Two Novel Myoviruses from the North of Iraq Reveal Insights into *Clostridium difficile* Phage Diversity and Biology

Srwa J. Rashid [1,†], Jakub Barylski [2,†], Katherine R. Hargreaves [3], Andrew A. Millard [4], Gurinder K. Vinner [5] and Martha R. J. Clokie [1,*]

[1] Department of Infection, Immunity and Inflammation, Medical Sciences Building, University of Leicester, University Road, Leicester, LE1 9HN, UK; saroa.rashid@gmail.com

[2] Department of Molecular Virology, Faculty of Biology, Adam Mickiewicz University in Poznan, Poznan 61-712, Poland; erytropoeta@gmail.com

[3] Department of Microbiology, The Ohio State University, Columbus, OH 43201, USA; krhargr@gmail.com

[4] Microbiology & Infection Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK; andrew.millard@warwick.ac.uk

[5] Department of Chemical Engineering, Loughborough University, Loughborough, LE11 3TU, UK; G.Vinner@lboro.ac.uk

[*] Correspondence: mrjc1@le.ac.uk; Tel.: +44-116-252-2959

[†] These authors contributed equally to this work.

Abstract: Bacteriophages (phages) are increasingly being explored as therapeutic agents to combat bacterial diseases, including *Clostridium difficile* infections. Therapeutic phages need to be able to efficiently target and kill a wide range of clinically relevant strains. While many phage groups have yet to be investigated in detail, those with new and useful properties can potentially be identified when phages from newly studied geographies are characterised. Here, we report the isolation of *C. difficile* phages from soil samples from the north of Iraq. Two myoviruses, CDKM15 and CDKM9, were selected for detailed sequence analysis on the basis of their broad and potentially useful host range. CDKM9 infects 25/80 strains from 12/20 *C. difficile* ribotypes, and CDKM15 infects 20/80 strains from 9/20 ribotypes. Both phages can infect the clinically relevant ribotypes R027 and R001. Phylogenetic analysis based on whole genome sequencing revealed that the phages are genetically distinct from each other but closely related to other long-tailed myoviruses. A comparative genomic analysis revealed key differences in the genes predicted to encode for proteins involved in bacterial infection. Notably, CDKM15 carries a clustered regularly interspaced short palindromic repeat (CRISPR) array with spacers that are homologous to sequences in the CDKM9 genome and of phages from diverse localities. The findings presented suggest a possible shared evolutionary past for these phages and provides evidence of their widespread dispersal.

Keywords: bacteriophage; *Clostridium difficile*; phylogenetic analysis; CRISPR/Cas system; genome evolution; endolysin; large terminase gene

1. Introduction

Clostridium difficile is a Gram-positive, spore-forming, anaerobic bacterium that can cause infection (termed *C. difficile* infection: CDI) and, in severe cases, lead to pseudomembranous colitis [1,2]. CDI usually occurs as a result of antibiotic treatment that decreases the diversity of the intestinal microbiota, allowing *C. difficile* to proliferate and cause infection [3]. In the last 15 years the

epidemiology of CDI has changed considerably due to the rapid emergence of hypervirulent strains, which has raised concerns over its evolution, pathogenicity and antibiotic resistance [4,5].

The natural resistance of *C. difficile* to multiple antibiotics has prompted researchers to investigate phage therapy to combat CDI [6–8]. Several *C. difficile* phages from both the *Myoviridae* and *Siphoviridae* families have already been characterised, including by genome sequencing [9–17]. These phages are currently classified into two genera, phicd119virus [18] and the proposed phiMMP04virus [19] (both in the *Myoviridae* family), however there are other *C. difficile* phages that do not fall into either genera [13,16,17]. Previous attempts to classify *C. difficile* myoviruses suggested that they can be grouped according to their particle morphology (by tail length and capsid diameter). This groups the long tailed myoviruses, medium myoviruses and small myoviruses, where each group has a shared gene content and genome architecture [11,19].

The aim of this work was to expand our knowledge of existing *C. difficile* phage diversity by isolating and characterising phages from the little studied soil and sediment ecosystems of the north of Iraq. Two myoviruses, CDKM9 and CDKM15, were isolated and characterised according to their host range and genome sequences. They both infect several clinically relevant strains, and were selected for further investigation as candidate therapeutic phages from a larger set of viruses based on their host range. Their genomes were analysed with known *C. difficile* phages to determine their overall relatedness and to inform their potential inclusion within phage cocktails developed for therapeutic use.

2. Materials and Methods

2.1. Phage Isolation

To isolate phages, soil and sediment samples were taken from sites across mountains and river banks in the north of Iraq. The samples were collected in March 2013 from a depth of 10–20 cm, and stored at 4 °C. The method for phage isolation was followed as described previously [20]. Briefly, the samples were suspended in 10 mL of fastidious anaerobic broth (FA: Bioconnections, Leeds, UK) supplemented with 250 $\mu g \cdot mL^{-1}$ cycloserine and 8 $\mu g \cdot mL^{-1}$ cefoxitin (as selective agents) (Bioconnections, Leeds, UK) and 0.1% sodium taurocholate (Sigma-Aldrich, Dorset, UK) for spore germination. These enrichment cultures were incubated for 10 days in a MiniMACS anaerobic chamber (Don Whitley Scientific, West Yorkshire, UK) at 37 °C under anaerobic conditions (10% H_2, 10% CO_2 and 80% N_2). Following incubation, the cultures were centrifuged for 10 min at 3398 × *g*. The supernatants were filtered through 0.22 μm filters and the phages were isolated using 15 *C. difficile* indicator strains (Table S1). Phages were purifed using standard plaque assays after resuspension in Oxoid Brain Heart Infusion broth (BHI; Oxoid Ltd, Basingstoke, UK) and stored in SM buffer (0.1 M NaCl, 1 mM $MgSO_4$, 0.2 M Tris-HCl, pH 7.5) [21] 50% glycerol (*v/v*) at −80 °C.

The morphology was determined by transmission electron microscopy (performed by Stefan Hyman and Natalie Allcock, Core Biotechnology Services, University of Leicester, UK). Briefly, samples were placed on individual glow discharged pioloform/carbon coated copper grids (Athene type 3 mm: Agar Scientific Ltd, Stansted, UK). Samples were negatively stained with 0.1% uranyl acetate and examined with a a JEOL JEM-1400 electron microscope (JEOL UK Ltd, Welwyn Garden, UK) with an accelerating voltage of 80 kV. Digital images were captured using an SIS Mega view III Digital camera with associated analysis software (Olympus Soft Imaging Solutions, Muenster, Germany).

2.2. Phage Host Range Assay

A standard spot test method was used to determine phage host range using phage stocks of 10^8 plaque-forming units (PFU) mL^{-1} as follows. 250 μL of overnight culture of each of the bacterial strains was mixed with 3 mL of BHI 0.5% agar supplemented with salt solution (0.4 M $MgCl_2$ and 0.01 M $CaCl_2$) and then poured onto a BHI 1% agar plate [22]. When the agar solidified, 10 μL undiluted drops of each phage were spotted onto its surface, performed in triplicate. Spots were

inspected for lysis (a clearing of the bacteria) after 24 h incubation at 37 °C under anaerobic conditions. The spot tests were performed in triplicate.

2.3. Purification of Phage Genomic DNA

The two phages, CDKM15 and CDKM9, were isolated on *C. difficile* strain CD105HE1 (R076, equine isolate) [21], and propagated in a liquid culture to obtain a high titre phage stock (10^9 PFU mL^{-1}). Phage genomic DNA was extracted for sequencing using standard phenol chloroform extraction and isopropanol precipitation methods [23] with modifications as follows. The crude lysate of 10^9 PFU mL^{-1} was centrifuged and filtered through a 0.22 μm filter, treated with 1.4 μg·μL^{-1} DNase, 3 μg·μL^{-1} RNase (Sigma-AldrichTh) and 12.5 μL 1 M MgCl$_2$ (Acros Organics, Morris Plains, NJ, USA), and incubated overnight at 37 °C. Proteinase K (Fisher Scientific UK Ltd, Loughborough, UK), ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich, Dorset, UK) and sodium dodecyl sulfate (SDS) (Sigma-Aldrich) were added to a final concentration of 0.5 mg·mL^{-1}, 20 mM and 0.5%, respectively. This was incubated at 55 °C for one hour. To obtain purified DNA, three rounds of phenol:chloroform:isoamyl alcohol (25:24:1) purifications were performed. The resulting fraction was treated with 0.3 M sodium acetate (Fisher Scientific UK Ltd) and two volumes of ice-cold 95% ethanol to precipitate the DNA, followed by a 10 min incubation on ice. The DNA was pelleted by centrifugation at 21,000× *g* for 20 min and the pellet was washed once with 0.5 mL of 70% ethanol before resuspension in an elution buffer (5 mM TrisCl, pH 8.5). DNA quantity and quality were measured using a Nanodrop 2000 and Qubit Fluorimeter (Thermo Scientific, Loughborough, UK) as described here [6].

2.4. Phage Genome Sequencing

The genome of CDKM15 was sequenced at Beijing Genomics Institute (BGI; Shenzhen, China). A paired-end library was prepared using 3 μg of DNA with an insert size of 170 bp and sequenced using an Illumina HiSeq 2000 (San Diego, CA, USA).

