UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO STATE UNIVERSITY

DEFINING THE ROLE OF BACTERIOPHAGE IN MUCOSAL IMMUNITY

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology

by

Sean Benler

Committee in charge:

University of California San Diego

Professor Partho Ghosh
Professor Joseph Pogliano

San Diego State University

Professor Forest Rohwer, Chair
Professor Anca Segall
Professor Roland Wolkowicz

2018

The Dissertation of Sean Benler is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

San Diego State University

2018

Dedication

To everyone who pushed me a little further.

Epigraph

And then they ask you, ‘what is the significance of DNA’?

-Francis Crick

Table of contents

List of figures

List of tables

Acknowledgements

Thanks, everyone!

Chapter 1, in part, is currently being prepared for submission for publication of the material. Sean Benler, Forest Rohwer. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, has been submitted for publication. Sean Benler, Shr-Hau Hung, Yan Wei Lim, Katelyn McNair, Jeremy Barr, Forest Rohwer, and Anca Segall. 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication. Sean Benler, Ana Cobian-Guemes, Katelyn McNair, Shr Hau Hung, Kyle Levi, Robert Edwards, and Forest Rohwer. 2018. The dissertation author was the primary investigator and author of this paper.

Vita

2018	Doctor of Philosophy in Biology, University of California San Diego and San Diego State University
2010	Bachelor of Science in Biology, University of California Davis

Publications

Benler, S., Hung, S., Lim, Y., McNair, K., Barr, J., Segall, A., Rohwer, F., (2018) 'Gp4 is a nuclease required for morphogenesis of T4-like bacteriophages', In review.

Benler, S., Cobian-Guemes, A., McNair, K., Hung, S., Levi, K., Edwards, R., Rohwer, F., (2018) 'A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage', In review.

Cobian-Guemes, A., **Benler, S.,** Brettin, T, Cantu, V.A., Conrad, D., Cuevas, D., Dorrestein, P., Edwards, R., Hamidi, R., Lim, Y.W., Maughan, H., Quinn, R., Rohwer, F. (2018) 'Cystic Fibrosis Rapid Response: translating multi-omics data into clinically relevant information', In review.

Schooley, R. T., Biswas, B., Gill, J. J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J. J., Reed, S. L., Rohwer, F., **Benler, S.**, Segall, A. M., Taplitz, R., Smith, D. M., Kerr, K., Kumaraswamy, M., Nizet, V., Lin, L., McCauley, M. D., Strathdee, S. A., Benson, C. A., Pope, R. K., Leroux, B. M., Picel, A. C., Mateczun, A. J., Cilwa, K. E., Regeimbal, J. M., Estrella, L. A., Wolfe, D. M., Henry, M. S., Quinones, J., Salka, S., Bishop-Lilly, K. A., Young, R. and Hamilton, T. (2017) 'Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant Acinetobacter baumannii infection', Antimicrobial Agents and Chemotherapy, 61, pp. e00954-17.

Quinn, R. A., Vermeij, M. J. A., Hartmann, A. C., Galtier d'Auriac, I., **Benler, S.,** Haas, A., Quistad, S. D., Lim, Y. W., Little, M., Sandin, S., Smith, J. E., Dorrestein, P. C. and Rohwer, F. (2016) 'Metabolomics of reef benthic interactions reveals a bioactive lipid involved in coral defence', Proceedings of the Royal Society B, 283, p. 20160469.

ABSTRACT OF THE DISSERTATION

DEFINING THE ROLE OF BACTERIOPHAGE IN MUCOSAL IMMUNITY

by

Sean Benler

Doctor of Philosophy in Biology

University of California San Diego

San Diego State University

2018

Professor Forest Rohwer, Chair

Since their discovery over one hundred years ago, bacteriophages (phages) play a central role in the ongoing molecular biology revolution. As such, model phages like Lambda and T4 are perhaps the most genetically well-understood biological entities. Yet, critical aspects regarding the life histories of these phages remain unaddressed. This dissertation examines the life history of T4 and non-model phages. In chapter 2, the steps involved in the morphogenesis of T4-like phages were reevaluated. During the final stages of morphogenesis, T4 packages DNA into an empty head that is subsequently joined with a tail to make a mature virion. Bioinformatic analysis of one gene involved in this head-to-tail joining process, gene 4 (gp4), identified a previously unrecognized nuclease motif. Biochemical analysis confirmed the DNA-cutting capabilities of

gp4. The nuclease function of gp4 is shown to be conserved in T4-like phages. In aggregate, the data contained in chapter 2 of this dissertation suggest gp4 plays a role in the termination of genome packaging to enable head-tail joining. Chapter 3 is dedicated to the analysis of non-model phages, focusing on those that contain diversity-generating retroelements (DGRs). DGRs are genetic cassettes that introduce sequence variation into target genes. Few isolated phages possess a DGR, yet metagenomic analyses reveal their presence in uncultured viral communities. In chapter 3, a survey identified 92 DGRs that were only found in phages exhibiting a temperate lifestyle. One novel temperate phage that possesses a DGR cassette targeting a gene of unknown function was isolated from *Bacteroides dorei*. This phage, here named Hankyphage, exhibits broad host-range and is capable of infecting at least 13 different *Bacteroides* species. Sequencing reads from whole-community metagenomes and viromes were recruited to the Hankyphage genome, highlighting the global distribution of Hankyphage. The results in chapter 3 suggest that targeted hypervariation by temperate phages, such as Hankyphage, may be a ubiquitous mechanism underlying phage-bacteria interaction in the human microbiome.

Chapter 1 : Introduction

Bacteriophage Adherence to Mucus (BAM): How ecological and evolutionary dynamics create a mucosal immune system

Mucosal layers are the primary region of contact between animals and the environment. As such, mucosal layers are densely populated by viruses, bacteria and other microbes. Mucus is a dynamic niche for these inhabitants, due to genetic polymorphisms in mucin genes, highly variable post-translational modifications and continual outward flux. The viruses and microbes in these layers evolved diverse solutions to negotiate their mucosal niche. Examples include the bacterial and viral lectins that enable adherence to mucin glycoproteins, which result in a spatially-structured community. The spatial structuring of mucus exerts density-dependent effects on populations of bacteriophages and bacteria. Specifically, at higher bacterial densities bacteriophage (phage) infection predominantly yields lysogens, i.e. bacteria stably infected with a phage. During lysogeny, a tripartite symbiosis between phage, bacteria and metazoans may emerge in the form of mucosal immunity. In this chapter I review the adaptations by phage to predate upon bacteria in mucus, considering mucus as a variable environment in which density-dependent factors drive phage infection outcomes. I present a model how mucus-adapted phage confer mucosal immunity in chapter 4.

Mucus is a rapidly changing environment. Metazoan mucosal surfaces mark the boundary between animal and environment. Mucosal layers are complex, stratified and transient environments that are repeatedly cleared from the underlying epithelium (Button et al., 2012). The principal macromolecular components of mucus are large mucin glycoproteins

secreted by specialized goblet cells (Johansson and Hansson, 2013). Mucin proteins contain tandemly repeated domains that are substrates for extensive O-linked glycosylation (Rose and Voynow, 2006). Addition of branched glycans results in a 'bottle-brush' shaped protein with glycan side-chains extending 0.5-5 nm from the peptide core (Cone, 2009; Linden et al., 2008). Continual secretion of mucin glycoproteins causes the mucosal layer to move outward from the epithelium, where proteolytic cleavage results in volume expansion and eventual sloughing to the environment (Johansson et al., 2011). In mice, the mucosal layer above the distal colon is renewed every 1 to 2 hours (Johansson and Hansson, 2013). The surface mucosal layer thereby functions as robust barrier against diverse microorganisms (Cone, 2009).

The barrier activity of mucus is augmented by its variable presentation to viruses and microbes. Humans possess 17 different mucin genes that are polymorphic in the number of tandemly repeated domains (Johansson et al., 2011; Vinall et al., 2002). Decoration of these domains with glycans can yield a protein that is more than 80% carbohydrate by mass (McGuckin et al., 2011). In the case of MUC2, greater than 100 different glycan side chains are present on the mature glycoprotein (Larsson et al., 2009). Addition of glycans to mucins is executed by glycosyltransferase enzymes. Genetic polymorphisms in glycosyltransferases can determine the type of glycan modification (e.g., Lewis blood group antigens on red blood cells) (McGuckin et al., 2011). Furthermore, the expression pattern of glycosyltransferases and their location in the Golgi apparatus changes glycosylation patterns (Gagneux and Varki, 1999). As a result, the same mucin proteins can be glycosylated differently between adjacent intestinal crypts in humans (Matsuo et al., 1997). Acute stressors such as disease or intestinal infection can initiate rapid release of mucus with altered glycosylation profiles (Lindén et al.,

2008; Pickard et al., 2014). Environmental changes in pH, ionic conditions, mucin concentration and hydration impact the viscoelastic properties of mucus (Bansil et al., 1995; Lai et al., 2009). Collectively, the mucosal layer is a dynamic environment that requires adaptive solutions by viruses and microbes for successful persistence.

Despite rapid turnover and variable composition, mucus supports dense populations of viral and microbial communities. In the human gastrointestinal tract, there are approximately 10^{12} viruses and 10^{13} bacteria (Cobián Güemes et al., 2016; Sender et al., 2016). The mucus associated community includes members who are stable in time and generally resilient to displacement (Faith et al., 2013; Hester et al., 2015; Lozupone et al., 2012; Minot et al., 2013). Their persistence can partly be explained by adaptations to negotiate their mucosal environment. Viruses and microbes often employ glycan-binding proteins (lectins) to adhere to mucin glycoproteins (Etzold and Juge, 2014; Johansson et al., 2011; Juge, 2012; Kankainen et al., 2009; Marionneau et al., 2002; Pang et al., 2014; Stevens et al., 2006). For example, the *Vibrio cholerae* lectin enables attachment to intestinal mucus (Wong et al., 2012). Lectins bind to specific glycan structures, which can target viruses and bacteria to different mucosal niches (Graaf and Fouchier, 2014; Naughton et al., 2013). Despite the recognized importance of mucus in structuring bacterial communities (Earle et al., 2015; Mark Welch et al., 2017), little work has considered mucus in structuring viral communities and their interaction with bacterial hosts (discussed below).

Lysogeny and superinfection-exclusion in viruses. The life history of a virus can be classified into two categories: lytic or temperate. Infection by lytic viruses rapidly leads to production of progeny and host cell death (e.g., T4, influenza). In contrast, temperate viruses are able to establish and maintain a stable relationship with their host (e.g., lambda, HIV).

