' instead.

However, why a phage encodes genes to produce its own signal is unknown. Some speculative explanations are offered below. As no response regulator gene was identified in the phage genome the response is likely to involve a host effector protein to elicit a response. Three scenarios as to when a prophage may evoke a QS coordinated response as bacterial cell density reaches a particular point are 1) playing a role in niche construction so as a weapon in intermicrobial wars, 2) as a population density dependant lysogen conversion factor, enhancing its hosts own fitness or 3) protection against secondary phage infection. All three strategies would promote phiCDHM1 and its' hosts' survival and replication.

In the first scenario, the phage-encoded signal could be released as an antagonist to reduce microbial competition for resources by causing lysis of neighbouring cells via phage induction. Depending on whether the signal is working on its own induction or unrelated phage induction, the phage may be co-ordinating its own release, or clearing unrelated lysogens which then become a food source for the phage host. The induced phages would also then be free to propagate and infect new hosts, also known as "kill the relatives" or lysogen alleopathy (Paul 2008, Abedon and LeJeune 2005, Abedon 2011).

Secondly, the phage may be eliciting a response in its own host to promote fitness, one that is dependent on cell density such as toxin or spore production as the *agr* locus regulates toxin production and sporulation in *C. perfringens* types A, B, C and D and (Li *et al* 2011, Chen *et al* 2011, Chen and McClane 2012, Vidal *et al* 2012) as well as the neurotoxin of *C. botulinum* (Cooksley *et al* 2010). Interestingly in *C. botulinum* also encodes two *agr* loci, and each evokes a different response, *agr*-1 modulates sporulation and *agr*-2 toxin production which supports the hypothesis the different *agr* loci types in *C. difficile* may elicit different responses.

Lastly, as previously discussed, the *lux* QS has been found to prevent phage infection in *E. coli*, as the signal decreases the number of phage lambda receptors on its cell surface protein (Høyland-Kroghsbo *et al* 2013). The phage encoded *agr* signal may be involved in inhibiting super

infection or secondary infection by down regulating cell surface proteins that other phages use as receptors.

Results from the bioinformatic approach to identify potential protein interactions using suggested that the QS system in *C. difficile* involves the global regulator CdtR and interacts with both host and other phage genes. However, these results are bioinnformatically predicted based on genome arrangement and co-carriage of genes in related species and experimental work to identify the targets of the *C. difficile* QS is needed. To determine whether the phage QS genes are transcribed, and investigate potential signal pathway, transcriptomic profiling of phiCDHM1 infection during lytic and lysogenic cycles is currently underway in the laboratory.

5.3.12. Detection of the phage encoded histidine kinase in environmental *C. difficile* isolates.

While it is clear there are highly related prophages present within *C. difficile* population, a screen based on the *agrB* revealed that while some genes are highly conserved between phages such as the capsid and holin, others appear highly phage specific. The *agrB* product was present in two other strains than the phiCDHM1 lysogen detected by the internal primer set, but not external, indicating this gene not encoded by the same/similar prophage.

5.4. Summary

The genome sequencing and annotation of novel *C. difficile* phages has revealed their genetic relatedness and diversity in terms of gene content. Their genome architecture is characteristic of phage genomes and follows a modular arrangement according to gene function. They encode several highly conserved ORFs and the combination of the presence of these ORFs, their shared genome architecture and individual gene identities to known *C. difficile* phages suggest

that the MMs can be grouped with the other phiCD119-like *C. difficile* myoviruses and that the SMVs form a new group, in which phiMMP04 also belongs.

While genetically similar to one another, the MMs encode several genes that are specific to their own genomes, and the distribution of these provide insight into the potential evolution during co-infection as well as across different host species. They may also serve as ecological markers in instances where the phages encode genes possessing unusual protein domains. Several genes that are predicted to be potential lysogen conversion factors were discovered in their genomes. These include for the first time genes involved in the *agr* QS (phiCDHM1) and synthesis of vitamin B12 (the SMVs).

Phylogenetic analysis of one of the most conserved genes, the capsid gene, was performed and revealed that the myoviruses can be split into four types, based on the resulting clades. In general these clades correspond to the phage morphology but the MMs fall into two groups. Primers to detect the different genetic groups of *C. difficile* myoviruses showed that lysogeny by different phages and co-carriage of morphologically similar but genetically distinct MMs is common in environmental strains of *C. difficile*.

The presence of the *agr* QS genes in a phage genome is highly novel, and to investigate their potential evolutionary origin, phylogenetic analysis was also performed on these genes and their bacterial homologs. This showed that the phage encoded agr genes from a thirdt ype of agr loci to be encoded within *C. difficile*, as is present in three other strains examined. This third group appears to be amenable to transfer via HGT. However, the phage encoded *agr* genes are genetically distinct from their closest bacterial homologs and their distribution was determined in environmental isolates. The HK gene was detected in two isolates belonging to different ribotypes although do not appear to be in phiCDHM1. This suggests they have either undergone HGT or are present on a different phage genome.

Both genomic annotation and phylogenetic analysis indicate the phage genomes are highly mosaic, which the next chapter investigates in more detail. Further work is being conducted in the Laboratory to establish if and when the phage encoded QS genes are transcribed.

6. Comparative genomes and CRISPR analysis of *C. difficile* phages.

6.1. Introduction

Currently there is no working definition of a phage species which is, in part, due to the fact that the majority of phages present in the biosphere have not be isolated or sequenced. The proposal that "all the world's a phage" and suggestion that there is continual gradient of phage phylogeny (with limits) arose because genetic exchange is highly prevalent across phage genomes and blurs the distinction of separate lineages (Hendrix *et al* 1999, Casjens, 2003, Hendrix 2003). An example of this gradual progression of genome relatedness has been observed in phages which infect *Listeria spp* (Dorscht *et al* 2009), but, in contrast, extensive work on mycophages maintains that phages infecting this bacteria can be classified into subtypes (Hatfull 2012). To date, a single effort has been made to assign phylogeny to *C. difficile* phages which rely on the presence of specific genes, but continued sequencing has shown that not all *C. difficile* phages encode these genes (Lavigne *et al* 2009). There is a clear need for an updated and comprehensive examination of the genetic diversity and grouping for current *C. difficile* phages. In this chapter, three different approaches will be used to identify the lineages and genome diversity of the 6 phage genomes sequenced in this project and also those of known *C. difficile* phages.

The mosaic nature of phage genomes means that performing meaningful comparisons between them is challenging (Hendrix 2003, Casjens and Thuman-Commike, 2011). Homologous recombination is considered to be the major mechanism of phage genome re-assortment, but also non-homologous recombination, point mutations and gene duplication all contribute to phage genome diversity (Hendrix 2008). The impact of each mechanism in shaping phage genome evolution has been found to vary between different phages. For example, T4-like phages form a distinct group with only non-essential genes undergoing lateral exchange, whereas lambdoid-like phages follow a modular theory of evolution through homologous evolution (Botstein 1980,

McAuliffe *et al* 2007). The modular theory of genome evolution is supported by the suggestion that some genes will be less susceptible to re-assortment because of their function, and adjacent genes have co-evolved with one another (Hendrix 2008). The mosaic nature of phage genomes has led to the development of methods to assess the phylogeny of related and unrelated phages based on more than one gene such as multilocus typing or whole genome comparisons (e.g. Rowher and Edwards 2002, Smith *et al* 2007, Hatfull *et al* 2008, Lavigne *et al* 2009, Smith *et al* 2012). In other systems the comparison of multiple phages sequences has revealed the genetic diversity within their genomes, as well as specific regions which undergo rapid evolution (e.g. Pedulla *et al* 2003, Smith *et al* 2012). Few studies have focused on determining the diversity present within *C. difficile* phage genomes or comparing overarching patterns in their genome architectures (Goh *et al* 2006, Fortier *et al* 2007, Meessen-Pinard *et al* 2012). One method for grouping *C. difficile* phages to their family relies on the presence of certain genes encoded in their genomes (Lavigne *et al* 2009), however since this method was proposed several more phage genomes have been sequenced that do not encode all of these genes. There is therefore a clear need to examine the diversity across the known *C. difficile* phages to provide insight into their evolution and identify distinct lineages for taxonomic purposes.

To examine the relatedness and diversity within *C. difficile* phage genomes three approaches were used and in all, the six genomes sequenced in this project were analysed as well as a representative selection of publically available phage genomes.

Firstly, whole genome comparisons in a pairwise manner were performed using the software Artemis Comparison Tool (ACT: Carver *et al* 2005). This software tool blasts two genomes against one another at the nucleotide level and identifies regions of homologous sequence which share an identity of >80%. Secondly, multiple genome alignments were performed using Mauve software (Darling *et al* 2010). Mauve permits the alignment of multiple mosaic genomes and identifies regions of conserved sequence which is free from homologous recombination. Comparing the presence or

absence of these regions shows both vertical and horizontal patterns of evolution as well as reveals genome lineages. A third approach examines how the genome relatedness of the phages could impact on phage-host interactions by analysis of the *C. difficile* CRISPR system. The CRISPR/*cas* system is a form of adaptive immunity that confers phage resistance to the bacterial cell. This is comprised of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) that contain Direct Repeat (DR) regions flanking variable 'spacers' of 26 - 72 nt in arrays, and CRISPR-Associated (Cas) proteins. The system recognises and degrades foreign DNA that is homologous to a spacer (Jansen *et al* 2002, Pourcel *et al* 2005, Bolotin *et al* 2005, Barrangou *et al* 2007, Sorek *et al* 2008). Spacers are heritable and can also be acquired through the incorporation of foreign DNA sequences and the CRISPR/cas system is a form of bacterial adaptive immunity. CRISPR arrays can therefore serve as sequence records of past infections (Barrangou *et al* 2007).

Analysis of the CRISPR/cas system has been applied to investigate phage abundance and diversity within specific species (Bolotin *et al* 2005, van der Ploeg *et al* 2009, Kuno *et al* 2012, Stern *et al* 2012). This has been unexplored in *C. difficile*, although two prophages of strain CD630 are the first reported phages to encode CRISPR arrays, a discovery that is highly unusual (Sebaihia *et al* 2006). The genomes of the phages sequenced in this project and publically available sequences were searched for predicted CRISPR arrays using CRISPRfinder (Grissa *et al* 2007a). Also, each genome was then searched against a database of CRISPR spacers, CRISPRdb (Grissa *et al* 2007b). If phage immunity in *C. difficile* is controlled by the CRISPR system then the identified matches between spacers and phage sequences predict their potential host ranges. An analysis of the number of shared protospacers and the genetic relatedness of phages was performed to determine if, according to the CRISPR system, similar phages would have similar host ranges. By searching for the locations of the protospacers in the phage genomes, it was determined whether the *C. difficile* CRISPR system targets specific genes and how phages evade the system via nucleotide mutations.

Table 18. Phage and prophages used in the genome and CRISPR analyses.
Abbreviations are MM = medium myovirus, LTM= long tailed myovirus, SMV= small myovirus.

Name	Type	Strain	Ribotype	Accession
phiCDHM1	MM			This study
phiCDHM19	MM			This study
phiC2	MM			NC_009231.1
phiCD119	MM			NC_007917.1
phiCD27	LTM			NC_011398.1
phiMMP02	LTM			NC_019421.1
phiCDHM14	SMV			This study
phiCDHM11	SMV			This study
phiCDHM13	SMV			This study
phiMMP04	SMV			NC_019422.1
phiCD38-2	SV			NC_015568.1
phiCD6356	SV			NC_015262.1
phiT6	prophage	T6	R076	This study
phiCD630_1	prophage	CD630	R012	NC_009089.1
phiCD630_2	prophage	CD630	R012	NC_009089.1
phiCF5	prophage	CF5	R017	NC_017173.1
phiM68	prophage	M68	R017	FN668375.1
phiCD196	prophage	CD196	R027	NC_013315.1
phiR20291	prophage	R20291	R027	NC_013316.1
phiBI-L	prophage	BI-L	R027	NC_017179.1
phiQCD-66c26	prophage	QCD-66c26	R027	NZ_CM000441.1
phiQCD-76w55	prophage	QCD-76w55	R027	NZ_CM000661.1
phiQCD-37x79	prophage	QCD-37x79	R027	NZ_CM000658.1
phiQCD-97b34	prophage	QCD-97b34	R027	NZ_ABHF00000000.2
phiCIP 107932	prophage	CIP 107932	R027	NZ_CM000659.1
phi207855	prophage	207855	R027	NC_017178.1

To summarise, the three approaches are used on the phages and prophages shown in Table 18. The first two approaches (using ACT and Mauve) show the mosaic nature of these phage genomes and provide evidence of horizontal and vertical genome evolution occurring between these phages. They also reveal regions that are commonly conserved and those which are more variable or unique to specific phages. Importantly, despite the high levels of observed mosaicism, the analyses also identify distinct genome lineages within the *C. difficile* phages. The results of the CRISPR analysis shows how the genetic diversity within *C. difficile* phages may be shaped by the host from selective pressure posed by the CRISPR system, as well as the impact that similar phages genetics could have on phage infection. Notably the CRISPR analysis also showed that prophage encoded CRISPR arrays

can target other phage genomes, and this provides evidence of a highly novel way for phages in to inhibit secondary phage infection.

6.2. Results

6.2.1. Comparative genomics of *C. difficile* phages reveal their relatedness to one another as well as extensive mosaicism.

The genome annotations performed in Chapter 5 show that for each ORF the identities of the top blastp hit differ for each phage and also by their position along the genome. In order to determine which phages are most closely related and where related sequence is most commonly present, pairwise comparisons of each phage sequence were performed using the Artemis Comparison Tool (Carver *et al* 2005). ACT identifies sequences in the two genomes that share greater than 80% identity, which are regions shown in red. Blue regions are where sequences are homologous but the sequence is inverted in one genome, and regions of white share less than 80% identity. First, phiCDHM1, phiCDHM19 and phiT6 were each compared in pairwise analyses to one another, as well as to seven publically available sequences, phiC2 and phiCD119 (MMs), phiCD27 (a LTM), phiCD630_1 and phiCD196 (two prophages encoded in R012 and R027 strains respectively). Secondly, phiCDHM13, phiCDHM11 and phiCDHM14 were compared to one another, and phiCDHM13 was then compared to the genomes used in the previous comparisons. Lastly, the genomes of the two sequenced *C. difficile* siphoviruses, phiCD6356 and phiCD38-2 were compared each of the phage genomes sequenced in this project.

The genome of the MM phiCDHM1 shares the greatest homology to the genome of phiC2, then to the three prophages phiT6, phiCD196 and phiCD630_1 (Figure 27). The shared regions of homology are present throughout the packaging, structural, lysis, lysogeny control and DNA replication parts of the genomes between these phages. Although there are also homologous regions between phiCDHM1 and phiCDHM19, phiCD119 and phiCD27, they are mainly restricted to the lysis, lysogeny control and DNA replication regions of the genomes. However, in the DNA replication region phiCDHM1 has greater similarity to the genomes of phiCDM19, phiCD119 and phiCD27 than those of phiT6 and phiCD196, which highlights the mosaic nature of the phiCDHM1 genome. As expected, there are regions of the genome which do not have identity to any of the phages in the analysis. Of particular note is a region that encodes the *agrD*/*agrB*/HK gene cluster and another region that encodes the NTPase gene. None of these genes are encoded by the other phages.
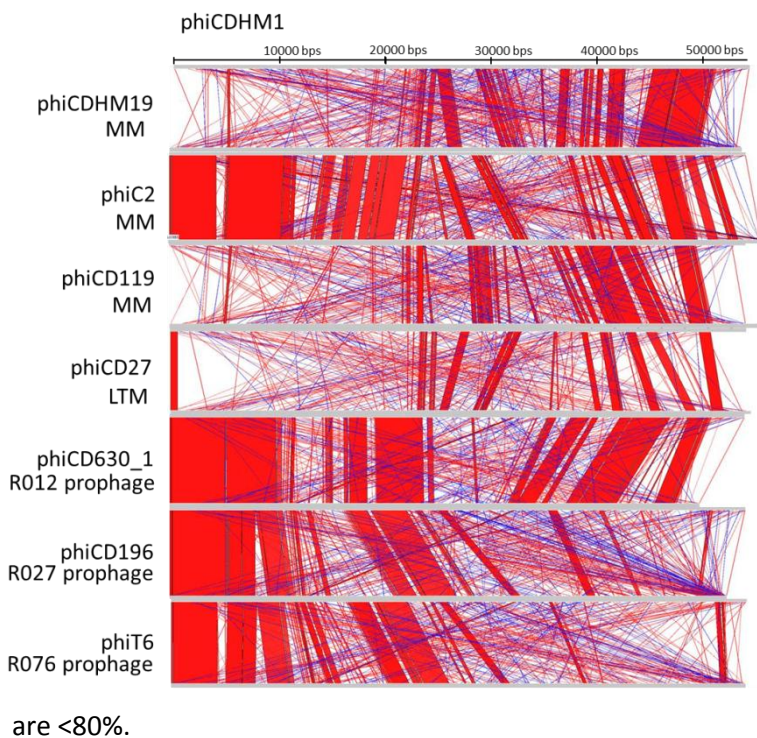

Figure 27. Whole genome comparisons of phiCDHM1 and *C. difficile* myoviruses. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of phiCDHM1 to three MMs (phiCDHM19, phiC2 and phiCD119), a LTM (phiCD27) and three prophage sequences (phiCD630_1, phiCD196 and phiT6). The prophages are from strains belonging to R012, R027 and R076 respectively. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

The second MM, phiCDHM19, shows the most overall genome similarity to phiCD119 and homologous regions are present throughout the structural, lysis, lysogeny, and DNA replication regions (Figure 28). The number and length of regions shared between phiCDHM19 and the other five phages vary, but are mainly located in the lysis, lysogeny control and DNA replication regions. Also, there are short lengths of homologous sequence in ORFs that do not cover the gene in the entirety, for example in the tail fiber and baseplate protein genes. This pattern of homology highlights the areas of the gene that are highly conserved, which suggests these sequences determine strain specificity/function. There are also parts of the phiCDHM19 genome which do not share identity to any of the examined phages, including two hypothetical proteins located in the tail assembly region and another hypothetical protein in the DNA replication region.
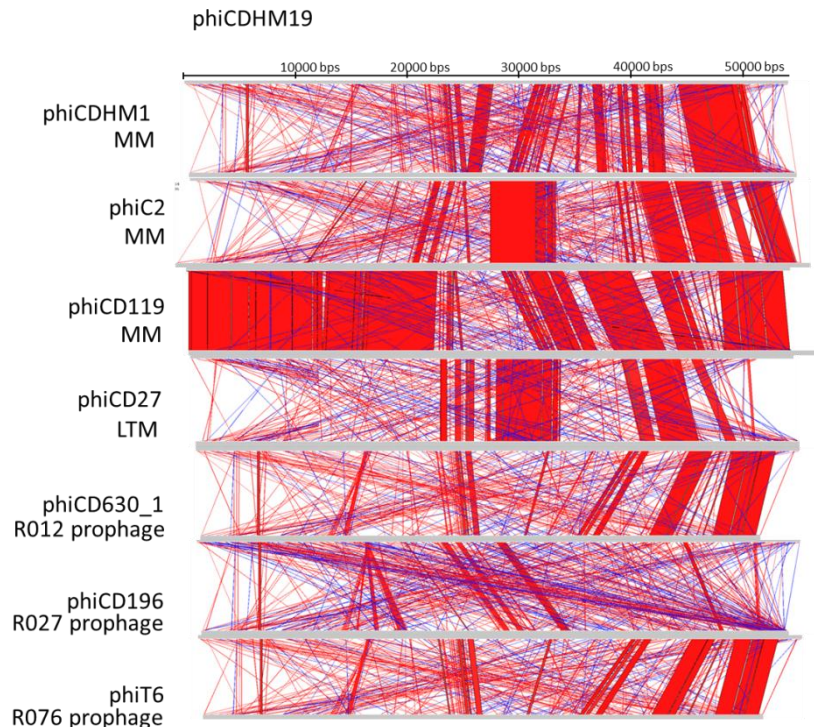

Figure 28. Whole genome comparisons of phiCDHM19 and *C. difficile* myoviruses. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of phiCDHM19 to three MMs (phiCDHM1, phiC2 and phiCD119), a LTM (phiCD27) and three prophage sequences (phiCD630_1, phiCD196 and phiT6). The prophages are from strains belonging to R012, R027 and R076 respectively. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

174

The genome comparisons for the prophage phiT6 reveal that its genome shares the greatest identity with phiCD196; then with phiCD630_1, phiCDHM1 and phiC2, and the homologous sequences are primarily located in the structural, lysis and lysogeny control regions (Figure 29). PhiT6 has few regions of homology to phiCDHM19, phiCD119 or phiCD27 and where found these are restricted to the lysis and tail assembly region, including the tail fiber gene. Also, regions of homology were found that are present in all of the paired comparisons in two of the most highly conserved genes, the endolysin and holin genes. In contrast, as expected, there is a homologous region which is only present in the phiC2 and phiCD196 comparisons, which encodes the AbiF gene and is not encoded in the other phage genomes. Also identified are regions in the phiT6 genome which do not share homology with any of the phages in the analysis and include the DNA methylase gene, the restriction enzyme gene and the zf-ribbon_3 gene, all of which may be involved in the prevention of secondary phage infection.
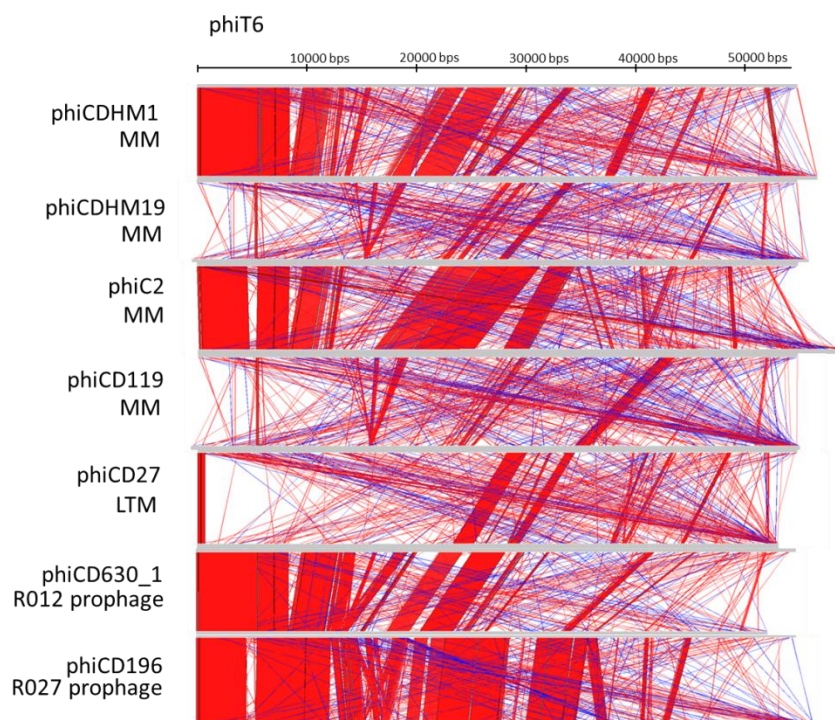

Figure 29. Whole genome comparisons of phiT6 and *C. difficile* myoviruses. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of phiT6 from a R076 *C. difficile* strain to three MMs (phiCDHM1, phiCDHm19, phiC2 and phiCD119), a LTM, (phiCD27) and two prophage sequences (phiCD630_1 and phiCD196) The prophages are from strains belonging to R012 and R027 respectively. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

Finally, the genome of the SMV, phiCDHM13, was compared to those of the two other SMVs sequenced in this project, phiCDHM14 and phiCDHM11 (Figure 30). The comparisons show that these genomes are highly similar to one another, but there is evidence of loss or gain of genome sequence. These regions encode the hypothetical proteins as discussed in section 5.2.4.
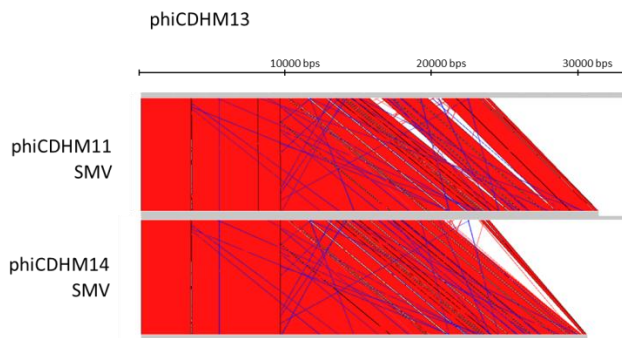

Figure 30. Whole genome comparisons of the SMVs, phiCDHM13 to phiCDHM11 and phiCDHM14. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of the three SMVs. Sequences of phiCDHM11 and phiCDHM14 are truncated due to exclusion of contigs <1000 nt. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

Further comparisons were performed to the other myovirus sequences using only the phiCDHM13 genome (Figure 31). The regions of sequence identity are restricted to the lysis region of the genomes but are present in all the comparisons. These regions include the LysM gene and the endolysin gene and again, there are shorter regions of homology within individual genes that are suggestive of host/strain specificity and function.
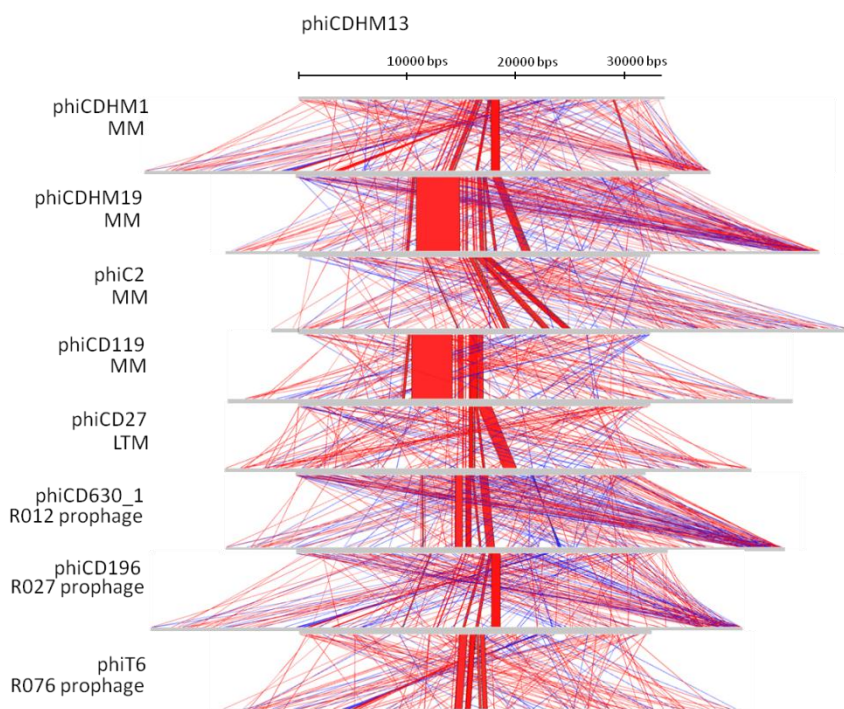

Figure 31. Whole genome comparisons of phiCDHM13 and *C. difficile* myoviruses. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of phiCDHM13 to four MMs, phiCDHM1, phiCDHm19, phiC2 and phiCD119, a LTM, phiCD27 and three prophage sequences, phiCD630_1 and phiCD196. The prophages are from strains belonging to R012 and R027 respectively. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

The paired whole genome comparisons reveal the relatedness of the genomes sequenced in this project varies by phage. PhiCDHM1 is most similar to phiC2, phiCDHM19 to phiCDHM19, phiT6 to phiCD196 and phiCDHM13 to the other SMVs phiCDHM11 and phiCDHM14. All the genomes display a highly mosaic pattern of synteny, which varies both between phages and across the functional regions of each genome, making whole genome comparisons necessary to determine their overall relatedness. Variable regions of the genomes were identified that contain sequences shared between only one or two phages, as well as those which lack homology to any of the examined phage genomes and correspond to genes that are specific to individual phage genomes. Also, where identity was detected between genomes, the lengths of the homologous regions vary considerably; in some instances regions can encompass several genes, individual genes or only part of genes. Several regions of homology are shared in many or all the comparisons and highlight which genes, or parts of a gene, are likely to be involved in host specificity. The different patterns of homology observed in this analysis provide evidence of where vertical and horizontal evolution has occurred in these genomes.