CDKM9 was sequenced at Warwick University, UK. An amount of 1 ng of input DNA was used to prepare a paired-end library using an Illumina Nextera XT DNA sample kit according to the manufacturer's protocol. Sequencing was performed on an Illumina MiSeq using the paired-end 2 × 250 bp protocol (version 2, 500 cycles).

Sequencing read quality was checked with FastQC version 0.11.3 [24] and reads were trimmed using sickle [25]. The reads were assembled using SOAP denovo 2.04 [26,27], SPades 3.1 [28] and Geneious 9.0.5 [29]. Phage sequences assembled into a single contig/gapless scaffold each time, except one case using SOAP which generated two contigs for CDKM9. Independent assemblies were compared and their quality was assessed by mapping reads back to each contig in Geneious using its read mapping algorithm (using "medium" and "medium–low" setting packages). Uncertain or ambiguous regions were resolved by manual inspection of the read mapping and, if needed, by PCR amplification and Sanger sequencing (carried out at GATC Biotech Ltd., London, UK). For tool settings and assembly statistics, see Table S2.

Protein coding genes were predicted using GeneMarkS, GeneMark.hmm [30], Glimmer 3 [31], RAST [32], FGENESV (Softberry, Inc., Mount Kisco, NY, USA) and Prodigal 1.20 [33]. Coding DNA sequences (CDSs) with no overlapping BLASTx hits (against the National Center for Biotechnology Information (NCBI) non-redundant (nr) database) and predicted by only a single tool were discarded. Conflicting start codons were resolved based on BLAST alignment and ribosome binding site (RBS) positions (located by scanning the whole genome with find individual motif occurrences (FIMO) tool using consensus RBS motif found in 5′ untranslated regions (UTRs) of uncontested genes with Multiple Em for Motif Elucidation (MEME) [34] (see Table S2). Results of the BLASTx analyses were also used for the functional annotation of CDSs (we manually assessed the top 50 hits against nr and RefSeq databases to find the most probable function of each protein). Predicted CDS were translated and their initial annotation was re-assessed using BLASTp, InterProScan 5 and CD-Search [35–37]. tRNA genes

were predicted by tRNAscan-SE version 1.21 and other non-coding RNAs by Infernal 1.1.1 [38,39]. Clustered regularly interspaced short palindromic repeats (CRISPR) arrays located by Infernal were confirmed by PILER-CR 1.06 and CRISPRFinder [40,41]. CRISPRTarget was used to identify matches to the spacers in the array (match reward +1, mismatch penalty −5, minimum score 25) and identify protospacer adjacent motifs (PAMs) [42].

Each genome was oriented to start at the terminase small subunit gene to be consistent with previously sequenced *C. difficile* phage genomes [9,10,12] and deposited in GenBank under the accession numbers KX228399 (CDKM9) and KX228400 (CDKM15).

2.5. Phylogenetic Analyses and Comparative Genomics

To determine the taxonomic relatedness of CDKM15 and CDKM9, genome comparisons were performed using Gegenees 2.2.1 [43] to other phages infecting *C. difficile* (reference phage genomes used are listed in Table S3). Gegenees calculates global similarity between pairs of sequences based on BLAST local alignments (we used both the BLASTn and BLASTx method with a fragment size of 200, sliding window size of 100). The resulting BLASTx similarity matrix was used to construct BioNJ phylograms with SplitsTree 4.13.1 [44].

A phylogenetic tree was generated using maximum likelihood (ML) analysis of the endolysin genes at the amino acid level to determine whether phage genes involved in host cell lysis share the same evolutionary history of the phage genomes overall. Homologous endolysin sequences were retrieved from GenBank using BLASTp and scanned for relevant domains using InterProScan in Geneious (Table S3) [29]. Verified sequences were aligned using ClustalW in Geneious, evolution models were selected with ProtTest 3.2.1 and ML analysis performed using PhyML 3.0 [45].

To predict the packaging strategy of analysed phages, we followed the method proposed by Casjens and Gilcrease [46]. The terminase large subunit genes from 24 *C. difficile* phages and 68 reference phages (Table S3) were aligned at the amino acid level using ClustalW in Geneious. FastTree 2.1.7 was used to generate approximate ML tree, run with the Whelan and Goldman (WAG) substitution model (selected using of ProtTest 3.4) and the computation of gamma likelihoods enabled and the Shimodaira–Hasegawa test used to calculate support values for the nodes [47]. This method is much faster than classic ML with only a negligible loss of topological accuracy [48]. Trees were visualised in Geneious.

Genome comparisons were performed using Blast Ring Image Generator (BRIG) v.0.95 [49] and EasyFig 2.2.2 [50] which displays results of BLASTn-based sequence comparisons based on pairwise similarity of the matches in circular and linear representations, respectively.

2.6. Protein Analysis

Protein cluster analysis was performed to determine the fraction of shared proteins between CDKM9, CDKM15 and 22 *C. difficile* phages. Protein clusters were created using CD-HIT [51] and protein cluster statistics were generated in Microsoft Excel.

3. Results

3.1. Phage Isolates and Host Range Analysis

To explore the biology of *C. difficile* phages in the north of Iraq, soil and sediment samples were collected from different sites which were then used for phage isolation using 15 indicator strains from four PCR ribotypes: R027, R078, R010 and R076 (Table S1). Fourteen phages were isolated and belonged to either the *Siphoviridae* (12/14) or *Myoviridae* (2/14) based on particle morphology as observed using transmission electron microscopy (TEM). All the *Myoviridae* displayed the characteristically long tails of the previously categorised long tailed myoviruses [29] (see Figure 1 for example). The phage host ranges were tested on 80 *C. difficile* isolates (Figure 2, Table S4). The panel of strains represented 20 ribotypes and originated from four countries: Kurdistan in the north of Iraq (*n* = 22), the UK

(n = 55), France (n = 1), Switzerland (n = 1) and USA (n = 1). The strain panel included clinical isolates (n = 31), environmental isolates (n = 46), asymptomatic infant isolates (n = 2) and a single bovine isolate. Two myoviruses, CDKM15 and CDKM9 (Figure 1), were selected for further characterisation as they had the broadest host ranges from the new set of phages and, in particular, could lyse *C. difficile* isolates from the ribotype R027.
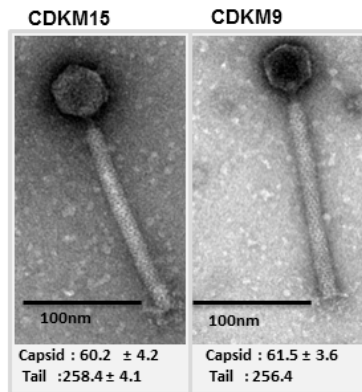

Figure 1. Transmission electron micrographs of phages CDKM15 and CDKM9. The sizes of the capsids and tails were measured for four particles and the mean values in nm with the standard deviations are reported below each phage.

Ribotypes	Strains	CDKM9	CDKM15	Ribotypes	Strains	CDKM9	CDKM15
R001	CD105HS23			R027	17L		
	CD105HS25				7L		
	CD105HS24				14L		
	CDNCTC11204				28L		
	R4001				22L		
	CD105KSE4				15L		
	F7			R031	CD105HS35		
	CD105KSE3			R035	CD105KSE11		
	F5				CD105KSE5		
	F2			R076	CD105HE1		
	F1			R078	CD105HS27		
R002	CD105HS17				CD105HS26		
R005	CD105HS20				CD105HS44		
	CD105HS10				CD105LC22		
R010	CD105HS15				CD105LC6		
	CD105HS16				CD105LC23		
	CD105HS9				CD105LC26		
	CD105LI07				CD105LC9		
	CD105HS33				M120		
	CD105HS14				CD105LC24		
	CD105KSE9				CD105LC25		
	M7			R085	R8		
R011	F9			R091	CD105KSE1		
	CD105KSE6				CD105KSE12		
R012	CD630				CD105KSE13		
	CD105HS1				CD105KSE14		
R014	CD105HS4				CD105KSE2		
	CD105HS42				CD105KSO10		
R015	CD105HS39			R106	CD106		
R014/020	CD105LC27				R40V0106		
	CD105LC278				R6106		
R027	CD105LC1				CD105HS46		
	CD017				R39V106		
	CD105HS8			R220	CD105HS22		
	CD105LI02				CD105HS6		
	BI-9				CD105HS12		
	CD0027			R604	CD105KSO7		
	CD196				CD105KSO8		
	R20291				CD105KSO15		
	16L			R691	CD105KSE16		

Figure 2. Heat map representation of lysis profiles of CDKM9 and CDKM15 for 80 *Clostridium difficile* strains. Colours indicate different phage infection parameters observed in the spot tests: red is clearing of the lawn, orange is clearing with turbidity and yellow was no clearing.

The host range analysis showed that CDMK15 infected 20/80 strains (25% tested) from 9/20 ribotypes, and CDMK9 infected 25/80 strains (31% tested) from 12/20 ribotypes. Also, a broad phage host range overlap was evident, as 16 isolates were infected by both CDKM9 and CKM15, which

were represented in nine ribotypes. Notably, isolates belonging to R015, R031 and R035 were infected exclusively by CDKM9 (Figure 2). Overall, both phages infected strains from different environments and different sources. Furthermore, both phages infected strains isolated from Kurdistan, the 'local isolates'. Of particular interest, due to their potential clinical utility, the phage CDKM9 infected four strains from R001 and R010, and CDKM15 infected two strains of R001 (Figure 2, Table S4).