This state is termed lysogeny, in which expression of genes for lysis of the host cell is blocked by the action of a repressor. Generally, viral DNA is integrated into the host chromosome and is replicated with the host DNA using machinery of the cell. Under different circumstances for different viruses, a switch to the lytic state occurs. Expression of lytic genes leads to replication of the viral DNA and production of progeny that burst from the cell into the environment, completing the life cycle of the virus.

Integration of viral DNA offers powerful advantages to the host cell. In particular, an integrated virus prevents infection by related viruses, providing immunity to the host cell. Immunity can occur through diverse molecular mechanisms, a well-studied example being superinfection exclusion. Superinfection exclusion of related viruses can occur by modifying the host cell surface to prevent adsorption or by degradation of incoming viral DNA (Bondy-Denomy et al., 2016; Boyd, 1954; Lindahl et al., 1970; Susskind and Botstein, 1978). Theoretical models predict that superinfection exclusion becomes progressively advantageous as viral and host densities rise.

Upon encountering a host cell, temperate viruses decide to coexist stably with their host cells (lysogeny) or decimate them (lysis). The genetic circuitry governing this decision is best understood in bacteriophage lambda and is reviewed elsewhere (Echols, 1986; Little, 2005; Ptashne, 1992). The most well-defined input that dictates the lysis-lysogeny decision the multiplicity of infection (MOI), i.e. the number of phage per cell. High MOI drives lambda towards lysogeny (Kourilsky, 1973). Other inputs, such as host cell physiology (e.g., growth rate) are less well understood (Echols, 1986). The circuitry of the lysis-lysogeny decision in other bacteriophages is less clear. In bacteriophage Mu, the frequency of

lysogenization is approximately 0.1% per phage particle (Leach and Symonds, 1979), whereas the remaining infections result in lytic development.

The effect of density on the lysis-lysogeny decision, as exemplified by phage lambda, extends to communities of viruses. For example, experimental manipulation of the growth rate of an uncultured microbial community results in fewer virus-like particles (VLPs) per cell as microbe concentrations increase (Knowles and Silveira et al., 2016). The same trend of 'more microbes, fewer viruses' can be observed by performing direct counts of viruses and cells in the environment. Counts from 223 environmental samples demonstrate that as the concentration of microbes increases, fewer VLPs are observed per cell (Knowles and Silveira et al., 2016). Metagenomic analysis of uncultured viral communities from environments with high host concentrations, including the human gut, discloses that the majority of viruses are temperate bacteriophage (Breitbart et al., 2003, 2008; Knowles et al., 2016; Manrique et al., 2016; Minot et al., 2011; Reyes et al., 2010). In summary, density-dependent dynamics influence the outcome of viral infections, driving temperate viruses to lysogenize at high-host densities.

Integrated phages can alter the population dynamics of their microbial hosts. For example, a common gram-positive symbiont of the human intestine, *Enterococcus faecalis*, harbors a prophage that provides a colonization advantage in gnotobiotic mice over strains lacking the prophage (Duerkop et al., 2012). The authors argue that *E. faecalis* utilizes the prophage as a 'weapon' against related strains and reduces competition for nutrients. Similar phenomena are observed with *Escherichia coli*, *Bordetella bronchisepta* and *Salmonella enterica* (Bossi et al., 2003; Brown et al., 2006; Gama et al., 2013; Joo et al., 2006; De Paepe et al., 2016). In contrast, harboring a prophage can be disadvantageous. One study

documented that the induction of a prophage in *Staphylococcus aureus* triggered by a competitor leads to displacement of the lysogen *in vivo* (Selva et al., 2009). These studies highlight the complex role of prophages in bacterial niche defense and microbial population ecology.

Directed evolution by diversity-generating retroelements: Variation is an essential substrate for the process of evolution. In prokaryotes, diversity-generating retroelements (DGRs) are genetic cassettes that introduce massive DNA sequence variation through targeted mutagenesis (Liu et al., 2002). Analogous to the mammalian immune system, DGRs create combinatorial diversity in target proteins that are subsequently purified by selection for optimal function (Medhekar and Miller, 2007). First identified in bacteriophage BPP-1, the DGR creates variable tail-fiber proteins that enable host-tropism switching among strains of *Bordetella bronchisepta* (Doulatov et al., 2004; Liu et al., 2002). To date, the BPP-1 tail fiber is the only hypervariable protein with a known function.

The mechanism of DGR function has been analyzed extensively in BPP-1. The BPP-1 tail fiber gene possesses a 134bp locus that exhibits high variability, designated the variable repeat (VR). Downstream of this locus is a second copy of the repeat that does not vary, named the template repeat (TR). Nearby the TR/VR pair are genes encoding an 'accessory variability determinant' and a reverse transcriptase (RT). These two proteins interact to generate cDNA from TR that contains point mutations, followed by stable incorporation of the mutagenized cDNA into the phage tail fiber gene (Handa et al., 2018). This process, termed 'mutagenic retrohoming', yields a VR that is distinct from the TR exclusively at adenine bases. Additional components downstream of the VR/TR are regions of DNA that regulate the directionality of information transfer and efficiency (Doulatov et al., 2004; Guo et

al., 2008, 2011; Naorem et al., 2017). At characteristic frequencies, the mutagenic process yields a population of progeny with tail fibers recognizing unique bacterial receptors and enable infection of *B. bronchisepta* in different physiological states. The cumulative function of the DGR is analogous to phage display: a population of phage possessing heterogeneous tail fibers are produced and those that confer binding to host bacteria receptors are 'panned' out from the population.

The function of the hypervariable proteins targeted by DGRs in other phages and their role in phage evolution remain to be explored. Based on the BPP-1 paradigm, phage DGRs are predicted to target tail fibers and enable tropism switching (Liu et al., 2002). Indeed, target genes with predicted structural similarity to the tail fiber of BPP-1 have been identified (Hannigan et al., 2017; Minot et al., 2012; Wu et al., 2017, Paul et al., 2015). In other phages, the function of the DGR-targeted protein is less clear as no sequence similarity to any characterized proteins can be assigned. Despite the lack of functional characterization, there is remarkable conservation of C-type lectin and Ig-like folds in DGR-target proteins (Le Coq and Ghosh, 2011; Wu et al., 2017). Both folds exhibit remarkable stability despite tremendous sequence variation (Chothia and Lesk, 1987; McMahon et al., 2005). Overall, some phages use a similar strategy as BPP-1 to mutagenize their tail fibers and broaden host range, whereas the functions and ligands of other phage hypervariable proteins remains to be determined.

Recent experimental evidence suggests that the ligands for some phage Ig-like domains may be glycans present on mucus glycoproteins. The Ig-like domain-containing protein *hoc* is displayed on the capsid of T4 and is dispensable for *E. coli* infection under standard laboratory conditions (Sathaliyawala et al., 2010). Using phage T4, deletion of *hoc* reduces the number of phage recoverable from mucosal surfaces (Barr et al., 2013a). A glycan

microarray revealed specific glycan structures that accrue more T4 *hoc*$^{+}$ virions than T4 *hoc*$^{-}$, suggesting the *hoc*-glycan interaction is responsible for the observed mucus-adherent phenotype (Barr et al., 2013a). Ig-like and lectin domains are encoded by approximately 25% of isolated phages (Fraser et al., 2006), hinting that that additional phages may have a mucus-adherent phenotype. In support of the generalizability of this phenotype, virus-like particles are enriched in mucosal layers compared to their adjacent environment across diverse metazoans, including coral mucus and surrounding seawater (Nguyen-Kim et al., 2014). The role of DGRs in mediating a mucus-adherence through targeted mutagenesis of glycan-binding Ig-like domains remains to be tested.

Bacteriophage adherence to mucus (BAM) via Ig-like domains has been shown to protect the underlying metazoan epithelium from bacterial infection (Barr et al., 2015). In the presence of mucus, T4 *hoc*$^{+}$ encounters *E. coli* more frequently and adsorbs more rapidly onto cells than T4 *hoc*$^{-}$. As a result of the increased frequency of bacterial encounters, significantly less *E. coli* accumulate in the mucosal layer of an epithelium experiencing constant fluid flow and mucin secretion dynamics (Barr et al., 2015). Collectively, these experiments demonstrate how mucus-adherent phage can impact the commensal microbiome. The BAM model posits that phage provide mucosal immunity by protecting the metazoan epithelium from bacterial colonization (Barr et al., 2013b).

Experiments demonstrating BAM immunity were achieved with the strictly lytic phages T4 and T7. How temperate phage contribute to BAM immunity is currently unknown. Given that most viruses in the human gastrointestinal tract are temperate, incorporating the lytic-lysogeny decision in the BAM model is warranted. One version of the model has been proposed previously, theorizing that a temperate phage carried by a mucus-invading

pathogenic bacteria could lysogenize commensals (Barr et al., 2013b). Spontaneous induction of commensals would result in progeny virions dispersed throughout the mucus-layer, preventing re-infection by that pathogen and constituting a 'phage-memory' of previous infections (Barr et al., 2013b). In support of this hypothesis, chapter 3 presents a broad-host range temperate *Bacteroides* phage capable of lysogenizing both commensals and pathogens. This phage, named Hankyphage, possesses both a DGR and an Ig-like domain on its HK97-like major capsid protein, suggesting a mucus-adherent phenotype. A longitudinal analysis demonstrates Hankyphage persists in the human microbiome for >2 years, consistent with the concept of a 'phage memory' of infection. In chapter 4, a model of how temperate, DGR-possessing phages generate mucosal immunity is proposed.

Genome packaging in T4-like bacteriophages: corking the champagne bottle

The introduction of "conditional lethals" to genetics enabled the complete developmental analysis of sophisticated biological systems, starting with bacteriophage T4 (Epstein et al., 1963; Stahl, 1995). Conditional lethal mutants of T4, named "amber mutants", were quickly exploited to demonstrate steps in phage morphogenesis (Edgar and Wood, 1966). Among the genes identified as essential for morphogenesis, amber mutations in gene product 4 (gp4) produce unattached phage heads and tails in a restrictive host (Edgar and Wood, 1966). Closer examination of the heads and tails by electron microscopy revealed that approximately half of the heads contain DNA and the other half are empty (Granboulan et al., 1971). In cases where an assembled phage particle could be observed, the tail appears at an angle relative to the head (i.e., kinked) (Granboulan et al., 1971). Based on these studies gp4 plays a role in head-tail joining of T4, though the mechanism by which gp4 functions is

unknown. A hint comes from the work of Snustad et. al, where gp4 was demonstrated to code for an enzyme (Snustad, 1968). In this study, a restrictive host cell was coinfected with a wild-type phage and a gp4 amber mutant phage and the number of progeny virions measured. Upon increasing the multiplicity of infection (MOI) of the amber mutant phage, the number of progeny produced from the coinfection remained stable. This result demonstrated that expression of a single wild-type allele of gp4 was sufficient to yield a normal burst size. In contrast, expression of wild-type genes encoding a structural proteins were insufficient (Snustad, 1968).