There is one region of the genome where homology has been observed for all comparisons, which is the lysis and attachment region. This region encodes highly conserved genes whose functions require host specificity. To test whether homology in these regions extends beyond the taxonomic boundaries, the genomes of two publically available *C. difficile* siphovirus sequences, phiCD63563 and phiCD38-2 were compared to four of the myoviruses sequenced in this project. The results show that there are multiple regions of homology in each of the comparisons (Figure 32).
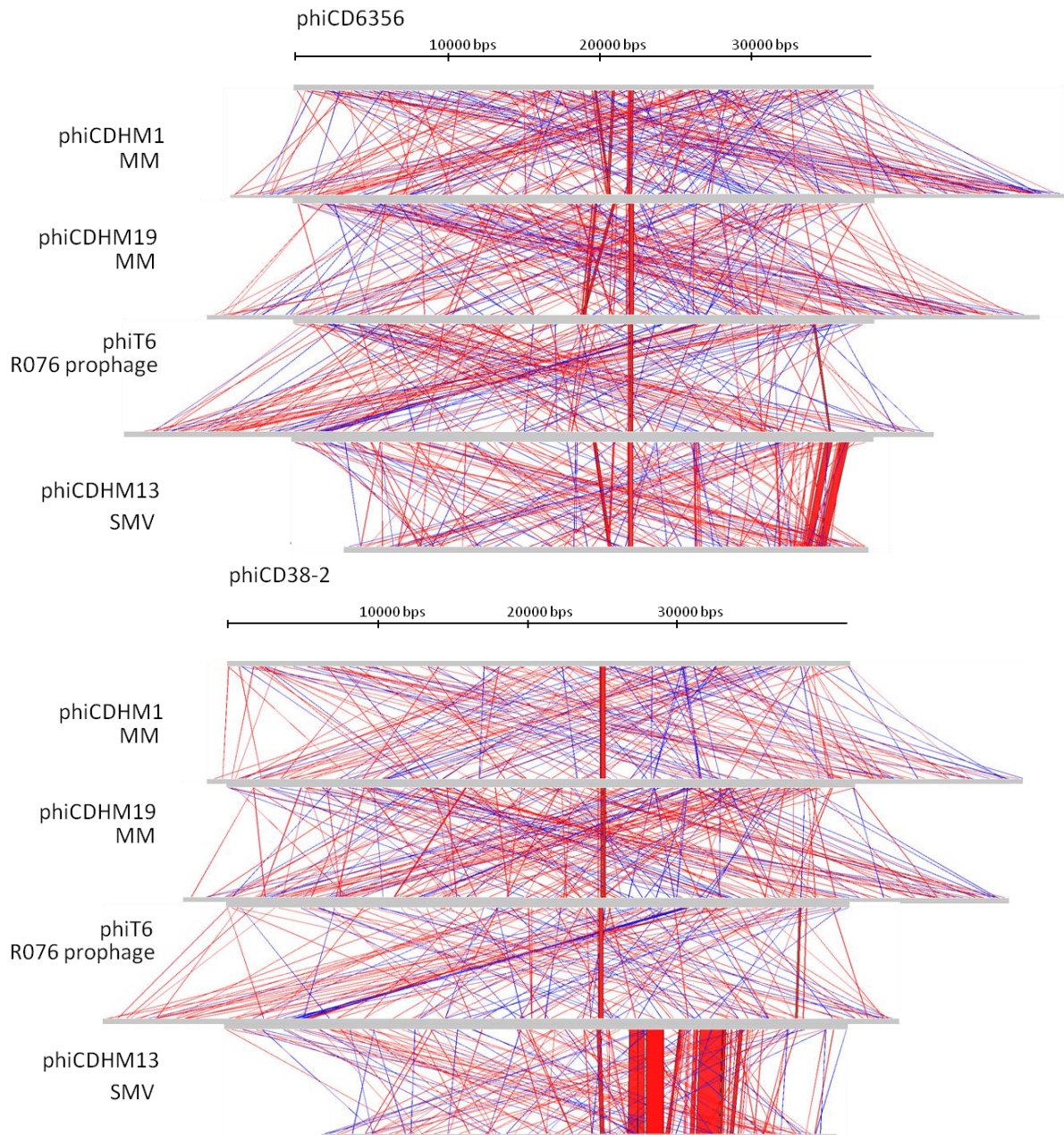
Figure 32. Whole genome comparisons of phiCD6356 and phiCD38-2 to *C. difficile* myoviruses. Pairwise comparisons were performed using ACT (Carver *et al* 2005) of phiCD6356 (top) and phiCD38-2 (bottom) to two MMs, phiCDHM1, phiCDHm19, the prophage phiT6 and SMV phiCDHM13. Red indicates homology of >80% between sequence; blue is homology of >80% identity when sequence is inverted and white indicates regions of homology that are <80%.

In general, the regions of identity are few and short; for example, the largest region of homology of the phiT6 genome to phiCD6356 is 170 bp and homology is restricted to sequences in the lysis and tail assembly modules including in the tail protein and tail fiber genes. However, in the comparisons of phiT6 and each siphovirus there is an additional region of homology which encodes the ssDNA binding protein. Also, there are more regions of homology between phiCDHM13 and each siphovirus, but these are restricted to within the lysis and DNA replication regions. The homology of SMV is greatest to phiCD38-2, with homologous sequences in several genes: the endolysin, six hypothetical proteins; the parA/amidase and the deoxyribonuclease. These results show that there is a high degree of sequence conservation present within genes that are predicted to be involved in host specific functions and that there may be some cross-over of genetic material between the genomes of these morphologically distinct phages.

6.2.2. Multiple alignments of *C. difficile* phage genomes reveal distinct genomic lineages.

In order to determine the overall genetic diversity between the new and known *C. difficile* phages, and to further investigate processes shaping their evolution, the genomes of the four phages sequenced in this project and 22 publically available phage sequences were used in whole genome alignments in Mauve software (Darling *et al* 2010). Mauve identifies conserved regions which are considered free from homologous recombination, termed Locally Colinear Blocks (LCBs). The conservation of the LCBs between genomes indicates different evolutionary patterns of vertical and horizontal acquisition. Mauve analysis permits the alignment of closely and distantly related genomes, even with highly mosaic sequences, making it useful for the analysis of phage genomes. The phage sequences used in the analysis include those of 26 phages that represent all known morphologies and prophage encoded in isolates belonging to different ribotypes.

The analysis reveals distinct genomic lineages between the phages and prophage sequences on the basis of the presence of several conserved LCBs, which are represented by differently

coloured blocks along each genome (Figure 33). These genome lineages correspond to the phage particle morphology and genome sizes, where known. There are also different LCBs present within these lineages. One example is the MMs which can be further sub-grouped in two types, phiCDHM1 and phiC2, and phiCDHM19 and phiCD119. The SMVs however have highly similar LCB profiles. The most diversity exists within the SV group which share only a few LCB between the genomes. The prophage genomes appear highly similar; for example the 027 group of prophages share nearly all their LCBs, with two genomes having slightly different profiles (phiCD196 and phiCIP1007932). The prophages that are both encoded in *C. difficile* strain CD630, phiCD630_1 and phiCD630_2 share LCBs throughout their structural and lysis regions but their genomes become more divergent in the DNA replication regions. Of the prophages, the two encoded in strains of R017, phiCF5 and phiM68, contain the most diversity.Interestingly, two LCBs were identified in all genomes examined. The lengths of these LCBs vary and encompass genes predicted to interact with the *C. difficile* cell (e.g. baseplate-J (PF04865) and holin (PF05105) genes), and are likely to be involved in host specificity.

While the LCB profiles highlight the vertical evolution of related phage genomes and enable identification of specific lineages, they also provide evidence which shows where homologous recombination events have occurred between these lineages. An example is found in the SMVs (which have the fewest LCBs shared with the other myoviruses) where LCBs are conserved with the SVs. Thus, the whole genome alignments permit grouping of the *C. difficile* phages, despite their mosaicism and horizontal genetic transfer events, which also reflects the results of the ACT comparisons. Furthermore, these lineages are supported by the morphological and genome size classfication of the phages determined in previous chapters.
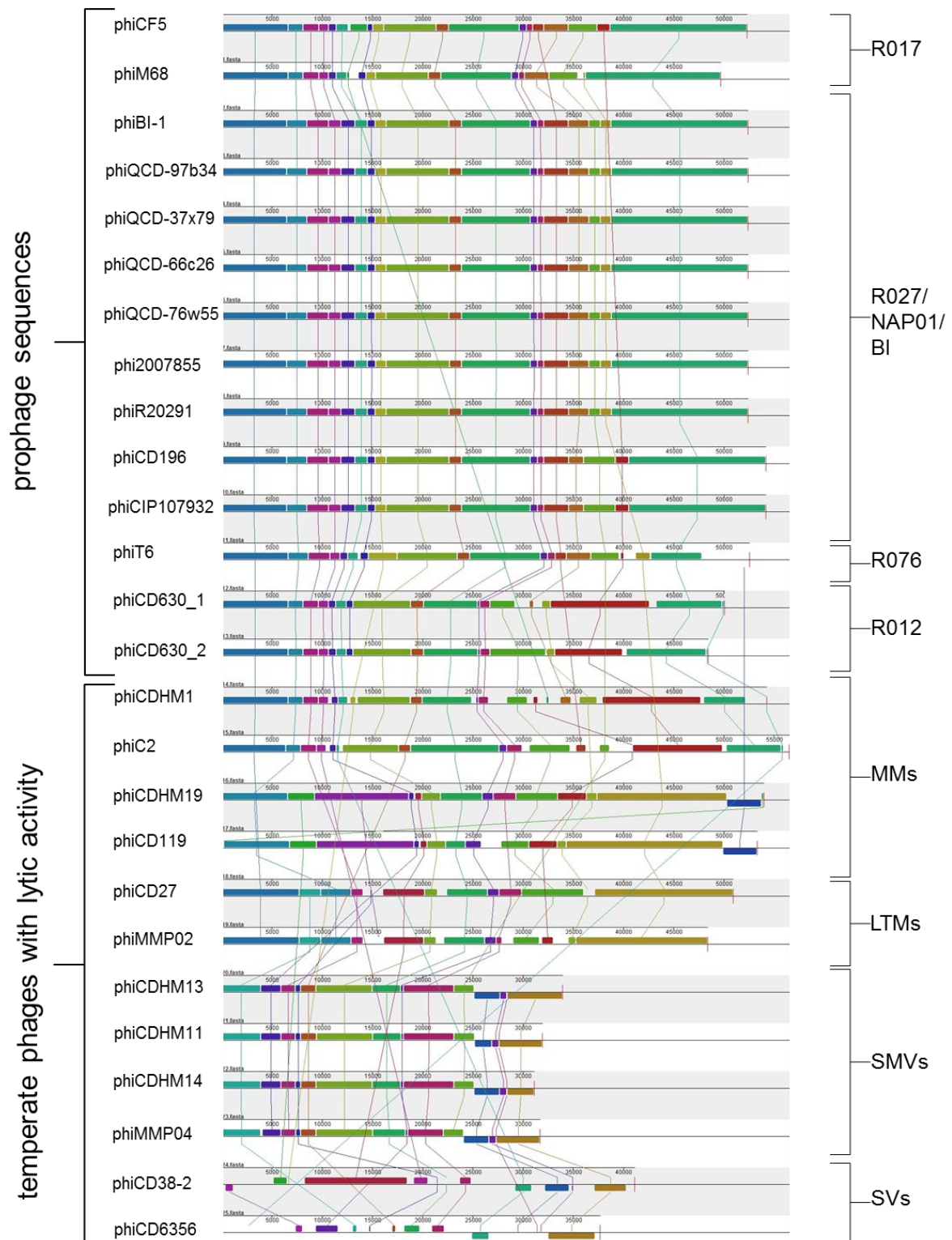
Figure 33. Whole genome alignments of 26 *C. difficile* phages. Nucleotide sequence alignments were performed in Mauve using the progressiveMauve algorithm (Darling *et al* 2010). Regions of conserved sequences which lack homologous recombination are termed Locally Collinear Blocks (LCB) and represented by differently coloured sections. Included in the analysis are the phages shown in Table 8. The phages have been grouped according to if they are prophage or been isolated through lytic propagation, and their host strain ribotype and morphologies, respectively.

6.2.3. The distribution of the *C. difficile* CRISPR system is widespread and diverse.

In other systems, the analysis of the CRISPR/*cas* system has provided insight into host-phage dynamics, phage diversity and ecology (e.g. Horvath *et al* 2008, van der Ploeg 2009, Stern *et al* 2012, Kuno *et al* 2012). Different strains of *C. difficile* encode multiple CRISPR arrays (Sebaihia *et al* 2006, Stabler *et al* 2009) but no work has investigated its CRISPR system, whether it appears active and if it could impact on phage infection.

To establish how diverse the *C. difficile* CRISPR system is, the CRISPR of the 7 *C. difficile* strains which are in the CRISPRdb were examined using the CRISPRcompar tool (Grissa *et al* 2008). These 7 strains (CD630, R20291, CD196, CF5, M120, BI-1 and 2007855) include clinical, bovine, historic and recent isolates which fall into 4 ribotypes groups: R012 (CD630), R017 (CF5), R078 (M120) and R027 (R20291, CD196, BI-1 and 2007855). All 7 *C. difficile* strains encode multiple CRISPR arrays with variable spacer content which indicates that the system is active (Table 19). 5 CRISPR arrays are shared between the 3 strains (based on the conserved direct repeat region sequences) but there are several polymorphisms in the spacer content. There are 7 'CD630-specific' arrays and 4 'R027-specific' arrays. The number of spacers differs in 2 of the 4 'R027-specific' arrays (data not shown) which suggests that these arrays have undergone a rapid evolution within a ribotype. The differences found in the CRISPR array number and content between ribotypes can be explained by their differential infection by temperate phages, as, in another species, *Yersinia pestis,* it has been shown that prophages are a source of newly acquired spacer regions (Pourcel *et al* 2005). Because of the high rate of lysogeny of *C. difficile* strains, it seems likely that phage infection will exert a strong pressure and could account for the high number of CRISPR arrays and spacers in this species.

Table 19. Diversity of CRISPR arrays encoded in seven *C. difficile* strains included in the CRISPRdb.

Strain	Ribotype	NCBI accession	CRISPR arrays	numbers of spacers in each CRISPR array	Genome length (nt)	% of genome
CD630	012	AM180355.1	13	3, 3,3, 5, 6, 6, 6, 8, 12, 13, 14, 15, 19,	4290252	0.18102
CD196	027 'historic'	FN538970.1	14	3,3, 4, 4,4, 4, 5, 5, 9, 9, 11, 14, 18, 23,	4110554	0.19326
R20291	027 'recent'	FN545816.1	14	3, 3, 4, 4,4 4, 5, 5, 9, 9, 11, 14, 14, 26,	4191339	0.18817
BI-1	027	FN668941.1	14	3, 3, 4, 4, 4, 4, 5, 5, 9, 9, 11, 14, 18, 24,	4118573	0.19451
2007855	N/A	FN665654.1	14	3, 3, 4, 4, 4, 4, 5, 5, 9, 9, 11, 14, 18, 24	4179867	0.19166
CF5	'historic' 017	FN665652.1	11	3, 3, 3, 3, 3, 4, 5, 7, 10, 14, 24,	4159517	0.13237
M120	078	FN665653.1	9	3, 4, 8, 9, 13, 17, 32, 38, 44	4047729	0.27991

6.2.4. *C. difficile* prophages encode multiple CRISPR arrays.

The *C. difficile* CRISPR system is unusual as the two prophages in *C. difficile* strain CD630 encode CRISPR arrays (Sebaihia *et al* 2006). In order to determine whether all *C. difficile* phages encode CRISPR arrays, the 26 genomes used in the Mauve alignments were searched using CRISPRfinder (Grissa *et al* 2007). Multiple CRISPR arrays were identified in all the prophage sequences examined, but not in the genomes of phages which have been isolated following lytic replication (Appendix Tables 3. to 7).

In all the prophage genomes the CRISPR arrays are located in the structural region, downstream of the XkdN gene and upstream of the tape measure protein (Figure 34). The number of CRISPR arrays, the DR sequences, spacer number and content varies by prophage but in the eight prophages examined, five main CRISPR types were identified (based on CRISPR array size, spacer content, DR sequence and presence of specific adjacent genes). Both prophages encoded in strain CD630 have identical CRISPR content and the prophages encoded by the R027 strains also share identical CRISPR content. However, the CRISPR content may not relate to ribotype as two prophages

which are encoded by two strains belonging to R017 (phiCF5 and phiM68) differ from one another in their CRISPR content.

However, all the prophages encode at least one conserved CRISPR array that has an identical DR sequence. The conservation of this CRISPR array indicates its possible vertical transmission through the phage population, but its spacer content varies between prophages and indicates subsequent divergence.
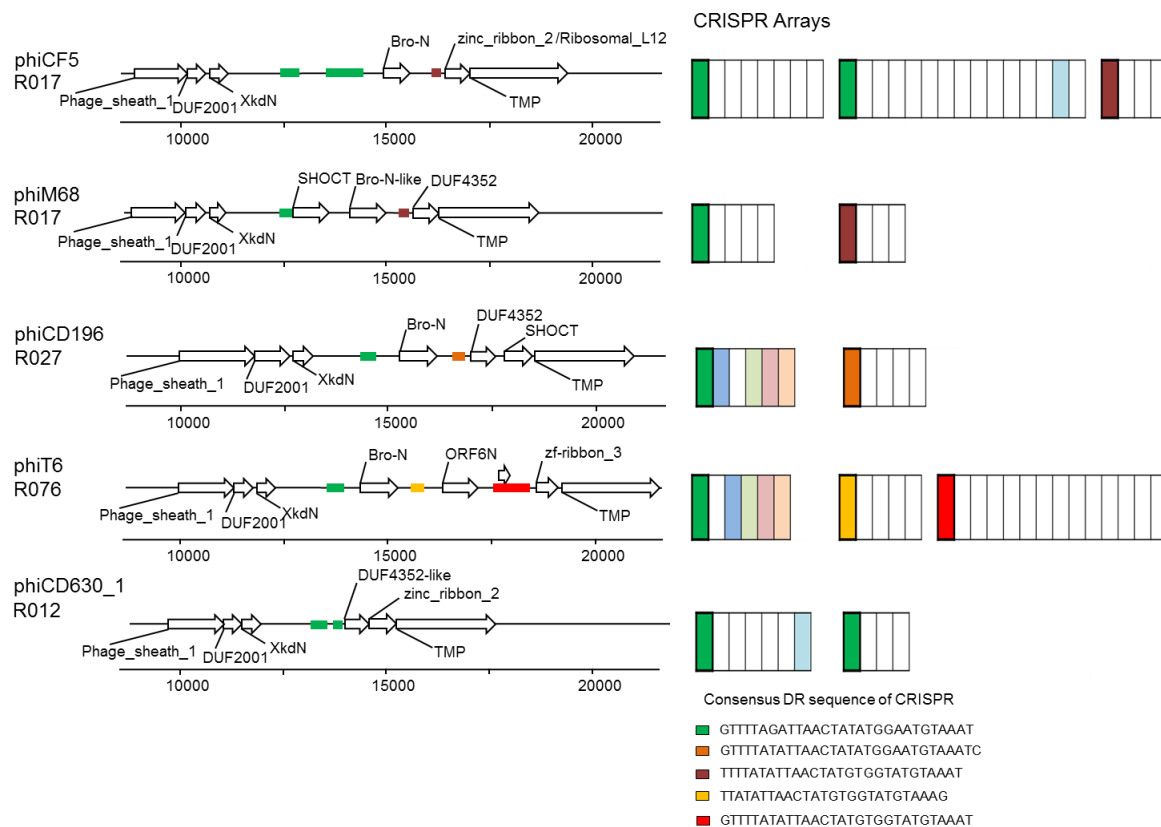
Figure 34. Diversity of *C. difficile* phage encoded CRISPR arrays and associated ORFs. CRISPR arrays were identified in 14 prophage genomes. The locations and content of the CRISPR arrays are shown and based on their DR sequence, spacer content and adjacent genes they can be grouped into five types. Homologous DR sequence and spacer sequences are indicated by matching colours and adjacent genes are annotated. Both prophage in *C. difficile* strain CD630 have identical CRISPR arrays and are represented by phiCD630_1, all of the 9 prophage from R027 strains examined have the same CRISPR arrays and are represented by phiCD196. The remaining prophages, phiCF5, phiM68 and phiT6 are distinct in their CRISPR content. For CRISPR DR and spacer sequences see Appendix Tables 3 to 7.

Other CRISPR arrays appear to have undergone horizontal transfer indicated by conservation of the DR sequence, spacer content and adjacent genes; for example, the first CRISPR array encoded by phiT6 and phiCD196 are closely related but their second and third arrays are different. There is also evidence that loss of CRISPR arrays from prophage genomes may occur, as phiCF5 and phiM68 share two homologous CRISPR arrays, but phiCF5 has an additional CRISPR array which is flanked by these two conserved arrays. This suggests that the array has either been lost from the genome of phiM68, or phiCF5 has gained it, through homologous recombination events.

No *cas* genes were identified in the prophage genomes, however there are several genes adjacent to the CRISPR arrays which contain protein domains predicted to have regulatory roles or which are involved in DNA binding such as ORF6N (Iyer *et al* 2002b), Bro-N (Zemskov *et al* 2000) and ribosomal_L12 (Leuonmarck and Liljas 1987). It is acknowledged that *cas* genes are very diverse and there may be more types than those currently recognised (Haft *et al* 2005, Makarova *et al* 2011); it may be that these genes are involved in the processing of the prophage encoded CRISPR arrays.

6.2.5. The CRISPR system of *C. difficile* targets known phages.

To determine whether the *C. difficile* CRISPR system could impact phage infection, the genomes of 26 phages and prophages previously characterised in the Mauve alignments were searched against the CRISPRdb (Grissa *et al* 2007b) in order to identify whether there were matches between spacers and phage sequences. These 26 phages and prophage sequences represent both genetically related as well as genetically distinct types and were considered to form a comprehensive set to test the specificity of the *C. difficile* CRISPR system.

The searches generated CRISPR profiles for each phage (and *C. difficile* strain) where phage sequence matched CRISPR spacers (Figure 35).
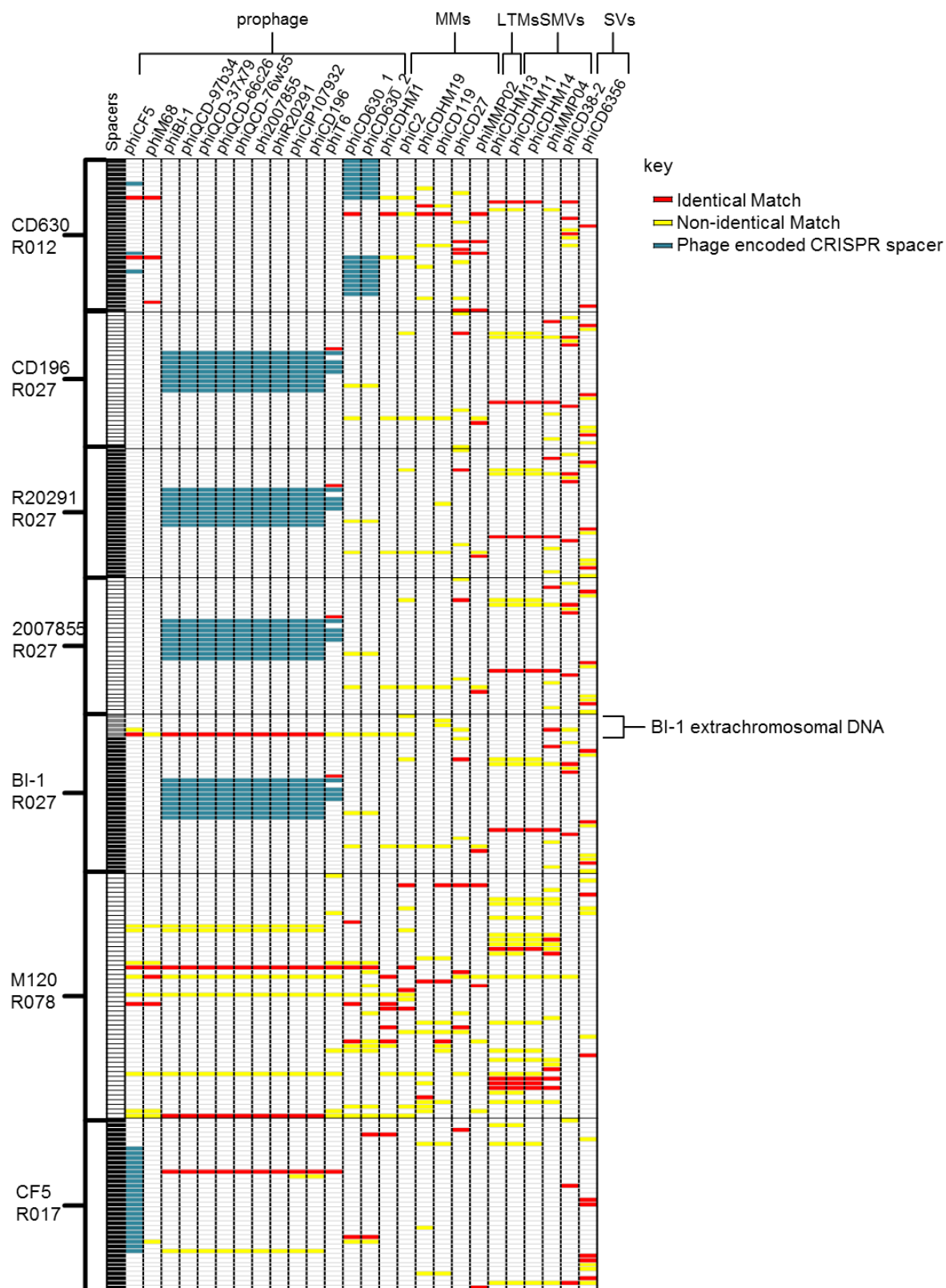
Figure 35. Homologous matches between spacers encoded in 7 *C. difficile* **strains and 26 phage sequences.** CRISPR profiles were generated for 26 phage and prophage (columns) by searching against the CRISPRdb which contains 7 *C. difficile* strains (rows). All matches between phage sequence and spacers are shown (score >0.005). Red: identical matches. Yellow: non-identical matches. Blue: "self" match between CRISPR spacers encoded on prophages.

The results of the searches showed that multiple matches were identified for individual phage genomes (n=10-26), as well as for individual *C. difficile* strains (n=21-53). Importantly, all phages had an identical match to at least one spacer and all *C. difficile* strains had an identical match to at least one phage sequence. These results strongly suggest that the CRISPR system in *C. difficile* is active as it can target known phages.

6.2.6. The CRISPR system predicts patterns of host specificity observed in host range analysis.

The CRISPR profiles generated can explain patterns of phage-host infections observed in the host range analysis. Several of the phages in the CRISPR analysis had matches to only one or two strains; for example, phiC2 has an identical match only to one strain, which predicts the phage to have a wide host range and is consistent with the pattern of generalist infection as seen in the host range analysis.

However, other phages have identical matches to multiple strains; for example phiCD27, phiMMP02 and phiCD6356 each have identical matches to all 7 *C. difficile* strains, which predict that these phages would have a much narrower host range and is consistent with the observation that some phages infect in a specialist manner with a restricted host range.

Also, similar patterns in CRISPR profiles can be observed for the individual strains, as some have identical matches to multiple phage genomes and thus have a narrow immunity to phage infection, and other strains to only one or two phages which would predict a wider immunity. An example of a strain with a predicted wide immunity is M120, which encodes spacers that have identical matches to all phage sequences except one. An example of strain with greater phage susceptibility is CF5 which, according to CRISPR, could be infected by 9 phages. These findings support the hypothesis that the CRISPR system of *C. difficile* impacts on phage infection.

6.2.7. The CRISPR analysis provides insight into past and future infection by related phages.

For most of the phage-strain pairings there are multiple matches (either identical or non-identical) identified. However, in some pairings there were only single identical matches identified, and in one there were no matches at all, phiC2 and strain CF5. The absence of a match (either identical or non-identical) suggests that there has been no previous exposure to this phage, or its relatives, as needed for the bacteria to acquire and incorporate spacers that match it. The presence of only one identical match between phage and CRISPR system implies that there has been a relatively recent exposure to this phage or a related one, as there has been no time for the phage to evolve resistance to it (which would then result in non-identical matches).

6.2.8. CRISPR evidence suggests cross-infection of related bacteria by *C. difficle* phages.

There are also other bacterial species in the CRISPRdb and matches were identified to spacers encoded by five bacteria: *C. botulinum*; *C. kluyveri*; *Caldicellulosiruptor kristjanssonii*; *Thermoanaerobacterium xylanolyticum* and *Porphyromonas gingivalis*. These matches suggest that cross-infection of the *C. difficile* phages occurs, or that these bacterial species are infected by closely related phages. Previously, cross-infection of a phage induced from *C. sordellii* on *C. difficile* was found (Schallehn 1985), which could explain these results as are most are related bacterial species belonging to the *Clostridia*. However, *P. gingivalis* belongs to the *Bacteroidia* and is gram-negative, but can be isolated from the gastrointestinal tract. Whilst phages have been found that infect across genera (e.g. Chen and Novick 2009, Deng and Hayes 2008), the CRISPR matches perhaps result from similar lifestyle pressures in this ecological niche than cross-infection.

6.2.9. *C. difficile* **phage encoded CRISPR spacers target known phages.**

The 7 *C. difficile* strains in the CRISPRdb encode 8 of the prophages that have CRISPR arrays on their genomes. It can be seen in the CRISPR analysis that some of these phage encoded spacers have matches to the phages. Spacers in the CRISPR arrays encoded by phiCF5 have identical matches to both prophages of strain CD630, phiCD38-2, phiCD6356 and the R027s group of prophages. This group of prophages also encodes spacers in their CRISPR arrays which match to phiCD119, phiCD630_1 and phiCD630_2. Finally, the spacers encoded by these two prophages, phiCD630_1 and phiCD630_2 match to the genomes of phiCF5 and phiM68. The presence of multiple identical matches between several different prophage encoded spacers to other phage genome sequences strongly suggests that these prophage encoded CRISPR arrays are active and have a role in host immunity. The carriage of CRISPR arrays by prophage is a novel way in which they could inhibit secondary phage infection and provides another aspect to the ongoing phage wars in this species.