3.2. Genome Features of CDKM15 and CDKM9

The dsDNA genomes of CDKM15 and CDKM9 are 50,606 bp and 49,822 bp, respectively. Both phages have a GC content of 28.98%, which is similar to that of published *C. difficile* phages, and to the reference *C. difficile* strain CD630 (29.06%) [52]. In total, the genome of CDKM15 had 79 predicted CDSs, with 73 on the sense strand and six on the antisense. Of these 79 CDSs, 34 (43%) had a predicted function assigned (with a BLASTp e-value of 1×10^{-10} as a cut-off) and 45 (57%) encode for genes with an unknown function (Figure 3 and Table S2).
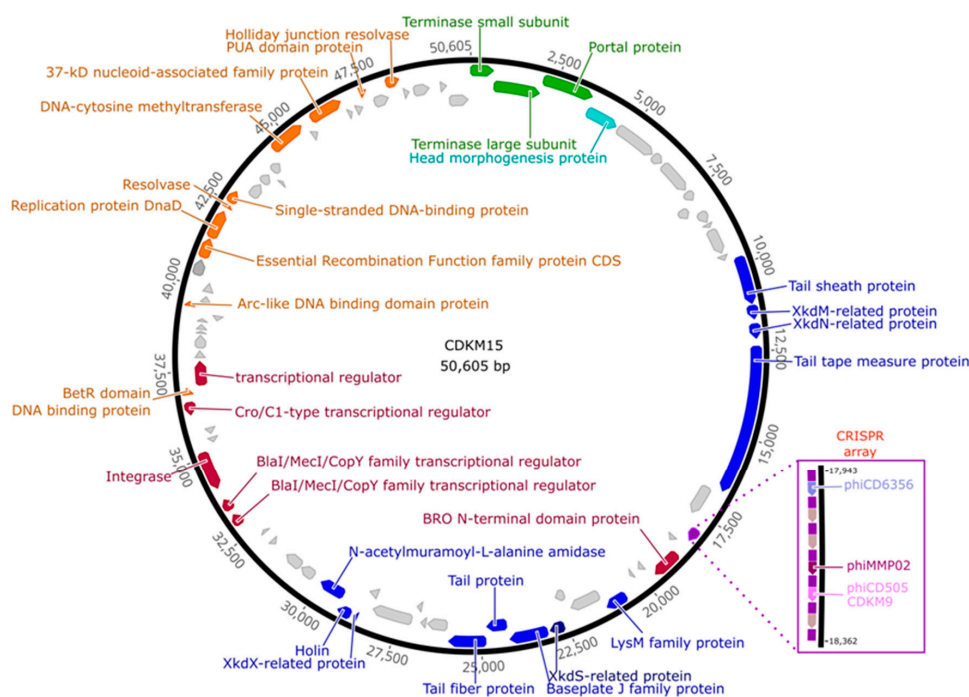

Figure 3. Genome organisation of CDKM15. The 50,605 bp genome assembled as a circle, shown here oriented to start at the terminase small subunit gene. Predicted coding DNA sequences (CDSs) are marked with arrows and colours indicate functional modules: head packaging (green), head (aquamarine), tail and lysis (blue), lysogenic conversion (purple) and DNA replication (orange). CDSs with no function assigned are light grey. Functional annotations are labelled. The clustered regularly interspaced short palindromic repeats (CRISPR) array is marked with a mauve arrow. The CRISPR array and its spacers are highlighted with spacer matches indicated.

The genome of CDKM9 had 75 predicted CDSs, with 66 on the sense strand and nine on the antisense strand. Of the 75, 32 (42.7%) could be assigned putative functions, and 43 (57.3%) could not (Figure 4, Table S2).

Following the genome annotations, it was apparent that the phage genomes displayed a clear modular organisation. There are distinguishable gene modules whose products are predicted to be involved in DNA packaging, virion assembly, host cell lysis, lysogeny control and DNA replication. No tRNA genes were identified in either genome. Each genome had a lysogeny control gene module with a predicted integrase, and two copies of *repR*, a predicted regulatory protein containing a penicillinase repressor family protein domain (Pfam: PF03965).
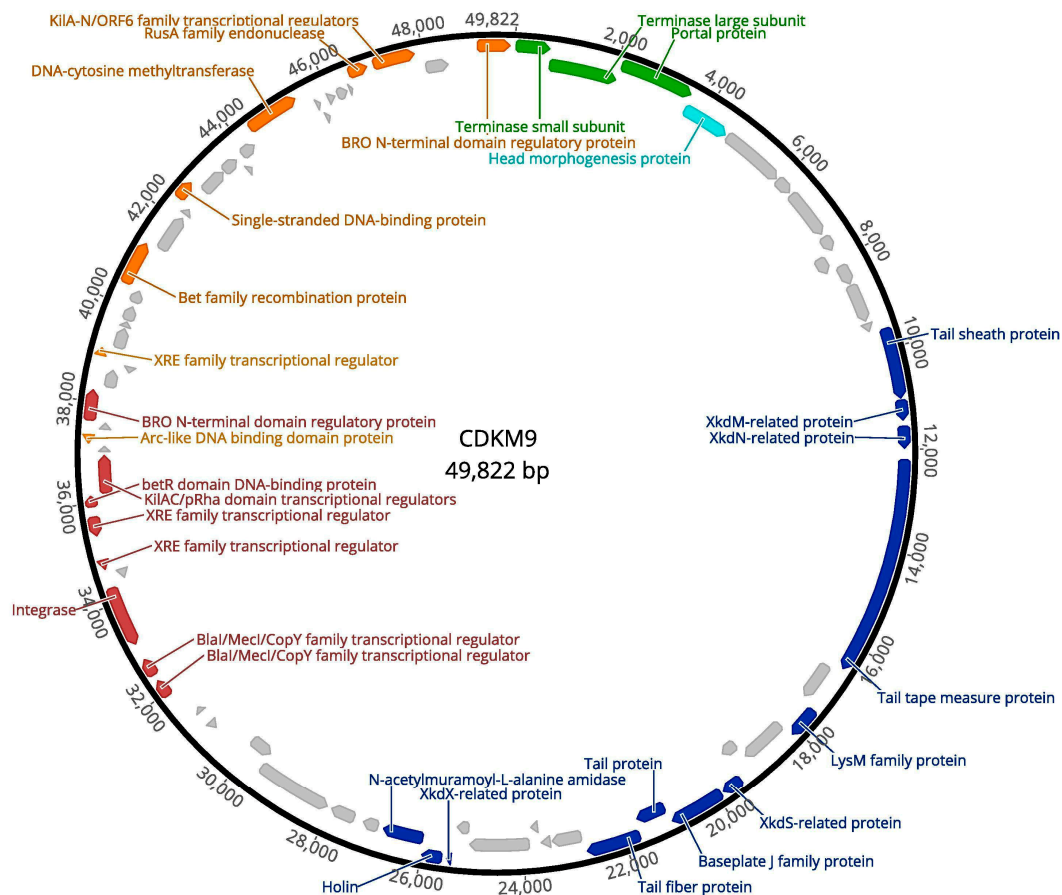
Figure 4. Genome organisation of CDKM9. The 49,822 bp genome assembled as a circle, shown here oriented to start at the terminase small subunit gene. The predicted CDSs are marked with arrows and colours indicate functional modules: head packaging (green), head (aquamarine), tail and lysis (blue), lysogenic conversion (purple) and DNA replication (orange). CDSs with no function assigned are light grey. Those with functional annotations are labelled.

Interestingly, CDMK15 has a CRISPR array located adjacent to a cluster of genes with unknown function preceding the tail morphogenesis module, including a CDS with a predicted baculovirus repeated open reading frame (Bro) N-terminal domain protein (Pfam: 02498) (Figure 3). The array contains six 34–37 bp long spacers and seven 29 bp direct repeats (DRs). There are no *cas* genes in the genome, a finding which is consistent with previously analyses of CRISPR arrays in *C. difficile* phages [11]. The first five DRs (from the 5′ end of the array) are identical, while the last two harbour two and seven mutations (compared to the consensus sequence), respectively. The array is preceded by a 215 bp leader region that shares similarity to that of CRISPR array no. 14 in *C. difficile* R20291 (71% similarity in pairwise ClustalW alignment, 100% in conserved 13-bp 3′ motif) [53,54]. The consensus DR sequence (ATTTTATATTAACTATGTGGTATGTAAAT) differs by four nucleotides to that of the DRs in prophage 1 from CD630 (GTTTTAGATTAACTATATGGAATGTAAAT). All spacer sequences are unique to CDKM15 when searched against *C. difficile* isolate sequences in NCBI. Of the six spacers, spacer 1 was found to perfectly match to the genome sequence of ΦCD6356, spacer 5 imperfectly matched to ΦCD505 and CDKM9 (97% identity), and spacer 4 imperfectly matched to ΦMMP02 (94.6% identity). A protospacer adjacent motif (PAM) of CCN (A or T) has been predicted for *C. difficile* [11,53]. We observed CCN PAMs in the phage genomes for all matches with the exception of the non-identical match to ΦMMP02. The PAM sequence is required for functional targeting, and our findings support the hypothesis that the phage carried spacers may be able to provide functional immunity against the corresponding phages.