Despite the 50 years since the demonstration that gp4 is an enzyme involved in head-tail joining of phage T4, the mechanism of function remained unknown. In chapter 2, evidence is presented to demonstrate that gp4 functions as a nuclease during genome packaging of T4-like phages. T4-like phages package their genomes using a headful mechanism, in which DNA is enzymatically pumped into the head until full (Black and Silverman, 1978). In the last stage of filling, the internal pressure inside the head exceeds that of bottled champagne (~6 MPa) (Lander et al., 2006). The termination of genome packaging thus requires a complex series of events to maintain the tightly packaged DNA, including (a) packaging holoenzyme reorganization, (b) terminating DNA cleavage and (c) holoenzyme dissociation, whose temporal order is not known (Rao and Feiss, 2015). Future progress in elucidating the mechanism of genome packaging in viruses requires examination of this dynamic transition.

Acknowledgements

Chapter 1, in part, is currently being prepared for submission for publication of the material. Sean Benler and Forest Rohwer. The dissertation author was the primary investigator and author of this paper.

References

Bansil, R., Stanley, E., and LaMont, J.T. (1995). Mucin biophysics. Annu. Rev. Physiol. *57*, 635–657.

Barr, J.J., Auro, R., Furlan, M., Whiteson, K.L., Erb, M.L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A.S., Doran, K.S., et al. (2013a). Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc. Natl. Acad. Sci. U. S. A. *110*, 10771–10776.

Barr, J.J., Youle, M., and Rohwer, F. (2013b). Innate and acquired bacteriophage-mediated immunity. Bacteriophage *3*, e25857-4.

Barr, J.J., Auro, R., Sam-Soon, N., Kassegne, S., Peters, G., Bonilla, N., Hatay, M., Mourtada, S., Bailey, B., Youle, M., et al. (2015). Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. Proc. Natl. Acad. Sci. *112*, 13675–13680.

Black, L.W., and Silverman, D.J. (1978). Model for DNA packaging into bacteriophage T4 heads. J. Virol. *28*, 643–655.

Bondy-Denomy, J., Qian, J., Westra, E.R., Buckling, A., Guttman, D.S., Davidson, A.R., and Maxwell, K.L. (2016). Prophages mediate defense against phage infection through diverse mechanisms. ISME J. *22*, 1–13.

Bossi, L., Fuentes, J.A., Mora, G., and Figueroa-bossi, N. (2003). Prophage Contribution to Bacterial Population Dynamics. J. Bacteriol. *185*, 6467–6471.

Boyd, J.S.K. (1954). Bacteriophage and heredity. Nature *173*, 1050–1051.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic Analyses of an Uncultured Viral Community from Human Feces. J. Bacteriol. *185*, 6220–6223.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B., Mahaffy, J.M., Mueller, J., Nulton, J., Rayhawk, S., et al. (2008). Viral diversity and dynamics in an infant gut. Res. Microbiol. *159*, 367–373.

Brown, S.P., Chat, L. Le, Paepe, M. De, and Taddei, F. (2006). Ecology of Microbial Invasions : Amplification Allows Virus Carriers to Invade More Rapidly When Rare. Curr. Biol. 2048–2052.

Button, B., Cai, L.-H., Ehre, C., Kesimer, M., Hill, D.B., Sheehan, J.K., Boucher, R.C., and Rubinstein, M. (2012). A periciliary brush promotes the lung health by separating the mucus layer from airway epithelia. Science (80-.). *337*, 937–941.

Chothia, C., and Lesk, A.M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins Canonical Structures for the Hypervariable Regions of Immunoglobulins. J. Mol. Biol. *196*, 901–917.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. Annu. Rev. Virol. *3*, 197–214.

Cone, R.A. (2009). Barrier properties of mucus. Adv. Drug Deliv. Rev. *61*, 75–85.

Le Coq, J., and Ghosh, P. (2011). Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. Proc. Natl. Acad. Sci. *108*, 14649–14653.

Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R., Zimmerly, S., and Miller, J. (2004). Tropism Switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476–481.

Duerkop, B. a, Clements, C. V, Rollins, D., Rodrigues, J.L.M., and Hooper, L. V (2012). A composite bacteriophage alters colonization by an intestinal commensal bacterium. Proc. Natl. Acad. Sci. U. S. A. *109*, 17621–17626.

Earle, K.A., Billings, G., Sigal, M., Lichtman, J.S., Hansson, G.C., Elias, J.E., Amieva, M.R., Huang, K.C., and Sonnenburg, J.L. (2015). Quantitative Imaging of Gut Microbiota Spatial Organization. Cell Host Microbe *18*, 478–488.

Echols, H. (1986). Bacteriophage λ development: temporal switches and the choice of lysis or lysogeny. Trends Genet. *2*, 26–30.

Edgar, R.S., and Lielausis, I. (1968). Some steps in the assembly of bacteriophage T4. J. Mol. Biol. *32*, 263–276.

Edgar, R.S., and Wood, W.B. (1966). Morphogenesis of bacteriophage T4 in extracts of mutant-infected cells. Proc. Natl. Acad. Sci. *55*, 498–505.

Epstein, R.H., Bolle, A., Steinberg, C.M., Kellenberger, E., Boy de la Tour, E., and Chevalley, R. (1963). Physiological studies of conditional lethal mutants of bacteriophage T4D. Cold Spring Harb. Symp. Quant. Biol. 375–394.

Etzold, S., and Juge, N. (2014). Structural insights into bacterial recognition of intestinal mucins. Curr. Opin. Struct. Biol. *28*, 23–31.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., et al. (2013). The Long-Term Stability of the Human Gut Microbiota. Science (80-.). *341*, 1237439.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507.

Gagneux, P., and Varki, A. (1999). Evolutionary considerations in relating oligosaccharide diversity to biological function. Glycobiology *9*, 747–755.

Gama, J.A., Reis, A.M., Domingues, I., Mendes-Soares, H., Matos, A.M., and Dionisio, F. (2013). Temperate Bacterial Viruses as Double-Edged Swords in Bacterial Warfare. PLoS One *8*, e59043.

Graaf, M. De, and Fouchier, R.A.M. (2014). Role of receptor binding specificity in influenza A virus transmission and pathogenesis. EMBO J. *33*, 823–841.

Granboulan, P., Sechaud, J., and Kellenberger, E. (1971). On the fragility of phage T4-related particles. Virology *46*, 407–425.

Guo, H., Tse, L. V., Barbalat, R., Sivaamnuaiphorn, S., Xu, M., Doulatov, S., and Miller, J.F. (2008). Diversity-Generating Retroelement Homing Regenerates Target Sequences for Repeated Rounds of Codon Rewriting and Protein Diversification. Mol. Cell *31*, 813–823.

Guo, H., Tse, L. V., Nieh, A.W., Czornyj, E., Williams, S., Oukil, S., Liu, V.B., and Miller, J.F. (2011). Target site recognition by a diversity-generating retroelement. PLoS Genet. *7*.

Hannigan, G.D., Zheng, Q., Meisel, J.S., Minot, S.S., Bushman, F.D., and Grice, E.A. (2017). Evolutionary and functional implications of hypervariable loci within the skin virome. PeerJ *5*, e2959.

Hester, E.R., Barott, K.L., Nulton, J., Vermeij, M.J., and Rohwer, F.L. (2015). Stable and sporadic symbiotic communities of coral and algal holobionts. ISME J. *10*, 1–13.

Johansson, M.E. V, and Hansson, G.C. (2013). Mucus and the goblet cell. Dig. Dis. *31*, 305–309.

Johansson, M.E. V, Ambort, D., Pelaseyed, T., Schutte, A., Gustafsson, J.K., Ermund, A., Subramani, D.B., Holmen-Larsson, J.M., Thomsson, K.A., Bergstrom, J.H., et al. (2011). Composition and functional role of the mucus layers in the intestine. Cell. Mol. Life Sci. *68*, 3635–3641.

Joo, J., Gunny, M., Cases, M., Hudson, P., Albert, R., and Harvill, E. (2006). Bacteriophage-mediated competition in Bordetella bacteria. Proc. R. Soc. B *273*, 1843–1848.

Juge, N. (2012). Microbial adhesins to gastrointestinal mucus. Trends Microbiol. *20*, 30–39.

Kankainen, M., Paulin, L., Tynkkynen, S., von Ossowski, I., Reunanen, J., Partanen, P., Satokari, R., Vesterlund, S., Hendrickx, A.P.A., Lebeer, S., et al. (2009). Comparative

genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. Proc. Natl. Acad. Sci. *106*, 17193–17198.

Knowles, B., Silveira, C.B., Bailey, B. a, Barott, K., Cantu, V. a, Cobián-Güemes, a G., Coutinho, F.H., Dinsdale, E. a, Felts, B., Furby, K. a, et al. (2016). Lytic to temperate switching of viral communities. Nature *531*, 466–470.

Kourilsky, P. (1973). Lysogenization by bacteriophage lambda. Mol. Gen. Genet. *122*, 183–195.

Lai, S.K., Wang, Y.Y., Cone, R., Wirtz, D., and Hanes, J. (2009). Altering mucus rheology to "solidify" human mucus at the nanoscale. PLoS One *4*, 1–6.

Lander, G., Tang, L., Casjens, S., Gilcrease, E., Prevelige, P., Poliakov, A., Potter, C., Carragher, B., and Johnson, J. (2006). The Structure of an Infectious P22 Virion Shows the Signal for Headful DNA Packaging. Science (80-.). *312*, 1791–1796.

Larsson, J.M.H., Karlsson, H., Sjövall, H., and Hansson, G.C. (2009). A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MS. Glycobiology *19*, 756–766.

Leach, D., and Symonds, N. (1979). The isolation and characterisation of a plaque-forming derivative of bacteriophage Mu carrying a fragment of Tn3 conferring ampicillin resistance. MGG Mol. Gen. Genet. *172*, 179–184.