6.2.10. *C. difficile* **mobile genetic elements may explain intra-ribotype variations in phage susceptibility.**

The prophages can confer phage immunity depending on their CRISPR content. In ribotypes with isolates that have the same prophage encoded with the same CRISPR content, the conferred immunity will be indentical. In ribotypes with isolates that encode prophages with different CRISPR content, the conferred immunity will be different. For example, all of the R027 prophages have identical CRISPR content and would therefore provide an identical immunity to their host bacterial cell. However, the prophages encoded by two isolates belonging to R017, phiM68 and phiCF5, have different CRISPR content and would provide immunity to different phages. The carriage of different prophages, encoding different CRISPR content, could therefore explain the intra-ribotype variation observed in the host range analyses for some specific ribotype groups.

Also, the R027 strains all had the same CRISPR profiles except for BI-1, as this strain has an extrachromosomal piece of DNA which encodes CRISPR arrays. These arrays encode 5 spacers that account for a further 23 matches to the examined phages and prophages, and results in this strain having a different predicted immunity to the other R027 strains. Oddly, a spacer on this extrachromosomal DNA targets the encoded prophage sequence in the R027 strains, and the CRISPR system may be a mechanism by which the megaplasmid promotes its maintenance in the host cell as it would prevent the excision and replication of the prophage and resultant cell lysis. The method of using CRISPR to maintain prophage may also be used by the host cell, as one spacer encoded on the bacterial chromosome of strain CD630 targets 1 of its 2 encoded prophages. The presence of spacers on this extrachromosomal piece of DNA appears to also contribute to host immunity, as well as provide an example of the CRISPR system to maintain mobile elements within a host cell.

6.2.11. The most genetically related phages share the most protospacers.

From the CRISPR profiles generated for each phage, it can be seen that some matches to the same spacer (the protospacers) are conserved between phages. If phage immunity is CRISPR controlled then the presence of shared protospacers indicates the phages would have similar host ranges. The number of shared protospacers appeared to be greatest between related phages. For example, the two MMs which are genetically similar, phiCDHM1 and phiC2, share 11 protospacers, but phiCDHM1 and phiCD6356, which are genetically distinct, do not share any protospacers. In order to test whether immunity controlled by the *C. difficile* CRISPR system could result in genetically similar phages having a similar host range, the number of shared matches was plotted to a whole genome blastn score in pairwise comparisons (Figure 36). To calculate this score, a whole genome blastn pairwise comparison was performed using ACT and the relative proportion of sequenced shared between the two genomes obtained (Appendix Table 8 and 9). These two values positively correlate and imply that the more genetically related a phage is to another, the more protospacers they will share, and thus have a similar host range based on their CRISPR profiles.
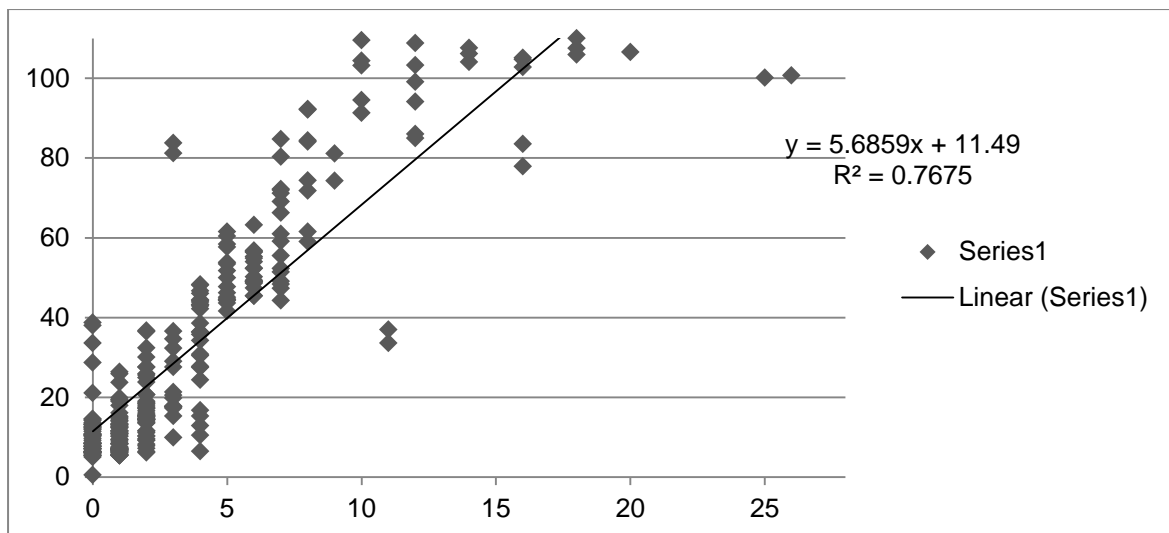
Figure 36. The number of shared protospacers positively correlates to the relative genetic relatedness between phages in pairwise comparisons. All combinations of phage genomes used in the CRISPR analysis were compared in a pairwise manner, with exception of the 027 prophages, as the two prophages phiCD196 and R20291 were included as representatives for all and phiCDHM13 as representative of phiCDHM11 and phiCDHM14. The number of shared spacers and relative genetic relatedness values were plotted from Appendix Tables 8 and 9. The two values positively correlate with an R^2 value of 0.7675.

However, related phages also have distinct protospacers which are not shared, as well as those which are conserved, with matches that are identical in one phage, but non-identical in another phage. An example of this is phiCD119 and phiC2, which share a protospacer located in a conserved DNA methylase (CDBPCV119_gp61 and phiC2p73 respectively). The protospacer sequences differ by two nucleotides, but these are both synonomous and result in the same amino acid sequence when translated. A second example is a shared protospacer sequence that is an identical match in phiCF5 but non-identical in phiM68. The protospacer sequences differ by a single nucleotide which results in a single non-synonomous substitution in the predicted endolysin gene, changing a serine to a proline residue, but both serine and proline have a polar, uncharged R group so this substitution may not impact on the endolysin function. This highlights the importance of retaining essential gene function while evading the CRISPR immune system.

6.2.12. The CRISPR system targets phage genes with respect to gene function and conservation.

To determine which genes are commonly targeted by the CRISPR system, the locations of each protospacer were identified for every phage. Protospacers were found to be in both functional and hypothetical protein genes. Of the functional genes, these included those involved in most aspects of the temperate phage lifecycle, including structural, lysis, lysogeny control and DNA replication genes (Figure 37). Notably, no protospacers were identified in genes that are putative lysogeny conversion factors, such as the quorum sensing genes in phiCDHM1, virulence factor VirE (e.g. phiC2p37 in phiC2) and phage immunity AbiF gene (e.g. CD_1450 in phiCD196). The relative proportions of the protospace containing genes compared to their proportion in the example genome of phiCDHM1 are different according to function; for example a larger proportion of lysis and attachment genes contain protospacers than is seen in phiCDHM1 genome. This could reflect the essential functions of these genes and the defree of conservation present in such genes.
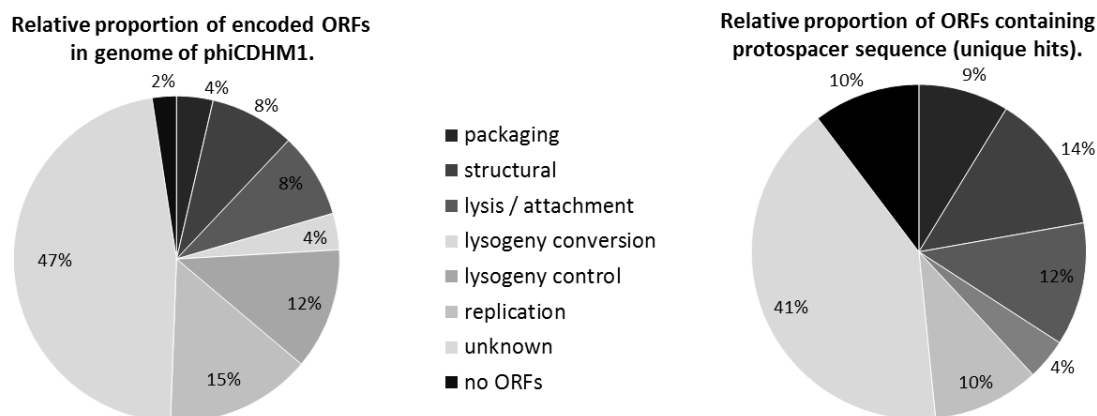
Figure 37. Functional bias in gene targeting by the *C. difficile* CRISPR system. Left: pie chart showing the relative proportion the phiCDHM1 genome in genes involved in different aspects of the temperate phage lifecycle, packaging, structural, lysis/attachment, lysogeny conversions, replication, unknown (hypothetical) and not in an identified coding sequence. Right: pie chart showing proportion of ORFs containing protospacers with unique hits. They show that protospacers are relatively more common in genes involved in packaging, structural, lysis and attachment functions than their reflects their relative abundance in the example genome of phiCDHM1. In contrast the proportion of protospacers in genes involved in lysogeny control, hypothetical proteins and DNA replication proteins is smaller than the abundance of these genes in the phiCDHM1 genome. There were no protospacers detected in the potential lysogen conversion factors.

From the CRISPR analysis it was clear that some genes had multiple matches to different spacers (Figure 38). These multiple matches suggest that these genes are targeted more frequently by the CRISPR system. Which genes are most commonly targeted, as well as the diversity and abundance of protospacers for specific genes, may reflect their conservation within phage genomes and their ability to mutate whilst retaining function.

An example of a gene that has multiple and diverse protospacers is the TMP gene which is encoded by all the phages in the analysis and in total has 16 unique protospacers across the different phage genomes, including both identical and non-identical matches. Several of these protospacers are conserved between phages, for example, a protospacer is shared between phiCDHM1 and phiC2 and a second protospacer is shared between phiCD119 and phiCDHM19. Also, the gene in a single phage genome can contain multiple protospacers, for example the TMP in phiCD6356 has five. Despite all phages encoding a version of the TMP gene, in the ACT and Mauve analyses it was seen that this gene in the structural region was highly conserved only between related phages. Therefore it appears the CRISPR system does target this essential gene but, due to its variability, targeting this gene usually provides only narrow immunity and generates a diverse collection of protospacers that match it.

An example of a conserved gene with abundant, but less diverse protospacers is the endolysin gene, which is also encoded by all the phage genomes used in the analysis. This gene is targeted by 3 spacers encoded by 2 *C. difficile* strains; the first spacer matches to protospacers in 19 phage and prophages; the second matches to 18 phage and prophages; the third matches to 4 phages. The spacers each target different sequences in the gene, 2 match to sequences in the C-terminus region of the gene, and the other is in the Amidase_3 protein domain (PF01520). The structure of the N-terminus region of CD27L, the endolysin of phiCD27, has been solved so it is possible to determine the third protospacer sequence corresponds to amino acid residues 56-67 which are located in a variable loop extension and alpha-2 helix of the protein (Meyer *et al* 2011).

This suggests that, by targeting a highly conserved gene, the CRISPR system can confer immunity to multiple phages resulting in less diverse spacers accumulated.
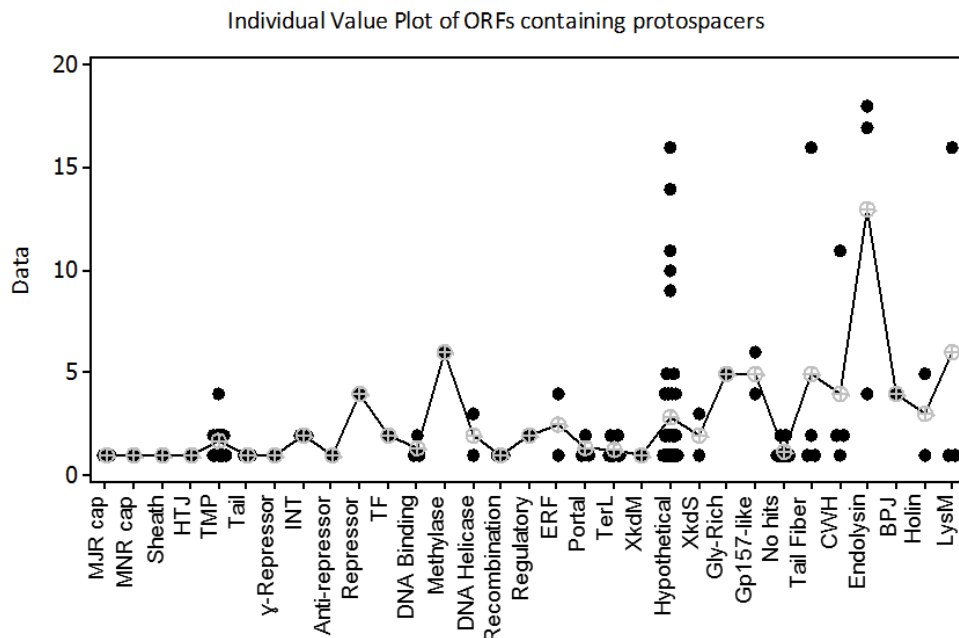
Figure 38. Multiple protospacers are present in individual genes and their conservation in multiple phage genomes. All locations of protospacers were identified and the numbers of phages sharing the protospacer plotted in an individual value plot. Symbols show the frequency (black) and mean (grey) of ORFs that contain protospacers. Abbreviations are: MJR = major; MNR = minor, HTJ = Head Tail Joining protein; TMP = Tape Measure Protein; INT = Integrase; TF = Transcriptional Factor; ERF = Essential Recombination Factor; TerL = Terminase Large SU; CWH = Cell Wall Hydrolase; and BPJ = Baseplate J protein.

Finally, another example is the holin gene, which is also conserved in all the phage genomes examined. It is targeted by two spacers encoded by the same strain, M120, which match to different phages, the first to the holin gene of phiMMP02 and the second matches to the holin gene of 7 phages and prophages. Previously, the sequence similarity between the *tcdE* gene and phage encoded holin led to the suggestion that there is a phage origin of the toxin encoding PaLoc in *C. difficile* (Goh *et al* 2007). To determine if the CRISPR system differentiates between these two genes we searched the *tcdE* gene from strain CD630 (CD630_06610) against the CRISPRdb. A non-identical match was identified to the first of these spacers which supports the theory that *tcdE* has a

xenologous origin. This presents interesting evidence that the CRISPR system in *C. difficile* could impact the acquisition of phage encoded genes by the bacteria.

From the results of the analysis it appears that the *C. difficile* CRISPR system targets specific genes according to their essential function and their conservation and as genes that can confer wide immunity most commonly contain protospacers. However, despite the apparent targeting effort of the CRISPR system, the protospacers examined show how the phages can evade the CRISPR system, whilst retaining essential protein functions.

6.3. Discussion

6.3.1. Genomic comparison of *C. difficile* phages.

Phage genomes are often characterised by having a distinct modular arrangement and evidence of rampant mosaicism occurs throughout (Hendrix 2003, Hendrix 2008). Results of the ACT and Mauve analyses show that the examined *C. difficile* phages have highly mosaic genomes, but despite this the phages can be separated into at least four distinct lineages based on their overall genome similarity. The proposed lineages correspond to phage genome size and particle morphology which is similar to the ability to subgroup phages infecting another species, the mycophages (Hatful 2012). However, within lineages there can be considerable diversity, especially in the prophage and MM genomes, which is consistent to a gradual scale of genome relatedness observed in genomes of *Listeria* phages (Dorscht *et al* 2009). Also, the analyses show that there are distinct differences between the groups in their genome architectures. Indeed, all the phages examined suggest they have undergone function module evolution, whereby genome evolution occurs through the exchange of entire functional modules (Botstein 1980). The order and content of these functional modules can differ between the identified lineages, for example for the SMV and SV groups. But as both analyses show the genetic "resolution" is smaller, as in some cases individual genes may have been exchanged rather than whole functional modules. A similar varying size of

genetic material exchange has been detected in other species, for example, in *S. pneumoniae* temperate phage genomes (Romaro *et al* 2009). Interestingly the ACT comparison displays a more sensitive approach to sequence homology and highlights single or partial gene sequences that may have been exchanged or co-evolved. The Mauve analysis provides a more comprehensive overview of the genome diversity and highlights which regions are conserved in all phage genomes examined; both have therefore been of use in probing the genetic diversity and relationships of these phage genomes.

The evolutionary patterns in the genomes identified in the ACT and Mauve analyses provide evidence of extensive vertical and horizontal gene transfer. Papke and Doolitte (2003) proposed that in order to make evolutionary sense of phage genomes, theoretical models should view evolution of these genomes as a web-like pattern rather than a tree, weighting lateral transfer of genes to their vertical evolution in genomes. The phage genomes presented in this thesis would likely fit best into this web-like pattern of evolution due to the extent of apparent lateral gene exchange. There are multiple regions and genes that are conserved differentially between the phages but, across these there appears to be two major patterns of homology throughout their genomes. These two types are seen in genes where small sections are homologous and the rest are highly variable such as the tape measure protein and the tail fiber proteins, or where genes as a whole are homologous and are conserved in fewer phage genomes, such as the holin, integrase, associated repressor or terminase genes. Previous work shows there may be a functional bias to the acquisition of genes between phage genomes, for example genes forming the tail structures are amongst those exchanged the most (Casjens and Thuman-Commike, 2011). Also observed was the conservation of genes between the otherwise genetically distant genomes of myoviruses and siphoviruses in the lysis and DNA replication regions. Similarly, there is homology in ORFs encoding lysis and DNA replication genes between myoviruses and siphoviruses that infect *Listeria* (Dorscht *et al* 2009). Furthermore, evidence of past recombination events between closely related lambdoid phages of *Salmonella* imply that exchange of replication proteins has occurred between genetically distinct phages

(Villafane *et al* 2008). While there is the possibility that these genes are present through convergent evolution, it is more likely that their exchange has occurred during co-infection of the same cell.

6.3.2. Identification of CRISPR protospacers and CRISPR loci in *C. difficile* phages and implication in phage evolution and phylogeny.

In other systems, analysis of the CRISPR system has revealed the diversity, relative abundance and impact of phages for specific bacterial species. No phage resistance mechanisms for *C. difficile* have been characterised before and an analysis of its CRISPR system with regards to the impact of phage infection was performed, which is summarised in Figure 39. In the CRISPR analysis the results showed that several different *C. difficile* strains encode multiple and diverse CRISPR arrays. Interestingly, the presence of numerous CRISPR has been linked to a high incidence of temperate phage infection (Reyes *et al* 2010). The results of the analysis show that the genomic relatedness of phage genomes can directly impact on the ability of their host *C. difficile* strains to resist infection by them, as well as highlight the evolution of key genes to evade this defence.

Results of the analysis showed that the different *C. difficile* strains encode multiple spacers that match to all of the examined phage genomes. The resulting CRISPR profiles can explain the observed general patterns of phage-host infections, if immunity is controlled by the CRISPR system. Notably, analysis of the *C. difficile* CRISPR system also provides evidence of previous exposure to specific phages through absence or presence of spacers to specific phages. Whilst Díez-Villaseñor *et al* (2010) found that the most recently incorporated spacers in *E. coli* had more matches to phage genomes, this is not the case in *C. difficile,* as spacers with matches were positioned throughout the bacterial CRISPR arrays. This could mean there is a relatively slow turnover of CRISPR spacers in *C. difficile* or that the *C. difficile* strains examined have been more recently challenged by unknown phages or plasmids indicating there is remains an undiscovered pool of phages or plasmids in the environment.
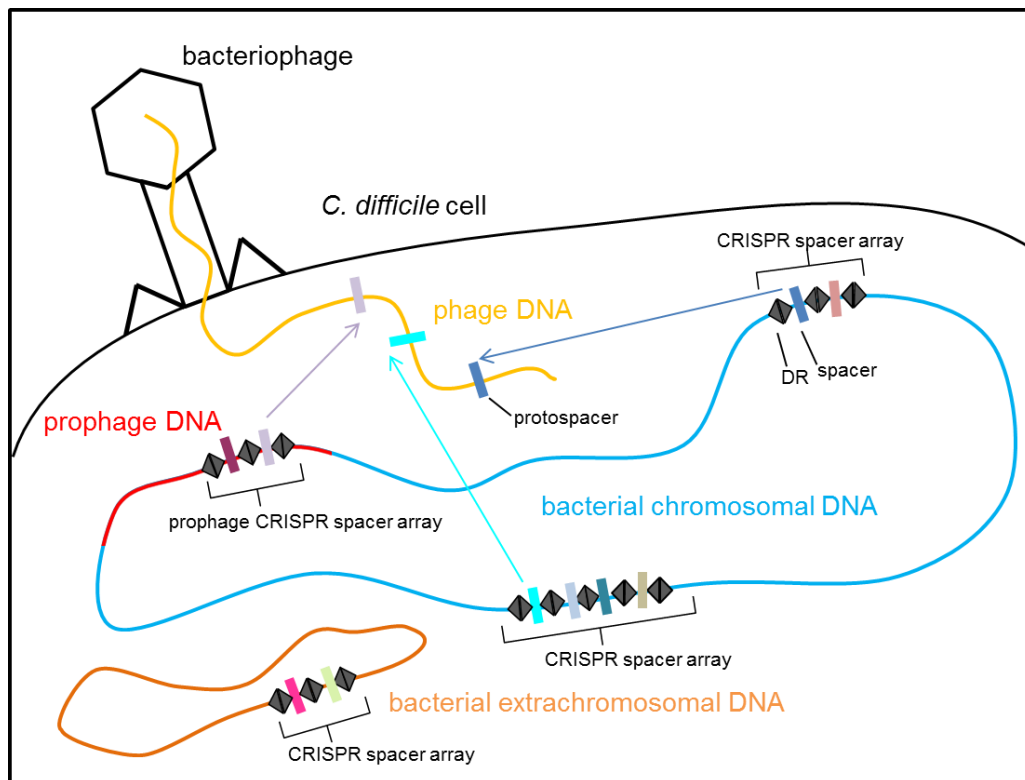
Figure 39. Locations of CRISPR arrays in the *C. difficile* cell. Multiple CRISPR arrays are encoded on the bacterial chromosome, extrachromosomal DNA and prophage DNA, all of which can have spacers that are homologous to phage genome sequence, as shown in this diagram by matching colour.

Unusually, two prophages infecting strain CD630 encode CRISPR arrays (Sebaihia *et al* 2006). The results of the analysis carried out in this project show that other *C. difficile* prophages encode CRISPR arrays and that the number and content of arrays vary. All of the *C. difficile* prophages examined were found to encode a conserved CRISPR array which shares an identical DR sequence. This CRISPR array may have evolved vertically with unique spacer content evolving subsequently. Other CRISPR arrays may have been transferred between prophage genomes through homologous recombination events during co-infection. In the CRISPR system spacers are typically acquired in a single direction, from the leader region, but other mechanisms of incorporation have been described in some CRISPR systems (Erdmann *et al* 2012). Also, while CRISPR evolution is considered to be a rapid mechanism of adaptive immunity, the R027 group of prophages have identical spacer content,

despite the fact that their host strains were isolated over a decade (Stabler *et al* 2010). Other systems have detected a slower rate of CRISPR acquisition in other bacterial species (Touchon and Rocha). Also detected was the possible loss of CRISPR arrays from phiM68. Mechanisms for the loss of CRISPR content are not well known, but in one study researchers found that a pathogenic host shift in *Mycoplasma gallisepticum* resulted in a significantly reduced CRISPR system (Delaney *et al* 2012). Interestingly, two prophages encoded in an environmental isolate (T6: this study) and an isolate from asymptomatic human (CF5: Stabler *et al* 2009) encode more spacers in their CRISPR arrays, than the prophages from clinical isolates, which have on average 38% fewer spacers. Just as loss of CRISPR arrays may reflect a shift in the host physiology, the loss of CRISPR arrays in the temperate phage genomes that are known to replicate in a lytic manner could reflect the shift of the phage to a lytic lifecycle.

This is the first time phage encoded CRISPR have been found that can target other phages. In another species, *Vibrio spp,* phages have been found which encode their own CRISPR system that acts against their bacterial host in order to produce a successful infection (Seed *et al* 2013) and phages infecting *Pseudomonas aeruginosa* encode proteins that inhibit the CRISPR system (Bondy-Denomy *et al* 2012). It seems likely that the CRISPR encoded on the *C. difficile* prophage could play a role in phage defence which is a highly novel mechanism of secondary phage inhibition. The differential carriage of prophage (and their encoded CRISPR arrays) in isolates of the same ribotype could explain the observed intra-ribotype variation observed in the host range analyses. The discovery of CRISPR in *C. difficile* prophages that target known phages introduces further complexity to the role of CRISPR systems, as well as adds to the genetic arsenal phages use in phage wars.

The CRISPR profiles generated from the analysis show that the *C. difficile* CRISPR system could account for similar host ranges if phages are related as indicated by the presence of shared protospacers, as well as account for individual phage differences arising from presence of 'unshared' or variable protospacer sequences. In addition, the *C. difficile* CRISPR system appears to target genes

by their function and conservation, which is in contrast to the results of Rho *et al* (2012) who report that protospacers occur randomly throughout the phage genome. However, the researchers also found regions that do not contain protospacer sequences which is similar to the findings of this project, in the lysogeny control region. Furthermore, the CRISPR system could be a selective pressure on the phage genomes from evidence of acquired mutations that would allow a phage to evade targeting through point mutation or gene loss/gain via homologous recombination. This is consistent with the patterns of genome homology observed in the genome comparison and alignments analyses using ACT and Mauve, and I propose that the *C. difficile* CRISPR system is a major influence on phage genomic diversity in this species. The novel findings that prophage encoded CRISPR systems can target phage genomes introduces an additional way for phages to prevent secondary phage infection, suggesting that *C. difficile* phages can influence each another's genome evolution as well as that of their bacterial hosts.

6.4. Summary

Distinct lineages can be determined in the *C. difficile* phage population, with highly conserved and variable regions identified. Comparative genomics of the *C. difficile* phage genomes revealed that distinct lineages can be determined, as inferred from homologous sequence and patterns of vertical evolution. The genomic analyses also showed that there are specific regions of the phage genomes that are likely to have undergone homologous recombination, as show patterns of horizontal genetic transfer, as well as highly conserved regions that are associated with essential functions especially host interaction.

An analysis of the CRISPR system has revealed that known phages are targeted by multiple and diverse arrays, and that the system appears to impact phage infection and genome evolution. One driving force of new spacer acquisition is phage predation and *C. difficile* has been shown to carry prophage throughout its population however the discovery of prophage encoded CRISPR that

target known phages present a novel way in which prophages could inhibit secondary phage infection, as well as disseminate phage immunity throughout the *C. difficile* population, impact HGT, population dynamics and the genome evolution of this pathogen.

Conclusions

The aim of this project was to isolate and characterise phages associated with environmental and non-clinical strains of *C. difficile.* In order to do so, and to investigate active temperate phage carriage in these strains, a collection of *C. difficile* isolates was obtained by sampling infants, animals and natural environments. One such envrionment, the Hampshire estuaries, was sampled over two years, and the results of *C. difficile* isolation, and characterisation of these isolates, showed that there is a diverse and dynamic population within the estuarine sediments. This has implications for commercial and recreational use of these environments, but also presents a reservoir of mixed strains in which strain evolution may occur.

The environmental isolates possess physiological traits that are susceptible to horizontal gene transfer, as evidenced by variation between isolates in motility and antibiotic resistance which occurred within single ribotypes. The prevalence and diversity of active temperate phages carried by these strains present a clear mechanism for such genetic exchange within this reservoir. Also, their spontaneous release from cultures, demonstrates exposure to inducing factors is not required for prophage excision and may be a natural process ongoing within the environment.

The isolation of temperate phages which can access the lytic cycle demonstrates that they can switch between lytic and lysogenic lifestyles on different host strains in a strain specific manner. Host range analysis of these, and potentially virulent phages (found in enrichment cultures), confirmed this tight strain specificity for most of the phages. For example, the efficacy of plaquing was highly variable for phages on related strains. The host range analysis also showed that while some phages appear to be specialists, others are generalists and could infect a different spectrum of hosts. The characterisation of the isolated phages showed that all were either myoviruses or siphoviruses and belonged to the Caudivirales. Based on their sizes, the myoviruses could be subgrouped into three types.

Whole genomic sequencing of six of the phage genomes showed they encode conserved genes, which support the subgrouping on the basis of morphology. Interestingly, it has been revealed that these phages have a number of highly novel features. This includes homologs of bacterial genes encoded by the phages. PhiCDHM1 encodes putative quorum sensing genes that present a highly novel way in which a phage could communicate to its bacterial host. PhiCDHM13 encodes a putative cobyrinic acid synthase, which is predicted to be involved in host metabolism. Interestingly, phiT6 encodes three putative AbiF proteins which are involved in abortive infection and a potential restriction enzyme system which could inhibit secondary phage infection of related or unrelated phages.

Comparative genomic analyses to one another and publically available phage and prophage sequences identified distinct genomic lineages with regions that are conserved between the specific subgroups of the phages, for example sequences in the structural regions. Also, all the phages have conserved sequences in the lysis region, presumably involved in host specificity. The analyses also provided evidence of horizontal gene transfer between these lineages, via homologous recombination that could have occurred during co-infection which is common within environmental *C. difficile* isolates*.* An analysis of the CRISPR system showed that the genetically related phages are likely to have similar host ranges, and that the CRISPR system could account for the patterns of interactions observed in the host range analysis, such as specialist phages, generalist phages, (and strains), and intra-ribotype variation.

Also, the CRISPR analysis revealed that several of *C. difficile* prophages, including phiT6 encode CRISPR arrays. The system appears to be active and several of these spacers target known *C. difficile* phage genomes. This is the first example showing phages could use their own CRISPR system to inhibit secondary phage infection and the differential carriage of prophages within *C. difficile* population explain phage-host interactions, observed in this project.