3.3. Phylogenetic Analyses

In order to determine how the newly isolated phages are related to previously described *C. difficile* phages, we constructed a phylogenomic tree based on values of nucleotide sequence pairwise similarity between CDKM9, CDKM15 and 22 *C. difficile* phages. The resulting tree revealed that CDKM9 and CDKM15 group with the long tailed myoviruses ΦCD27, ΦCD505 and ΦMMP02 (Figure 5). There may be a distant evolutionary relationship between the two clusters of medium myoviruses (one of them corresponding to the currently accepted genus phicd119virus), but overall similarity between genomes of these groups fell below 40% on the nucleotide level and 50% for translated comparison (Table S5). The "jumbo" myoviruses (ΦCD211 and ΦCDIF1296T) were even more divergent and could not be classified into any of these groups. As might be expected, siphoviruses clustered together, but sub-clusters were observed for the phages that split into a group containing ΦCD38-2, ΦCD111 and ΦCD146, and two singletons ΦCD24-1 and ΦCD6356.
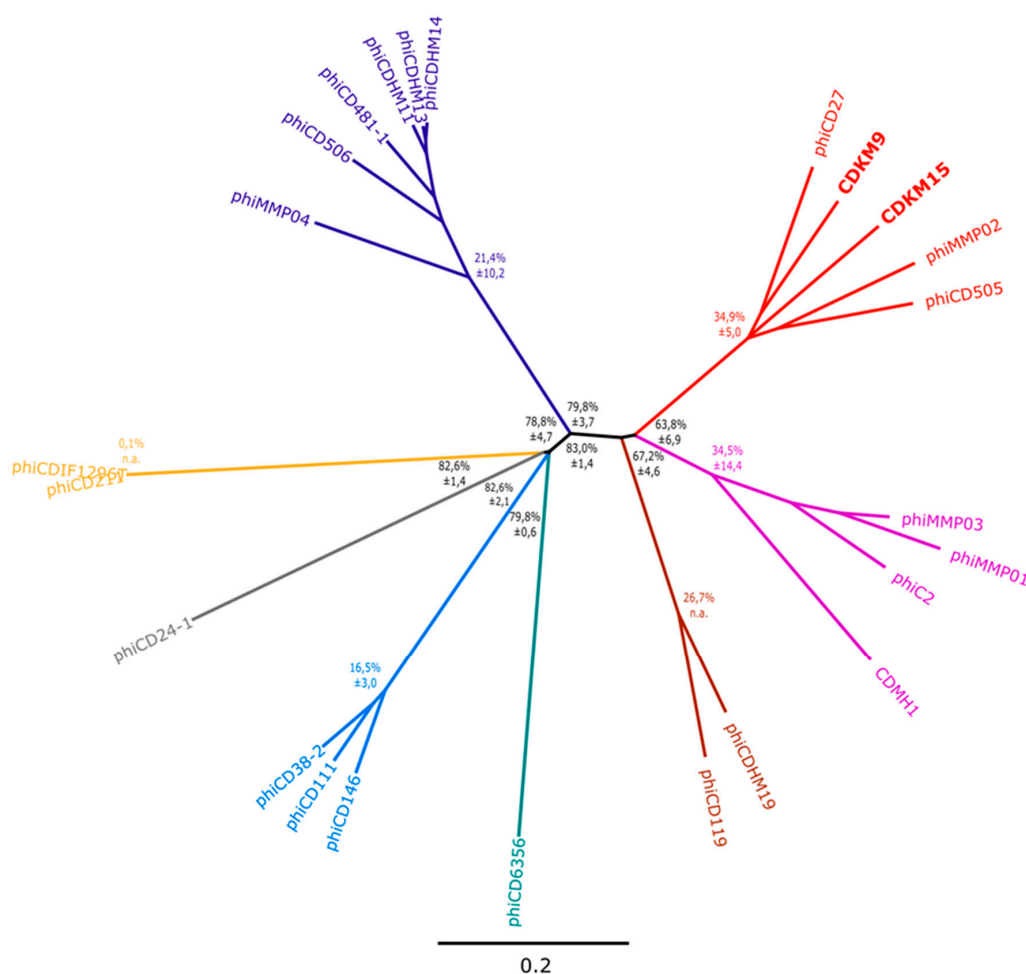

Figure 5. Phylogenetic tree based on whole genome comparison of *C. difficile* phage genomes. The similarity values were calculated based on a translated pairwise comparison of the analysed sequences using Gegenees software. The phylogenetic tree was constructed with SplitsTree using the neighbor joining method. The scale bar represents a 20% difference in average tBLASTx score. Branch colours correspond to the colours in Figure 6. Black node markers represent mean percentage distance between the clades (calculated by averaging each distance between group 1 member and group 2 member) ± standard deviation. Coloured node markers represent the mean percentage distance within the clade (calculated by averaging distances between group members) ± standard deviation. Distance is defined as 100% similarity in translated whole genome comparison.

3.4. Phylogeny of the Endolysin Genes

Phage endolysins are required for lysis of the cell wall, and all sequenced *C. difficile* phages encode an endolysin containing *N*-acetylmuramoyl-L-alanine amidase domains. Phylogenetic analysis was performed on endolysin sequences to establish whether this gene is subject to horizontal gene transfer (HGT), or follows the same evolutionary trajectory as the phage genome. To do this, ML analysis was applied to endolysin sequences from 24 phages (Figure 6). The resulting tree showed that taxa grouped into clades largely reflect the taxonomic division between siphoviruses and myoviruses, in which 15 taxa are myoviruses, and six are siphoviruses clade. However, the topology of the endolysin tree is incongruent with that of the whole genome based tree, as the myoviruses were clustered in such a way that did not reflect their particle morphology (unlike the case for the whole genome phylogeny). Evidence of HGT of the endolysin can be seen for ΦCD506 which clustered with the siphoviruses highlighting this mechanism as facilitating phage genome evolution.
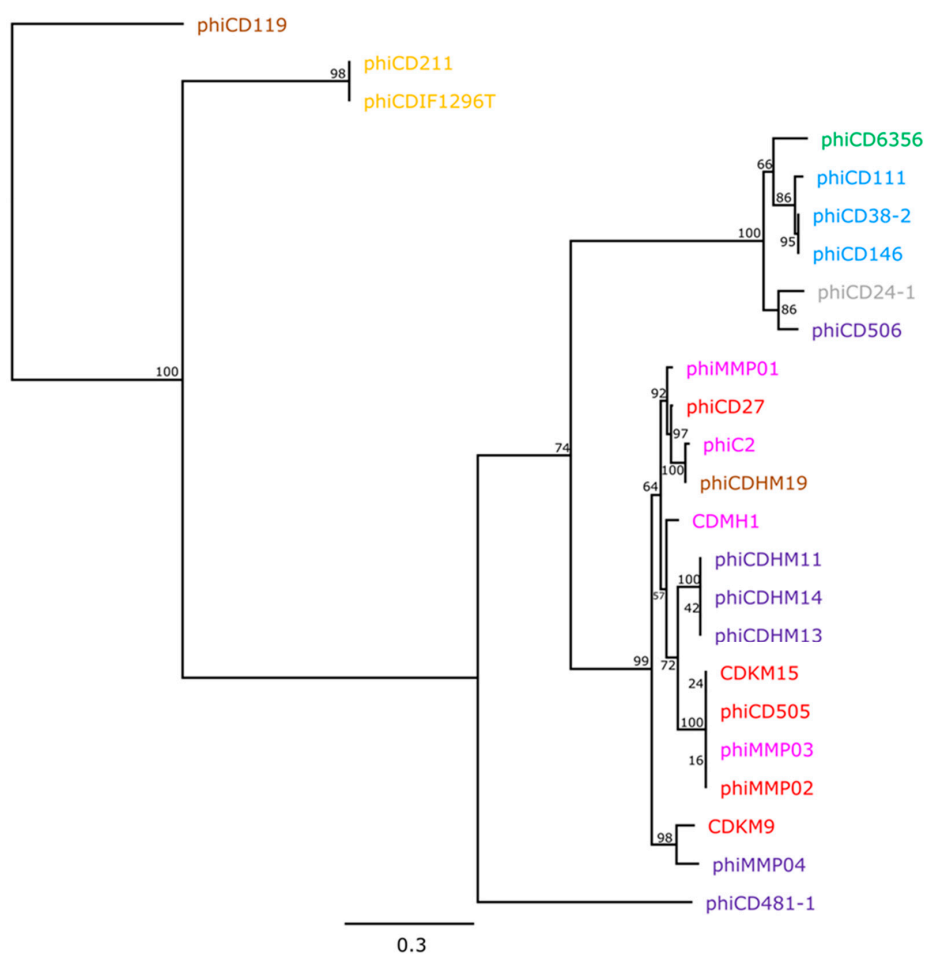

Figure 6. Maximum likelihood phylogenetic analysis of *C. difficile* phage endolysins. Tree node labels represent bootstrap values.