Lindahl, G., Sironi, G., Bialy, H., and Calendar, R. (1970). Bacteriophage lambda; abortive infection of bacteria lysogenic for phage P2. Proc. Natl. Acad. Sci. U. S. A. *66*, 587–594.

Linden, S.K., Sutton, P., Karlsson, N.G., Korolik, V., and McGuckin, M.A. (2008). Mucins in the mucosal barrier to infection. Mucosal Immunol. *1*, 183–197.

Lindén, S.K., Florin, T.H.J., and McGuckin, M.A. (2008). Mucin dynamics in intestinal bacterial infection. PLoS One *3*.

Little, J.W. (2005). Lysogeny, Prophage Induction, and Lysogenic Conversion. In Phages: Their Role in Bacterial Pathogenesis and Biotechnology, M.K. Waldor, D.I. Friedman, and S. Adhya, eds. (Washington, D.C.: ASM Press), pp. 37–54.

Liu, M., Deora, R., Doulatov, S.R., Gingery, M., Eiserling, F.A., Preston, A., Maskell, D.J., Simons, R.W., Cotter, P.A., Parkhill, J., et al. (2002). Reverse Transcriptase – Mediated Tropism Switching in Bordetella. Science (80-.). *295*, 2091–2095.

Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature *489*, 220–230.

Manrique, P., Bolduc, B., Walk, S.T., Oost, J. Van Der, Vos, W.M. De, and Young, M.J. (2016). Healthy human gut phageome. Proc. Natl. Acad. Sci. *113*, 10400–10405.

Marionneau, S., Ruvo??n, N., Le MoullacVaidye, B., Clement, M., CailleauThomas, A., RuizPalacois, G., Huang, P., Jiang, X., and Le Pendu, J. (2002). Norwalk Virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals. Gastroenterology *122*, 1967–1977.

Mark Welch, J.L., Hasegawa, Y., McNulty, N.P., Gordon, J.I., and Borisy, G.G. (2017). Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. Proc. Natl. Acad. Sci. 201711596.

Matsuo, K., Ota, H., Akamatsu, T., Sugiyama, A., and Katsuyama, T. (1997). Histochemistry of the surface mucous gel layer of the human colon. Gut 782–789.

McGuckin, M.A., Lindén, S.K., Sutton, P., and Florin, T.H. (2011). Mucin dynamics and enteric pathogens. Nat. Rev. Microbiol. *9*, 265–278.

McMahon, S. a, Miller, J.L., Lawton, J. a, Kerkow, D.E., Hodes, A., Marti-Renom, M. a, Doulatov, S., Narayanan, E., Sali, A., Miller, J.F., et al. (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. Nat. Struct. Mol. Biol. *12*, 886–892.

Medhekar, B., and Miller, J.F. (2007). Diversity-generating retroelements. Curr. Opin. Microbiol. *10*, 388–395.

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

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2012). Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. *109*, 3962–3966.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. *110*, 12450–12455.

Naorem, S.S., Han, J., Wang, S., Lee, W.R., Heng, X., Miller, J.F., and Guo, H. (2017). DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. Proc. Natl. Acad. Sci. *114*, E10187–E10195.

Naughton, J.A., Mariño, K., Dolan, B., Reid, C., Gough, R., Gallagher, M.E., Kilcoyne, M., Gerlach, J.Q., Joshi, L., Rudd, P., et al. (2013). Divergent Mechanisms of Interaction of Helicobacter pylori and Campylobacter jejuni with Mucus and Mucins. *81*, 2838–2850.

Nguyen-Kim, H., Bouvier, T., Bouvier, C., Doan-Nhu, H., Nguyen-Ngoc, L., Rochelle-

Newall, E., Baudoux, A.C., Desnues, C., Reynaud, S., Ferrier-Pages, C., et al. (2014). High occurrence of viruses in the mucus layer of scleractinian corals. Environ. Microbiol. Rep. *6*, 675–682.

De Paepe, M., Tournier, L., Moncaut, E., Son, O., Langella, P., and Petit, M.A. (2016). Carriage of Lamda Latent Virus Is Costly for Its Bacterial Host due to Frequent Reactivation in Monoxenic Mouse Intestine. PLoS Genet. *12*, e1005861.

Pang, S.S., Nguyen, S.T.S., Perry, A.J., Day, C.J., Panjikar, S., Tiralongo, J., Whisstock, J.C., and Kwok, T. (2014). The three-dimensional structure of the extracellular adhesion domain of the sialic acid-binding adhesin SabA from Helicobacter pylori. J. Biol. Chem. *289*, 6332–6340.

Pickard, J.M., Maurice, C.F., Kinnebrew, M. a, Abt, M.C., Schenten, D., Golovkina, T. V, Bogatyrev, S.R., Ismagilov, R.F., Pamer, E.G., Turnbaugh, P.J., et al. (2014). Rapid fucosylation of intestinal epithelium sustains host – commensal symbiosis in sickness. Nature *514*, 638–641.

Ptashne, M. (1992). A genetic switch: phage Lambda and higher organisms (Cambridge, Mass.: Blackwell Scientific Publications).

Rao, V.B., and Feiss, M. (2015). Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses. Annu. Rev. Virol. *2*, 351–378.

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

Rose, M.C., and Voynow, J.A. (2006). Respiratory Tract Mucin Genes and Mucin Glycoproteins in Health and Disease. 245–278.

Sathaliyawala, T., Islam, M.Z., Li, Q., Fokine, A., Rossmann, M.G., and Rao, V.B. (2010). Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. Mol. Microbiol. *77*, 444–455.

Selva, L., Viana, D., Regev-Yochay, G., Trzcinski, K., Corpa, J.M., Lasa, I., Novick, R.P., and Penadés, J.R. (2009). Killing niche competitors by remote-control bacteriophage induction. Proc. Natl. Acad. Sci. *106*, 1234–1238.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. *14*, 1–14.

Snustad, D.P. (1968). Dominance interactions in Escherichia coli cells mixedly infected with bacteriophage T4D wild-type and amber mutants and their possible implications as to type of

gene-product function: catalytic vs. stoichiometric. Virology *35*, 550–563.

Stahl, F.W. (1995). Anecdotal, historical and critical commentaries on genetics. Genetics 439–442.

Stevens, J., Blixt, O., Glaser, L., Taubenberger, J.K., Palese, P., Paulson, J.C., and Wilson, I.A. (2006). Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. J. Mol. Biol. *355*, 1143–1155.

Susskind, M.M., and Botstein, D. (1978). Molecular Genetics of Bacteriophage P22. Microbiol. Rev. *42*, 385–413.

Vinall, L.E., King, M., Novelli, M., Green, C.A., Daniels, G., Hilkens, J., Sarner, M., and Swallow, D.M. (2002). Altered expression and allelic association of the hypervariable membrane mucin MUC1 in Helicobacter pylori gastritis. Gastroenterology *123*, 41–49.

Wong, E., Vaaje-Kolstad, G., Ghosh, A., Hurtado-Guerrero, R., Konarev, P. V., Ibrahim, A.F.M., Svergun, D.I., Eijsink, V.G.H., Chatterjee, N.S., and van Aalten, D.M.F. (2012). The Vibrio cholerae colonization factor GbpA possesses a modular structure that governs binding to different host surfaces. PLoS Pathog. *8*, 1–12.

Wu, L., Gingery, M., Abebe, M., Arambula, D., Czornyj, E., Handa, S., Khan, H., Liu, M., Pohlschroder, M., Shaw, K.L., et al. (2017). Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res. 1–14.

Chapter 2 : gp4 is a nuclease required for morphogenesis of T4-like bacteriophages

Abstract

Bacteriophage YC is a lytic marine phage that infects *Vibrio coralliilyticus,* an opportunistic pathogen of the coral holobiont. Here, sequencing of the complete YC genome identifies it as a new member of the T4-like bacteriophages. One of the genes encoded by YC is a homolog of T4 gene product 4 (gp4). Through an unknown mechanism, gp4 plays an essential role in head-tail joining during T4 morphogenesis. Alignment of gp4 homologs from T4, YC and other T4-like bacteriophages identifies a conserved type II restriction endonuclease motif. Testing both purified gp4 proteins of T4 and YC demonstrates that both proteins retain nuclease activity *in vitro*. Mutation of a single conserved amino acid residue in the endonuclease fold of T4 and YC gp4 abrogates nuclease activity. Expression of T4 gp4 *in trans* rescues a T4 gp4 amber mutant phage, while the mutated T4 allele or YC homolog does not. These results indicate that gp4 nuclease activity is essential for morphogenesis of T4-like bacteriophages and suggests close interactions between gp4 and a partner protein. Together, the respective protein pairs may play a role in termination of genome packaging.

Introduction

Species of the bacterial genus *Vibrio* are ubiquitously found in aquatic environments and in association with metazoans (Arias et al., 1995; Heidelberg et al., 2002; Kushmaro et al., 1997; Ruby and Asato, 1993). *Vibrio* spp. are the etiological agents of numerous human diseases, including cholera, gastroenteritis and septicemia, as well as diseases of other metazoans (Thompson et al., 2004). The coral symbiont *Vibrio coralliilyticus* uses chemotaxis to target the mucus of its coral host (Garren et al., 2014) and can become pathogenic in elevated seawater temperatures (Ben-Haim et al., 2003). *V. coralliilyticus* harbors several

putative prophages that may confer virulence via lysogenic conversion (Weynberg et al., 2015). Since corals lack an adaptive immune system, bacteriophage (phage) therapy has been proposed as a method of treatment to protect corals from bacterial challenge (Efrony et al., 2006).

Five phages infecting *V. coralliilyticus* have been isolated to date including YC, a lytic myophage retrieved from the water above coral colonies of the Great Barrier Reef (Cohen et al., 2013). *In vitro* experiments demonstrate that YC interferes with the interaction between *V. coralliilyticus* and juvenile *Acropora millepora* corals. Pretreatment of *V. coralliilyticus* cultures with YC successfully prevents bacterial-induced photoinactivation of symbiotic algae and coral tissue lysis. The ability of YC to attenuate bacterial disease likely depends on its high reported replication rate and burst size (Cohen 2013).