The conclusion of this project is that phages are likely to strongly impact upon *C. difficile* biology, in several aspects. The phages associated with environmental strains are numerous and characterisation of their genomes, has led to the discovery of several highly unusual genetic features that could influence their bacterial hosts, as well as other phages evolution, with evidence of inter-phage wars. From the results presented in this thesis, the *C. difficile* phages facilitate horizontal gene transfer significantly in this species. Whilst the environmental phages are related to phages associated with clinical strains, there may be some individual differences, as detected in the protein domain analysis. This could be due to the fact there is cross-over between environmental and clinical ribotypes. The three *C. difficile* isolates which could not be assigned to a known ribotype may be exclusively environmental strains. From this project several more phages are currently undergoing sequencing, in addition to over 40 *C. difficile* isolates, which include these three environmental strains. Future work will establish whether these are exclusive environmental isolates, and work underway in the laboratory is investigating the potential application of these phages in therapeutic and diagnostic settings. Experiments are underway to investigate more fundamental biology of the phage-host interactions including transcriptomic profiling to establish whether the QS genes and prophage CRISPR are expression by the phages.

References

Abedon, S. T. (2010) Lysis from without. Bacteriophage 1(1):46-9

Abedon, S.T. (2011) Communication Among Phages, Bacteria and Soil Environments. In: Biocomminucations in Soil Microorganisms. Editor Witzuny, G. Springer, London. pp 37-66.

Abedon, S.T. and LeJeune, J.T. (2005) Why bacteriophage encode exotoxins and other virulence factors. Evolutionary Bioinformatics Online 1:97-110

Abedon, S.T. and Yin, J. (2009) Bacteriophage plaques: theory and analysis. Methods of Molecular Biology 501:161-74

Ackermann, H-W. *(*2007) 5500 Phages examined in the electron microscope. Archives of Virology 152:227–43.

Ackermann, H-W. *(*2009*)* Phage classification and characterisation. In: Bacteriphages: Methods and Protols. Editors Clokie, M.J. and Kropinski, A.M. *Humana Press,* New York. pp 127-140

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

Adlerberth, I., Strachen, D.P. and Matricardi, P.M. (2007) Gut Microbiota and Development of Atopic Eczema in 3 European Birth Cohorts. Journal of Allergy and Clinical Immunology 120(2):343-50

Adlerberth, I. and Wold, A.E. (2009) Establishment of the gut microbiota in Western infants. Acta Pædiatrica 98(2):229038

Adriaenssens, E.M., Mattheus, W., Cornelissen, A., Shaburova, O., Krylov, V.N., Kropinski, A.M. and Lavigne, R. (2012). Complete Genome Sequence of the Giant *Pseudomonas* Phage Lu11. Journal of Virology 86:6369-70.

Aertsen, A. and Michiels, C.A. (2005) Mrr instigates the SOS response after high pressure stress in *Escherichia coli*. Molecular Microbiology 58(5):1381-91

Al Saif, N. and Brazier, J.S. (1996) The distribution of *Clostridium difficile* in the environment of South Wales. Journal of Medical Microbiology 45:133-7

Al Saif, N.M., O'Neill, G.L., Magee, J.T., Brazier, J.S. and Duerden, B.I. (1998) PCR-ribotyping and pyrolysis mass spectrometry fingerprinting of environmental and hospital isolates of *Clostridium difficile*. Journal of Medical Microbiology 47(2):117-21.

Allen, H.K., Loof, T., Bayles, D.O., Humphrey, S., Levine, U.Y., Alt, D. and Stanton, T.B. (2011) Antibiotics in Feed Induce Prophages in Swine Fecal Microbiomes. mBio 2:6

Alonso, R., Peláez, T., González-Abad, M. J., Alcalá, L., Muñoz, P., Rodríguez-Créixems, M. and Bouza, E. (2001) *In vitro* activity of new quinolones against *Clostridium difficile.* Journal of Antimicrobical Chemotherapy 47(2): 195-7

Anantharaman, V. and Aravind, L. (2003) Evolutionary history, structutral features and biochemical diversity of the Nlp/P60 superfamily of enzymes. Genome Biology 4(2):R11

Antunes, L.C.M., Ferreira, R.B.R., Buckner, M.M.C. and Finlay, B.B. (2010) Quorum sensing in bacterial virulence. Microbiology 156(8):2271-82

Araujo, R., Muniesa, M., Méndez, J., Puig, A., Queralt, N., Lucena, F. and Jofre, J. (2001) Optimisation and standardisation of a method for detecting and enumerating bacteriophages infecting *Bacteroides fragilis.* Journal of Virological Methods 93(1-2):127-36

Aspevall, O., Lundberg, A., Burman, L.G., Åkerlund, T. and Svenungsson, B. (2012) Antimicrobial susceptibility pattern of *Clostridium difficile* and its relation to PCR ribotypes in a Swedish university hospital. Antimicrobial Agents and Chemotherapy 50(5):1890-92

Avbersek, J., Janezic, S., Pate, M., Rupnik, M., Zidaric, V., Logar, K., Vengust, M., Zemljic, M., Pirs, T. and Ocepek, M. (2009) Diversity of *Clostridium difficile* in pigs and other animals in Slovenia. Anaerobe 15(6):252-5

Bakri, M.M., Brown, D.J., Butcher, J.P. and Sutherland, A.D. (2009) *Clostridium difficile* in ready-to-eat salads, Scotland. Emerging Infectious Diseases 15(5):817-818

Banerjee, P., Merkel, G. and Bhunia, A. (2009) *Lactobacillus delbrueckii* ssp. *bulgaricus* B-30892 can inhibit cytotoxic effects and adhesion of pathogenic *Clostridium difficile* to Caco-2 cells. Gut Pathogens 1(1):8

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709-12.

Barbut, F., Mastrantonia, P., Delmée, M., Brazier, J., Kuijper, E. and Poxton, I. (2007) Prospective study of *Clostridium difficile* infections in Europe with phenotypic and genotypic characterisation of the isolates. Clinical Microbiology and Infection 13(11):1048-57

Bartlett, J.G., Moon, N., Chang, T.W., Taylor, N. and Onderdonk, A.B. (1978) Role of *Clostridium difficile* in antibiotic-associated pseudomembranous colitis. Gastroenterology 75(5):778-82

Båverud, V., Gustafsson, A., Franklin, A., Aspán, A. and Gunnarsson, A. (2003) *Clostridium difficile*: prevalence in horses and environment, and antimicrobial susceptibility. Equine Veterinary Journal 35(5):465-71

Be'er, A., Zhang, H.P., Florin, E-L. Payne, S.M., Ben-Jacob, E. and Swinney, H.L. (2009) Deadly competition between sibling bacterial colonies. PNAS 106(2):428-33

Berthe, T., Toruon, A., Leloup, J., Deloffre, J. and Petit, F. (2008) Faecal-indicator bacteria and sedimentary processes in estuarine mudflats (Seine, France). Marine Pollution Bulletin 57(1-5):59-67

Besemer, J. and Borodovsky, M. (1999) Heuristic approach to deriving models for gene finding. Nucleic Acids Research 27(19):3911-20

Birge, E.A. (2006) Bacterial and Bacteriophage Genetics. In: Genetics of Temperate Bacteriophages Spring, New York, pp. 253- 92

Bojesen, A.M., Olsen, K.E. and Bertelsen, M.F. (2006) Fatal enterocolitis in Asian elephants (*Elephas maximus*) caused by *Clostridium difficile*. Veterinary Microbiology 116(4):329-35

Bolotin, A., Quinquis, B., Sorokin, A. and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151(8):2551-61

Bolton, R.P., Tait, S.K., Dear, P.R. and Losowsky, M.S. (1984) Asymptomatic neonatal colonisation by *Clostridium difficile*. Archives of Diseases in Childhood. 59(5):466-72

Bondy-Denomy, J., Pawluk, A., Maxwell, K. and Davidson, A. (2012). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 493:429-32

Borrell, B. (2012) Phage Factor. Scientific American 307:80-3

Borriello, S. (1998) Pathogenesis of *Clostridium difficile* Infection. Journal of Antimicrobial Chemotherapy 41:13-9

Borriello, S. P. and Brazier, J.S. (2000) Microbiology, Epidemiology and Diagnosis of *Clostridium difficile* infection. In: *Clostridium difficile* Editor Aktories, K. Springer, U.K. pp. 1-33

Borriello, S.P. and Honour, P. (1983) Concomitance of cytotoxigenic and non-cytotoxigenic *Clostridium difficile* in stool specimens. Journal of Clinical Microbiology 18(4):1006-7

Borriello, S.P., Honour, P., Turner, T. and Barcley, F. (1983) Household pets as a potential reservoir for *Clostridium difficile* infection. Journal of Clinical Pathology 36:84-7

Botstein, D. (1980) A theory of modular evolution for bacteriophages. Annuals of the New York Academy of Science 354**:**484-90

Bouza, E. (2012) Consequences of *Clostridium difficile* infection: understanding the healthcare burden. Clinical Microbiology and Infection 18(S6):5-12

Boyer, M., Haurat, J., Samain, S., Segurens, B., Gavory, F., Gonzalez, V., Mavingui, P., Rohr, R., Bally, R. and Wisniewski-Dye, F. (2008). Bacteriophage prevalence in the genus *Azospirillum* and analysis of the first genome sequence of an *Azospirillum brasilense* integrative phage. Applied Environmental Microbiology 64(3):861-74

Braun, V., Hundsberger, T., Leukel, P., Sauerborn, M. and von Eichel-Streiber, C. (1996) Definition of the single integration site of the pathogenicity locus in *Clostridium difficile.* Gene 181(1-2):29-38

Brazier, J.S., Fawley, W., Freeman, J. and Wilcox, M.H. (2001) Reduced susceptibility of *Clostridium difficile* to metronidazole. Journal of Antimicrobial Chemotherapy 48:735-48

Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? Trends in Microbiology 13(6):278-84

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Meads, D., Azam, F. and Rohwer, F. (2002) Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Science USA 99(22):14250-5

Breitbart, M., Thompson, L.R., Suttle, C.A. and Sullivan M.B. (2007). Exploring the vast diversity of marine viruses. Oceanography 20:135-9

Breitbart, M., Miyake, J.H. and Rohwer, F. (2006) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiology Letters 236(2):249-56

Brown, S.P., Inglis, R.F. and Taddei, F. (2009) Evolutionary ecology of microbial wars: within-host competition and (incridental) virulence. Evolutionary Applications 2(1):32-9

Brouwer, M.M.S., Allan, E., Mullany, P. and Roberts, A.P. (2012) Draft Genome Sequence of the Nontoxigenic *Clostridium difficile* Strain CD37. Journal of Bacteriology 194(8):2125-6

Bruntun, L.A., Knapp, J.S., Heritage, J. and Miller, H.M. (2010) Asymptomatic carriage of *Clostridium difficile* PCR ribotype 078 in pigs. Advances in Animal Biosciences 1:118

Brüssow, H. (2005) Phage therapy: the *Escherichia coli* experience. Microbiology 151:2133-40

Brüssow, H., Canchaya, C. and Hardt, W-D. (2004) Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. Microbiology and Molecular Biology Review 68(3): 560-602

Buist, G., Steen, A., Kok, J. and Kuijpers, O.P. (2008) LysM, a widely distributed protein motif for binding to (peptide)glycans. Molecular Microbiology 68(4):838–47

Caetano, L., Antunes, M., Ferreira, B.R., Buckner, M.M.C. and Finlay, B.B. (2010) Quorum sensing in bacterial virulence. Microbiology 156:2271-82

Cady, K.C. and O'Toole, G.A. (2011) Non-identity Targeting of Yersinia-Subtype CRISPR-Prophage Interaction Requires the Csy and Cas3 Proteins. Journal of Bacteriology 193(14):3433-45

Cairns, M.D., Stabler, R.A., Shetty, N. and Wren, B.W. (2012) The continually envolving *Clostridium difficile* species. Future Microbiology 7(8):945-57

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. and Brüssow, H. (2003) Prophage Genomics. Microbiology and Molecular Biology Review 67(2):238-76

Cantón, R. (2009) Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. Clinical Microbiological Infection 15(1):20-5

Carey, P.G. and Maeda, M. (1985) Horizontal distribution of Psammophilic Ciliates in Fine Sediments of the Chichester Harbor Area. Journal of Natural History 19(3):555-74

Carrolo, M., Frias, M.J., Pinto, F.R., Melo-Cristino, J. and Ramirez, M. (2010) Prophage Spontaneous Activation Promotes DNA Release Enhancing Biofilm Formation in *Streptococcus pneumonia.* PLoS One 5(12):e15678.

Carter, G.P., Lyras,D., Allen, D.L.,Mackin, K.E., Howarth, P.M., O'Connor, J.R. and Rood, J.I. (2007) Binary toxin production in *Clostridium difficile* is regulated by CdtR, a LytTR family response regulator. Journal of Bacteriology 189(20):290-301

Carter, G.P., Purdy, D., Williams, P. and Minton, N.P. (2005) Quorum sensing in the *Clostridium difficile*: analysis of a luxS-type signalling system. Journal of Medical Microbiology 54(2):119-27

Cartwright, C.P., Stock, F., Beekman, S.E., Williams, E.C. and Gill, V.J. (1995) PCR amplification of rRNA intergenic spacer regions as a method for epidemiologic typing of *Clostridium difficile*. Journal of Clinical Microbiology 33(1):184-7

Carver, T.J., Thomson, N., Bleasby, A., Berriman, M. and Parkhill, J. (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25(1):119-20

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21(16):3422-3

Casjens, S.R. (2005) Comparative genomics and evolution of the tailed-bacteriophages. Current Opinion in Microbiology 8(4):451-8

Casjens, S.R. (2011) The DNA-packaging nanomotor of tailed bacteriophages. Nature Reviews Microbiology 9:647-57

Casjens, S.R. and Thuman-Commike, P.A. (2011) Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. Virology 411:393-415

Castagliuolo, I., Riegler, M.F., Valenick, L., LaMont, J.T. and Pothoulakis, C. (1999) *Saccharomyces boulardii* protease inhibits the effects of *Clostridium difficile* toxins A and B in human colonic mucosa. Infection and Immunology 67(1):302-7

Cerquetti, M., Biol, D., Luzzi, I., Caprioli, A., Sebastianelli, A. and Mastrantonio, P. (1995) Role of *Clostridium difficile* in Childhood Diarrhea. Pediatric Infectious Disease Journal 14(7):598-603

Ceyssens, P-J., Glonti, T., Kropinski, A.M., Lavigne, R., Chanishvili, N., Kulakov, L., Lashkhi, N., Tediashvili, M. and Merabishvili, M. (2011) Phenotypic and genotypic variations within a single bacteriophage species. Virology Journal 8:134

Chang, J., Kalra, A., Tonelli, A., Khalife, W. and Schmidt, T. (2008) Decreased diversity of the faecal microbiome in recurrent *Clostridium difficile*-associated diarrhoea. Journal of Infectious Diseases 197(3):435-8

Chang, J.Y., Shim, J.O., Ko, J.S., Seo, J.K., Lee, J.A., Kim, H.S., Choi, J.H., Shin, S. and Shin, S.M. (2012) Monitoring of *Clostridium difficile* Colonization in preterm infants in neonatal intensive care units. PGHN 15(1):29-37

Cheung, J.K., Keyburn, A.L., Carter, G.P., Lanckriet, A.L., Immerseel, F.V., Moore, R.J. and Rood, J.I. (2010) The VirSR Two-Component Signal Transduction System Regulates NetB Toxin Production in *Clostridium perfringens*. Infection and Immunity 78(7):3064-72

Chen, F. and Lu, J. (2002) Genomic Sequence and Evolution of Marine Cyanophage P60L a New Insight on Lytic and Lysogenic Phages. Applied Environmental Microbiology 68(5):2589-94

Chen, J. and McClane, B.A. (2012) Role of the Agr-Like Quorum-Sensing System in Regulating Toxin Production by *Clostridium perfringens* Type B Strains CN1793 and CN1795. Infection and Immunity 80(9):3008-17

Chen, J. and Novick, R.P. (2009) Phage-mediated Intergenic Transfer of Toxin Genes. Science 323(5910):139-41

Chen, J., Rood, J.I. and McClane, B.A. (2011) Epsilon-Toxin Production by *Clostridium perfringens* Type D Strain CN3718 Is Dependent upon the *agr* Operon but Not the VirS/VirR Two-Component Regulatory System. mBio 2:6

Chen, F., Wang, K., Stewart, J. and Belas, R. (2006) Induction of Multiple Prophages from a Marine Bacterium: a Genomic Approach. Applied Environmental Microbiology 72(7):4995-5001

Chopin, M-C., Chopin, A. and Bidnenko, E. (2005) Phage abortive infection in lactococci: variations on a theme. Current Opinion in Microbiology 8(4):473-9

Christie, G.E., Allison, H.A., Kuzio, J., McShan, M., Waldor, M.K. and Kropinski, A.M. (2012) Prophage-induced changes in cellular cytochemistry and virulence. In: Bacteriophages in Health and Disease. Editors Abedon, S. and Hyman, P. CABI Press, Oxfordshire, U.K. pp. 33-60

Colomer-Lluch, M., Jofre, J. and Muniesa, M. (2011b) Antibiotic Resistance Genes in the Bacteriophage DNA Fraction of Environmental Samples. PLoS one 6(3): e17549

Colomer-Lluch, M., Imamovic, L., Jofre, J. and Muniesa, M. (2011a) Bacteriophages Carrying Antibiotic Resistance Genes in Fecal Waste from Cattle, Pigs, and Poultry. Antimicrobial Agents and Chemotherapy 55(10): 4908-11

Comeau, A.M., Tremblay, D., Moineau, S., Rattei, T., Kushkina, A.I., Tovkach, F.I., Krisch, H.M. and Ackermann, H.W. (2012) Phage morphology recapitulates phylogeny: the comparative genomes of a new group of myoviruses. PLoS One 7(7):e40102

Cooksley, C.M., Davis, I.J., Winzer, K., Chan, W.C., Peck, M.W. and Minton, N.P. (2010) Regulation of Neurotoxin Production and Sporulation by a Putative agrBD Signaling System in Proteolytic *Clostridium botulinum*. Applied Environmental Microbiology 76(13):4448-60

Cooperstock, M.S., Steffen, E., Yolken, R. and Onderdonkl, A. (1982) *Clostridium difficile* in Normal Infants and Sudden Infant Death Syndrome: An Association with Infant Formula Feeding. Pediatrics 70(1):91-5

Costa, M.C., Stämpli, H.R., Arroyo, L.G., Pearl, D.L. and Weese, J.S. (2010) Epidemiology of *Clostridium difficile* on a veal farm: Prevalence molecular characterization and tetracycline resistance. Microbiological Drug Resistance 10(1):57-63

Curry, S.R., Marsh, J.W., Schlackman, J.L. and Harrison, L.H. (2012) Prevalence of *Clostridium difficile* in uncooked ground meat products from Pittsburgh, Pennsylvania. Applied and Environmental Microbiology 78(12):4183-6

Dallas, S. (1998) Binding of *Clostridium difficile* toxin A to human milk secretary component. Journal of Medical Microbiology 47(10):879-88

Darling, A.R., Mau, B. and Perna, N.T. (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Plos One 5(6):e11147

Davies, J. (1994) Inactivation of antibiotics and the dissemination of resistance genes. Science 264(5157):375-82

Davies, H.A. and Borriello, S.P. (1990) Detection of capsule in strains of *Clostridium difficile* of varying virulence and toxigenicity. Microbial Pathogenesis 9: 141–6

D'Costa, V.M., McGrann, K.M., Hughes, D.W., Wright, G.D. (2006) Sampling the Antibiotic Resistome. Science 311(5759):374-7

Dei, R. (1989) Observations on phage-typing of *Clostridium difficile*: Preliminary evaluation of a phage panel. European Journal of Epidemiology 5(3):351-4

del Casale, A., Flanagan, P.V., Larkin, M.J., Allen, C.C.R. and Kulakov, L.A. (2011) Analysis of transduction in wastewater bacterial populations by targeting the phage-derived 16S rRNA gene sequences. FEMS Microbiology Ecology 76(1):100-8

del Mar Gamboa, M., Rodríguez, E. and Vargas, P. (2005) Diversity of mesophilic *Clostridium* in Costa Rican soils. Anaerobe 11(6):322-6

Delaney, N., Balenger, S., Bonneaud, C., Marx, C., Hill, G., Ferguson-Noel, N., Tsai, P., Rodrigo, A. and Edwards, S. (2012). Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*. PloS Genetics 8:e1002511

Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Research 27(23):4636-41

Delmée, M. and Avesani, V. (1990) Virulence of ten serogroups of *Clostridium difficile* in hamsters. Journal of Medical Microbiology 33:85-90

Dempsey, R.M., Carroll, D., Kong, H., Higgins, L., Keane, C.T. and Coleman, D.C. (2005) Sau421, a BcgI-like restriction-modification system encoded by the *Staphylococcus aureus* quadruple-converting phage Phi42. Microbiology 151(4):1301-11

Denève, C., Janoir, C., Poilane, I., Fantinato, C. and Collignon, A. (2009) New trends in *Clostridium difficile* virulence and pathogenesis. International Journal of Antimicrobial Agents 33(1):24-8

Deng, L. and Hayes, P.K. (2008) Evidence for cyanophages active against bloom-forming freshwater cyanobacteria. Freshwater Biology 53(6):1240-52

Dethlefsen, L., Huse, S., Sogin, M. and Relman, D. (2008) The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. PLOS Biology 6(11):2383-400

Dingle, K.E., Griffiths, D., Didelot, X., Evans, J., Vaughan, A., Kachrimanidou, M., Stoesser, N., Jolley, K.A., Golubchik, T., Harding, R.M., Peto, T.E., Fawley, W., Walker, A.S., Wilcox, M. and Crook, D.W. (2011) Clinical *Clostridium difficile*: Clonality and Pathogenicity Locus Diversity PLoS One e19993

Dingle, T.C., Mulvey, G.L. and Armstrong, G.D. (2011) Mutagenic analysis of the *Clostridium difficile* flagellar proteins, FliC and FliD and their contribution to virulence in hamsters. Infection and Immunology 79(10):4061-67

Donlan, R. (2009) Preventing biofilms of clinically relevant organisms using bacteriophage. Trends in Microbiology 17(2):66-72

Dorscht, J., Klumpp, J., Beilmann, R., Schmelcher, M., Born, Y., Zimmer, M., Calendar, R. and Loessners, M.J. (2009) Comparative Genome Analysis of *Listeria* Bacteriophages Reveals Extensive Mosaicism, Programmed Translational Frameshifting, and a Novel Prophage Insertion Site. Journal of Bacteriology 191(23):7206-15

Drlica, K. and Zhao, X.L. (1997) DNA gyrase, topoisomerase IV and the 4-quinolones. Microbiology and Molecular Biology Reviews 61(3): 377

Echols, H. (1972) Developmental pathways for the temperate phage: lysis vs lysogeny. Annual Reviews Microbiology 6:157-90

Edwards, P. and Smit, J. (1991) A transducing Bacteriophage for *Caulobacter crescentus* Uses the Paracrystallin Surface Layer Protein as a Receptor. British Journal of Bacteriology 173(17):5568-72

Eidhin, D. N., Ryan, A.W., Doyle, R.M., Walsh, J.B. and Kelleher, D. (2002) Sequence and phylogenetic analysis of the gene for surface layer protein *slpA* from 14 PCR ribotypes of *Clostridium difficile*. Journal of Medical Microbiology 55(1):69-83

Environmental Agency (2010) Fact Sheet: An environmental summary of Portsmouth (U.K. Environment Agency).

Erdmann, S. and Garrett, R. (2012). Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Molecular Microbioliology 85: 1044-56.

Fagan, R. and Fairweather, N. (2010) Dissecting the Cell Surface. Methods of Molecular Biology 646:117-34.

Fallani, M., Rigottier-Gois, L., Aguilera, M., Bridonneau, C., Collignon, A., Edwards, C., Corthier, G. and Doré, J. (2006) *Clostridium difficile* and *Clostridium perfringens* species detected in infant faecal microbiota using 16S rRNA targeted probes. Journal of Microbiological Methods 67(1):150-61

Farrow, K.A., Lyras, D. and Rood, J.I. (2001) Genomic analysis of the erythromycin resistance element Tn5398 from *Clostridium difficile*. Microbiology 147(10):2717-28

Fawley, W. N., Freeman, J., Smith, C., Harmanus, C., van den Berg, R.J., Kuijper, E.J. and Wilcox, M.H. (2008) Use of highly discriminatory fingerprinting to analyse clusters of *Clostridium difficile* infection cases due to epidemic ribotype 027 isolates. Journal of Clinical Microbiology 46(3):954-60.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-91.

Fakaety, R., Kim, K-H., Brown, D., Batts, D.H., Cudmore, M. and Silva, J. (1981) Epidemiology of antibiotic-associated colitis: Isolation of *Clostridium difficile* form the hospital environment. The American Journal of Medicine. 70(4):906-908

Fiorentin, L., Vieira, N.D. and Barioni, W. J. (2005) Oral treatment with bacteriophages reduces the concentration of *Salmonella* Enteritidis PT4 in caecal contents of broilers. Avian Pathology 34(3):258-63

Fischetti, V.A. (2008) Bacteriophage Lysins as Effective Antibacterials. Current Opinion in Microbiology 11(5):393-400.

Fischer, M.G., Allen, M.J., Wilson, W.H. and Suttle, C.A. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. PNAS 107(45):19508-13

Flores, C.O., Meyer, J.R., Valverde, S., Farr, L. and Weitz, J.S. (2011) Statistical structure of host-phage interactions. Proceedings of the National Academy of Sciences U.S.A. 108(28):E288-97

Fortier, L-C. and Moineau, S. (2007) Morphological and Genetic Diversity of temperate phages in *Clostridium difficile*. Applied and Environmental Microbiology 73(22):7358-66

Fortuna, W., Miedzybrodzki, R., Weber-Dabrowska, B. and Górski, A. (2008) Bacteriophage therapy in children: facts and prospects. Medical Science Monitor 14(8):126-32

Franks, A.H., Harmsen, H.J., Raangs, G.C., Jansen, G.J., Schut, F. and Welling, G.W. (1998) Variations of bacterial populations in human feces measured by fluorescent *in situ* hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. Applied Environmental Microbiology 64(9):3336-45

Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. Nature 3:722-32

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

Gao, E.B., Yuan, X.P., Li, R.H. and Zhang, Q.Y. (2009) Isolation of a novel cyanophage infectious to the filamentous cyanobacterium *Planktothrix agardhii* (Cyanophyceae) from Lake Donghu, China. Aquatic Microbiology Ecology 54:163-70

Ghosh, D., Roy, K., Williamson, K.E., Srinicasiah, S., Wommack, K.E. and Radosvich, M. (2009) Acyl-Homoserin Lactones Can Induce Virus Production in Lysogenic Bacteria: an Alternative Paradigm for Prophage Induction. Applied Environmental Microbiology 75(22):7142-52

Goh, S., Riley, T.V. and Chang, B.J. (2005a) Isolation and Characterization of Temperate Bacteriophages of *Clostridium difficile*. Applied Environmental Microbiology 71(2):1078-83

Goh, S., Chang, B.J. and Riley, T.V. (2005b) Effect of phage infection on toxin production by *Clostridium difficile*. Journal of Medical Microbiology 54:129-35.