3.5. Phylogenetic Analysis of terL and the Packaging Strategy of the Isolated Phages

To determine the packaging mechanism for CDKM15 and CDKM9, we followed a method described by Casjens and Gilcrease [46]. A phylogenetic tree was generated for *terL* which encodes the terminase large subunit (Figure 7). CDMK9 and CDMK15 form a clade with ΦCD505, ΦCD27 and ΦMMP02 which is consistent with the whole genome tree results. These phages do not cluster with any phages that have a predicted or experimentally confirmed packaging strategy. In an attempt to identify a predicted packaging strategy, we performed read mapping to identify distinct genome termini but

did not detect a signal indicating the location of these termini (e.g., *cos* sites). While we did not detect these termini it is important to remember that fragments generated during library preparation may not be entirely random and their uneven distribution may hinder end analysis (this may be especially true for the Nextera libraries used to sequence CDKM9).
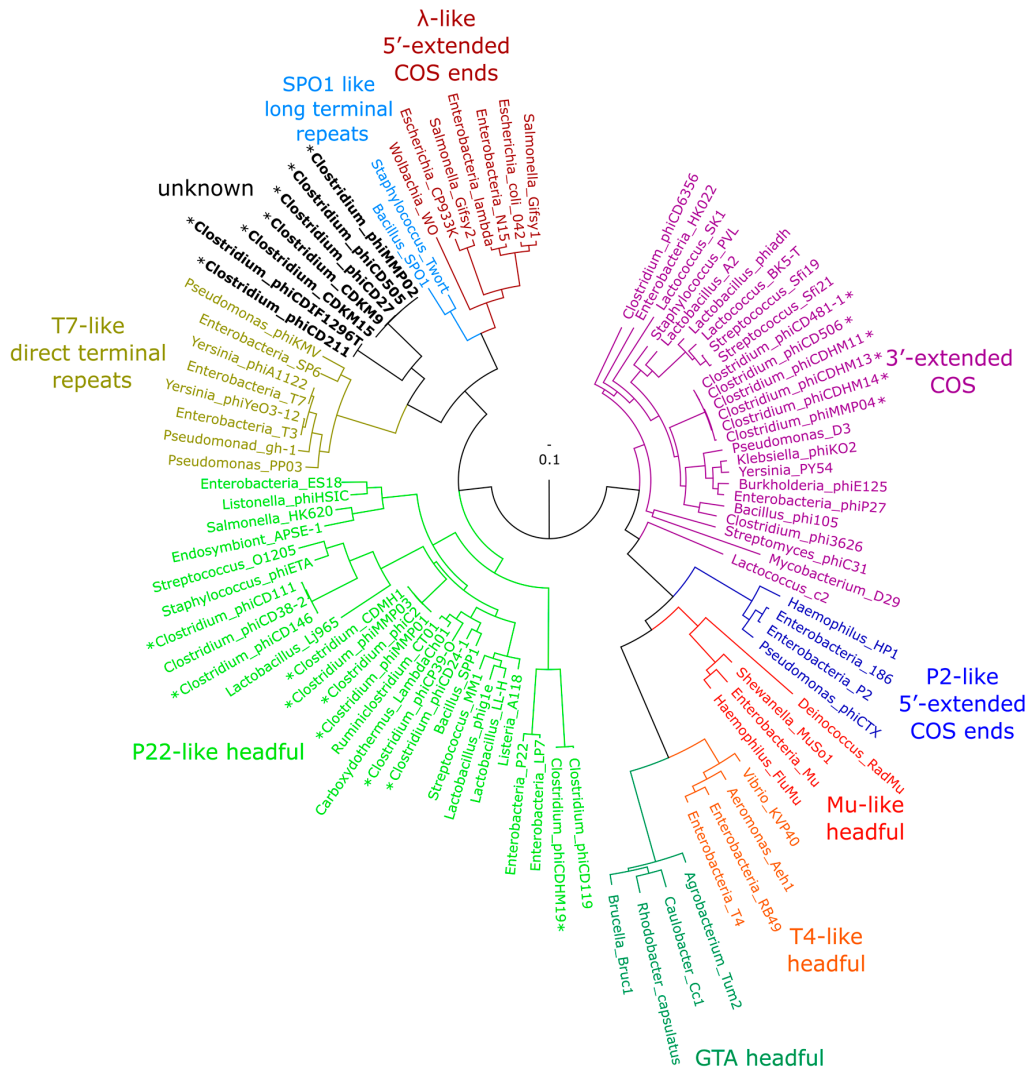
Figure 7. Phylogenetic tree of phage terminase large subunit (*terL*) gene. The name of the phage or prophage is shown at each terminal node and the packaging strategy for each group is labelled where known. The branches are coloured according to the DNA packaging strategy: purple (3′-cohesive ends), blue (5′-cohesive ends), red (Mu-like headful), orange (T4-like headful), jade (GTA headful), green (P22-like headful), olive (T7-like direct terminal repeats), black (unknown), light blue (SPO1 long terminal repeat) and maroon (λ-like 5′-extended COS ends). An asterisk (*) next to a taxon label indicates phages with a packaging strategy predicted during this analysis or earlier predictions using similar methodology, but without experimental evidence (to the best of our knowledge).

3.6. Comparative Genomics

CDKM9 and CDKM15 were each used as a reference to which 22 phage sequences were compared by BLASTn in order to visualise regions sharing similarity across the *C. difficile* phage genomes (Figure 8, Figure S1 and S2). In the resulting maps, the three long-tailed myoviruses (LTM) (ΦCD505, ΦCD27 and ΦMMP02) display the most similarity to CDMK9 and CDKM15. Regions of similarity were observed in the packaging and structural modules, but to a lesser extent in the lysogeny control

and DNA replication modules. Instead, here, the viruses are more similar to the medium myoviruses (ΦC2, ΦCDHM1, ΦMMP03, ΦMMP01, ΦCDHM19 and ΦCD119). As might be expected, the genetic variability of the modules varied to different degrees; the lysogeny control region is divergent across the genomes, while the cluster of genes responsible for lysis is conserved across the 24 phages, which can be expected considering they all infect the same host species.
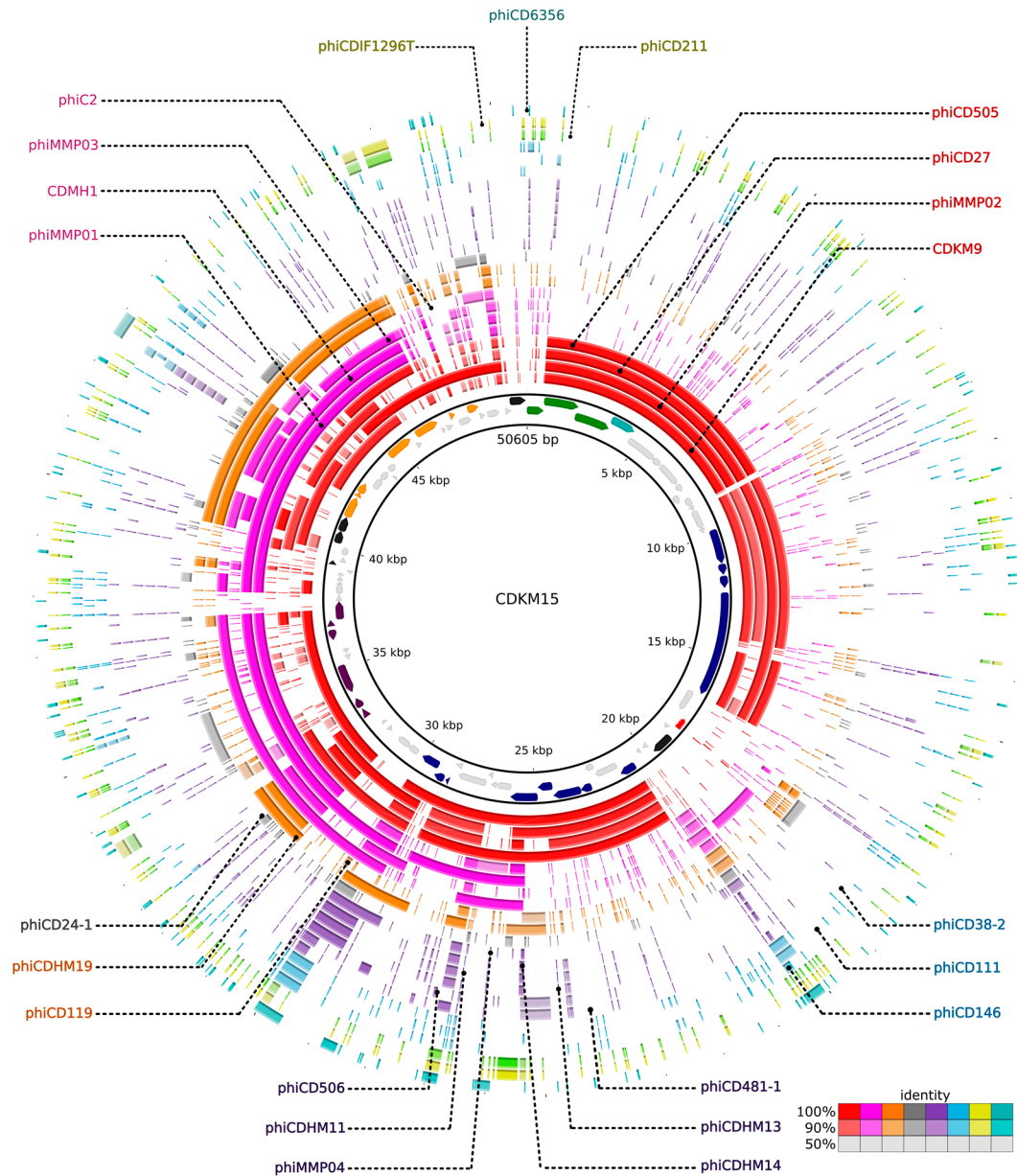

Figure 8. Whole genome comparison of CDKM15 and *C. difficile* phages. The local similarity of each phage is calculated based on BLASTn high scoring pairs and plotted against a circular map of the reference genome represented as the inner circle (in this case genome of CDKM15). Similarity to each of the 23 *C. difficile* phages is shown as colouring intensity in consecutive rings. Outer rings are coloured consistent to the scheme in Figure 6.