An essential step in the life cycle of phages is encapsulation of viral DNA into the capsid. Tailed phages accomplish this feat by using two different packaging strategies, termed headful and cohesive end packaging. The headful mechanism has been analyzed extensively in T4, whereby phage genomic DNA is pumped enzymatically into the phage prohead until full (Black and Silverman, 1978). The protein responsible for packaging the DNA is the large subunit of terminase (TerL), which serves as a motor enzyme (Oram et al., 2008; Sun et al., 2008). The large subunit of terminase binds a concatemer of phage DNA (i.e., a covalently linked end-to-end repetition of the phage genome) in the presence of a second protein, the small subunit of terminase (Rao and Feiss, 2015). Together, small and large subunits constitute the packaging holoenzyme. A double-stranded break is introduced by TerL at a packaging signal sequence in the DNA. The holoenzyme then docks with an awaiting phage

prohead and begins translocating the DNA (Casjens, 2011). Upon packing approximately 102% of the phage genome, TerL receives a 'headful' signal to make a double-stranded cut in the DNA and stop packaging (Dixit et al., 2013). The holoenzyme undocks from the nascent virion and neck proteins assemble on the capsid to retain the DNA packaged at near-crystalline density (Rao and Feiss, 2008).

Assembly of neck proteins on the virion facilitates the attachment of the tails and yields mature viral particles (Fokine and Rossmann, 2014). In phage T4, neck proteins gp13 and gp14 serve this function (Akhter et al., 2007; Hamilton and Luftig, 1972). A third protein, gp4, is required for head-tail attachment, but is not a structural component of the virion (Fokine et al., 2013; Granboulan et al., 1971). The mechanism of gp4 function during T4 morphogenesis is not yet known.

In this work, the genome of a lytic phage infecting *Vibrio coralliilyticus* was sequenced. Phage YC has been identified as a new member of T4-like phages and carries several open reading frames similar in sequence to bacterial genes, with most of the annotated genes related to T4-like myophages. Specifically, gp4 is conserved among T4-like phages. This gene contains a conserved catalytic motif found in type II restriction endonucleases. To confirm this prediction, *in vitro* biochemical assays demonstrate that both T4 and YC gp4 possess nuclease activity. The activity of gp4 may be required for the termination of genome packaging in T4-like phages.

Materials and Methods

Phage propagation and sequencing

High titer phage YC lysates (provided by Dr. David Bourne, Australian Institute of Marine Science) were prepared by infecting log-phase *Vibrio coralliilyticus* strain P1 grown

in marine broth with YC phage as described in (Cohen et al., 2013). Cleared lysates were centrifuged at 5000 g to remove bacterial debris, followed by chloroform treatment of the supernatant and filtration through a 0.22 µm pore size Sterivex filter (Millipore). The filtered YC lysate (20 mL) was subjected to cesium chloride gradient ultracentrifugation. The viral fraction corresponding to 1.5 g cm^{-3} buoyant density was retrieved (Thurber et al., 2009), treated with 100 U mL^{-1} of DNAse I for one hour at 37 °C, and inactivated for 10 minutes at 70 °C. PCR with universal 16S primers (amplifying variable regions 1-6) was performed to test for bacterial DNA contamination. Viral nucleic acid was extracted using the formamide/CTAB method, yielding 18.6 ng of viral DNA (Thurber et al., 2009). Illumina sequencing libraries were prepared and sequenced on a MiSeq instrument with 2 x 300 bp paired-end chemistry (Illumina, CA).

Genome assembly and annotation

Sequence reads were filtered using PrinSeq to remove duplicate reads, low complexity reads, low quality bases, poly-A/T tails and sequences less than 60 bp (Schmieder and Edwards, 2011). Contigs were assembled using SPAdes version 3.5 with default parameters, namely k-mer lengths of 21, 33, 55, 77, 99 and 127 (Bankevich et al., 2012). Reads were mapped onto the contigs using the Burrows-Wheeler Aligner (bwa-0.7.12) (Li and Durbin, 2009) and the resulting alignments were visualized using Tablet (v. 1.15) (Milne et al., 2013). BLASTn comparison of 5' and 3' contig ends identified one long circular contig that represented the complete YC phage genome. The YC phage genome was deposited in Genbank under accession number MH375644. Gene prediction was performed using GenemarkS (Besemer et al., 2001). ORF homologs were identified using BLASTp against the nr database (Benson et al., 2015) and RPS-BLAST against the Conserved Domain Database

(Marchler-Bauer et al., 2013). To uncover highly divergent homologs, ORFs were queried with HMMER v. 3.1 (Finn et al., 2011) against the Pfam-A database and subjected to Phyre2 structural analysis (Kelly et al., 2015). Additional phage structural proteins were identified using iVIREONS (Seguritan et al., 2012). The genome was graphically represented using CGView (Stothard and Wishart, 2005). To identify conservation of catalytic sites in gp4, selected gp4 homologs from T4-like phages were aligned to Tn7 transposase A protein using ClustalOmega with default parameters. Alignments were visualized using Jalview and WebLogo (Crooks et al., 2004; Waterhouse et al., 2009).

Phage proteomic tree

The phage proteomic tree was constructed as previously described (Rohwer and Edwards, 2002). Briefly, all phage proteins from the PhAnToMe database (n = 432,158) were compared against each other using BLASTp with an e-value cutoff of 0.1. Hits were aligned using ClustalW and protein distances calculated with PROTDIST (Felsenstein et al. 1990). The resulting distance matrix was used to construct a neighbor-joining tree.

Identification of YC in metagenomes

To uncover the geographic distribution of YC, publicly available metagenomes were competitively searched with the YC genome, as in (Dutilh et al., 2014). Briefly, a database of phage genomes was created by adding YC to all complete genomes from the PhAnToMe database (n = 2,996, http://www.phantome.org). All publicly available metagenomes from MG-RAST were then downloaded and used as queries against the phage database. Hits were considered if reads aligned with 85% identity over 50 bp.

Cloning of wild-type gene 4

Gene product 4 (gp4) from T4 and YC was PCR amplified using primers containing BamHI, NcoI and HindIII restriction enzyme sites (Table S4). Sticky ends of the amplicons were generated by restriction enzyme digestion, ligated into a modified pET vector and transformed into *E. coli* BL21 (DE3) pLysS (Table S3). The *E. coli* transformants were selected on LB plates containing 50 μg mL^{-1} kanamycin and 34 μg mL^{-1} chloramphenicol. Colonies were picked, plasmid DNA extracted using the alkaline lysis procedure, and inserts verified using Sanger sequencing (Retrogen, CA).

Isolation and cloning of mutations in gp4

Non-specific nucleases with high basal expression are toxic to *E. coli* and can produce mutations in the encoding gene (Davidson and Gold, 1992; Kuebler and Rao, 1998; Murialdo, 1988). This approach was used generate mutations gp4. The vector containing untagged wild-type T4 gp4 was transformed into *E. coli* BL21 (DE3) lacking the pLysS plasmid. Transformants were plated on LB plates containing 50 μg mL^{-1} kanamycin, selected colonies picked, plasmid DNA extracted and sequenced (Retrogen, CA). A mutated gp4 allele (T4 E29G) was isolated and subcloned into a vector encoding a N-terminal, hexahistidine-Smt tag for protein purification. A construct of YC gp4 containing a point mutation (YC E23G) was synthesized (IDT, CA) and cloned into the tagged vector.

Protein expression and purification

Proteins were expressed by inoculating engineered strains into LB broth containing selective antibiotics and incubated overnight at 37 °C. Overnight cultures were diluted at 1:100 and grown to an A_{600} of ~ 0.5. Protein expression was induced by adding of 0.5 mM IPTG and then transferring the culture to a 30 °C incubator. After three hours the cells were

harvested by centrifugation at 5000 x g for 15 minutes at 4 °C. Next, the cell pellet was re-suspended in a buffer containing 20 mM sodium phosphate, 500 mM NaCl, and 20 mM imidazole at pH 7.4. Re-suspended cells were sonicated, then centrifuged to separate soluble and insoluble fractions. The soluble fraction was purified using His SpinTrap columns (GE Healthcare) and eluted in buffer containing 20 mM sodium phosphate, 500 mM NaCl, and 500 mM imidazole at pH 7.4. The chromatography fractions were analyzed by SDS-PAGE on a 4-12 % bis-tris gel (ThermoFisher Scientific, CA).

***In vitro* nuclease assays**

The nuclease activity of gp4 proteins was assayed as described in (Ghosh-Kumar et al., 2011). Briefly, purified gp4 protein (350 ng, 10 pM) was incubated with DNA (100 ng Taq amplified dsDNA, 570 bp [0.3 pM] or 1,465 bp [0.1 pM]) in a reaction mixture of 10 µL containing 4 mM Tris-HCl at pH 7.8, 5 mM $MgCl_2$ and 6 mM NaCl for 5 - 60 minutes at 37 °C. The reaction was stopped by addition of EDTA to a final concentration of 50 mM. The nuclease-digested DNA products were analyzed by electrophoresis on a 1 % agarose gel followed by staining with ethidium bromide.

***In vivo* complementation assay**

A T4 phage containing amber mutations in gene 2 and gene 4 was obtained (T4 $gp2^-4^-$, provided by Dr. Dominique Belin) to assess the ability of plasmid-expressed gp4 to function *in trans*. Amber mutations in gp2 yield T4 progeny that are susceptible to *E. coli recBCD*-mediated DNA digestion (Appasani et al., 1999). Therefore, an *E. coli* MG1655 *recB* background was employed to suppress mutations in T4 gp2 (Table 2-3). *E. coli* MG1655 *recB* was transformed with plasmids expressing T4 and YC gp4 alleles. Overnight cultures of these strains were used as hosts to quantify viable particle production from infection with 10^8

PFU/mL of T4 gp2$^-$4$^-$ at an MOI of 0.1. Phage infectivity was determined by counting plaques on lawns in top-agar. Lysates of phage T4 gp2$^-$4$^-$ were further purified as described in (Bonilla et al., 2016). Phage titering was performed by diluting phage lysates and adding to 1 mL of bacteria and 3 mL LB top agar and enumerating plaques on LB plates after overnight incubation at 37 °C.

Transmission electron microscopy

Phage lysates of T4 gp4$^-$2$^-$ were grown on cultures of *E. coli* supD or *E. coli* MG1655 *recB* hosts expressing gp4 *in trans* (Table 2-3). Phages were stained for transmission electron microscopy as described in (Bonilla et al., 2016). Briefly, glow-discharged 300-mesh copper grids coated with formvar were overlaid with drops (30 µl) of purified phage samples (0.2 mg/ml) for 3 minutes. To remove the salts from the buffer, the grids were rinsed 3 times with 20 µl drops of water. Then, grids were negatively stained with uranyl acetate (0.5 %) for 15 seconds, dried, and examined using a FEI Tecnai T12 TEM (FEI, Hillsboro, OR) at the SDSU Electron Microscopy Facility, operating at 120 kV. Micrographs were taken with an AMT HX41 side mounted digital camera (Advanced Microscopy Technique, Woburn, MA).