Goh, S., Ong, P.F., Song, K.P., Riley, T.V. and Chang, B. J. (2007) The complete genome sequence of *Clostridium difficile* phage φC2 and comparisons to φCD119 and inducibile prophages of CD630. Microbiology 153(3):676-85

Goorhuis, A., Bakker, D., Corver, J., Debast, S.B., Harmanus, C., Notermans, D.W., Bergwerff, A.A., Dekker, F.W. and Kuijper, E.J. (2008) Emergence of *Clostridium difficile* Infection Due to a New Hypervirulent Strain, Polymerase Chain Reaction Ribotype 078. Clinical Infectious Diseases 47 (9):1162-70

Gordon, J.I., Ley, R.E., Wilson, R., Mardis, E., Xu, J., Fraser, C. and Relman, D.A. (2005) Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI). Available: http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf. Accessed: 13th December 2012.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic acids Research 38 Supplement: W695-9

Gould, L.H. and Limbago, B. (2010) *Clostridium difficile* in Food and Domestic Animals: A New Foodborne Pathogen? Clinical Infectious Diseases 51(5):577-82

Govind, R., Fralick, J.A. and Rolfe, R. D. (2006) Genomic Organization and Molecular Characterization of *Clostridum difficile* Bacteriophage phiCD119. Journal of Bacteriology 188(7):2588-77

Govind, R., Vediyappan, G., Rolfe, R.D., Dupuy, B. and Fralick, J.A. (2009) Bacteriophage-Mediated Toxin Gene Regulation in *Clostridium difficile*. Journal of Virology 83(23):12037-45

Grissa, I., Vergnaud, G. and Pourcel, C. (2007a) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research 35 (2):52-7

Grissa, I., Vergnaud, G. and Pourcel, C. (2007b) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and re CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. BMC Bioinformatics 8:172

Grissa, I., Vergnaud, G. and Pourcel, C. (2008) CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Research 36(W): 145–8

Garvey, P., Fitzgerald, G.F. and Hill, C. (1995) Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40. Applied Environmental Microbiology 61(12):4321-8

Groth, A.C. and Calos, M. P. (2004) Phage Integrases: Biology and Applications. Journal of Molecular Biology 335(3):667-78

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Research 38(W): 695-9

Govind, R., Fralick, J.A. and Rolfe, R. D. (2006) Genomic Organization and Molecular Characterization of *Clostridum difficile* Bacteriophage phiCD119. Journal of Bacteriology 188(7):2588-77

Gu, J., Liu, X., Li, Y., Han, W., Lei, L., Yang, Y., Zhao, H., Gao, Y., Song, J., Lu, R., Sun, C. and Feng, X. (2012) A method for generation phage cocktail with great therapeutic potential. PLoS One 7(3):e31698

Gürtler, V. (1993) Typing of *Clostridium difficile* strains by PCR-amplification of variable length 16S-23S rDNA spacer regions. Microbiology 139(12):3089-97

Hächler, H. and Wüst, J. (1984) Reexamination by bacteriophage typing of *Clostridium difficile* strains isolated during a nosocomial outbreak. Journal of Clinical Microbiology 20(3):604

Hafiz, S. (1974) *Clostridium difficile* and its toxins (thesis, Ph.D.) Department of Microbiology, University of Leeds

Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. (2005) A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. PLoS Computational Biology 1(6):474-83

Hatfull, G.F., Creswan, S.G. and Hendrix, R.W. (2008) Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. Research in Microbiology 159(5):332-9

Hall, I.C. and O'Toole, E. (1935) Intestinal flora in new-born infants with a description of a new pathogenic anaerobe, *Bacillus difficilis.* American Journal of Diseases of Children 49:390-402

Hatfull, G.F. (2012) Complete Genome Sequences of 138 Mycobacteriophages. Journal of Virology 86(4):2382-4

Hatfull, G. F. and the Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science Program, the KwaZulu-Natal Research Institute for Tuberculosis and HIV Mycobacterial Genetics Course Students and the Phage Hunters Integrating Research and Education Program (2012) Complete Genome Sequences of 138 Mycobacteriophages. Journal of Virology 86(4):2382-4

Hatfull, G.F., Creswan, S.G. and Hendrix, R.W. (2008) Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. Research in Microbiology 159(5):332-9

Hawkins, C.C., Buggy, B.P., Fekety, R. and Schaberg, D. R. (1984) Epidemiology of Colitis Induced by *Clostridium difficile* in Hamsters: Application of a Bacteriophage and Bacteriocin Typing System. The Journal of Infectious Diseases 149(5):775-80

He, M., Sebaihia, M., Lawley, T.D., Stabler, R.A., Dawson, L.F., Martin, M.J., Holt, K.E., Seth-Smith, H.M., Quail, M.A., Rance, R., Brooks, K., Churher, C., Harris, D., Bentley, S.D., Burrows, C., Clark, L., Corton, C., Murray, V., Rose, G., Thurston, S., van Tonder, A., Walker, D., Wren, B.W., Dougan, G. and Parkhill, J. (2010) Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proceedings of the National Academy of Sciences USA 107(16)7527-32

He, M., Miyajima, F., Roberts, P., Ellison, L., Pickard, D.J., Martin, M.J., Connor, T.R., Harris, S.R., Fairley, D., Bamford, K.B., D'Arc, S., Brazier, J., Brown, D., Coia, J.E., Douce, G., Gerding, D., Kim, H.J., Koh, T.H., Kato, H., Senoh, M., Louie, T., Michell, S., Butt, E., Peacock, S.J., Brown, N.M. *et al* (2012) Emergence and global spread of epidemic healthcare-associated *Clostridium difficile.* Nature Genetics doi:10.1038/ng.2478

Hendey, N.I. (1951) Littoral diatoms of Chichester harbour with special reference to fouling. Journal of the Royal Microscopical Society, Great Britain 71(1):1-86

Hendry, G.S. and Fitz-James, P.C. (1974) Exlusion of Induced Bacteriophage from Cells of a Lysogenic *Bacillus megaterium* Committed to Sporulation. Journal of Bacteriology 11(1):295-303

Hendrix, R. W. (2003) Bacteriophage genomics. Current Opinion in Microbiology 6(5):506-11

Hendrix, R.W. (2008) Phage Evolution. In: Bacteriophage Ecology. Editor Abedon, S.T. Cambridge University Press, U.K. pp 177-194.

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E. and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proceedings of the National Academy of Sciences USA 96(5):2192-7

Hensgens, M.P.M., Keessen, E.C., Squire, M.M., Riley, T.V., Koene, M.G.J., de Boer, E., Lipman, L.J.A. and Kuijper, E.J. (2012) *Clostridium difficile* infection in the community: a zoonotic disease? Clinical Microbiology and Infection 18(7):635-45

Higazi, T.B., Mohannad, AL. S., Burkett, M. and Pusok, R. (2011) PCR detection of *Clostridium difficile* and its Toxigenic Strains in Public Places in Southeast Ohio. International Journal of Microbiological Research 2(2):105-11

Ho, T.D. and Ellermeier, C.D. (2011) PrsW is required for colonization, resistance to antimicrobial peptides, and expression of extracytoplasmic function σ factors in *Clostridium difficile*. Infection and Immunology 79(8):3229-38

Hopman, N.E.M., Keesen, E.C., Harmanus, C., Sanders, I.M.J.G., van Leengoed, L.A.M.G., Kuijper, E.J. and Lipman, L.J.A. (2011b) Acquisition of *Clostridium difficile* by piglets. Veterinary Microbiology 149(1-2):186-92

Hopman, N.E.M., Sanders, I.M.J.G., Kuijper, E.J. and Lipman, L.J.A. (2011a) Low risk of transmission of *Clostridium difficile* to humans at petting farms. Veterinary Microbiology 150(3-4):416-7

Horgan, M., O'Sullivan, O., Coffey, A., Fitzgerald, G.F., van Sinderen, D., McAuliffe, O. and Ross, R.P. (2010) Genome analysis of the *Clostridium difficile* phage phiCD6356, a temperate phage of the *Siphoviridae* family. Gene 462(1-2):34-43

Horvath, P., Coûté-Monvoisin, A.C., Romero, D.A., Bovaval, P., Fremaux, C. and Barrangou, R. (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. International Journal of Food Microbiology 131(1):62-70

Høyland-Kroghsbo, N. M., Mærkedahl, R.B. and Svenningsen, S.L. (2013) A Quorum-Sensing-Induced Bacteriophage Defense Mechanism, mBio 4(1):e00362-12

HPS (2013) HPS Weekly Report 47(2013/18) ISSN 1753-4224 Available: http://www.documents.hps.scot.nhs.uk/ewr/pdf2013/1318.pdf. Accessed 5th May 2013

Huang, H., Fang, H., Weintraub, A. and Nord, C.E. (2009) Distinct ribotypes and rates of antimicrobial drug resistance in *Clostridium difficile* from Shanghai and Stockholm. Clinical Microbiology and Infection 15(12):1170-3

Huang, L.H., Farnet, C.M., Ehrlich, K.C. and Ehrlich, M. (1982) Digestion of highly modified bacteriophage DNA by restriction endonucleases. Nucleic Acid Research 10(5):1579-91

Huang, H., Weintraub, A., Fang, H. and Nord, C.E. (2009) Antimicrobial resistance in *Clostridium difficile*. International Journal of Antimicrobial Agents 34(6):516-22

Indra, A., Lassnig, H., Baliko, N., Much, P., Fiedler, A. and Huhulescu, S. (2009) *Clostridium difficile*: a new zoonotic agent? Wiener Klinische Wochenschrift 121(3-4):91-5

Indra, A., Blaschitz, M., Kernbichler, S., Resichl, U., Wewalka, G. and Allerberger, F. (2010) Mechanisms behind variation in the *Clostridium difficile* 16S-23S rRNA intergenic spacer region. Journal of Medical Microbiology 59(11):1317–23.

Indra, A., Huhulescu, S., Schneeweis, M., Hasenberger, P., Kembichler, S., Fiedler, A., Wewalker, G., Allerberger, F. and Kuijper, E.J. (2008) Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis based PCR Ribotyping. Journal of Medical Microbiology 57(11):1377-82

Iyer, L.M., Koonin, E.V. and Aravind, L. (2002a) Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC Genomics 3:8

Iyer, L., Koonin, E. and Aravind, L. (2002b) Extensive domain shuffling in transcription regulators of NA viruses and implications for the origin of fungal APSES transcription factors. Genome Biology 3(3):R0012.1

Janezic, S., Ocepek, M., Zidaric, V. and Rupnik, M. (2012) *Clostridium difficile* genotypes other than ribotype 078 that are prevalent among human, animal and environmental isolates. BMC Microbiology 12:48

Jank, T. and Aktories, K. (2008) Structure and mode of action of clostridial glucosylating toxins: the ABCD model. Trends in Microbiology 16(5):222-229

Jang, S.H., Yoon, B.H. and Chang, H.I. (2011) Complete nucleotide sequence of the temperate bacteriophage LBR48, a new member of the family *Myoviridae*. Archives of Virology 156(2):319-22

Jansen, R., van Embden, J., Gaastra, W. and Schouls, L. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology 43:1565-75.

Jasni, A. S., Mullany, P., Hussain, H. and Roberts, A.P. (2010) Demonstration of Conjugative Transposon (tn5397)-Mediated Horizontal Gene Transfer between *Clostridium difficile* and *Enterococcus faecalis.* Antimicrobial Agents Chemotherapy 54(11):4924-6

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S,m Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. and von Mering, C. (2009) STRING 8—a global view on proteins and their functional interactions. Nucleaic Acids Research 37(D): 412-6

Jhung, M.A., Thompson, A.D., Killgore, G.E., Zukowski, W.E., Songer, G., Warny, M., Johnson, S., Gerding, D.N., McDonald, L.C. and Limbago, B.M. (2008) Toxinotype V *Clostridium difficile* in Humans and Food Animals. Emerging Infectious Diseases 14(7):1039-45

Jiang, Z.D., Garey, K.W., Price, M., Graham, G., Okhuysen, P., Dao-Tran, T., LaRocco, M., DuPont, H.L. (2007) Association of interleukin-8 polymorphism and immunoglobulin G anti-toxin A in patients with *Clostridium difficile*-associated diarrhoea. Clinical Gastroenterology and Hepatology 2007(8):964-8.

Jiang, S.S. and Paul, J.H. (1998) Gene Transfer by Transduction in the Marine Environment. Applied Environmental Microbiology 64(8):2780-7

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8:275-82.

Lavigne, R., Darius, P., Summer, E.J., Seto, D., Mahadevan, P., Nilsson, A.S., Ackermann, H.W. and Kropinski, A.M. (2009) Classification of *Myoviridae* bacteriophages using protein sequence similarity. BMC Microbiology 9:224

Kaatz, G.W., Gitlin, S.D., Schaberg, D.R., Wilson, K.H., Kauffman, C.A., Seo, S.M. and Fekety, R. (1988) Acquisition of *Clostridium difficile* from the hospital environment. American Journal of Epidemiology. 127(6):1289-94

Kalchayanand, N., Arthur, T.M., Bosilevac, J.M., Brichta-Harhav, D.M., Shackelford, S.D., Wells, J.E., Wheeler, T.L. and Koohmaraie, M. (2013). Isolation and characterization of *Clostridium difficile* associated with beef cattle and commercially produced ground beef. Journal of Food Protection 76(2):256-64

Kang, J., Lee, G., Jeung, E. and Yang, M. (2009) Trans-10, cis-12 conjugated linoleic acid modulates phagocytic responses of canine peripheral blood polymorphonuclear neutrophilic leukocytes exposed to *Clostridium difficile* toxin B. Veterinary Immunology and Immunopathology 130(3-4):178-86

Karginov, F.V. and Hannon, G.J. (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. Molecular Cell 37(1):7-19

Karlsson, S., Dupuy, B., Mukherjee, K., Norin, E., Burman, L.G. and Åkerland, T. (2003) Expression of *Clostridium difficile* Toxins A and B and their sigma factor TcdD is controlled by temperature. Applied and Environmental Microbiology 71(4):1784-93

Kato, H., Kato, N., Watanabe, K., Iwai, N., Nakamura, H., Yamamoto, T., Suzuki, K., Kim, S-M., Chong, Y. and Wasito, E.B. (2008) Identification of Toxin A-Negative, Toxin B-Positive *Clostridium difficile* by PCR. Journal of Clinical Mirobiology 3(8):2178-82

Keel, K., Brazier, J.S., Post, K.W., Weese, S. and Songer, J.G. (2007) Prevalence of PCR Ribotypes among *Clostridium difficile* Isolates from Pigs, Calves and Other Species. Journal of Clinical Microbiology 45(6):1963-4

Keel, M.K. and Songer, J.G. (2006) The Comparative Pathology of *Clostridium difficile*–associated Disease. Veterinary Pathology 43:225–40

Kelly, D., McAuliffe, O., Ross, R.P., O'Mahony, J. and Coffey, A. (2011) Development of a broad-host-range phage cocktail for biocontrol. Bioengineered Bugs 2(1):31-7

Kenzaka, T., Tani, K. and Nasu, M. (2010) High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. ISME Journal 4(5):648-59

Kidambi, S.P., Ripp, S. and Miller, R.V. (1994) Evidence for phage-mediated gene transfer among *Pseudomonas aeruginosa* strains on the phylloplane. Applied and Environmental Microbiology 60:496-500.

Killgore, G., Thompson, A., Johnson, S., Brazier, J., Kuijper, E., Pepin, J., Frost, E.H., Savelkoul, P., Nicholson, B., can den Berg, R.J., Kato, H., Sambol, S.P., Zukowski, W., Woods, C., Limbago, B., Gerding, D.N. and McDonald, L.C. (2008) Comparison of seven techniques for typing international epidemic strains of *Clostridium difficile*: Restriction endonuclease analysis, pulsed-field gel electrophoresis, variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. Journal of Clinical Microbiology 46(2):431-7

Kim, K.H., Fekety, R., Batts, D.H., Brown, D., Cudmore, M., Silva, J. J. and Waters, D. (1981) Isolation of *Clostridium difficile* from the environment and contacts of patients with antibiotic-associated colitis. Journal of Infectious Disease 143(1):42-50

Kim, J., Kang, J.O., Pai, H. and Choi, T.Y. (2012) Association between PCR ribotypes and antimicrobial susceptibility among *Clostridium difficile* isolates from healthcare associated infections in South Korea. International Journal of Antimicrobial Agents 40(1):24-9

Kleppen, H.P., New, I.F. and Holo, H. (2012) Characterisation of a *Leuconostoc* bacteriophage infecting flavour producers of cheese starter cultures. Applied and Environmental Microbiology 78(18):6769-72

Konstantinjdis, K.T. and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences, USA, 102(7):2567-72

Krogh, S., Jørgensen, S.T. and Devine, K.M. (1998) Lysis Genes of the *Bacillus subtilis* Defective prophage PBSX. Journal of Bacteriology 180(8):2110-7

Kropinski, A.M. (2000). Sequence of the genome of the temperate, serotype-converting, *Pseudomonas aeruginosa* bacteriophage D3. Journal of Bacteriology 182:6066-6074.

Kuehne, S.A., Cartman, S.T. and Minton, N.P. (2011) Both, toxin A and toxin B, are important in *Clostridium difficile* infection. Gut Microbes 2(4):252-5

Kuijper, E.J., Coignard, B. and Tüll, P. (2006) Emergence of *Clostridium difficile*-associated disease in North America and Europe. ESCMID Study Group for *Clostridium difficile*; EU Member States; European Centre for Disease Prevention and Control. Clinical Microbiology Infection 12(6):2-18

Kuijper, E. J. and Wilcox, M.H. (2008) Decreased Effectiveness of Metronidazole for the Treatment of *Clostridium difficile* Infection? Clinical Infectious Diseases 47:63–5

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinformatics 9:299-306.

Kuno, S., Yoshida, T., Kaneko, T. and Sako, Y. (2012). Intricate Interactions between the Bloom-Forming Cyanobacterium *Microcystis aeruginosa* and Foreign Genetic Elements, Revealed by Diversified Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Signatures. Applied and Environmental Microbiology 78:5353-60.

Labrie, S.J., Samson, J.E. and Moineau, S. (2010) Bacteriophage resistance mechanisms. Nature Reviews Microbiology 8(5):317-27

Lan, S-F., Huang, C-H., Chang, C-H., Liao, W-C., Lin, I.H., Jian, W-N., Wu, Y-G., Chen, S-Y. and Wong, H-C. (2009). Characterization of a New Plasmid-Like Prophage in a Pandemic *Vibrio parahaemolyticus* O3:K6 Strain. Applied and Environmental Microbiology 75:2659-67.

Larkin, M.A., Blacksheilds, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) ClustalW and ClustalX version 2. Bioinformatics 23(21):2947-8

Lawley, T.C., Croucher, N.J., Yu, L., Clare, S., Sebaihia, M., Goulding, D., Pickard, D.J., Parkhill, J., Choudhary, J. and Dougan, G. (2009) Proteomic and Genomic Characterization of Highly Infectious *Clostridium difficile* 630 Spores. Journal of Bacteriology 181(17):5377-86

Lawrence, J.G., Hatfull, G.F. and Hendrix, R.W. (2002) Imbroglios of Viral Taxonomy: Genetic Exchange and Failings of Phenetic Approaches. Journal of Bacteriology 184(17):4891-905

LeBlanc, C., Caumont-Sarcos, A., Comeau, A.M. and Krisch, H.M. (2009) Isolation and genomic characterization of the first phage infecting *Iodobacteria*: phiPLPE, a myovirus having a novel set of features. Environmental Microbiology Reports 1(6):499–509

Lee, A.S. and Song, K.P. (2005) LuxS/autoinducer-2 quorum sensing molecule regulates transcriptional virulence gene expression in *Clostridium difficile*. Biochemical and Biophysical Research Communications 335(3):656-66

Lee, T., Schwartz, C. and Guo, P. (2009) Construction of bacteriophage phi29 DNA packaging motor and its applications in nanotechnology and therapy. Annals in Biomedical Engineering Online 37(10):2064-81

Leijonmarck, M. and Liljas, A. (1987) Structure of the C-terminal domain of the ribosomal protein-L7 protein-L12 from *Escherichia coli* at 1.7 A. Journal of Molecular Biology 195:555-80.

Lemire, S., Figueroa-Bossi, N. and Bossi, L. (2011) Bacteriophage Crosstalk: Coordination of Prophage Induction by Trans-Acting Antirepressors. PLoS Genetics 7(6):e1002149.

Lepage, P., Leclerc, M.C., Joossens, M., Mondot, S., Blottière, H.M., Raes, J., Ehrlich, D. and Doré, J. (2012) A metagenomic insight into our gut's microbiome. Gut 23:doi:10.1136/gutjnl-2011-301805

Li, J., Chen, J., Vidal, J.E. and McClane, B.A. (2011) The Agr-like quorum –sensing system regulates sporulation and production of enterotoxin and beta2 toxin by *Clostridium perfringens* type A non-food-bourne human gastrointestinal disease strain F5603. Infection and Immunology 79(6):2451-9

Lin, Y.C., Huang, Y.T., Tsai, P.J., Lee, T.F., Lee, N.Y., Liao, C.h., Lin, S.Y., Ko, W.C. and Hsueh, P.R. (2011) Antimicrobial Susceptibilities and Molecular Epidemiology of Clinical Isolates of *Clostridium difficile* in Taiwan. Antimicrobial Agents and Chemotherapy 55(4):1701-5

Linares, J.F., Gustafsson, I., Baqueros, F. and Martinez, J.L. (2006) Antibiotics as intermicrobial signalling agents instead of weapons. PNAS 103(51):19484-89

Linevsky, J.K., Pothoulakis, C., Keates, S., Warny, M., Keates, A.C., Lamont, J.T. and Kelly, C.P. (1997) IL-8 release and neutrophil activation by *Clostridium difficile* toxin-exposed human monocytes. American Journal of Physiology 273(6):1333-40.

Loc Carrillo, C., Atterbury, R.J., El-Shibiny, A., Connerton, P.L., Dillon, E., Scott, A. and Connerton, I.F (2005) Bacteriophage Therapy To Reduce *Campylobacter jejuni* Colonization of Broiler Chickens. Applied and Environmental Microbiology 71(11):6554-63

Loo, V.G., Poirier, L., Miller, M.A., Oughton, M, Libman, M.D., Michaud, S., Bourgault, A.M., Nguyen, T., Frenette, C., Kelly, M., Vibien, A., Brassard, P., Fenn, S., Dewar, K., Hudson, T.J., Horn, R., René, P., Monczak, Y. and Dascal, A. (2005) A Predominantly Clonal Multi-Institutional Outbreak of *Clostridium difficile*–Associated Diarrhea with High Morbidity and Mortality. The New England Journal of Medicine 353:2442-9

Lood, R. and Collin, M. (2011) Characterization and genome sequencing of two *Propionibacterium acnes* phages displaying pseudolysogeny. BMC Genomics 12:198

Los, J., Golec, P., Wegrzyn, G., Wegrzyn, A. and Los, M. (2008) Simple method for *plating Escherichia coli* bacteriophages forming very small plaques or no plaques under standard conditions. Applied and Environmental Microbiology 74(16):5113-20

Lyerly, D. M., Saum, K. E., MacDonald, D. K. and Wilkins, T. D. (1985) Effects of *Clostridium difficile* toxins given intragastrically to animals. Infection and Immunology 47(2):349–52

Lyras, D., O'Connor, J.R., Howarth, P.M., Sambol, S.P., Carter, G.P., Phumoonna, T., Poon, R., Adams, V., Vedantam, G., Johnson, S., Gerding, D.N. and Rood, J.I. (2009) Toxin B is essential for virulence of *Clostridium difficile*. Nature 458(7242):1176-9

Mahida, Y.R., Galvin, A., Makh, S., Hyde, S., Sanfilippo, L., Borriello, S.P. and Sewell, H.F. (1998) Effect of *Clostridium difficile* Toxin A on Human Colonic Lamina Propria Cells: Early Loss of Macrophages Followed by T-Cell Apoptosis. Infection and Immunology 66(11):5462-69

Mahida, Y.R., Makh, S., Hyde, S., Gray, T. and Borriello, S.P. (1996) Effect of *Clostridium difficile* toxin A on human intestinal epithelial cells: induction of interleukin 8 production and apoptosis after cell detachment. Gut 38(3):337-47

Mahoney, D.E., Bell, P.D. and Easterbrook, K.B. (1985) Two bacteriophages of *Clostridium difficile*. Journal of Clinical Microbiology 21(2):251-4

Mahoney, D. E., Clow, J., Atkinson, L., Vakharia, N. and Schlech, W.F. (1991) Development and application of a multiple typing system for *Clostridium difficile.* Applied and Environmental Microbiology 57(7):1873-9

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica., F.J.M., Wolf, Y.I., Yakunin, A.F., van der Oost., J. and Koonin, E.V. (2011) Evolution and classification of the CRISPR–Cas systems. Nature Reviews Microbiology 9:467-77

Mani, N. and Dupuy, B. (2001) Regulation of toxin synthesis in *Clostridium difficile* by an alternative RNA polymerase sigma factor. Proceedings of the National Academy of Science USA 98:5844–9.

Mani, N., Lyras, D., Barroso, L., Howarth, P., Wilkins, T., Rood, J.I., Sonenshein, A.L. and Dupuy, B. (2002) Environmental response and autoregulation of *Clostridium difficile* TxeR, a sigma factor for toxin gene expression. Journal of Bacteriology 184(21):5981-8

Mann, N.H., Clokie, M.R.J., Millard, A., Cook, A., Wilson, W., Wheatley, P.J., Letarov, A. and Krisch, H.M. (2005) The genome of the T4-type bacteriophage S-PM2 that infects the oceanic picoplankton. Journal of Bacteriology 187:3188-200.

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer., L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczyzki, J., Leibert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Tamashita, R.A., Zhang, D., Zhang, N. and Bryant, S.H. (2011) CDD:a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 39(D):225-9.

Marsh, P. and Wellington, E.M.H. (1994) Phage host interactions in soil. FEMS Microbiology Ecology 15:99-108

Matamouros, S., England, P. and Dupuy, B. (2007) *Clostridium difficile* toxin expression is inhibited by the novel regulator TcdC. Molecular Microbiology 64(5):1274–88

Matsuki, S., Ozaki, E., Shozu, M., Inoue, M., Shimizu, S., Yamaguchi, N., Karasawa, T., Tamagishi, T. and Nakamura, S. (2005) Colonization by *Clostridium difficile* of neonates in a hospital, and infants and children in three day-care facilities of Kanazawa, Japan. International Microbiology 8(1):43-8

Matsushita, I. And Yanase, H. (2008) A novel thermophillic lysozyme from bacteriophage phiIN93. Biochemical and Biophysical Research Communications 377(1):89-92

Maurice, C.F., Mouillot, D., Bettarel, Y., De Wit, R., Sarmento, H. and Bouvier, T. (2010) Disentangling the relative influence of bacterioplankton phylogeny and metabolism on lysogeny in reservoirs and lagoons. The ISME Journal 5(5):831-42

Mayer, M., Garefalaki, V., Spoel, R., Narbad, A. and Meijers, R. (2011) Structure-based modification of a *Clostridium difficile*-targeting endolysin affects activity and host range. Journal of Bacteriology 193(13):5477-86

Mayer, M.J., Narbad, A. and Gasson, M.J. (2008) Molecular Characterization of a *Clostridium difficile* Bacteriophage and Its Cloned Biologically Active Endolysin. Journal of Bacteriology 190(20):6734-40

Mayer, M.J., Payne, J., Gasson, M.J. and Narbad, A. (2010). Genomic Sequence and Characterization of the Virulent Bacteriophage phiCTP1 from *Clostridium tyrobutyricum* and Heterologous Expression of Its Endolysin. Applied and Environmental Microbiology 76:5415-22.

Mayor, D.J., Thornton, B. and Zuurs A. F. (2012) Resource Quanitity Affects Benthic Microbial Community Structure and Growth Efficiency in a Temperate Intertidal Mudflat. PloS One 7(6):e38582.

McAuliffe, O., Ross, R.P. and Fitzgerald, G.F. (2007) The New Phage Biology: From Genomics to Applications. In: Bacteriophage Genetics and Molecular Biology. Editors McGrath, S. and van Sinderen, D. Caister Academic Press, U.K. pp.1-42

McBee, R.H. (1960) Intestinal Flora of Some Antarctic Birds and Mammals. Journal of Bacteriology 79(2):311-2

McClane, B.A., Lyerly, D.M. and Wilkins, T.D. (2006) Enterotoxic Clostridia: *Clostridium perfringens* Type A and *Clostridium difficile*, In: Gram-positive Pathogens, 2nd Edition, Editors Fischetti, A.V., Novick, P.A., Ferretti, J.J., Portnoy, D.A. and Rood, J.I. ASM Press, Washington D.C. pp. 703-14

McCusker, M.E., Harris, A.D., Perencevich, E. and Roghmann, M-C. (2003) Fluoroquinolone Use and *Clostridium difficile*-Associated Diarrhea. Emerging Infectious Diseases 9(6):730-3

McDonald, L.C., Killgore, G.E., Thompson, A., Owens, R.C., Kazakova, S.V., Sambol, S.P., Johnson, S. and Gerding, D.N. (2005) An Epidemic, Toxin Gene-Varient Strain of *Clostridium difficile*. The New England Journal of Medicine 353:2433-41

Meader, E., Mayer, M.J., Gasson, M.J., Steverding, D., Carding, S.R. and Narbad, A. (2010) Bacteriophage treatment significantly reduces viable *Clostridium difficile* and prevents toxin production in an *in vitro* model system. Anaerobe 16(6):549-54

Mebrhatu, M.T., Wywial, E., Ghosh, A., Michiels, C.W., Lindner, A.B., Taddei, F., Bujnicji, J.M., Melderen, L.V. and Aertsen, A. (2011) Evidence for an evolutionary antagonism between Mrr and Type III modification systems, Nucleic Acids Research 39(14):5991-6001

Medina-Torres, C.E., Weese, J.S., Staempfli, H.R. (2011) Prevalence of *Clostridium difficile* in horses. Veterinary Microbiology 152(1-2):212-5

Meijer, W. J.J., Castilla-Llorente, V., Villar, L., Murray, H., Errington, J. and Salas, M. (2005) Molecular basis for the exploitation of spore formation as survival mechanism by virulent phage 29. The EMBO Journal 24(20):3647–57.

Metcalf, D.S., Avery, B.P., Janecko, N., Matic, N., Reid-Smith, R. and Weese, J.S. (2011) *Clostridium difficile* in seafood and fish. Anaerobe 17(2):85-6.

Matcalf, D.S., Costa, M.C., Dew, W.M. and Weese, J.S. (2010) *Clostridium difficile* in vegetables, Canada. Letters in Applied Microbiology 51(5):600-2

Miller, M.A., Byrne, B.A., Jang, S.S., Dodd, E.M., Dorfmeier, E., Harris, M.D., Ames, J., Paradies, D., Worcester, K., Jessup, D.A. and Miller, W.A. (2010) Enteric bacterial pathogen detection in southern sea otters (*Enhydra lutris nereis*) is associated with coastal urbanization and freshwater runoff. Veterinary Research 41(1):1–13

Monot, M., Boursaux-Eude, C., Thibonnier, M., Vallenet, D., Moszer, I., Medigue, C., Martin-Verstraete, I. and Dupuy, B. (2011) Re-annotation of the genome of *Clostridium difficile* strain 630. Journal of Medical Microbiology 60(8):1193-9

Moreau, H., Piganeau, G., Desdevises, Y., Cooke, R., Derelle, E. and Grimsely, N. (2010) Marine Prasinovirus Genomes Show Low Evolutionary Divergence and Acquisition of Protein Metabolism Genes by Horizontal Gene Transfer. Journal of Virology 84(24):1255-63

Morelli, L. and Callegari, M.L. (1997) Surface layer of *Lactobacillus helveticus* CRNZ 892. FEMs Microbiology Reviews 20(2):99-149

Muniesa, M., García, A., Miró, E., Mirelis, B., Prats, G., Jofre, J. and Navarro, F. (2004) Bacteriophages and Diffusion of β-Lactamase Genes. Emerging Infectious Diseases 10(6):1134-7

McCoubrey, J. and Poxton, I.R. (2001) Variation in the surface layer proteins of *Clostridium difficile*. FEMS Immunology and Medical Microbiology 31(2):131-5

McDonough, M.A. and Butterton, J.R. (1999) Spontaneous tandem amplification and deletion of the Shiga toxin operon in *Shigella dysenteriae* 1. Molecular Microbiology 34(5):1058-69

McFarland, L.V., Mulligan, M.E., Kwok, R.Y. and Stamm, W.E. (1989) Nosocomial acquisition of *Clostridium difficile* infection. New England Journal of Medicine 320(4):204-10.