3.7. Protein Cluster Analysis

Cluster analysis was performed on 1850 predicted proteins from 24 phage genomes to identify the core genes shared by all *C. difficile* phages and those which are specific to groups or particular phages. The protein sequences were grouped into 479 clusters, with 229 singletons (Figure 9, Table S5).

Interestingly, the most prevalent cluster (cluster no. 0 with sequences from 19 phages) was comprised entirely of protein sequences with no known function. The gene was conserved amongst 18/19 myoviruses and a single siphovirus, ΦCD6356 and usually located between putative tail fiber and endolysin genes. Thus, we hypothesise that the CDS product may be involved in phage-host attachment or interaction. Almost all endolysin sequences grouped into two clusters that corresponded to a family level division for the phages (14/19 myoviruses had endolysins from cluster no. 1, while all five siphovirus endolysins grouped to cluster no. 47). Exceptions to this were the endolysin of myovirus ΦCD506 which clustered with the siphoviruses, (which is consistent with results of the phylogenetic analysis) and the endolysins of "jumbo" myoviruses (ΦCDIF1296T and ΦCD211) which formed a separate cluster.
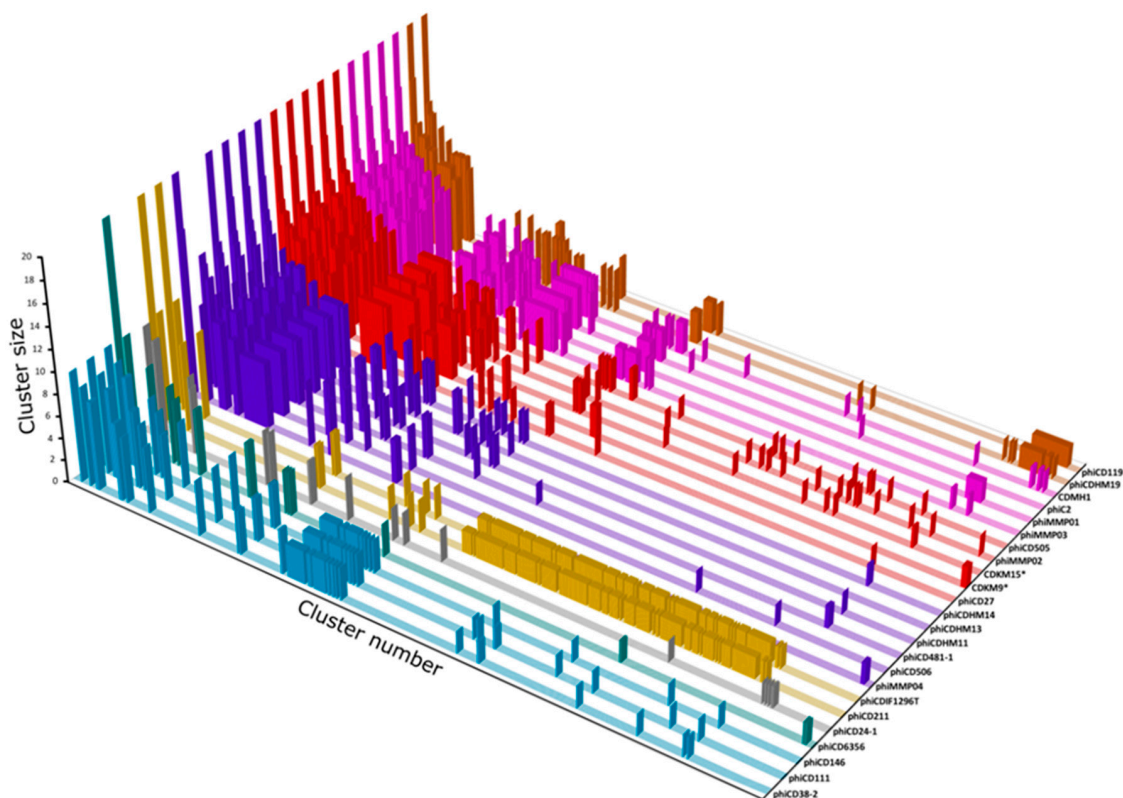

Figure 9. Plot of protein clusters shared by the 24 *C. difficile* phages. The *y*-axis represents the cluster size (number of phages encoding homologues) and clusters are arranged along the *x*-axis by size. Colouring is consistent with Figure 6.

Generally, the protein clusters were confined to phage family (siphovirus or myovirus) or a subgroup, but CDSs in 41/708 clusters were not. These include the above-mentioned cluster no. 0 and a further five clusters that contain CDSs from the siphoviruses and the short tailed myoviruses (clusters no. 5, 7, 17, 21and 25 which are carried in the replication modules of at least eight phages).

Lastly, this analysis revealed four proteins unique to CDKM9, and 10 to CDKM15. For CDMK9, these are three proteins of unknown function (CDS 59, 67, and 68) and a predicted recombination protein with a lambda Bet-like RecT superfamily domain [55,56] encoded in the DNA replication region. For CDKM15, the unique CDS encodes eight hypothetical proteins (CDS 11, 18, 19, 21, 22, 50, 78, and 79), a predicted anti-repressor (CDS 49) and its terminase small subunit (CDS 1).

4. Discussion

Before this work all phages that infect *C. difficile* were isolated from sources within Europe and North America. In this study, we extend the global picture of their diversity by isolating phages from northern Iraq. In total, 14 phages were isolated and two phages further characterised. CDKM9 and CDKM15 could infect 18% (4/22) (CDKM9) and 9% (2/22) (CDKM15) 'local' isolates from Kurdistan (north of Iraq), and 36% (21/58) (CDKM9) and 31% (18/58) (CDKM15) of 'global' isolates from the UK. Their host ranges include isolates from ribotypes associated with major epidemics: R027, R001, R014, R014/020 and R005 [57–59], however, no isolates of R078 were sensitive to either phage. An overlap in each phage's host range was observed, with 16 of 25 and 20 infected by CDKM9 and CDKM15, respectively, suggesting they likely use the same receptor [22]. Both the genome organisation and particle morphology of CDKM9 and CDKM15 resemble that of the long tailed myoviruses ΦCD27, ΦCD505 and ΦMMP02. Interestingly, however, CDKM9 has the broadest reported host spectrum compared to the reported host ranges of other long tail myoviruses (ΦCD27, ΦCD505, ΦCD508, ΦCDHM2, ΦCDHM4, ΦCDHM5 and ΦCDHM6) as these infected 13% (4/30), 11% (5/47), 4/47, (28%) 22/80, 4/80 (5%), 20/80 (25%), (29%) 23/80 isolates, respectively [6,14,16]. Further, no other long tailed myoviruses have been shown to lyse R027 isolate [6,14,16,60], while siphoviruses infecting R027 strains have been identified [6,16,17] and the only other report in the literature of myoviruses with this ability are the medium myoviruses ΦCD481-1 and ΦCDHM3, both of which caused turbid plaques on a single strain of R027 each [6,16].

The genomes of CDKM9 and CDKM15 were sequenced to determine their taxonomic relationships to other *C. difficile* phages. While the genome lengths and GC contents are similar to those values specified for inclusion into the species phicd119virus [18], at just ~50 kb their genomes are at the lower end of the 51–60 kbp range. To further characterise these phages, their shared protein content was analysed (see below).

Annotation of the two phage genomes revealed a notable feature of phage CDKM15; it encoded a CRISPR array. *C. difficile* strains carry multiple CRISPR arrays including those which, based on the *cas* gene content, belong to type I-B/Tneap classification [11,52,61]. In addition to arrays on the main chromosome, arrays have been identified on mobile genetic elements including the skin^cd, prophages and a plasmid [11]. The prophage carried arrays do not have obvious *cas* genes but do encode proteins predicted to have DNA binding ability [11]. Here, we identified arrays on a phage that had been propagated via lytic replication on its indicator host strain. The two prophages found in CD630 both contain CRISPR arrays and can be propagated through lytic infections [9], however whether the CRISPR arrays were maintained during these cycles was unknown, but CRISPR loci had been discovered on a related phage from free viruses in the human gut [62] and in the induced viral particle of prophage in CD105HSE1 [11]. Consistent with the previously characterised phage arrays, the array in CDKM15 is located in the region involved in tail morphogenesis near the gene encoding Bro N-terminal protein [11]. The structure of the array is similar to those found to be active and expressed arrays in *C. difficile* in terms of crucial leader motifs, direct repeat sequences and spacer lengths [11,53] but it has unique spacer content. As CDKM15 was isolated from a poorly studied environment, we wanted to determine if its spacer sequences might be derived from phages isolated in the UK and USA. Indeed, one spacer perfectly matches to ΦCD6356, and other spacers imperfectly match to sequences from phages ΦCD505, CDKM9 and ΦMMP02 which could suggest that there is a global dissemination of strains and phages. The finding of a CRISPR array in CDKM15 expands our understanding of how phages might transfer CRISPR arrays (and resulting immunity) to sensitive cells. Furthermore, the finding that spacers match to other phage genomes, including the 'neighbouring' (CDMK9) and globally distributed phages, such as ΦCD6356 isolated in the Ireland, are evidence of past co-infection events which resulted in acquisition of these spacers. These findings suggest that either long standing evolution and/or subsequent dispersal of similar phages have occurred at a global scale within *C. difficile*.