Results

Bacteriophage YC is a novel member of T4-like bacteriophages

Vibrio coralliilyticus is an opportunistic pathogen of the scleractinian coral *Pocillipora damicornis* (Ben-Haim and Rosenberg, 2002) and is susceptible to infection by a lytic phage named YC (Cohen et al., 2013). To better understand how YC attenuates *V. coralliilyticus* colonization of *P. damicornis* (Cohen et al., 2013), we sequenced phage YC DNA and obtained a circular contig of 147,890 bp with an average coverage of 759X (Figure 2-1). The sequence has been submitted to GenBank (MH375644).

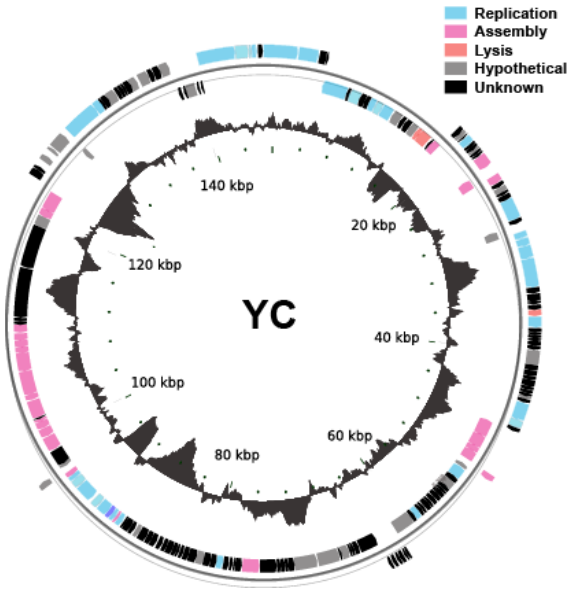

Figure 2-1: The YC genome. A) ORFs are represented as arrows and color coded according to their annotated function. Blue, replication; pink, assembly; red, lysis; grey, conserved hypothetical; black, unknown. The inner ring plots the deviation from average GC content across the entire genome, scaled from minimum to maximum.

Using a genome-based taxonomy approach, YC clustered with T4-like phages (Figure 2-5) (Rohwer and Edwards, 2002). Gene prediction using GenemarkS software identified 195 potential open reading frames (ORFs) greater than 139 nt, of which only about 35 % could be assigned a predicted function (Table 2-1). Some of the annotated ORFs have similar sequence to genes from *Vibrio* spp. as determined by BLASTp searches against the nr database. For example, an ORF encoding thymidylate synthase was most closely related to *Vibrio crassostreae*. An additional four ORFs in the YC genome are most closely related to genes found in *V. parahaemolyticus, V. vulnificus, V. ER1A* and *V. cholera*. ORFs from bacterial species other than the genus *Vibrio* were also identified. These encode a GroEL/ES protein chaperone (ORFs 109 and 134, respectively), the DNA binding protein IHF (ORF 169) and RNA polymerase sigma factor D (ORF 170) (Table 2-1).

The presence of YC in the environment was determined by searching for metagenomic reads that align uniquely to the YC genome. Out of all publicly available metagenomes on the MG-RAST database (Meyer et al., 2008), seven metagenomes contain reads that are greater than 85 % identical over a length of at least 50 bp (Table 2-2). These metagenomes are from terrestrial soils, marine and freshwater biomes. Given the paucity of reads mapping to YC, this phage is a rare member of the virosphere.

Gene 4 is a putative nuclease found in T4-like phages

Gene 4 (gp4) is a gene conserved in both YC and T4 that mediates the attachment of phage heads to tails (Edgar and Wood, 1966). Bioinformatic examination of this gene identifies a region with sequence similarity to the transposase A gene of Tn7 (Conserved Domain Database expect value of 10^{-45}). Tn7 is a mobile genetic element that transposes via a cut and paste mechanism (Bainton et al., 1991). The mechanism involves transposase proteins A and B, which together mediate DNA cleavage at the 5' and 3' end of the transposon, respectively (Sarnovsky et al., 1996). Transposase A (*tnsA*) contains a catalytic triad of one basic and two acidic residues that form a restriction endonuclease-like active site (Hickman et al., 2000). Alignment of *tnsA* with gp4 proteins from T4-like phages indicates conservation of these catalytic residues (Figure 2-2).

Figure 2-1: A) Alignment of T4 and YC gp4 proteins with Tn7 transposase A (*tnsA*). Experimentally determined catalytic residues of *tnsA* are shaded in grey. B) Alignment of 50 gp4 proteins from T4-like phages represented as a sequence logo format, with letter heights proportional to the degree of conservation (Schneider and Stephens, 1990). The region corresponding to residues 25-93 of T4 gp4 are shown, with the predicted catalytic residues highlighted in red.

To test whether gp4 possesses nuclease activity, T4 and YC gp4 were cloned into an expression vector and the proteins purified using affinity chromatography (Figure 2-6). Incubation of purified T4 gp4 protein with a 570 bp PCR fragment of dsDNA resulted in a smear of fragments on 1 % agarose gels (Figure 2-3). Similar results were observed with purified YC gp4 protein, with greater than 70 % of digestion occurring within 20 minutes (Figure 2-3). This assay demonstrates that gp4 functions as a nuclease *in vitro*.

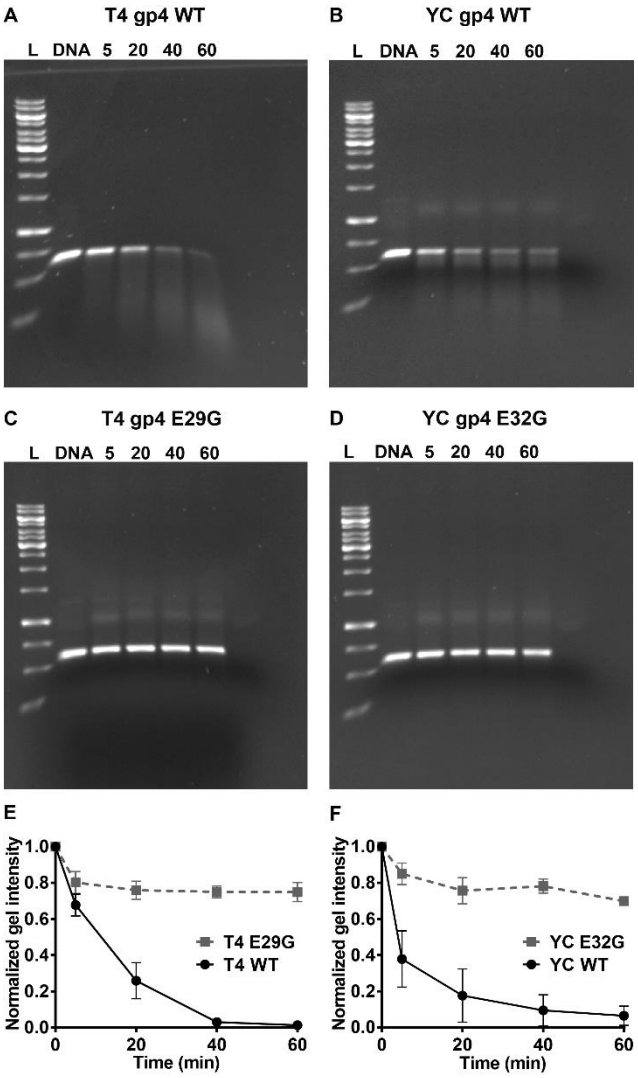

Figure 2-2: Purified gp4 digests DNA *in vitro*. Time course of DNA cleavage by wild-type T4 gp4 (A), wild-type YC gp4 (B), T4 E29G (C), or YC E32G (D). Gp4 protein (350 ng, 10 pM) was incubated with dsDNA (100 ng, 0.3 pM) for 5 – 60 minutes. Reactions were stopped with 50 mM EDTA and electrophoresed on a 1 % agarose gel. Lane 1, 10 kb ladder (L); Lane 2, DNA only control. Lanes 3-5, incubation time (minutes). The amount of undigested DNA was quantified by laser densitometry in each lane (panels E, F) (n = 3).

To exclude the possibility that the nuclease activity observed was due to contaminating cellular nucleases, we engineered a mutation in the predicted active site of gp4 and tested the activity of the resulting mutated protein. A previously established method to screen for mutations in phage genome packaging nucleases was leveraged to mutate residues

necessary for gp4 functional activity (Davidson and Gold, 1992; Kuebler and Rao, 1998; Murialdo, 1988). Transformation of a plasmid encoding T4 gp4 into a host with high basal expression (*E. coli* BL21 DE3) resulted in lower transformation efficiency, reduced growth and poor recovery of plasmid DNA (data not shown). Surviving transformed colonies were picked and the inserts sequenced (Retrogen, CA). We suspected that these colonies had gained mutations that abrogated the nuclease activity. An insert with a single point mutation causing a glutamate to glycine substitution at position 29 (T4 gp4 E29G) was recovered. The substituted amino acid replaces the conserved glutamate residue that aligns with the *tnsA* catalytic motif. The mutated T4 gp4 was purified and the protein incubated with dsDNA. DNA digestion by T4 gp4 E29G is attenuated compared to digestion by the wild type T4 gp4 (Figure 2-3). The corresponding glutamate to glycine mutation was introduced into YC gp4 (YC gp4 E32G). Similarly, this mutation of the YC gp4 glutamate residue reduced nuclease activity (Figure 2-3). These experiments verify the presence of a conserved nuclease fold in gp4 and demonstrate that mutating a single catalytic residue reduces enzymatic activity.

To address the possibility that DNA cleavage by gp4 is sequence-specific, the *in vitro* nuclease assays were repeated with an alternative dsDNA fragment. In this assay, a 16S rDNA PCR amplicon was used as substrate for digestion. As above, gp4-cleaved DNA migrated as a smear throughout the lane (Figure 2-6). Smearing of DNA is indicative of random cleavage (Prasad Bhattacharyya and Basaveswara Rao, 1993), suggesting gp4 cleaves DNA non-specifically.