Miller, R.V. (2001) Environmental bacteriophage-host interactions: factors contribution to natural transduction. Antonie Van Leeuwenhoek 79(2):141-7

Miura, M., Kato, H. and Matsushita, O. (2011) Identification of a novel virulence factor in *Clostridium difficile* that modulates toxin sensitivity of cultured epithelia cells. Infection and Immunology 79(9):3810-20

Monot, M., Boursaux-Eude, C., Thibonnier, M., Vallenet, D., Moszer, I., Medique, C., Martin-Verstraete, I. and Dupuy, B. (2011) Reannotation of the genome sequence of *Clostridium difficile* strain 630. Journal of Medical Microbiology 60(8):1193-9

Moran, N., Degnan, P., Santos, S., Dunbar, H. and Ochman, H. (2005) The players in a mutualistic symbiosis: Insects, bacteria, viruses, and virulence genes. The Proceedings of the National Academy of Sciences of the United States of America 102(47):16919-26

Muniesa, M., Imamovic, L. and Jofre, J. (2011) Bacteriophages and genetic mobilization in sewage and faecally polluted environments. Microbial Biotechnology 4(6):725-34

Nagy, E. and Foldes, J. (1991) Electron-Microscopic investigation of lysogeny of *Clostridium difficile* stains isolated from antibiotic-associated diarrhea cases and from healthy carriers. APMIS 99(4):321-6

Nale, J.Y., Shan, J., Hickenbotham, P.T., Fawley, W.N., Wilcox, M.H. and Clokie, M.R.J. (2012) Diverse Temperate Bacteriophage Carriage in *Clostridium difficile* 027 Strains. PLoS One 7(5):e37263

Nam, H.J., Kang, J.K., Kim, S-K., Ahn, K.J., Seok, H., Park, S.J., Chang, J.S., Pothoulakis, C., Lamont, J.T. and Kim, H. (2010) Clostridium difficile Toxin A Decreases Acetylation of Tubulin, Leading to Microtubule Depolymerization through Activation of Histone Deacetylase 6, and This Mediates Acute Inflammation. Journal of Biological Chemistry 285(43):32888-96

Norman, K.N., Harvey, R.B., Scott, H.M., Hume, M.E., Andrews, K., Brawley, A.D. (2009) Varied prevalence of *Clostridium difficile* in an integrated swine operation. Anaerobe 15(6):256–60

Norman, K.N., Morgan Scott, H., Harvey, R.B., Norby, B., Hume, M.E. and Andrews, K. (2011) Prevalence and Genotypic Characteristics of *Clostridium difficile* in a Closed and Integrated Human and Swine Population. Applied Environmental Microbiology 77(16):5755-60

Novick, R.P. and Geisinger, E. (2008) Quorum Sensing in Staphylococci. Annual Review of Genetics 42:541-64

O'Flaherty, S., Ross, R.P. and Coffey, A. (2009) Bacteriophage and their lysins for elimination of infectious bacteria. FEMS Microbiology Reviews 33(4):801-19

O'Neill, G.L., Ogunsola, F.T., Brazier, J.S. and Duerden, B.I. (1996) Modification of a PCR Ribotyping Method for Application as a Routine Typing Scheme for *Clostridium difficile*. Anaerobe 2(4):205-9

Osbourne, M.S. and Soneneshein, A.L. (2076) Behaviour of a temperate bacteriophage in differentiating cells of *Bacillus subtilis*. Journal of Virology 19(1):26-35

Ohtani, K., Kawsar, H.I., Okumura, K., Hayashi, H. and Shimizu, T. (2003) The VirR/VirS regulatory cascade affects transcription of plasmid-encoded putative virulence genes in *Clostridium perfringens* strain 13. FEMS Microbiology Letters 222(1):137-41

Ozaki, E., Kato, H., Karasawa, T., Maegawa, T., Koino, Y., Matsumoto, K., Takada, T., Nomoto, K., Tanaka, R. and Nakamura, S. (2004) *Clostridium difficile* colonisation in healthy adults: transient colonization and correlation with enterococcal colonization. Journal of Medical Microbiology 53(2):167-72

Papke, R. T. and Doolitte, W.F. (2003) Phage evolution: New worlds of genomic diversity. Current Biology 13(15):606-7

Pasquale, V., Romano, V.J., Rupnik, M., Dumontet, S., Cizner, I., Aliberti, F., Mauri, F., Saggiomo, V., Krovacek, K. (2011) Isolation and characterization of *Clostridium difficile* from shellfish and marine environments. Folia Microbiologica 56(5):431-7

Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME Journal 2(6):579-89

Paul, J.H., Sullivan, M.B., Segall, A.M. and Rohwer, F. (2002) Marine phage genomics. Comparative biochemistry and physiology Part B Biochemistry and molecular biology 133(4):463-76

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J.G., Jacobs, W.R., Hendrix, R.W. and Hatfull, G.F. (2003) Origins of Highly Mosaic Mycobacteriophage Genomes. Cell 112(2):171-82

Peláez, T., Alcalá, L., Alonso, R., Rodríguez-Créixems, M., García-Lechuz, J.M. and Bouza, E. (2002) Reassessment of *Clostridium difficile* Susceptibility to Metronidazole and Vancomycin. Antimicrobial Agents Chemotherapy 45(6):1647-50

Peláez, T., Cercenado, E., Alcalá, L., Marín, M., Martín-López, A., Martínez-Alarcón, J., Catalán, P., Sánchez-Somolinos, M. and Bouza, E. (2008) Metronidazole Resistance in *Clostridium difficile* Is Heterogeneous. Journal of Clinical Microbiology 45(9):3028-32

Penders, J. Stobberingh, E.E., van Ree, R., van den Brandt, P.A. and Thijs, C. (2006) *Clostridium difficile* and *E. coli* colonization in early life are associated with an increased risk of allergy. European Journal of Epidemiology 21:86-86

Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I., van den Brandt, P.A. and Stobberingh, E.E. (2006) Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 118(2):511-21

Penders, J., Vink, C., Driessen, C., London, N., Thijs, C. and Stobberingh, E.E. (2005) Quantification of *Bifidobacterium spp., Escherichia coli* and *Clostridium difficile* in faecal smaples of breast-fed and formula-fed infants by real time PCR. FEMS Microbiology Letters 243(1):141-7

Perelle, S., Gibert, M., Bourlioux, P., Corthier, G. and Popoff, M.R. (1997) Production of a complete binary toxin (actin-specific ADPribosyltransferase) by *Clostridium difficile* CD196. Infection and Immunity 65:1402–7.

Picozzi, C., Volpini, G., Vigentini, I., Grassi, S. and Foschino, R. (2012) Assessment of transduction of *Escherichia coli* Stx2-encoding phage in dairy process conditions. International Journal of Food Microbiology 153(3):388-94

Popoff, M.R., Rubin, E., Gill, D.M. and Boquet, P. (1988) Actin-specific ADPribosyltransferase produced by a *Clostridium difficile* strain. Infection and Immunology 56:2299-306

Pollich, M. and Klug, G. (1995) Identification and sequence analysis of genes involved in late steps in cobalamin (vitamin B12) synthesis in *Rhodobacter capsulatus.* Journal of Bacteriology 177(15):4481-7

Pothoulakis, C. (2000) Effects of *Clostridium difficile* toxins on epithelial cell barrier. Annals of New York Academy of Sciences 915:347-56

Pourcel, C., Salvignol, G. and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151(3):653-63

Poxton, I.R., McCoubrey, J. and Blair, G. (2001) The pathogenicity of *Clostridium difficile*. Clinical Microbiology and Infection 7(8):421-7

Purcell, E.R., McKee, R.W., McBride, S.M., Waters, C.M. and Tamayo, R. (2012) Cycle Diguanylate Inversely Regulates Motility and Aggregation in *Clostridium difficile.* Journal of Bacteriology 194(13):3307-16

Rafferty, M.E., McCormick, M.I., Bopp, L.H., Baltch, A.L., George, M., Smith, R.P., Rheal, C., Ritz, W. and Schoonmaker, D. (1997) Vancomycin-resistant enterococci in stool specimens submitted for *Clostridium difficile* cytotoxin assay. Infection Control and Hospital Epidemiology 18(5):342-4

Ramesh, V., Fralick, J. and Rolfe, R. (1999) Prevention of *Clostridium difficile*-induced ileocecitis with bacteriophage. Anaerobe 5(2):69-78

Rajendran, R., Garva, R., Krstic-Demonacos, M. and Demonacos, C. (2011) Sirtuins: Molecular Traffic Lights in the Crossroad of Oxidative Stress, Chromatin Remodeling, and Transcription. Journal of Biomedicine and Biotechnology e368276.

Raux, E., Schubert, H.L. and Warren, M.J. (200) Biosynthesis of cobalamin (vitamin B_{12}): a bacterial conundrum. Cellular and Molecular Life Sciences 57:1880–93

Rea, M.C., O'Sullivan, O., Shanahan, F., O'Toole, P.W., Stanton, C., Ross, R.P. and Hill, C. (2012) *Clostridium difficile* carriage in elderly subjects and associated changes in the intestinal microbiota. Journal of Clinical Microbiology 50(3):867-75

Recsei, P., Kreiswirth, B., O'Reilly, M., Schlievert, P., Gruss, A. and Novicj, R.P. (1986) Regulation of exoprotein gene expression in *Staphlococcus aureus* by *agr*. Molecular and General Genetics 202(1):58-61

Reeves, A.E., Theriot, C.M., Bergin, I.L., Huffnagle, G.B., Schloss, P.D. and Young, V.B. (2011) The interplay between microbiome dynamics and pathogen dynamics in a murine model of *Clostridium difficile* Infection. Gut Micrbobes 2(3):145-158

Rho, M., Wu, Y., Tang, H., Doak, T.G. and Ye, Y. (2012) Diverse CRISPRs Evolving in Human Microbiomes. PloS Genetics 8(6):e1002441

Rice, S.A., Tan, C.H., Mikkelsen, P.J., Kung, V., Woo, J., Tay, M., Hauser, A., McDougald, D., Webb, J.S. and Kjelleberg, S. (2009) The biofilm life cycle and virulence of *Pseudomonas aeruginosa* are dependent on a filamentous prophage. IMSE Journal 3(3):271-282

Riegler, M., Sedivy, R., Pothoulakis, C., Hamilton, G., Zacherl, J., Bischof, G., Cosentini, E., Feil, W., Schiessel, R. and LaMont, J.T. (1995) *Clostridium difficile* toxin B is more potent than toxin A in damaging human colonic epithelium *in vitro*. Journal of Clinical Investigation 95(5):2004-2011

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

Rodriguez-Palacios, A., Koohmaraie, M. and Lejeune, J.T. (2011) Prevalence Enumeration, and Antimicrobial Agent Resistance of *Clostridium difficile* in Cattle at Harvest in the United States. Journal of Food Protection 74(10):1618-1624

Rodriguez-Palacios, A. and LeJeune, J.T. (2011) Moist-Heat Resistance, Spore Aging, and Super-dormancy in *Clostridium difficile.* Applied Environmental Microbiology 77(9):3085-3091

Rodriguez-Valera, F., Martin-Cuadrado, A.B., Rodriguez-Brito, B., Pasic, L., Thingstad, T.F., Rohwer, F. and Mira, A. (2009) Explaining microbial population genomics through phage predation. Nature Reviews Microbiology 7:828-836

Rodriguez-Palacios, A., Reid-Smith, R.J., Staempfli, H.R., Daignault, D., Janecko, N., Avery, B.P., Martin, H., Thomspon, A.D., McDonald, L.C., Limbago, B. and Weese, J.S. (2009) Possible seasonality of *Clostridium difficile* in retail meat, Canada. Emerging Infectious Diseases 15(5):802-805

Rodriguez-Palacios, A., Stämpfli, H.R., Duffield, T., Peregrine, A.S., Trotz-Williams, L.A., Arroyo, L.G., Brazier, J.S. and Weese, J.S. (2006) *Clostridium difficile* PCR ribotypes in calves, Canada. Emerging Infectious Diseases 12(11):1730-6.

Rodriguez-Palacios, A., Staempfli, H.R., Duffield, T. and Weese, J.S. (2007) *Clostridium difficile* in retail ground meat, Canada. Emerging Infectious Disease 13 (3):485-7.

Rohwer, F. and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phages. Journal of Bacteriology 184(16)L4529-4535

Romano, V., Pasquale, V., Krovacek, K., Mauri, F., Demarta, A. and Dumontet, S. (2012) Toxigenic *Clostridium difficile* PCR Ribotypes from Wastewater Treatment Plants in Southern Switzerland. Applied Environmental Microbiology 78(18):6643-6

Romaro, P., Croucher, N.J., Hiller, N.L., Hu, F.Z., Ehrlich, G.D., Bentley, S.D., Garcia, E. and Mitchell, T.J. (2009) Comparative Genomic Analysis of Ten *Streptococcus pneumonia* Temperate Bacteriophages. Journal of Bacteriology 191(15):4854-62

Ropp, P. (2011) Treatment of *Clostridium difficile* infection by use of novel phage therapy. Basic Biotechnology 7:1-5

Rousseau, C., Lemee, L., Le Monnier, A., Poilane, I., Pons, J.L. and Collignon, A. (2011a) Prevalence and diversity of *Clostridium difficile* strains in infants. Journal of Medical Microbiology 60(8):1112-8

Rousseau, S., Levenez, F., Fouqueray, C., Dore, J., Collignon, A. and Lepage, P. (2011b) *Clostridium difficile* colonization in early infancy is accompanied by changes in intestinal microbiota composition. Journal of Clinical Microbiology 49(3):3858-865

Rozen, S. and Skaletsky, H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics Methods and Protocols: Methods in Molecular Biology. Editors Krawetz, S. and Misener, S. Humana Press, Totowa, NJ, pp 365-86

Ruby, E., Henderson, B. and McFall-Ngai, M. (2004) We Get By with a Little Help from Our (Little) Friends. Science 303(5662):1305-7

Rupnik, M. (2001) How to detect *Clostridium difficile* variant strains in a routine laboratory. Clinical Microbiology 7(8):417-20

Rupnik, M. (2008) Heterogeneity of large clostridial toxins: importance of *Clostridium difficile* toxinotypes. FEMS Microbiology Reviews 32(3):541-55

Rupnik, M. (2010) *Clostridium difficile*: (Re)emergence of Zoonotic Potential. Clinical Infectious Diseases 51(5):583-4

Rupnik, M., Wilcox, M.H. and Gerding, D.N. (2009) *Clostridium difficile* infection: new developments in epidemiology and pathogenesis. Nature Reviews in Microbiology 7(7):526-36

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10):944-5

Queen, E.V., Marks, S.L. and Farver, T.B. (2012) Prevalence of Selected Bacterial and Parasitic Agents in Feces from Diarrheic and Healthy Control Cats from Northern California. Journal of Veterinary Internal Medicine 26(1):54-62

Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-25

Sakaguchi, Y., Hayashi, T., Kurokawa, K., Oshima, K., Fujinaga, Y., Ohnishi, M., Ohtsubo, E., Hattori, M. and Oguma, K. (2005)The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. PNAS 102(48):17472-7

Salzberg, S., Delcher, A., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Research 26(2):544-8

Samore, M.H., Venkatarman, L., DeGirolami, P.C., Arbeti, R.D. and Karchmer, A.W. (1996) Clinical and molecular epidemiology of sporadic and clustered cases of nosocomial *Clostridium difficile* diarrhea. American Journal of Medicine 100(1):32-40

São-José, C., Nascimento, J.G., Parreira, R. and Santos, M.A. (2007) Release of Progeny Phages from Infected Cells. In: Bacteriophage Genetics and Molecular Biology. Editors McGrath, S. and van Sinderen, D. Caister Academic Press, U.K. pp. 309-43

Saujet, L., Monot, M., Dupuy, B., Soutourina, O. and Martin-Verstraete, I. (2011) The Key Sigma Factor of Transition Phase, SigH, Controls Sporulation, Metabolism, and Virulence Factor Expression in *Clostridium difficile.* Journal of Bacteriology 193(13):3186-96

Scaria, J., Ponnala, L., Janvilisri, T., Tan, W., Mueller, L.A. and Chang, Y-F. (2010) Analysis of Ultra Low Genome Conservation in *Clostridium difficile*. PLoS One 5(12):e15147

Schallehn, G. (1985) *Clostridium sordellii* bacteriophage active on *Clostridium difficile*. Letters in Applied Microbiology 1(4):71-4

Schuch, R. and Fischetti, V.A. (2009) The Secret Life of the Anthrax Agent *Bacillus anthracis*: Bacteriophage-Mediated Ecological Adaptations. PLoS One 4(8):e6532.

Scott, C. Fletcher, R.L. and Bremer, G.B. (1996) Observations on the mechanisms of attachment of some marine fouling blue-green algae. Biofouling 10(1-3):161-73

Sears, C. L. (2005) A dynamic partnership: Celebrating our gut microbiota. Anaerobe 11(5):247-51

Sebaihia, M., Wren, B.W., Mullany, P., Fairweather, N.F., Minton, N., Stabler, R., Thomson, N.R., Roberts, A.P., Cerdeño-Tárraga, A. M., Wang, H., Holden, M.T.G., Wright, A., Churcher, C., Quail, M.A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., moule, S., Mungall, K., Price, C., Rabbinowitsch, E., Shapr, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B. and Parkhill, J. (2006) The multi-drug resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome, Nature Genetics 38(7):779-86

Seed, K.D., Lazinski, D.W., Calderwood, S.B. and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494:489-91

Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B and Parkhill, J. (2006) The multi-drug resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nature Genetics 38:779-86

Sekulovic, O., Meessen-Pinard, M. and Fortier, L-C. (2011) Prophage-stimulated toxin production in *Clostridium difficile* NAP1/027 lysogens. Journal of Bacteriology 193(11):2726-34

Serwer, P., Hayes, S.J., Thomas, J.A. and Hardies, S.C. (2007) Propagating the missing bacteriophages: a large bacteriophage in a new class. Virology Journal 4:21

Sell, T. L., Schaberg, D.R. and Fekety, F.R. (1983) Bacteriophage and Bacteriocin Typing Scheme for *Clostridium difficile*. Journal of Clinical Microbiology 17(6):1148-52

Shan, J., Jia, Y., Clokie, M.R.J. and Mann, N.H. (2008) Infection by the 'photosynthetic' phage S-PM2 induces increased synthesis of phycoerythrin in *Synechococcus* sp. WH7803. FEMs Microbiology Letters 283(2):154-1

Shan, J., Patel K. V., Hickenbotham, P.T., Nale, J.Y., Hargreaves, K.R. and Clokie, M.R. (2012) Prophage carriage and diversity within clinically relevant strains of *Clostridium difficile*. Applied Environmental Microbiology 78(17):6027-34.

Sharp, J. and Poxton, I.R. (1985) An immunochemical method for fingerprinting *Clostridium difficile.* Journal of Immunological Methods 83(2):241-8

Silver-Mysliwiec, T. H. and Bramucci, M.G. (1990) Bacteriophage-enhanced sporulation: comparison of spore-converting bacteriophages PMB12 and SP10. Journal of Bacteriology 172(4):1948-53

Simango, C. (2006) Prevalence of *Clostridium difficile* in the environment in a rural community. Transaction of the Royal Society of Tropical Medicine and Hygiene 100(12):1146-50

Simango, C. and Mwakurudza, S. (2008) *Clostridium difficile* in broiler chickens sold at market places in Zimbabwe and their antimicrobial susceptibility. International Journal of Food Microbiology 124(3):268-70.

Smith, J.A., Cooke, D.L., Hyde, S., Borriello, S.P. and Long, R.G. (1997) *Clostridium difficile* toxin A binding to human intestinal epithelial cells. Journal of Medical Microbiology 46(11):953-8

Smith, D.L., Rooks, D.J., Fogg, P.C.M., Darby, A.C., Thomson, N.R., McCarthy, A.J. and Allison, H.E. (2012) Comparative genomics of Shiga toxin encoding bacteriophages. BMC Genomics 13:311

Smith, D.L., Wareing, B.M., Fogg, P.C.M., Riley, L.M., Spencer, M., Cox, M.J., Saunders, J.R., McCarthy, A.J. and Allison, H.E. (2007) Multilocus Characterization Scheme for Shiga Toxin-Encoding Bacteriophages. Applied Environmental Microbiology 73(24):8032–40.

Solomon, K., Fanning, S., McDermott, S., Murray, S., Scott, L., Martin, A., Skally, M., Burns, K., Kuijper, E., Fitzpatrick, F., Fenelon, L. and Kyne, L. (2011) PCR ribotype prevalence and molecular basis of macrolide-lincosamide-streptogramin B (MLSB) and fluoroquinolone resistance in Irish clinical *Clostridium difficile* isolates. The Journal of Antimicrobial Chemotherapy 66(9):1976-82

Songer, J.G., Trinh, H.T., Killgore, G.E., Thompson, A.D., McDonald, L.C. and Limbago, B.M. (2009) *Clostridium difficile* in retail meat products, USA. Emerging Infectious Diseases. 15(5):819-21.

Sorek, R., Kunin, V. and Hugenholtz, P. (2008) CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology 6:181-86

Spigaglia, P. and Barbanti, F. (2011) Multidrug resistance in European *Clostridium difficile* clinical isolates. Journal of Antimicrobial Chemotherapy 66(10):2227-34

Stabler, R.A., Gerding, D.N., Songer, J.G., Drudy, D., Brazier, J.S., Trinh, H.T., Witney., A.A., Hind, J. and Wren, B.W. (2006) Comparative Phylogenomics of *Clostridium difficile* Reveals Clade Specificity and Microevolution of Hypervirulent Strains. Journal of Bacteriology 188(20):7297-305

Stabler, R.A., Dawson, L.F., Valiente, E., Cairns, M.D., Martin, M.J., Donahue, E.H., Riley, T.V., Songer, J.G., Kuijper, E.J., Dingle, K.E. and Wren, B.W. (2012) Macro and Micro Diversity of *Clostridium difficile* Isolates from Diverse Sources and Geographical Locations. PLoS One 7(3):e31559

Stabler, R.A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T.D., Sebaihia, M., Quail, M.A., Rose, G., Gerding, D.N., Gibert, M., Popoff, M.R., Parkhill, J., Dougan, G., Wren, B.W. (2009) Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biology 2009:10(9):102

Stark, P.L., Lee, A. and Parsonage, B.D. (1982) Colonization of the large bowel by *Clostridium difficile* in healthy infants: quantitative study. Infection and Immunity 35(3):895-899.

Stern, A., Keren, L., Wurtzel, O., Amitai, G. and Sorek, R. (2010) Self-targeting by CRISPR: gene reglation or autoimmunity? Trends in Genetics 26(8):335-40

Stern, A., Mick, E., Tirosh, I., Sagy, O. and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Research 22:1985-94.

Snyder, M.L. (1937) Further Studies on *Bacillus difficilis* (Hall and O'Toole.).The Journal of Infectious Diseases 60(2):223-31

Stoesser, N., Crook, D.W., Fung, R., Griffiths, D., Harding. R.M., Kachrimanidou, M., Keshav, S., Peto, T.E., Vaughan, A., Walker, A.S. and Dingle, K.E. (2011) Molecular Epidemiology of *Clostridium difficile* Strains in Children Compared with That of Strains Circulating in Adults with *Clostridium difficile*-Associated Infection. Journal of Clinical Microbiology 49(11):3994-6

Swithers, K.S., Petrus, A.K., Secinaro, M.A., Nesbo, C.L., Gogarten, J.P., Noll, K.M. and Butzin, N.C. (2012) Vitamin B_{12} Synthesis and Salvage Pathways Were Acquired by Horizontal Gene Transfer to the Thermotogales. Genome Biology Evolution 4(8):730-9

Sudarsan, N., Lee, E.R., Weinberg, Z., Moy, R.H., Kim, J.N., Link, K.H. and Breaker, R.R. (2008) Riboswitches in *Eubacteria* Sense the Second Messenger Cyclic Di-GMP. Science 321(5887):411-3

Sun, X., Savidge, T. and Feng, H. (2010) The Enterotoxicity of *Clostridium difficile* Toxins. Toxins 2(7):1848-80

Suttle, C.A. (2007) Marine Viruses – Major Players in the Global Ecosystem. Nature Reviews Microbiology 5:801-12

Szczepankowska, A. (2012) Bacteriophages Part A In: Advances in Virus Research. Editors Łockbocka, M. and Szybalski, W.T. Elsevier, U.K. pp. 2-396

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology Evolution 24:1596-9.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28(10):2731-9

Taori, S.K., Hall, V. and Poxton, I.R. (2010) Changes in antibiotic susceptibility and ribotypes in *Clostridium difficile* isolates from southern Scotland, 1979-2004. Journal of Medical Microbiology 59(3):338-4

Tasteyre, A., Barc, M., Collignon, A., Boureau, H. and Karjalainen, T. (2001) Role of FliC and FliD Flagellar Proteins of *Clostridium difficile* in Adherence and Gut Colonisation. Infection and Immunity 69(12):7937-40

Tasteyre, A., Karjalainen, T., Avesani, V., Delmee, M., Collignon, A., Bourlioux, P. and Barc, M. (2000) Phenotypic and Genotypic Diversity of the Flagellin Gene (*fliC*) among *Clostridium difficile* Isolates from Different Serogroups. Journal of Clinical Microbiology 38(9):3179-86

Taylor, R., Flecther, R.L. and Raven, J.A. (2001) Preliminary studies on the growth of selected 'Green tide' algae in laboratory culture: Effects of irradiance, temperature, salinity and nutrients on growth rate. Botanica Marina 44(4):327-36

Teasley, D.G., Olson, M.M., Gebhard, R.L., Gerding, D.N., Peterson, L.R., Schwartz, M.J. and Lee, J.T. Jr. (1983) Prospective Randomised Trial of Metronidazole versus Vancomycin for *Clostridium difficile*-Associated Diarrhoae and Colitis. The Lancet 322(8358):1043-6

Tenover, F.C., Äkerlund, T., Gerding, D.N., Goering, R.V., Boström, Jonsson, A-M., Wong, E., Wortman, A.T. and Persing, D.H. (2011) Comparison of Strain Typing Results for *Clostridium difficile* Isolates from North America. Journal of Clinical Microbiology 49(5):1831-7

Thean, S., Elliott, B. and Riley, T. V. (2011) *Clostridium difficile* in horse in Australia – a preliminary study. Journal of Medical Microbiology 60(8):1188-92

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J. and Chisholm, S.W. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences U.S.A. 108(39):e757-64

Touchon, M. and Rocha, E.P.C. (2010) The small, slow and specialized CRISPR and Anti-CRISPR of *Escherichia* and *Salmonella*. Plos One 5(6):e11126

Townend, I. (2007) ABPmer A Conceptual Model of Southampton Water. ABP Marine Environmental Research Ltd. Prepared for the Estuary Guide as a case study supporting document Available: http://www.estuary-guide.net/pdfs/southampton_water_case_study.pdf Accessed 13 December 2012.

Travis, J. (2012) Investigation into water company after Langstone Harbour sewage spill. In The News. Portsmouth, U.K. Johnston PressT

Twine, S. M., Reid, C. W., Aubry, A., McMullin, D.R., Fulton, K.M., Austin, J. and Loganm S.M. (2009) Motility and Flagella Glycosylation in *Clostridium difficile.* Journal of Bacteriology 191(22):7050-62

Tyson G.Q. and Banfeild, J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environmental Microbiology 10(1):200-7.