The results of the comparative genomic analysis of the 24 *C. difficile* phages highlight the mosaic nature of their genomes (Figure 8, Figure S1 and S2). All phage genomes contained a predicted endolysin gene and its phylogenetic analysis suggested that it has undergone HGT, exemplified by the *endolysin* gene of the myovirus ΦCD506 as it is in a clade with siphovirus *endolysin* genes. It seems likely that exchange of this gene can occur between the different phages and host genomes during co-infection, but stabilising selection, driven by the need to recognise the host wall, prevents protein sequences from diverging [63]. Polylysogeny in *C. difficile* is known, for example the sequenced genome of CD630 contains two related prophages, CD630-1 and CD630-2 [52] and is evident from PCR based screens and the observations of multiple different phage particle morphologies in culture lysates [16,21,64–66]. Co-infection with multiple phage types clearly presents the opportunity for gene exchange.

Phage *terL* sequences have been used previously to reconstruct phylogenic relationships [67]. Moreover, Casjens and Gilcrease [46] demonstrated that this sequence may be used to predict the phage's DNA packaging strategy. One of the considerations of using phages for therapeutic purposes is their ability to facilitate HGT via transduction [68]. The mechanism of DNA packaging into the virion determines how this might occur [69]. Headful packaging (*pac*) phages may perform generalised transduction, but phages with cohesive end (*cos* phages) require sequence recognition in the packaging process. In *C. difficile* phages, ΦCD38-2 has been found to contain a *pac* site, whereas ΦCD6356 has a *cos* site identified [13,17]. However, attempts failed to identify cohesive ends for both ΦCD119 and ΦCD27 [10,14], and no mechanism has been specified for ΦC2, although this phage is capable of performing transduction [70]. Phylogenetical analysis of *terL* showed that the genes from CDKM9 and CDKM15 clustered with phages whose packaging mechanism is unclassified. A second method to predict the DNA packaging mechanism is to identify the termini of the DNA molecule ends, as the *cos* and *pac* strategies result in different sequences [46]. No apparent termini could be located in either CDMK9 or CDMK15, a result which is consistent with the fact that no *cos* sites could be identified in the related phage ΦCD27 [14]. The mechanism of DNA packaging used by this group of related phages therefore remains unknown.

To conclude, phages represent a source of novel antimicrobials and the beneficial properties of 'therapeutic phages' include those that are efficient in lysing a wide range of bacterial isolates within the target species. *C. difficile* phages have been investigated from very few countries, and here we demonstrate that phages can be isolated from new sources. Host range analysis of two of these phages suggested they may be of use in the development of phage-based therapeutics. Genome sequencing and analysis revealed new insights into *C. difficile* phage phylogeny and identified signals of HGT. Furthermore, in this work we suggest that the taxonomic framework for classifying this phage group needs to be widened. With this in mind, we propose two genera in addition to phicd119virus, phiMMP04virus, "phiCD38-2virus" containing ΦCD38-2, ΦCD111 and ΦCD146, and "phiCD211virus" containing ΦCD211 and ΦCDIF1296T.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Surawicz, C.M.; McFarland, L.V. Pseudomembranous colitis: Causes and cures. *Digestion* **1999**, *60*, 91–100. [CrossRef] [PubMed]

2. Goudarzi, M.; Seyedjavadi, S.S.; Goudarzi, H.; Mehdizadeh Aghdam, E.; Nazeri, S. *Clostridium difficile* infection: Epidemiology, pathogenesis, risk factors, and therapeutic options. *Scientifica* **2014**, *2014*, 9. [CrossRef] [PubMed]

3. Buffie, C.G.; Jarchum, I.; Equinda, M.; Lipuma, L.; Gobourne, A.; Viale, A.; Ubeda, C.; Xavier, J.; Pamer, E.G. Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to *Clostridium difficile*-induced colitis. *Infect. Immun.* **2012**, *80*, 62–73. [CrossRef] [PubMed]

4. Knight, D.R.; Elliott, B.; Chang, B.J.; Perkins, T.T.; Riley, T.V. Diversity and evolution in the genome of *Clostridium difficile*. *Clin. Microbiol. Rev.* **2015**, *28*, 721–741. [CrossRef] [PubMed]

5. Hunt, J.J.; Ballard, J.D. Variations in virulence and molecular biology among emerging strains of *Clostridium difficile*. *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 567–581. [CrossRef] [PubMed]

6. Nale, J.Y.; Spencer, J.; Hargreaves, K.R.; Buckley, A.M.; Trzepinski, P.; Douce, G.R.; Clokie, M.R. Better together: Bacteriophage combinations significantly reduce *Clostridium difficile* growth in vitro and proliferation in vivo. *Antimicrob. Agents Chemother.* **2015**, *60*, 968–981. [CrossRef] [PubMed]

7. Meader, E.; Mayer, M.J.; Gasson, M.J.; Steverding, D.; Carding, S.R.; Narbad, A. Bacteriophage treatment significantly reduces viable *Clostridium difficile* and prevents toxin production in an in vitro model system. *Anaerobe* **2010**, *16*, 549–554. [CrossRef] [PubMed]

8. Ramesh, V.; Fralick, J.A.; Rolfe, R.D. Prevention of *Clostridium difficile*-induced ileocecitis with bacteriophage. *Anaerobe* **1999**, *5*, 69–78. [CrossRef]

9. Goh, S.; Ong, P.F.; Song, K.P.; Riley, T.V.; Chang, B.J. The complete genome sequence of *Clostridium difficile* phage ΦC2 and comparisons to ΦCD119 and inducible prophages of CD630. *Microbiology* **2007**, *153*, 676–685. [CrossRef] [PubMed]

10. Govind, R.; Fralick, J.A.; Rolfe, R.D. Genomic organization and molecular characterization of *Clostridium difficile* bacteriophage phicd119. *J. Bacteriol.* **2006**, *188*, 2568–2577. [CrossRef] [PubMed]

11. Hargreaves, K.R.; Flores, C.O.; Lawley, T.D.; Clokie, M.R. Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *mBio* **2014**, *5*, e01045–e01013. [CrossRef] [PubMed]

12. Hargreaves, K.R.; Kropinski, A.M.; Clokie, M.R.J. What does the talking? Quorum sensing signalling genes discovered in a bacteriophage genome. *PLoS ONE* **2014**, *9*, e85131. [CrossRef] [PubMed]

13. Horgan, M.; O'Sullivan, O.; Coffey, A.; Fitzgerald, G.F.; van Sinderen, D.; McAuliffe, O.; Ross, R.P. Genome analysis of the *Clostridium difficile* phage ΦCD6356, a temperate phage of the *Siphoviridae* family. *Gene* **2010**, *462*, 34–43. [CrossRef] [PubMed]

14. Mayer, M.J.; Narbad, A.; Gasson, M.J. Molecular characterization of a *Clostridium difficile* bacteriophage and its cloned biologically active endolysin. *J. Bacteriol.* **2008**, *190*, 6734–6740. [CrossRef] [PubMed]

15. Meessen-Pinard, M.; Sekulovic, O.; Fortier, L.-C. Evidence of in vivo prophage induction during *Clostridium difficile* infection. *Appl. Environ. Microbiol.* **2012**, *78*, 7662–7670. [CrossRef] [PubMed]

16. Sekulovic, O.; Garneau, J.R.; Neron, A.; Fortier, L.C. Characterization of temperate phages infecting *Clostridium difficile* isolates of human and animal origins. *Appl. Environ. Microbiol.* **2014**, *80*, 2555–2563. [CrossRef] [PubMed]

17. Sekulovic, O.; Meessen-Pinard, M.; Fortier, L.C. Prophage-stimulated toxin production in *Clostridium difficile* NAP1/027 lysogens. *J. Bacteriol.* **2011**, *193*, 2726–2734. [CrossRef] [PubMed]

18. Lavigne, R.; Darius, P.; Summer, E.J.; Seto, D.; Mahadevan, P.; Nilsson, A.S.; Ackermann, H.W.; Kropinski, A.M. Classification of myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.* **2009**, *9*, 224. [CrossRef] [PubMed]

19. Hargreaves, K.R.; Clokie, M.R.J. A taxonomic review of *Clostridium difficile* phages and proposal of a novel genus, "Phimmp04likevirus". *Viruses* **2015**, *7*, 2534–2541. [CrossRef] [PubMed]

20. Hargreaves, K.R. Isolation and Characterisation of Bacteriophages Infecting Environmental Strains of *Clostridium difficile*. Ph.D. Thesis, University of Leicester, Leicester, UK, 2012.