Complementation of amber mutant T4 by gp4 *in trans*

Phage T4 bearing unsuppressed amber mutations in gp4 yield non-viable progeny with detached heads unfilled with DNA and tails (Edgar and Wood, 1966; Granboulan et al.,

1971). To test whether gp4 can complement defects in head-tail joining, host bacteria expressing gp4 *in trans* were infected with a T4 gp4^{-}2^{-} double amber mutant. T4 gp2 binds double-stranded DNA termini to protect the phage genome from *E. coli recBCD* mediated digestion (Appasani et al., 1999). Mutations in *recB* suppress the gp2^{-} amber mutation. Therefore, an *E. coli recB* background was used for transformation with gp4-expressing plasmids. These strains were used as hosts to quantify viable particle production upon infection with T4 gp4^{-}2^{-} phage. Equivalent titers of phages were produced from the amber mutant-suppressing host *E. coli* SupD as well as *E. coli recB* expressing wild-type T4 gp4 (degrees of freedom = 6, unpaired t-test, $p = 0.34$) (Figure 2-4). In contrast, *E. coli recB* hosts expressing gp4 with a non-synonymous point mutation in the catalytic site produce significantly fewer phage plaques (degrees of freedom = 7, unpaired t-test, $p = 0.004$). This result indicates that T4 gp4 can function *in trans* to rescue T4 amber mutants, dependent on wild type nuclease activity.

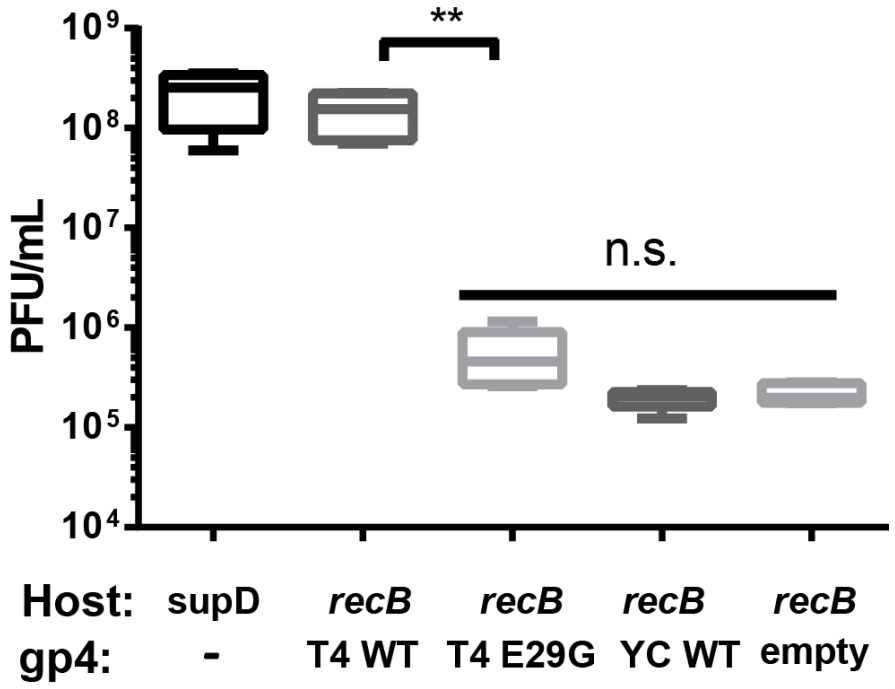

Figure 2-3: Complementation of amber mutant phage T4 $4^{-}2^{-}$ by gp4 *in trans*. *E. coli* strains with plasmid-expressed gp4 were infected with phage T4 $4^{-}2^{-}$ and plaques enumerated on top-agar. Plasmids expressing wild-type T4 gp4 (T4 gp4 WT), the T4 catalytic mutant (T4 gp4 E29G) or wild-type YC gp4 (YC gp4 WT). An amber mutation suppressing strain (SupD) or plasmid lacking any insert (empty) were included as positive and negative controls, respectively. Note that the *E. coli* host possesses a nonfunctional recB and suppresses mutations in T4 gp2.

The ability for YC gp4 to complement T4 gp4 was also tested. As above, YC gp4 was cloned into an expression vector and transformed into *E. coli recB*. In contrast to wild-type T4 gp4, expression of YC gp4 produced comparable numbers of phages as a plasmid backbone control (degrees of freedom = 8, unpaired t-test, $p = 0.46$) (Figure 2-4). In this experiment, YC gp4 did not complement the activity of T4 gp4.

To determine whether the same head-tail joining defect occurs in the presence of mutated or heterologous gp4, phage lysates were examined using transmission electron microscopy. Amber mutant phage lysates propagated on *E. coli* SupD or *E. coli recB*

harboring T4 gp4 WT have fully mature phage particles (Figure 2-8). *E. coli recB* expressing T4 gp4 E29G or YC gp4 WT yielded phage particles with empty heads and detached tails (Figure 2-8). The defective head-tail joining phenotype can thus be attributed to the lack of nuclease activity by gp4.

Discussion

The genome of YC represents another example of the mosaic nature of marine phages (Hendrix et al., 1999; Rohwer et al., 2000). The gp4 nuclease from YC cannot complement the defective gp4 in T4 $4^{-}2^{-}$ (Figure 2-4). The lack of complementation is probably not attributable to sequence specificity by gp4, as both proteins cleaved DNA non-specifically *in vitro* (Figure 2-3 and Figure 2-7). The overall shared amino acid identity between T4 gp4 and YC gp4 is low (36 %) compared to the ORF with the highest similarity to YC gp4 (58 %, Table 2-1). In aggregate, our data suggest that regions of gp4 flanking the conserved catalytic site are required for protein-protein interactions with other components of the packaging holoenzyme.

Amber mutants of gp4 cause defects in the joining of tails to heads in T4 and the progeny are non-viable (Granboulan et al., 1971). Based on this observation it was suggested that gp4 is a structural protein; however, an earlier study predicted gp4 to be an enzyme (Snustad, 1968). Moreover, structural analysis of the T4 virion demonstrates that gp4 is not part of the phage particle (Fokine et al., 2013). Alignment of YC gp4 with homologs from T4-like phages identifies a conserved nuclease catalytic motif consisting of two acidic and one basic residue. These residues match the canonical catalytic site found in type II restriction endonucleases (Kovall and Matthews, 1999) and in Tn7 transposase A (Hickman et al., 2000). The experiments presented here show that purified gp4 possesses nuclease activity,

confirming the enzymatic nature of gp4 and ascribing the activity as DNA cleavage, although the protein may have an additional role in scaffolding the packaging apparatus.

The nuclease function of gp4 could be required for the termination of genome packaging. Phage DNA is recognized by the packaging holoenzyme and a double stranded break is introduced into the DNA concatemer by TerL. The large subunit of terminase then acts as a molecular motor, using energy from ATP hydrolysis to translocate phage DNA into the phage capsid (Black, 2015). When the phage capsid is packed tightly with just over one copy of the phage genome, TerL is triggered to execute a terminating cleavage cut (Rao and Feiss, 2015). How terminase reorients itself to switch from a motor enzyme back into an endonuclease is not well understood. Gp4 may interact with the large terminase subunit, portal and/or other proteins to form an oligomeric complex that executes the terminating cleavage event. Subsequent studies identifying binding partners of gp4 may yield additional information regarding genome packaging in T4-like phages.

Acknowledgements

Chapter 2, in full, has been submitted for publication. Sean Benler, Shr-Hau Hung, Yan Wei Lim, Katelyn McNair, Jeremy Barr, Forest Rohwer, and Anca Segall. 2018. The dissertation author was the primary investigator and author of this paper.

Appendix

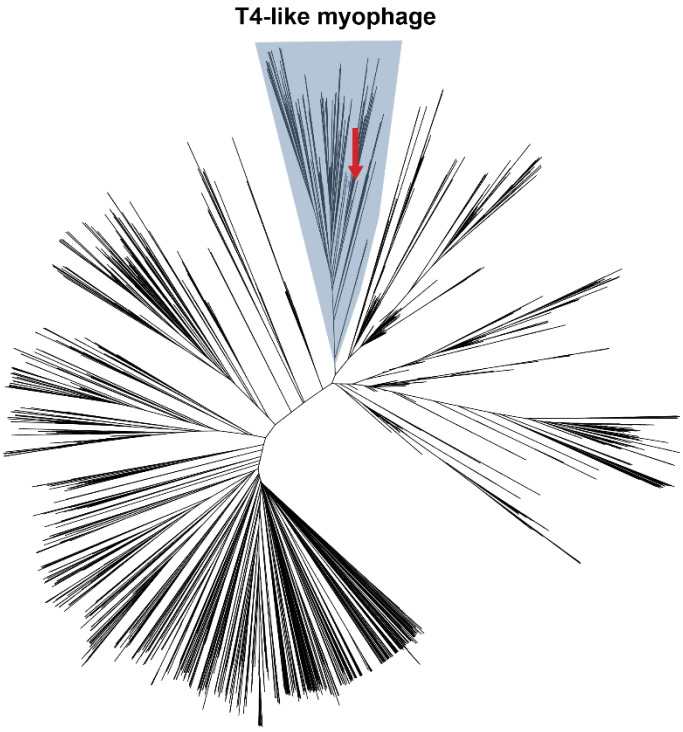

Figure 2-4: The phage proteomic tree of 2,997 phages, including YC. YC is indicated with an arrow in red and the T4-like branch is shaded in blue. Other branches are not labeled for clarity.