Vale, P. F. and Little, T. J. (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. Proceedings of the Royal Society Biological Sciences 277(1691):2097-103

van den Berg, R. J., Ameen, H.A., Furusawa, T., Claas, E. C., van der Vorm, E.R. and Kuijper, E.J. (2005) Coexistence of multiple PCR-ribotype strains of *Clostridium difficile* in faecal samples limits epidemiological studies. Journal of Medical Microbiology 54(2):173-9.

van der Ploeg, J. (2009). Analysis of CRISPR in *Streptococcus mutans* suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. Microbiology-SGM 155:1966-76.

van Elsas, J. D., Semeov, A.V., Costa, R. and Trevors, J.T. (2011) Survival of *Escherichia coli* in the environment: fundamental and public health aspects. The ISME Journal 5:173-83

van der Woude, M.W. and Baumler, A.J. (2004) Phage and Antigenic Variation in Bacteria. Clinical Microbiology Reviews 17(3):581-611

van der Wilk, F., Dullemans, A., Verbeek, M. and van den Heuvel, J. (1999) Isolation and characterization of APSE-1, a bacteriophage infecting the secondary endosymbiont of *Acyrthosiphon pisum*. Virology 262(1):104-13

Ventura, M., Callegari, M.L. and Morelli, L. (1999) Surface layer variations affecting phage adsorption on seven *Lactobacillus helveticus* strains. Annali di Microbiologia ed Enzimologia 49:45-53

Viau, E. and Peccia, J. (2009) Survey of Wastewater Indicators and Human Pathogen Genomes in Biosolids Produced by Class A and Class B Stabilization Treatments. Applied Environmental Microbiology 75(1):164-74

Vidal, J.E., Ma, M., Saputo, J., Garcia, J., Uzal, F.A. and McClane, B.A. (2012) Evidence that the Agr-like quorum sensing system regulates the toxin production, cytotoxicity and pathogenicity of *Clostridium perfringens* type C isolate CN3685. Molecular Microbiology 83(1):179-94

Villafane, R., Zayas, M., Gilcrease, E.B., Kropinski, A.M. and Casjens, S.R. (2008) Genomic analysis of bacteriophage ε[34] of *Salmonella enteric* serovar Anatum (15+). BMC Microbiology 8:227

Viscidi, R.P. and Bartlett, J.G. (1981) Anti-biotic pseudomembranous colitis in children. Pediatrics 67(3):381-6

Visser, M., Sepehrim, S., Olson, N., Du, T. Mulvey, M.R. and Alfa, M.J. (2012) Detection of *Clostridium difficile* in retail ground meat products in Manitoba. Canadian Journal of Infectious Diseases and Medical Microbiology 23(1):28-30

von Abercron, S.M. Karlsson, F., Wigh, G.T., Wierup, M. and Krovacek, K. (2009) Low occurrence of *Clostridium difficile* in retail ground meat in Sweden. Journal of Food Pathogens 72(8):1732-4

Voth, D.E. and Ballard, D.J. (2005) *Clostridium difficile* Toxins: Mechanism of Action and Role in Disease, Clinical Microbiology Reviews 18(2):247–63

Wagner, P.L. and Waldor, M.K. (2002) Bacteriophage control of bacterial virulence. Infection and Immunity 70(8):3985-93

Wagenaar, J. A., Van Bergen, M. A.P., Mueller, M. A., Wassenaar, T. M. and Carlton, R. M. (2005) Phage therapy reduces *Campylobacter jejuni* colonization in broilers. Veterinary Microbiology 109(3-4):275-83

Waligora, A-J., Barc, M-C., Bourlioux, P., Collignon, A. and Karjalinen, T. (1999) *Clostridium difficile* Cell Attachment in Modified by Environmental Factors. Applied Environmental Microbiology 65(9):4234-4238

Waligora, A.J., Henneguin, C., Mullany, P., Bourlioux, P., Collignon, A. and Karialainen, T. (2001) Characterization of a cell surface protein of *Clostridium difficile* with adhesive properties. Infection and Immunology 69(4):2144-53

Walker, N., Guptaa, R. and Cheesbrough, J. (2006) Blood pressure cuffs: friend or foe? Journal of Hospital Infection 63(2):167-9

Wang, I-N., Smith, D.L. and Young, R. (2000) Holins: The Protein Clocks of Bacteriophage Infections. Annual Review of Microbiology 54:799-82

Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J., Frost, E. and McDonald, L.C. (2005) Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. Lancet 366(9491):1079-84.

Warren, R.A.J. (1980) Modified bases in bacteriophage DNA. Annual Reviews Microbiology.34:137-58

Weese, J. S., Finley, R., Reid-Smith, R. R., Janecko, N. and Rousseau, J. (2010) Evaluation of *Clostridium difficile* in dogs and the household environment. Epidemiology and Infection 138(8):1100-4

Weese, J.S., Staempfli, H.R., Prescott, J.F., Kruth, S.A., Greenwood, S.J. and Weese, H.E. (2001b) The roles of *Clostridium difficile* and enterotoxigenic *Clostridium perfringens* in diarrhoea in dogs. Journal of Veterinary Internal Medicine 15(4):374-8

Weese, J.S., Avery, B.P., Rousseau, J. and Reid-Smith, R.J. (2009) Detection and Enumeration of *Clostridium difficile* Spores in Retail Beef and Pork. Applied Environmental Microbiology 75(15):5009-11

Weigele, P.R., Pope, W.H., Pedulla, M.R., Houtz, J.M., Smith, A.L., Conway, J.F., King, J., Hatfull, G.F., Lawrence, J.G. and Hendrix, R.W. (2003) Genomic and structural analysis of Syn9, a cyanophage infecting marine *Procholorococcus* and *Synechococcus.* Environmental Microbiology 9(7):1675–95

Weynberg, K.D., Allen, M.J., Ashelford, K., Scanlan, D.J. and Wilson, W.H. (2009) From small hosts come big viruses: the complete genome of a second *Osterococcus tauri* virus, OtV-1. Environmental Microbiology 11(11):2821-39

Weynberg, K.D., Allen, M.J., Gilg, I.C., Scanlan, D.J. and Wilson, W.H. (2011) Genome sequence of *Osterococcus tauri* virus OtV-2 thros light on the role of picoeukaryote niche separation in the ocean. Journal of Virology 85(9):4520-9

Wiegand, P.N., Nathwani, D., Wilcox, M.H., Stephens, J., Shelbaya, A. and Haider, S. (2012) Clinical and economic burden of *Clostridium difficile* infection in Europe: a systematic review of healthcare-facility-acquired infection. Journal of Hospital Infection 81(8):1-14

Wilcox, M. H., Fawley, W., Freeman, J. and Brayson, J. (2000) *In vitro* activity of new generation fluoroquinolones against genotypically distinct and indistinguishable *Clostridium difficile* isolates. The Journal of Antimicrobial Chemotherapy 46(4):551-6

Williamson, K.E. (2011) Soil Phage Ecology: Abundance, Distribution and Interactions with Bacterial Hosts. In: Biocommunication in Soil Microorganisms Editors Witzany, G. Springer, U.K. pp 113-36

Williamson, S.J. and Paul, J.H. (2006) Environmental factors that influence the transition from lysogenic to lytic existence in the phiHSIC/*Listonella pelagia* marine phage-host system. Microbioloigcal Ecology 52(5):217-25

Wilson, K. H., Kennedy, M. J. and Fekety, F. R. (1982) Use of sodium taurocholate to enhance spore recovery on a medium selective for *Clostridium difficile.* Journal of Clinical Microbiology 15:443–46

Withers, R.G. and Thorp, C.H. (1978) The macrobenthos inhabiting sandbanks in Langstone Harbour, Hampshire. Journal of Natural History 12(4):445-55

Wright, G.D. (2007) The antibiotic resistome: the nexus of chemical and genetic diversity. Nature Reviews Microbiology 5:175-86

Wright, A., Hawkins, C.H., Änggård, E.E. and Harper, D. (2009) A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; a preliminary report of efficacy. Clinical Otolaryngology 34(4):349-57

Wright, G.D. (2010) Antibiotic resistance in the environment: a link to the clinic? Current Opinion in Microbiology 13(5):589-94

Wommack, K.E., Bench, S.R., Bhavsar, J., Mead, D. and Hanson, T. (2009) Isolation independent methods of characterizing phage communities 2: characterizing a metagenome. Methods in Molecular Biology 502:279-89

Woodcock, A., Moradi, M., Smilie, F.E., Murray, C.S., Burnie, J.P. and Custovic, A. (2002) *Clostridium difficile*, atopy and wheeze during the first year of life. Paediatric Allergy and Immunology 13(5):357-60

Wuster, A. and Babu, M.M. (2008) Conservation and Evolutionary Dynamics of the *agr* Cell-to-Cell Communication System across *Firmicutes.* Journal of Bacteriology 109(2):743-6

Xiang, C., Chen, L., Huang, X., Luo, Y., She, Q. and Huang, L. (2005) *Sulfolobus tengchongensis* spindle-shaped virus STSV1: virus-host interactions and genomic features. Journal of Virology 79(14):8677-86

Xiao, M., Kong, F., Ping, J., Wang, Q., Xiao, K., Jeoffreys, N., James, G. and Gilbert, G.L. (2012) Comparison of Two Capillary Gel Electrophoresis Systems for *Clostridium difficile* Ribotyping, Using a Panel of Ribotype 027 Isolates and Whole-Genome Sequences as a Reference Standard. Journal of Clinical Microbiology 50(8):2755-80

Yamamotoosaki, T., Kamia, S., Sawamura, S., Kai, M. and Ozawa, A. (1994) Growth Inhibition of *Clostridium difficile* by Intestinal Microbiota of Infant Faeces in Continuous-Flow Culture. Journal of Medical Microbiology 40(3):179-87

Young, R. and Udo, Bläsi. (1995) Holins: form and function in bacteriophage lysis. FEMS Microbiology Reviews 17(1-2):191-205

Zedd, A., Sell, T., Scharberg, D., Fekety, F. and Cooperstock, M. (1984) Nosocomial *Clostridium difficile* Reservoir in a Neonatal Intensive-Care Unit. Paediatric Infectious Disease Journal 3(5):429-32

Zegans, M. E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H. and O'Toole, G.A. (2009) Interaction between bacteriophages DMS3 and host CRISPR region inhibits group behaviours of *Pseudomonas aeruginosa*. Journal of Bacteriology 191:210-9

Zeng, Q.L. and Chisholm, S.W. (2012) Marine Viruses Exploit Their Host's Two-Component Regulatory System in Response to Resource Limitation. Current Biology 22(2):124-8

Zemskov, E., Kang, W. and Maeda, S. (2000) Evidence for nucleic acid binding ability and nucleosome association of Bombyx mori nucleopolyhedrovirus BRO proteins. Journal of Virology 74:6784-9

Zhang, Y., Laing, C. Zhang, Z, Hallewell, J., Tou, C., Ziebel, K., Johnson, R.P., Kropinski, A.M., Thomas, J.E., Karmali, M. and Gannon, V.P. (2010) Lineage and host source are both correlated with levels of

Shiga toxin 2 production by *Esherichia coli* O157:H7 strains. Applied Environmental Microbiology 76(2):474-82

Zhang, X-X., Zhang, T. and Fang, H.H.P. (2008) Antibiotic resistance genes in water environment. Applied Microbiology and Biotechnology 82(3):397-414

Zidaric, V., Beigot, S., Lapaine, S. and Rupnik, M. (2010) The occurrence and high diversity of *Clostridium difficile* genotypes in rivers. Anaerobe 16(4):371-5

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In: Evolving Genes and Proteins Editors Bryson, V. and Vogel, H.J. Academic Press, New York. pp. 97-166.

Zuo, Y. and Deutscher, M.P. (1999) The DNase activity of RNase T and its application to DNA cloning. Nucleic Acids Research 27(20):4077-82

1.7. Appendices
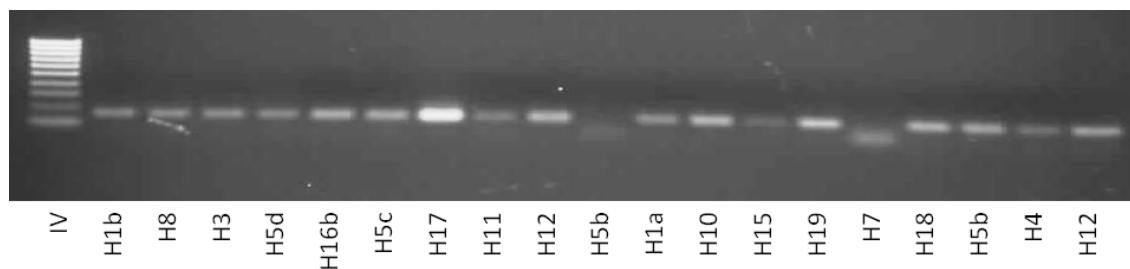
Appendix Figure 1. *C. difficile* **specific 16S rRNA gene PCR screen of Hampshire 2010 isolates.** 19 isolates were tested, with confirmation of 17 isolates as *C. difficile.* Although two isolates were negative for the 16S CD product, this acted in confirmation that these were not *C. difficile* as had been suspect colonies, despite growing in the selective conditions.
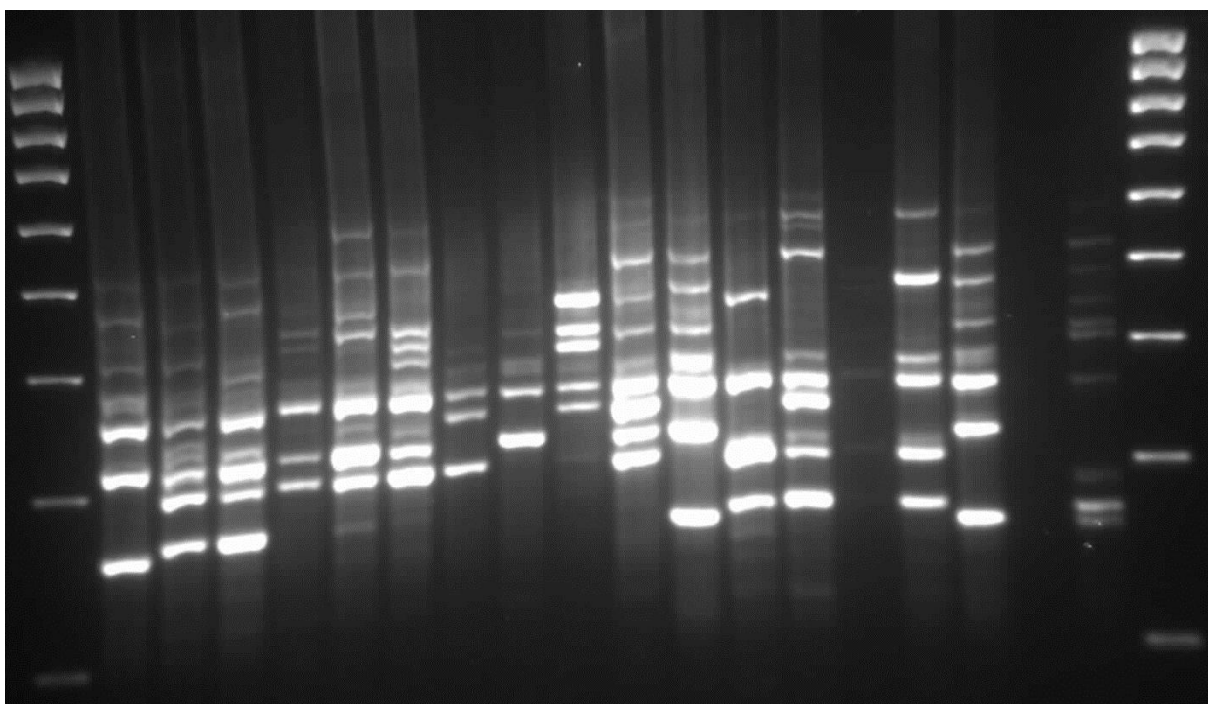
Appendix Figure 2. Ribotype band profiles for environmental *C. difficile* isolates. Lanes 1, 11 and 16 have the same band profile. Samples in lanes 14 and 17 do not have a profile, and despite using 100ng template DNA for each PCR reaction, followed by an evaporation step there remains problems with band resolution on the Hi Res 3% Agarose gel in several lanes.
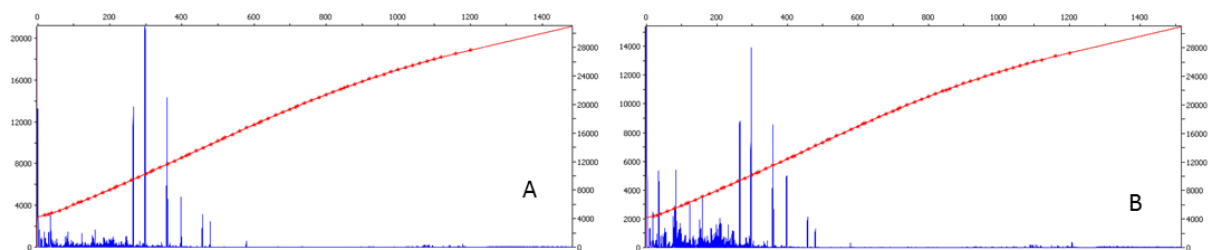
Appendix Figure 3. Capillary data profiles showing matches A: ICD02 and B: 027 ribotype. Peaks of fluorescence indicate presence of nucleotide fragments, their length as determine with a known marker run in the same capillary, with a different dye (not shown in graphs). Cut offs for analysis are fragments smaller than 250 nucleotide base pairs, and lower than 1000 fluorescence.


```
phiCDHM13  ATGGATAATAAAAAAGAAATGGTAACAATAGAGGATATTTTAAGAAGAAAAGAATATTTT 60
phiCDHM11  ATGGATAATAAAAAAGAAATGGTAACAATAGAGGATATTTTAAGAAGAAAAGAATATTTT 60
phiCDHM14  ATGGATAATAAAAAAGAAATGGTAACAATAGAGGATATTTTAAGAAGAAAAGAATATTTT 60
           ************************************************************

phiCDHM13  GCAAAGAAAAGTGAAGAAACCAAGCAATTATATATTCCTTCGCTTGATGGAAATATAGAG 120
phiCDHM11  GCAAAGAAAAGTGAAGAAACCAAGCAATTATATATTCCTTCGCTTGATGGAAATATAGAG 120
phiCDHM14  GCAAAGAAAAGTGAAGAAACCAAGCAATTATATATTCCTTCGCTTGATGGAAATATAGAG 120
           ************************************************************

phiCDHM13  ATTTCAAAGCCGGACAGGATGTTGTGTCTTGATGCAATAGAAATGGAAGACGCAGTCGAA 180
phiCDHM11  ATTTCAAAGCCGGACAGGATGTTGTGTCTTGATGCAATAGAAATGGAAGACGCAGTCGAA 180
phiCDHM14  ATTTCAAAGCCGGACAGGATGTTGTGTCTTGATGCAATAGAAATGGAAGACGCAGTCGAA 180
           ************************************************************

phiCDHM13  GGAGATAAATATTTTGTATATGAAATTGTTAAAAGTCCGAATTTAAAGAGCGAAAAATTG 240
phiCDHM11  GGAGATAAATATTTTGTATATGAAATTGTTAAAAGTCCGAATTTAAAGAGCGAAAAATTG 240
phiCDHM14  GGAGATAAATATTTTGTATATGAAATTGTTAAAAGTCCGAATTTAAAGAGCGAAAAATTG 240
           ************************************************************

phiCDHM13  CATGCTGAATTTGGATGTAAAGATAACCCACTTGATATAGTCGATG-------------- 286
phiCDHM11  CATGCTGAATTTGGATGTAAAGATAACCCACTTGATATAGTCGATGTATTAAAAGATAAC 300
phiCDHM14  CATGCTGAATTTGGATGTAAAGATAACCCACTTGATATAGTCGATG-------------- 286
           **********************************************

phiCDHM13  -------------------TATTATTTGAAGCAGGTGAAATTACTGATATTGTCAAGATT 327
phiCDHM11  CCACTTGATATAGTCGATGTATTATTTGAAGCAGGTGAAATTACTGATATTGTCAAGATT 360
phiCDHM14  -------------------TATTATTTGAAGCAGGTGAAATTACTGATATTGTCAAGATT 327
                              *****************************************

phiCDHM13  GCAACAAAGTTTGCAGGGTTTGGGGTTGTAGAGGAAATTGAAGACTTAAAAAACTAA 384
phiCDHM11  GCAACAAAGTTTGCAGGGTTTGGGGTTGTAGAGGAAATTGAAGACTTAAAAAACTAA 417
phiCDHM14  GCAACAAAGTTTGCAGGGTTTGGGGTTGTAGAGGAAATTGAAGACTTAAAAAACTAA 384
           ********************************************************
```

Appendix Figure 4. Nucleotide sequence alignments of Orf11 show an indel region. Alignments were performed using CLUSTALW. An indel region can be seen in the ORF encoded by phiCDHM11 which truncates gene by 33 bp.

```
phiCDHM14  ATGATTGTGTCATTAGAAGAAATAAAAGAATATTTGAGATTAGAGGCAGATTATAAAGAA 60
phiCDHM11  ATGATTGTGTCATTAGAAGAAATAAAAGAATATTTGAGATTAGAGGCAGATTATAAAGAA 60
phiCDHM13  ATGATTGTGTCATTAGAAGAAATAAAAGAATATTTGAGATTAGAGGCAGATTATAAAGAA 60
           ************************************************************

phiCDHM14  GATGATAATCTGCTTTTGTCTCTTCTTAAAGCAGCAGAGGAGGATTTAGAAAATCGAACA 120
phiCDHM11  GATGATAATCTGCTTTTGTCTCTTCTTAAAGCAGCAGAGGAGGATTTAGAAAATCGAACA 120
phiCDHM13  GATGATAATCTGCTTTTGTCTCTTCTTAAAGCAGCAGAGGAGGATTTAGAAAATCGAACA 120
           ************************************************************

phiCDHM14  GGAAAAGTTTTTAATGAAACTAACAAATCTGAACTTGTTAGTTTGTATGTAAAAATGTAT 180
phiCDHM11  GGAAAAGTTTTTAATGAAACTAACAAATCTGAACTTGTTAGTTTGTATGTAAAAATGTAT 180
phiCDHM13  GGAAAAGTTTTTAATGAAACTAACAAATCTGAACTTGTTAGTTTGTATGTAAAAATGTAT 180
           ************************************************************

phiCDHM14  GTTGCAGAGCAATATG-AAAAAAGAGGTGCAACAGAAAGTAATAGCGAAAAAGTTAGATT 239
phiCDHM11  GTTGCAGAGCAATATG-AAAAAAGAGGTGCAACAGAAAGTAATAGCGAAAAAGTTAGATT 239
phiCDHM13  GTTGCAGAGCAATATAGAAAAAAGAGGTGCAACAGAAAGTAA----------------- 222
           ***************. *************************

phiCDHM14  TGTTTTAGAAAGTATAATATCTCAAATTTCTATATGTAGTAGGTATTAA 288
phiCDHM11  TGTTTTAGAAAGTATAATATCTCAAATTTCTATATGTAGTAGGTATTAA 288
phiCDHM13  ------------------------------------------------
```

Appendix Figure 5. Nucleotide sequence alignments of Orf5 show a frameshift mutation due to a single point deletion. Alignments were performed using CLUSTALW. A single nucleotide deletion in the ORF encoded by phiCDHM13 which causes a frameshift mutation that results in the truncation of the gene by 66 bp.

```
phiCDHM13  GTGAAAAAGGAAGATAAAAGTTTGAATTGGGATAGTTATGAAAAATTCAATAAAGAAAAT 60
phiCDHM11  GTGAAAAAGGAAGATAAAAGTTTGAATTGGGATAGTTATGAAAAATTCAATAAAGAAAAT 60
phiCDHM14  GTGAAAAAGGAAGATAAAAGTTTGAATTGGGATAGTTATGAAAAATTCAATAAAGAAAAT 60
           ************************************************************

phiCDHM13  TTCGAAGAAATCGCTGAAAGAAATATAGATGATAGCTCATCAGCTTTTTT-AGAAAAATT 119
phiCDHM11  TTCGAAGAAATCGCTGAAAGAAATATAGATGATAGCTCATCAGCTTTTTT-AGAAAAATT 119
phiCDHM14  TTCGAAGAAATCGCTGAAAGAAATATAGATGATAGCTCATCAGCTTTTTTTAGAAAAATT 120
           ************************************************** ********

phiCDHM13  ACGTGAAGGCATTGAAAATTGTAAACAAGAATTACAAAATATAGTTGAAAATTAA 174
phiCDHM11  ACGTGAAGGCATTGAAAATTGTAAACAAGAATTACAAAATATAGTTGAAAATTAA 174
phiCDHM14  ACGTGA------------------------------------------------ 126
           ******
```

Appendix Figure 6. Nucleotide sequence alignments of Orf30 show a frameshift region resulting from a point insertion. Alignments were performed using CLUSTALW. A point insertion in the ORF encoded by phiCDHM14 which results in a frameshift mutation that causes the truncation of the sequence. As this occurs in a polyT string region it may be that this is a sequencing error rather than an existing mutation.
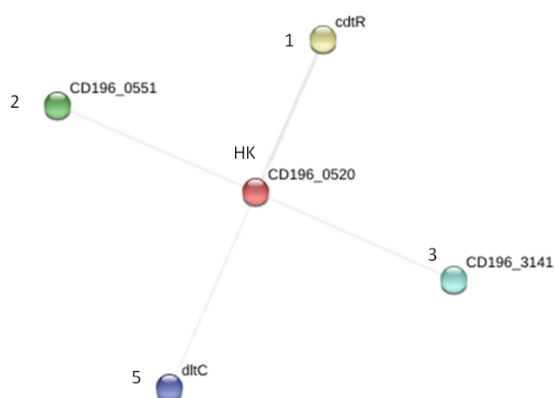
	Gene	Product	Score
1	*cdtR*	putative response regulator	0.696
2	*CD196_0551*	putative ABC transport/permease protein	0.472
3	*CD196_3141*	accessory gene regulator	0.445
4	*dltC*	D-alanine-poly(phosphoribitol) ligase subunit 2(D-alanyl carrier protein)	0.424

Appendix Figure 7. STRING prediction of functional associations of CD196_0520 a homolog of phiCDHM1 HK gene. Left: the network of interactions predicted for CD196_0520, a homolog of the QS HK (centre red circle) as generated by STRING (Jensen *et al* 2008). Right: table with the predicted genes and confidence scores (a score above 0.4 is moderate and a score of 0.6 is high).
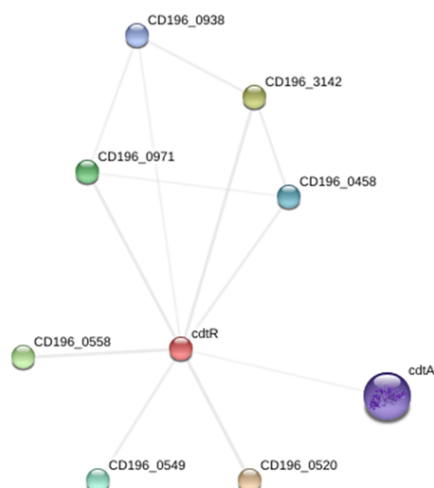
	Gene	Product	Score
1	*CD196_0520*	two-component sensor histidine kinase	0.696
2	*CD196_3142*	sensor histidine kinase VirS	0.662
3	*CD196_0558*	putative transcriptional regulator	0.652
4	*CD196_0971*	sensor histidine kinase VirS	0.650
5	*CD196_0549*	two-component sensor histidine kinase	0.567
6	*CD196_0458*	hypothetical protein	0.555
7	*CD196_0938*	hypothetical protein	0.438
8	*cdtA*	CdtA (adp-ribosyltransferase enzymatic component)	0.403

 Appendix Figure 8. STRING prediction of functional associations of CdtR, a putative response regulator. Left: the network of interactions predicted for the protein *cdtR* (centre red circle) as generated by STRING (Jensen *et al* 2008). Right: table with the predicted genes and confidence scores (a score above 0.4 is moderate and a score of 0.6 is high)

Appendix Figure 9. PCR screen of environmental *C. difficile* isolates for carriage using the internal primer set. Image of agarose gel showing screen of all Hampshire sediment isolates from 2009 and 2010 using the internal primer set, 003AR/004AR. Strains K12, H5c and H17 produced expected amplicon of ~456 bp.
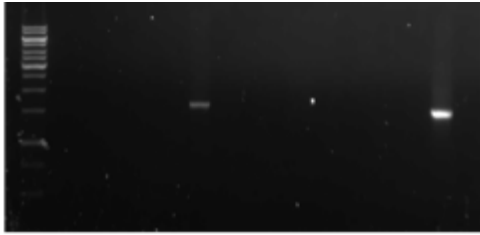
Appendix Figure 10. PCR screen of environmental *C. difficile* using the external primer set. Agarose gel pictured show screen of 14 isolates. Strain K12 and phiCDHM1 produced the expected amplicon product of 1657 bp using the external primer set, WHKF/WHKR.

Appendix Table 1. Accession numbers for genes used in phylogenetic analysis.

Strain	Loci type	AgrB	AgrD	HK/AgrC	LytTR/AgrA
QCD-66c26	1	ZP_05273396.1	ZP_05273397	ZP_05273398.1	ZP_05273399.1
CIP 107932	1	ZP_05323787.1	ZP_05323788	ZP_05323789.1	ZP_05323790.1
QCD-76w55	1	ZP_05357647.1	ZP_05357648	ZP_05357649.1	ZP_05357650.1
QCD-97b34	1	ZP_05386399.1	ZP_05386400	ZP_05386401.1	|ZP_05386402.1
QCD-37x79	1	ZP_05398745.1	ZP_05398746	ZP_05398747.1	ZP_05398748.1
CD196	1	YP_003216155.1	YP_003216156	YP_003216157.1	YP_003216158.1
R20291	1	YP_003219662.1	YP_003219663	YP_003219664.1	YP_003219665.1
QCD-63q42	1	ZP_05331470.1	ZP_05331471	ZP_05331472.1	ZP_05331473.1
QCD-32g58	1	ZP_07407976.1	ZP_07407977	ZP_07405550.1	
630	2	YP_001089263.1	YP_001089262		
QCD-63q42	2	ZP_05330899.1	ZP_05330898		
ATCC 43255	2	ZP_05351962.1	ZP_05351961		
QCD-66c26	2	ZP_05272818.1	ZP_05272817		
CIP 107932	2	ZP_05323209.1	ZP_05323208		
QCD-76w55	2	ZP_05357066.1	ZP_05357065		
QCD-97b34	2	ZP_05385822.1	ZP_05385821		
QCD-37x79	2	ZP_05398164.1	ZP_05398163		
CD196	2	YP_003215611.1	YP_003215610		
R20291	2	YP_003219119.1	YP_003219118		
QCD-32g58	2	ZP_07407451.1	ZP_07407450		
QCD-23m63	2	ZP_05402145.1	ZP_05402144		
NAP08	2	ZP_06894136.1	ZP_06894137		
NAP07	2	ZP_06901810.1	ZP_06901811		
NAP08	3	ZP_06890947.1	ZP_06890946.1	ZP_06890948.1	
NAP07	3	ZP_06904810.1	ZP_06904811.1	ZP_06904809.1	
QCD-23m63	3	ZP_05399847.1	ZP_05399846.1	ZP_05399848.1	
phiCDHM1	3	TBA	TBA	TBA	
Staphylococcus aureus subsp. aureus MRSA252	-	YP_041486.1	YP_041487.1	YP_041488.1B	YP_041489.1

Appendix Table 2. Accession numbers for genes in the phylogenetic analysis of TcdE.