21. Hargreaves, K.R.; Colvin, H.V.; Patel, K.V.; Clokie, J.J.; Clokie, M.R. Genetically diverse *Clostridium difficile* strains harboring abundant prophages in an estuarine environment. *Appl. Environ. Microbiol.* **2013**, *79*, 6236–6243. [CrossRef] [PubMed]

22. Goh, S.; Riley, T.V.; Chang, B.J. Isolation and characterization of temperate bacteriophages of *Clostridium difficile*. *Appl. Environ. Microbiol.* **2005**, *71*, 1079–1083. [CrossRef] [PubMed]

23. Sambrook, J.; Russell, D.W. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: New York, NY, USA, 2001.

24. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 13 November 2016).

25. Joshi, N.A.; Fass, J.N. Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for Fastq Files. 2011. Available online: https://github.com/najoshi/sickle (accessed on 1 October 2015).

26. Li, R.; Zhu, H.; Ruan, J.; Qian, W.; Fang, X.; Shi, Z.; Li, Y.; Li, S.; Shan, G.; Kristiansen, K.; et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **2010**, *20*, 265–272. [CrossRef] [PubMed]

27. Li, R.; Li, Y.; Kristiansen, K.; Wang, J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **2008**, *24*, 713–714. [CrossRef] [PubMed]

28. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef] [PubMed]

29. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [CrossRef] [PubMed]

30. Besemer, J.; Borodovsky, M. Genemark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **2005**, *33*, W451–W454. [CrossRef] [PubMed]

31. Delcher, A.L.; Bratke, K.A.; Powers, E.C.; Salzberg, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **2007**, *23*, 673–679. [CrossRef] [PubMed]

32. Overbeek, R.; Olson, R.; Pusch, G.D.; Olsen, G.J.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Parrello, B.; Shukla, M.; et al. The SEED and the Rapid Annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* **2014**, *42*, D206–D214. [CrossRef] [PubMed]

33. Hyatt, D.; Chen, G.L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010**, *11*, 119. [CrossRef] [PubMed]

34. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]

35. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]

36. Marchler-Bauer, A.; Bryant, S.H. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* **2004**, *32*, W327–W331. [CrossRef] [PubMed]

37. Jones, P.; Binns, D.; Chang, H.Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. Interproscan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]

38. Schattner, P.; Brooks, A.N.; Lowe, T.M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **2005**, *33*, W686–W689. [CrossRef] [PubMed]

39. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef] [PubMed]

40. Edgar, R.C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **2007**, *8*, 18. [CrossRef] [PubMed]

41. Grissa, I.; Vergnaud, G.; Pourcel, C. CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **2007**, *35*, W52–W57. [CrossRef] [PubMed]

42. Biswas, A.; Gagnon, J.N.; Brouns, S.J.J.; Fineran, P.C.; Brown, C.M. Crisprtarget. *RNA Biol.* **2013**, *10*, 817–827. [CrossRef] [PubMed]

43. Ågren, J.; Sundström, A.; Håfström, T.; Segerman, B. Gegenees: Fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS ONE* **2012**, *7*, e39107. [CrossRef] [PubMed]

44. Huson, D.H.; Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evolut.* **2006**, *23*, 254–267. [CrossRef] [PubMed]

45. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321. [CrossRef] [PubMed]

46. Casjens, S.R.; Gilcrease, E.B. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol. Biol.* **2009**, *502*, 91–111. [PubMed]

47. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef] [PubMed]

48. Liu, K.; Linder, C.R.; Warnow, T. RAxML and FastTree: Comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **2011**, *6*, e27731. [CrossRef] [PubMed]

49. Alikhan, N.-F.; Petty, N.K.; Ben Zakour, N.L.; Beatson, S.A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **2011**, *12*, 402. [CrossRef] [PubMed]

50. Sullivan, M.J.; Petty, N.K.; Beatson, S.A. Easyfig: A genome comparison visualizer. *Bioinformatics* **2011**, *27*, 1009–1010. [CrossRef] [PubMed]

51. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]

52. Sebaihia, M.; Wren, B.W.; Mullany, P.; Fairweather, N.F.; Minton, N.; Stabler, R.; Thomson, N.R.; Roberts, A.P.; Cerdeno-Tarraga, A.M.; Wang, H.; et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **2006**, *38*, 779–786. [CrossRef] [PubMed]

53. Boudry, P.; Semenova, E.; Monot, M.; Datsenko, K.A.; Lopatina, A.; Sekulovic, O.; Ospina-Bedoya, M.; Fortier, L.C.; Severinov, K.; Dupuy, B.; et al. Function of the CRISPR-Cas System of the Human Pathogen *Clostridium difficile*. *mBio* **2015**, *6*, e01112–e01115. [PubMed]

54. Soutourina, O.A.; Monot, M.; Boudry, P.; Saujet, L.; Pichon, C.; Sismeiro, O.; Semenova, E.; Severinov, K.; Le Bouguenec, C.; Coppee, J.Y.; et al. Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. *PLoS Genet.* **2013**, *9*, e1003493. [CrossRef] [PubMed]

55. Nafissi, N.; Slavcev, R. Bacteriophage recombination systems and biotechnical applications. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 2841–2851. [CrossRef] [PubMed]

56. Hillyar, C.R. Genetic recombination in bacteriophage lambda. *Biosci. Horiz.* **2012**, *5*, hzs001. [CrossRef]

57. McDonald, L.; Killgore, G.; Thompson, A.; Owens, R.; Kazakova, S.; Sambol, S.; Johnson, S.; Gerding, D. An epidemic, toxin gene-variant strain of *Clostridium difficile*. *N. Engl. J. Med.* **2005**, *353*, 2433–2441. [CrossRef] [PubMed]

58. Robinson, C.D.; Auchtung, J.M.; Collins, J.; Britton, R.A. Epidemic *Clostridium difficile* strains demonstrate increased competitive fitness compared to nonepidemic isolates. *Infect. Immun.* **2014**, *82*, 2815–2825. [CrossRef] [PubMed]

59. Tickler, I.A.; Goering, R.V.; Whitmore, J.D.; Lynn, A.N.; Persing, D.H.; Tenover, F.C. Strain types and antimicrobial resistance patterns of *Clostridium difficile* isolates from the United States, 2011 to 2013. *Antimicrob. Agents Chemother.* **2014**, *58*, 4214–4218. [CrossRef] [PubMed]

60. Fortier, L.-C.; Moineau, S. Morphological and genetic diversity of temperate phages in *Clostridium difficile*. *Appl. Environ. Microbiol.* **2007**, *73*, 7358–7366. [CrossRef] [PubMed]

61. Andersen, J.M.; Shoup, M.; Robinson, C.; Britton, R.; Olsen, K.E.; Barrangou, R. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biol. Evol.* **2016**, *8*, 2841–2855. [CrossRef] [PubMed]

62. Minot, S.; Sinha, R.; Chen, J.; Li, H.; Keilbaugh, S.A.; Wu, G.D.; Lewis, J.D.; Bushman, F.D. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res.* **2011**, *21*, 1616–1625. [CrossRef] [PubMed]

63. Oliveira, H.; Melo, L.D.; Santos, S.B.; Nobrega, F.L.; Ferreira, E.C.; Cerca, N.; Azeredo, J.; Kluskens, L.D. Molecular aspects and comparative genomics of bacteriophage endolysins. *J. Virol.* **2013**, *87*, 4558–4570. [CrossRef] [PubMed]

64. Nagy, E.; Foldes, J. Electron microscopic investigation of lysogeny of clostridium difficile strains isolated from antibiotic-associated diarrhea cases and from healthy carriers. *Acta Pathol. Microbiol. Immunol. Scand.* **1991**, *99*, 321–326. [CrossRef]

65. Nale, J.Y. Isolation and Characterisation of Temperate Bacteriophages of the Hypervirulent *Clostridium difficile* 027 Strains. Ph.D. Thesis, University of Leicester, Leicester, UK, 2013.

66. Shan, J.; Patel, K.V.; Hickenbotham, P.T.; Nale, J.Y.; Hargreaves, K.R.; Clokie, M.R. Prophage carriage and diversity within clinically relevant strains of *Clostridium difficile*. *Appl. Environ. Microbiol.* **2012**, *78*, 6027–6034. [CrossRef] [PubMed]

67. Casjens, S.R.; Gilcrease, E.B.; Winn-Stapley, D.A.; Schicklmaier, P.; Schmieger, H.; Pedulla, M.L.; Ford, M.E.; Houtz, J.M.; Hatfull, G.F.; Hendrix, R.W. The generalized transducing *Salmonella* bacteriophage ES18: Complete genome sequence and DNA packaging strategy. *J. Bacteriol.* **2005**, *187*, 1091–1104. [CrossRef] [PubMed]

68. Chen, J.; Carpena, N.; Quiles-Puchalt, N.; Ram, G.; Novick, R.P.; Penades, J.R. Intra- and inter-generic transfer of pathogenicity island-encoded virulence genes by *cos* phages. *ISME J.* **2015**, *9*, 1260–1263. [CrossRef] [PubMed]

69. Rao, V.B.; Feiss, M. Mechanisms of DNA packaging by large double-stranded DNA viruses. *Annu. Rev. Virol.* **2015**, *2*, 351–378. [CrossRef] [PubMed]

70. Goh, S.; Hussain, H.; Chang, B.J.; Emmett, W.; Riley, T.V.; Mullany, P. Phage ΦC2 mediates transduction of Tn6215, encoding erythromycin resistance, between *Clostridium difficile* strains. *mBio* **2013**, *4*, e00840-13. [CrossRef] [PubMed]