Phage	Holin gene accession
phiMMP04	gb|AFO72162.1
phiMMP02	gb|AFO72094.1
phiCD27	gb|ACH91324.1
phiCD38-2	gb|AEF56897.1
phiCD6356	gb|ADK37889.1
TcdE	emb|CAC19892.1
phiCD119	gb|AAZ32274.1
phiR20291	emb|CBE04036.1
phiC2	gb|ABE99497.1
phiCD196	emb|CBA62850.1
phiCD630_1	emb|CAJ67812.1

Appendix Table 3. Three CRISPR arrays encoded on the prophage of *C. difficile* strain T6 (R076).

CRISPR id : tmp_1_Crispr_1
CRISPR start position : 13543
CRISPR end position : 13966
CRISPR length : 423
DR consensus: GTTTTAGATTAACTATATGGAATGTAAAT
DR length: 29 Number of spacers : 6

13543	**GTTTTAGATTAACTATATGGAATGTAAAT**	TAACCTTTTTATTGTTGTTGTTATTCTTATTGTTATTC	13609
13610	**GTTTTAGATTAACTATATGGAATGTAAAT**	GTTGTAAGAAGTATCATTCTATTTTTTAATCTTTCT	13674
13675	**GTTTTAGATTAACTATATGGAATGTAAAT**	AGTACATATAATGAGTCTTTAACATCAGTTATGAAAG	13740
13741	**GTTTTAGATTAACTATATGGAATGTAAAT**	GATTGTACTTTAGCGTCTGCACTAGCTTTGTCTATC	13805
13806	**GTTTTAGATTAACTATATGGAATGTAAAT**	TATTTTACAGATGAACAATTACAGTTACTTCTTGAAT	13871
13872	**GTTTTATATTAACTAAGTGGTATGTAAAT**	AATGGAAGCAAGATAATTTTTAAAGGTATGGATAATT	13937
13938	**GTTTTATATTAACTATGTGGATTCAAAAT**		13966

CRISPR id : tmp_1_Crispr_2
CRISPR start position : 15574
CRISPR end position : 15861
 CRISPR length : 287
DR consensus : TTATATTAACTATGTGGTATGTAAAG
DR length : 26 Number of spacers : 4

15574	**TTATATTAACTATGTGGTATGTAAAG**	AATGTATAAGGCATCTTCTCAAATTCTGCGTTGGTCATT	15638
15639	**TTATATTAACTATGTGGTATGTAAAG**	TGCGTTAACTGCAATAATGCTGAACTCATTTCTTGATT	15702
15703	**TTATATTAACTAAGTGGTATGTAAAT**	GTTAATTCCTTCAACCTTCTAATAGAAGATGCTTTTAGGT	15768
15769	**TTATATTAACTAAGTGGTATGTAAAT**	GTTTCAACATTTGACATGTGCTGTGGAGTAATGTTTATGTT	15835
15836	**TTATATTAACTATGTGGATTCAAAAT**		15861

CRISPR id : tmp_1_Crispr_3
CRISPR start position : 17588
CRISPR end position : 18473
CRISPR length : 885
DR consensus: GTTTTATATTAACTATGTGGTATGTAAAT
DR length : 29 Number of spacers : 13

17588	**GTTTTATATTAACTATGTGGTATGTAAAT**	CTTTTTAAAAAAGCACGTCAATTACAATACTAAAATT	17653
17654	**GTTTTATATTAACTATGTGGTATGTAAAT**	TTGAAAACTGCATGTTCTTTAACAAAAGCATCTAATT	17719
17720	**GTTTTATATTAACTATGTGGTATGTAAAT**	GTATTACTCCCAATTATGAACATAGGATATAGTCTTT	17785
17786	**GTTTTATATTAACTATGTGGTATGTAAAT**	TATAGTTTGCCATATGTTCCTGGTTCGTATGATTGGG	17851
17852	**GTTTTATATTAACTATGTGGTATGTAAAT**	AAATTTAATATTACTGGAAAAGGAGAAATGGATATCG	17917
17918	**GTTTTATATTAACTATGTGGTATGTAAAT**	TTTGTTATAAGAATATCGCCACCTTCCTTTTCTGGT	17982
17983	**GTTTTATATTAACTATGTGGTATGTAAAT**	GCACATTTAACTTCAAGTCTTAGAGTATTATCTGCAA	18048
18049	**GTTTTATATTAACTATGTGGTATGTAAAT**	TTCGTTCAACTTTTATCTTAGCTTCTCCTGCTACTT	18113
18114	**GTTTTATATTAACTATGTGGTATGTAAAT**	GTACTAGAAAATGCACTACCTGCATTCATTCCTATACG	18180
18181	**GTTTTATATTAACTATGTGGTATGTAAAT**	GTGCTATTTCCAACTTTTTTTTCAAGTTCTGAATACT	18246
18247	**GTTTTATATTAACTATGTGGTATGTAAAT**	TAGTTAGTCCCATATCGTTATGGTACTGCATTAACGC	18312
18313	**GTTTTATATTAACTAAGTGGTATGTAAAT**	AGTGTAAAAGGCATTTTTTCAAATTGTGCATTGGTTT	18378
18379	**GGTTTATATTAACTAAGTGGTATGTAAAG**	GCTAATGTTTGAGTATTTAAAAGCATCTGAGCAAACA	18444
18445	**ATTTTATATTAACTATGTGGACTTAAAAT**		18473

Appendix Table 4. Two CRISPR arrays encoded on prophages in R027 *C. difficile* strains.

CRISPR id : tmp_1_Crispr_1
CRISPR start position : 14285
CRISPR end position : 14641
 CRISPR length : 356
DR consensus : GTTTTAGATTAACTATATGGAATGTAAAT
DR length : 29 Number of spacers : 5

14285	**GTTTTAGATTAACTATATGGAATGTAAAT**	GTTGTAAGAAGTATCATTCTATTTTTTAATCTTTCT	14349
14350	**GTTTTAGATTAACTATATGGAATGTAAAT**	TTCAGTGAGAATAAGCTTTATTGTCGATGTAACACTC	14415
14416	**GTTTTAGATTAACTATATGGAATGTAAAT**	AGTACATATAATGAGTCTTTAACATCAGTTATGAAAG	14481
14482	**GTTTTAGATTAACTATATGGAATGTAAAT**	GATTGTACTTTAGCGTCTGCACTAGCTTTGTCTATC	14546
14547	**GTTTTAGATTAACTATATGGAATGTAAAT**	TATTTTACAGATGAACAATTACAGTTACTTCTTGAAT	14612
14613	**GTTTTATATTAACTATGTGGATTCAAAAT**		14641

CRISPR id : tmp_1_Crispr_2
CRISPR start position : 16506
CRISPR end position : 16796
CRISPR length : 290
DR consensus : GTTTTATATTAACTATATGGAATGTAAATC
DR length : 30 Number of spacers : 4

16506	**GTTTTATATTAACTATATGGAATGTAAATC**	TTTGTAATGGTAGTGTATTTAAGATTGAAACATCAA	16571
16572	**GTTTTATATTAACTATATGGAATGTAAATC**	AACATTAGTAGTTGTCTTTATACACATAGCATCAC	16636
16637	**GTTTTATATTAACTATATGGAATGTAAATA**	CAGCTCCCAAGACATACAACGAATCTGTAACATCAGT	16703
16704	**GTTTTATATTAACTATGTGGTATGTAAAAG**	ACTTATTTACAGCTTTATTTGCTAAATCAGAAG	16766
16767	**GTTTTATATTAACTATGTGGACTTAAAATT**		

All examined prophages from *C. difficile* strains phiCD196, phiCD20291, phi2007855, phiCIP107932, phiQCD-97b34, phiQCD-37x79, phiQCD-66c26, phiQCD-76w55 and phiBI-1 belonging to R027 contain the same CRISPR content.

Appendix Table 5. Two CRISPR arrays encoded on the two prophages of *C. difficile* strain CD630 (R012).

CRISPR id : tmp_1_Crispr_1
CRISPR start position : 12384
CRISPR end position : 12810
 CRISPR length : 426
DR consensus : GTTTTAGATTAACTATATGGAATGTAAAT
DR length : 29 Number of spacers : 6

12384	**GTTTTAGATTAACTATATGGAATGTAAAT**	GTAGAGTCTTTATATGGTAGAGGTGGAATATATAAGT	12449
12450	**GTTTTAGATTAACTATATGGAATGTAAAT**	AACTCTTCAATATCTCCTAATCGTTCAATATAATTCTC	12516
12517	**GTTTTAGATTAACTATATGGAATGTAAAT**	ACTGTATGCCATCTAAATGCATCATACAAACTTATTT	12582
12583	**GTTTTAGATTAACTATATGGAATGTAAAT**	TATACTTCCTAATGCAATCAAATAAGTACCTAAAATC	12648
12649	**GTTTTAGATTAACTATATGGAATGTAAAT**	TTCGCAACTTATGATGGTGAAATGATTACATTAACAG	12714
12715	**GTTTTATATTAACTATGTGGTATGTAAAT**	AATATAACAAAGTGGATGTTCTCTAAAAATAAAGAGG	12780
12781	**GTTTTAGATTAACTATGTGGTATGTAAAT**		12810

CRISPR id : tmp_1_Crispr_2
CRISPR start position : 12954 ----------
CRISPR end position : 13180 ----------
CRISPR length : 226
DR consensus : GTTTTAGATTAACTATATGGAATGTAAAT
DR length : 29 Number of spacers : 3

12954	**GAGTTATATTAACTAAGTGGTATGTAAAG**	GTTTGAGAACGCTGTATAAAGCTTGGTAGCTAGTTCTT	13019
13020	**GTTTTATATTAACTAAGTGGTATGTAAAG**	GTTGCATCATTAAATATTTTCCCTACAAACTCTTTTC	13085
13086	**ATTTTATATTAACTATGTGGTATGTAAAT**	AGTACATACAAAGAGTCTATAAAATTATCTATAAAGG	13151
13152	**GTTTTATATTAACTATGTGGACTTAAAAT**		13180

Appendix Table 6. Three CRISPR arrays encoded on prophage of *C. difficile* strain CF5 (R017).

CRISPR id : tmp_1_Crispr_1
CRISPR start position : 12413
CRISPR end position : 12925
CRISPR length : 512
DR consensus : GTTTTAGATTAACTATATGGAATGTAAAT
DR length : 29 Number of spacers : 7

12413	**GTTTTAGATTAACTATATGGAATGTAAAT**	CACAAAAAATATTGTTTCGGTGATAAGGTTGACCA GCTTCAATTTCTTTAGCTCAAATAATACTAGATTCTTTAGCTCA	12476
12477	**GTTTTAGATTAACTATATGGAATGTAAAT**	AATAATACTAGAAAGTG	12566
12567	**GTTTTAGATTAACTATATGGAATGTAAAT**	ACATTTAAACCTGTTGCAAAAGCATTATCTGAATATG	12632
12633	**GTTTTAGATTAACTATATGGAATGTAAAT**	TAATTTAATATTCTCTGGATTTGGTTCTCCATTTATGC	12699
12700	**GTTTTAGATTAACTATATGGAATGTAAAT**	AGAACTTTTTTATATCAGAAGTACTTGACATGATTG	12764
12765	**GTTTTAGATTAACTATATGGAATGTAAAT**	ATAGATAGTTTTTGTTCTATCTTAACTCTTAGAGGA	12829
12830	**GTTTTAGATTAACTATATGGAATGTAAAT**	GCTTTTGGGGAAAAACTAGGTGTCAGTAAAGATGTTA	12895
12896	**GTTTTAGATTAACTATATGGACTTAAAAT**		12925

CRISPR id : tmp_1_Crispr_2
CRISPR start position : 13527
CRISPR end position : 14478
CRISPR length : 951
DR consensus : GTTTTAGATTAACTATATGGAATGTAAAT
DR length : 29 Number of spacers : 14

13527	**GTTTTAGATTAACTATATGGAATGTAAAT**	TTTTTTAGTTAGTATTATTTTTGCTTCAAGGGGCAATG	13593
13594	**GTTTTAGATTAACTATATGGAATGTAAAT**	ATAAAAAATGCAAATATAATCAAAATTATCATTCCTAA	13660
13661	**GTTTTAGATTAACTATATGGAATGTAAAT**	ATAGGTTACAGCAGTAAACAGACTAACTGGAACAAATC	13727
13728	**GTTTTAGATTAACTATATGGAATGTAAAT**	TCTGATGACATGAAAGAAATACTTGATATTTTTAGT	13792
13793	**GTTTTAGATTAACTATATGGAATGTAAAT**	ATTAAAAAAGAGTTAGAAGGTGGATTTGCAAAAATTA	13858
13859	**GTTTTAGATTAACTATATGGAATGTAAAT**	AGTTGTCACAATATTATTTTTAATTTCTCCTAACTT	13923
13924	**GTTTTAGATTAACTATATGGAATGTAAAT**	TTAGGTTTACAATTTAAACATCTAAAGTCAAGTAAAA	13989
13990	**GTTTTAGATTAACTATATGGAATGTAAAG**	CCAGAAGAAGCCGCATCTGCAATAAAGAAGTATGGA	14054
14055	**GTTTTAGATTAACTATATGGAATGTAAAG**	ATTGAAAAAACGTCCAAATAGCAATATAATTTAATTA	14120
14121	**GTTTTAGATTAACTATATGGAATGTAAAG**	ATTTATGTCTGATTTGTGTCCAAAAATAACCGATTTGT	14187
14188	**GTTTTAGATTAACTATGTGGTATGTAAAG**	ACACCTCCAAACACTGCAATTTCTCCTGTTTTCAATA	14253
14254	**GTTTTAGATTAACTATGTGGTATGTAAAG**	CCTATAGGAAGCTTAACTATATGTTCATAAAAAACA	14318
14319	**GTTTTAGATTAACTATGTGGTATGTAAAT**	AATATAACAAAGTGGATGTTCTCTAAAAATAAAGAGG	14384
14385	**GTTTTAGATTAACTATATGGAATGTAAAT**	TGCTGTGATTTCACTAATAAAACAATTAATAATTGT	14449
14450	**GTTATAGATTAACTATGTGGAGAAAAACT**		14478

CRISPR id : tmp_1_Crispr_3
CRISPR start position : 16132
CRISPR end position : 16355
CRISPR length : 223
DR consensus : TTTTATATTAACTATGTGGTATGTAAAT
DR length : 28 Number of spacers : 3

16132	**AATTATATTAACTAAGTGGTATGTAAAG**	GGAGTAGGCAAGTTTTATAAACAAATAACAGGAAGA	16195
16196	**TTTTATATTAACTATGTGGTATGTAAAT**	TATATGGTGCTTTCCAGTCTTGTGGTTTTACAGGTTCG	16261
16262	**TTTTATATTAACTATGTGGTATGTAAAT**	GCGTTTAGGAATATGACAGCATCTATTAATACGACAAG	16327
16328	**TTTTATATTAACTATGTGGACTTAAAAT**		16355

Appendix Table 7. Two CRISPR arrays encoded on prophage of *C. difficile* strain M68 (R017).

CRISPR id : tmp_1_Crispr_1
CRISPR start position : 12413
CRISPR end position : 12706
 CRISPR length : 293
DR consensus : GTTTTAGATTAACTATATGGAATGTAAA
DR length : 28 Number of spacers : 4

12413	**GTTTTAGATTAACTATATGGAATGTAAA**	TTTATCTGCGTCTTCTGAAATAGCATACTCTTCAAGAG	12478
12479	**GTTTTAGATTAACTATATGGAATGTAAA**	TGTTGTAAGAAGTATCATTCTATTTTTTAATCTTTCT	12543
12544	**GTTTTAGATTAACTATATGGAATGTAAA**	TCTTATTTATAGTTGGGAAAATGGTTATAAATGTGGTAC	12610
12611	**GTTTTAGATTAACTATATGGAATGTAAA**	GCAACAAGAACAACAGTTGATTTTACCTTTTCTAGCCAAC	12678
12679	**GTTTTATATTAACTATGTGGAGAAAAAC**		12706

CRISPR id : tmp_1_Crispr_2
CRISPR start position : 15308
CRISPR end position : 15532
CRISPR length : 224
DR consensus : TTTTATATTAACTATGTGGTATGTAAAT
DR length : 28 Number of spacers : 3

15308	**AATTATATTAACAATGTGGTATGTAAAT**	AATGAATTTTGAGAATCTACATAGGATTGAACTACTAA	15373
15374	**TTTTATATTAACTATGTGGTATGTAAAT**	ACTAAATTAGAGGTCGAGAAGGCTATTGAATCAAAAGG	15439
15440	**TTTTATATTAACTATGTGGTATGTAAAT**	TACACATTTTCTCAAATTTGGACTTCAATTAAAAATG	15504
15505	**TTTTATATTAACTATGTGGACTTAAAAT**		15532

Appendix Table 8. Blastn score of *C. difficile* phages in pairwise analysis.

Phage pairwise relative genome relatedness score

	phiCF5	phiM68	phiR20291	phiCD196	phiT6	phiCD630_1	phiCD630_2	phiCDHM1	phiC2	phiCDM19	phiCD119	phiCD27	phiMMP02	phiCDHM13	phiMMP04	phiCD38-2	phiCD6356
phiCF5	107.6	94.1	92.3	94.5	84.4	54	53.4	49.1	55,5	14.7	14.5	17.3	13.6	9.3	6.1	6.4	6.2
phiM68	99.1	104.1	84.7	81.1	74.4	49.3	48.9	48.4	55.3	17.3	14.6	17.5	14	7.8	6	5.6	6.3
phiR20291	92.1	80.3	109.5	104.4	72.2	43.9	43.2	45.1	49.1	13.1	12.1	15.1	13.1	7.2	6.8	5.4	7.7
phiCD196	91.3	74.3	101	108.8	69.1	44.4	43.2	44.6	47.4	13.2	10.5	13.5	12.5	9.6	6.7	5.5	7.9
phiT6	84.1	71.8	72	71.2	117.1	52.3	52.4	47.7	53.8	19.9	19.3	18.4	16.6	9.9	6.6	5.4	6.1
phiCD630_1	56.5	48.9	46	48.1	54.9	104.7	81.2	66.3	59.1	27.6	44.3	38.7	19.8	9	6.5	5	6.4
phiCD630_2	57.7	50.2	46.7	48.3	56.8	83.8	105.1	63.2	51.7	26.4	37	32.4	20.7	8.2	6.3	0.5	7.2
phiCDHM1	47.3	44.3	43.6	44.5	46.2	61	56.4	106.2	61.5	36.5	35.8	25.9	27.5	11.6	9.4	5.4	7.5
phiC2	51.4	48.6	45.5	45.4	50	52.3	60.4	59	110	42.2	34.2	38.6	28.7	9.2	7	5.5	8.4
phiCDHM19	14.3	16	12.7	13.3	19.4	25.6	23.7	36.7	44.2	106.6	85	34.6	29	15.3	15.2	6.3	7.1
phiCD119	14.2	13.6	11.9	10.7	19	41.6	33.6	36.4	36.3	86	107.5	23.8	27.7	17.6	12.5	5.4	5.9
phiCD27	17.8	17.1	15.6	14.4	19	38	30.1	27.6	42.9	36.6	24.9	105.9	58.4	10.3	6.5	5.6	7.7
phiMMP02	14.7	14.4	14.2	14	18	20.5	20.7	30.8	33.6	32.3	30.5	61.5	103.3	10.2	7	5.4	8.4
phiCDHM13	14.4	11.5	11.1	15.3	15.3	13.2	11.6	18.6	15.3	24.4	27.6	15.4	14.6	102.4	77.9	21.3	12.8
phiMMP04	10.2	9.4	11.2	11.5	10.9	10.2	9.7	16.1	12.5	25.9	21.1	10.5	10.8	83.5	100.1	16.7	15
phiCD38-2	8.2	8.4	6.9	7.2	6.9	6.1	6.1	7.1	7.6	8.3	7	7	6.4	17.6	12.9	102.8	12.3
phiCD6356	8.6	8.4	10.7	11.4	8.5	8.5	9.2	10.8	12.6	10.2	8.3	10.4	10.8	11.5	17.9	13.4	100.7

Appendix Table 9. Number of shared protospacers between *C. difficile* phages.

Number of shared protospacers between phage genomes

	phiCF5	phiM68	phiR20291	phiCD196	phiT6	phiCD630_1	phiCD630_2	phiCDHM1	phiC2	phiCDHM19	phiCD119	phiCD27	phiMMP02	phiCDHM13	phiMMP04	phiCD38-2	phiCD6356
phiCF5	14	12	8	10	8	6	5	7	7	2	0	3	2	2	2	2	0
phiM68	12	14	7	9	8	6	6	7	6	2	0	2	2	2	1	1	0
phiR20291	8	7	10	10	7	4	4	5	6	0	0	2	1	2	1	1	0
phiCD196	10	9	10	12	7	4	4	5	6	0	0	2	1	2	1	1	0
phiT6	8	8	7	7	13	6	6	5	5	1	1	2	2	3	1	1	1
phiCD630_1	6	6	4	4	6	16	3	7	7	2	5	0	3	1	0	0	0
phiCD630_2	5	6	4	4	6	3	16	6	5	1	11	2	2	1	0	0	0
phiCDHM1	7	7	5	5	5	7	6	14	8	2	4	2	4	2	1	0	1
phiC2	7	6	6	6	5	7	5	8	18	4	4	4	0	2	0	0	1
phiCDHM19	2	2	0	0	1	2	1	2	4	20	12	3	3	4	1	1	1
phiCD119	0	0	0	0	1	5	11	4	4	12	18	2	4	3	0	0	0
phiCD27	3	2	2	2	2	0	2	2	4	3	2	18	5	2	4	1	0
phiMMP02	2	2	1	1	2	3	2	4	0	3	4	5	12	1	1	1	0
phiCDHM13	2	2	2	2	3	1	1	2	2	4	3	2	1	29	16	3	1
phiMMP04	2	1	1	1	1	0	0	1	0	1	0	4	1	16	25	4	1
phiCD38-2	2	1	1	1	1	0	0	0	0	1	0	1	1	3	4	16	1
phiCD6356	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	26

Appendix Table 10. Strains used in this project.

Isolate	Strain	Origin
K10	CD105HS1	This study, Hampshire sediment
K3	CD105HS2	This study, Hampshire sediment
K4	CD105HS3	This study, Hampshire sediment
K6	CD105HS4	This study, Hampshire sediment
K9	CD105HS5	This study, Hampshire sediment
K12	CD105HS6	This study, Hampshire sediment
K14	CD105HS7	This study, Hampshire sediment
K15	CD105HS8	This study, Hampshire sediment
K16	CD105HS9	This study, Hampshire sediment
K18	CD105HS10	This study, Hampshire sediment
K20	CD105HS11	This study, Hampshire sediment
K25	CD105HS12	This study, Hampshire sediment
H9	CD105SH13	This study, Hampshire sediment
H1b	CD105HS14	This study, Hampshire sediment
H3	CD105HS15	This study, Hampshire sediment
H5d	CD105HS16	This study, Hampshire sediment
H15	CD105HS17	This study, Hampshire sediment
H11	CD105HS18	This study, Hampshire sediment
H1a	CD105HS19	This study, Hampshire sediment
H5b	CD105HS20	This study, Hampshire sediment
K25	CD105HS22	This study, Hampshire sediment
H8	CD105HS23	This study, Hampshire sediment
H4	CD105HS24	This study, Hampshire sediment
H18	CD105HS25	This study, Hampshire sediment
H17	CD105HS26	This study, Hampshire sediment
H5c	CD105HS27	This study, Hampshire sediment
H16	CD105HS28	This study, Hampshire sediment
H14	CD105HS29	This study, Hampshire sediment
OA	CD105LI01	This study, Infant
HB	CD105LI02	This study, Infant
SS1	CD105LI03	This study, Infant
B1	CD105LI04	This study, Infant
B2	CD105LI05	This study, Infant
R8	CD105LI06	This study, Infant
R10	CD105LI07	This study, Infant
RB1	CD105HS30	This study, Hampshire sediment
S1	ZZV09-562	Maja Rupnik, Slovenia
S2	ZZV09-633	Maja Rupnik, Slovenia
S3	ZZV09-634	Maja Rupnik, Slovenia
S4	ZZV09-646	Maja Rupnik, Slovenia
S5	ZZV09-1065	Maja Rupnik, Slovenia
S6	ZZV09-1623	Maja Rupnik, Slovenia
S7	ZZV09-1622	Maja Rupnik, Slovenia
S8	ZZV09-1647	Maja Rupnik, Slovenia
S9	ZZV09-1742	Maja Rupnik, Slovenia
S10	ZZV09-1789	Maja Rupnik, Slovenia
S11	ZZV09-1745	Maja Rupnik, Slovenia
S12	ZZV09-1653	Maja Rupnik, Slovenia

Appendix Table 11. Oligonucleotides used in this project.

Pair	Name	Forward Primer	Target	Expected product	Reference
1	g-cdif-F	5'-TTGAGCGATTTACTTCGGTAAAGA-3'	*C. difficile* specific 15S rRNA gene	157	Rintilla *et al* 2006
	g-cdif-R	5'-CCATCCTGTACTGGCTCACCT-3'			
2	8F	5'-AGAGTT TGATCCTGGCTCAG-3'	Universal bacterial primers 16S rRNA	1800	Edwards *et al* 1989 and Lane *et al* 1986
	1391R	5'-GACGGGCGGTGTGTRCA-3'			
4		5'-GTGCGGCTGGATCACCTCCT-3'	Intergenic regions between 16S rRNA gene and 23S rRNA gene	-	Bidet *et al* 2000
		5'-CCCTGCACCCTTAATAACTTGACC-3'			
5	NK2	5'-CCCAATAGAAGATTCAATATTAAGCTT-3'	partial *tcdA*	251	Kato *et al* 1991 and 1998
	NK3	5'-GGAAGAAAAGAACTTCTGGCTCACTCAGGT-3'			
6	NK9	NK9 (5'-CCACCAGCTGCAGCCATA-3'	repeat region of *tcdA*	1265	Kato *et al* 1991 and 1998
	NK11	5'-TGATGCTAATAATGAATCTAAAATGGTAAC-3'			
7	NK104	5'-GTGTAGCAATGAAAGTCCAAGTTTACGC-3'	*tcdB*	203	Kato *et al* 1991 and 1998
	NK105	5'-CACTTAGCTCTTTGATTGCTGCACCT-3'			
8	cdtAF	5'-TGAACCTGGAAAAGGTGATG-3'	*cdtA*	375	Stubbs *et al* 2000
	cdtAR	5'-AGGATTATTTACTGGACCATTTG-3'			
9	cdtBF	5'-CTTAATGCAAGTAAATACTGAG-3'	*cdtB*	510	Stubbs *et al* 2000
	cdtBR	5'-AACGGATCTCTTGCTTCAGTC-3'			
10	003AR	5'-TCACAAGCCTCAATTGCATTA-3'	internal section of histidine kinase phiCDHM1	456	This study
	004AR	5'-TGGCATTATTGTTAACAGCATCA-3'			
11	003BF	5'-TTTGATATGAACAATGAAAATGAACA-3'	NTPase internal section phiCDHM1	689	This study
	004BF	5'-TCCATATACTCATCGGAATTTTCA-3'			
12	WHKF	5'-AGGATTTGTAATCCATAGGAACAT-3'	histidine kinase of phiCDHM1	1657	This study
	WHKR	5'-TTTTCGTTCGTTTTATTATTACAGTTT-3'			
13	NTPaseF	5'-CGCAAGTTACTGAAAAACTCCA-3'	NTPase of phiCDHM1	840	This study
	NTPaseR	5'-TTTCTCCCAATTTTTACACTGTTGA-3'			
14	K12groupF	5'-AACTTCGGGGATTTGTATGC-3	MM phage capsid	814	This study
	K12groupR	5'-CAACAAATTGTATTGCATCAGC-3			
15	K6groupF	5'-TGGTTGGATGGATTCTAATGCT-3'	MM phage capsid	771	This study
	K6groupR	5'-GACCAAGCATTTGCTGTTTG-3'			
16	CD27groupF	5'-GAGGGCAGGAATAAGAAAAGC-3'	LTM specific structural protein	711	This study
	CD27groupR	5'-GATTCCCTATCCTCAACTACGC-3'			
17	K7groupF	5'-GAGCGGAAGTTCAAACAAGC-3'	SMV capsid protein	707	This study
	K7groupR	5'-AGCAAGAATCTCGCCATCTG-3'			