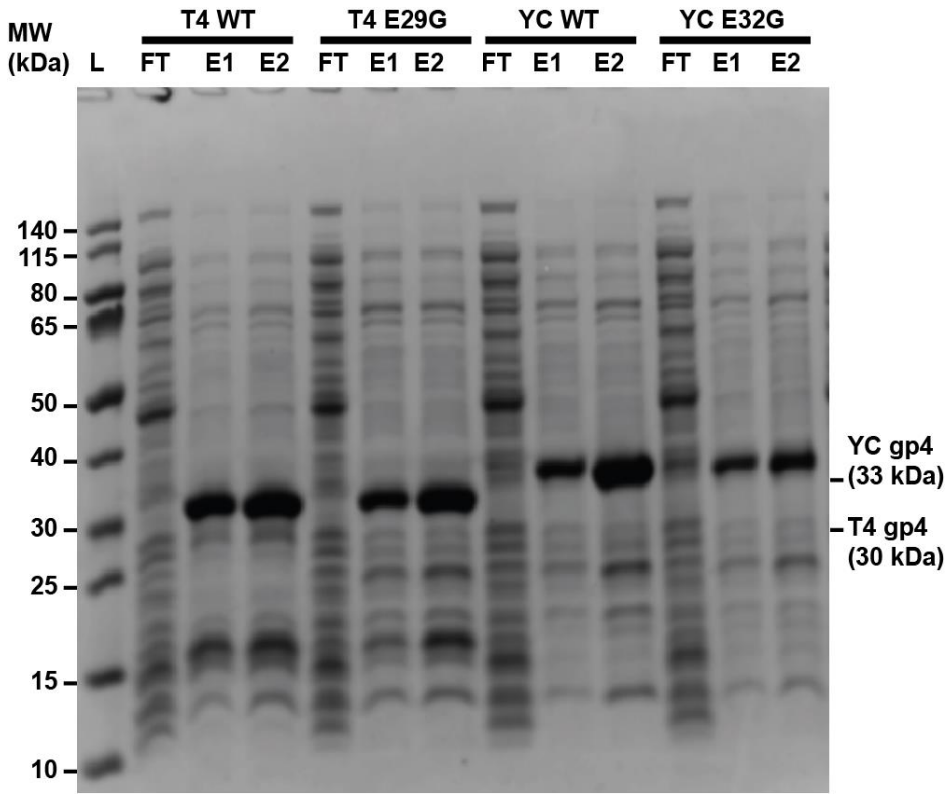

Figure 2-6: SDS-PAGE of chromatography fractions. Five nanograms of total protein in each fraction was loaded onto a 4–12 % bis-tris gel (ThermoFisher Scientific, CA). For each protein indicated, the flow-through (FT), elution 1 (E1) and elution 2 (E2) fractions were electrophoresed using MOPS running buffer. The gel was stained using Coomassie blue and the amount of gp4 protein relative to total protein was estimated using laser densitometry. The expected sizes of the ladder (L) and tagged gp4 proteins are labeled for clarity. Note that the tagged gp4 proteins migrate less than expected.

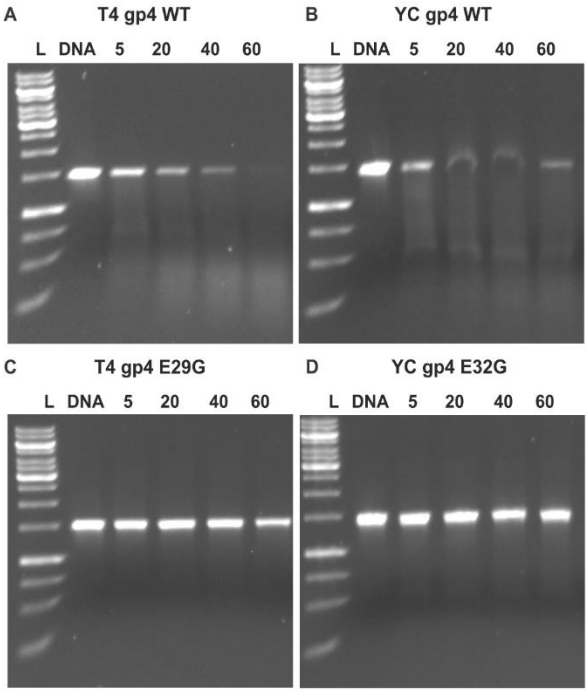

Figure 2-7: Digestion of 16S rDNA PCR amplicons by gp4. Time course of DNA cleavage by wild-type T4 gp4 (A), wild-type YC gp4 (B), T4 E29G gp4 (C), or YC E32G gp4 (D). Gp4 protein (350 ng, 10 pM) was incubated with a 16S rDNA PCR amplicon (100 ng, 0.1 pM) for 5 – 60 minutes. Reactions were stopped with 50 mM EDTA and electrophoresed on a 1 % agarose gel. Lane 1, 10 kb ladder (L); Lane 2, DNA only control. Lanes 3-5, incubation time (minutes).

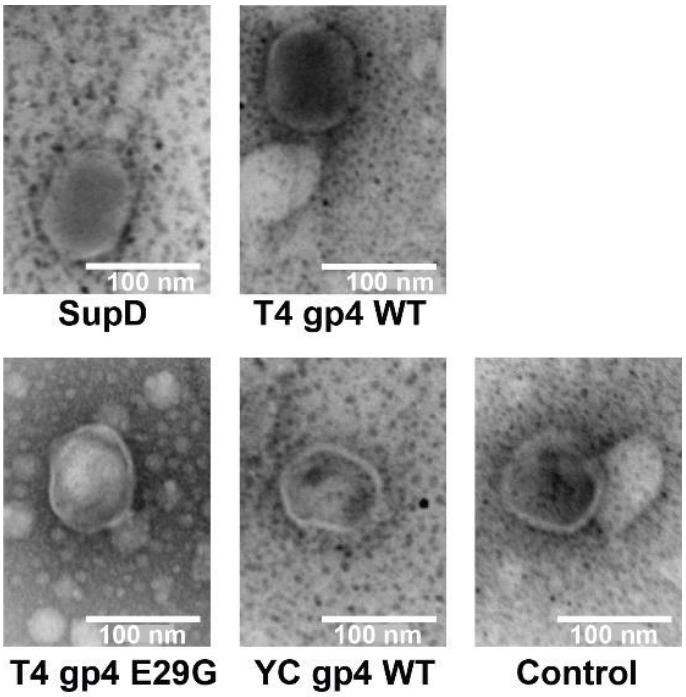

Figure 2-5: Transmission electron micrographs of T4 $4^{-}2^{-}$ progeny phages grown on *E. coli* host bacteria harboring different gp4 alleles.The genotype of the *E. coli* host is listed under each image, including *E. coli* suppressor D (SupD), *E. coli recB* T4 gp4 WT, *E. coli recB* T4 gp4 E29G, *E. coli recB* YC gp4 WT, or *E. coli recB* plasmid only (control). Scale bar = 100 nm.

Table 2-1: Open reading frames (ORFs) in the YC genome. The highest-ranking hit revealed by BLASTp, RPS-BLAST, or Phyre2 analysis is given for the appropriate ORF. ORFs with no annotation are not listed for clarity. The taxonomic information of the top hit is also provided, along with percentage of amino acid identity and expect value. [a]Function of ORF determined using RPS-BLAST against the NCBI Conserved Domain Database. Organism and percent identity determined using BLASTp against the NCBI nr database. [b]Function of ORF determined using Phyre2 protein fold recognition server. Organism, E-value (where applicable), and percent identity determined using BLASTp against the NCBI nr database.

ORF	*Strand*	*Coordinates*		*Significant hit*	*Organism*	*e-value*	*% id (shared aa)*
1	+	2	2989	RIIA	*Erwinia* phage phiEa2809	7.00E-72	40
2	+	3062	4819	RIIB	*Klebsiella* phage 0507-KN2-1	3.00E-58	41
5	-	5728	8148	DNA primase-helicase	*Klebsiella* phage 0507-KN2-1	7.00E-77	49
7[a]	-	8452	9525	RecA	*Salmonella* phage SKML-39	1.25E-06	60
10	-	10516	11136	dUTPase	*Salmonella* phage PhiSH19	2.00E-16	32
11[b]	-	11145	11807	Nucleoside triphosphate hydrolase	*Klebsiella* phage 0507-KN2-1	4.00E-28	39
12	-	11812	12945	Thymidylate synthase	*Vibrio crassostreae*	3.00E-76	41
13	-	13076	13915	Conserved hypothetical	*Salmonella* phage SKML-39	1.00E-05	27
16	-	15096	15770	Conserved hypothetical	*Vibrio parahaemolyticus*	5.00E-35	34
17[a]	-	15833	16222	Lysozyme, gp25	*Salmonella* phage vB_SalM_SJ2	3.46E-12	43
18[a]	-	16215	17279	Tail associated lysozyme, gp5	*Dickeya* phage RC-2014	2.70E-04	48
20	-	17682	18404	Baseplate hub subunit	*Serratia* phage phiMAM1	6.00E-20	23
22	+	18940	19287	Conserved hypothetical	*Vibrio vulnificus*	7.00E-06	48
24	+	19526	20188	RuvC	*Erwinia* phage phiEa2809	1.00E-25	35
27	+	20767	21015	Conserved hypothetical	*Erwinia* phage phiEa2809	6.00E-13	39
30[a]	+	21545	22615	ssDNA binding protein, gp32	*Salmonella* phage SKML-39	1.62E-55	45
31	-	22652	23560	tail tube, gp19	*Salmonella* phage ViI	8.00E-41	30
32[a]	+	23615	24289	DNA end protector, gp2	*Serratia* phage phiMAM1	1.82E-36	48
34	+	24759	25232	Conserved hypothetical	*Serratia* phage phiMAM1	2.00E-38	45

Table 2-1: Open reading frames in the YC phage genome, continued.

ORF	*Strand*	*Coordinates*		*Significant hit*	*Organism*	*e-value*	*% id (shared aa)*
37	+	25983	27818	DNA helicase	*Erwinia* phage phiEa2809	1.00E-118	37
39	-	28197	28958	Conserved hypothetical	*Salmonella* phage PhiSH19	9.00E-23	29
40	+	28993	29595	Ribonuclease H	*Dickeya* phage Limestone	5.00E-28	43
41[a]	+	29607	30281	RNA polymerase binding protein, gp55	*Salmonella* phage Maynard	4.42E-13	39
42[a]	+	30274	31401	DNA endonuclease, gp47	*Salmonella* phage Marshall	6.92E-65	48
43[b]	+	31398	33926	DNA endonuclease, recN	*Serratia* phage phiMAM1	3.00E-138	35
48	+	35880	36395	Lysozyme	*Brevundimonas diminuta*	3.00E-38	46
49[b]	+	36447	37412	Adenine alpha hydrolase	*Ralstonia* phage RSL2	6.00E-10	26
56	+	39393	40628	Conserved hypothetical	*Pseudomonas mendocina S5.2*	1.00E-23	27
66[a]	+	44017	45579	DNA ligase, gp30	*Klebsiella* phage 0507-KN2-1	2.80E-60	34
67[a]	+	45583	46257	DNA helicase loader, gp59	*Erwinia* phage phiEa2809	5.82E-15	31
69	-	46590	48125	Tail length tape measure, gp29	*Erwinia* phage phiEa2809	9.00E-20	30
70[b]	-	48118	49320	Baseplate hub, gp27	*Shigella* phage Ag3	4.00E-28	26
71[a]	-	49313	49858	Baseplate wedge, gp53	*Erwinia* phage phiEa2809	1.55E-16	40
72	-	49858	50808	Baseplate cap, gp48	*Klebsiella* phage 0507-KN2-1	4.00E-80	44
73	+	50852	51379	Head completion, gp4	*Serratia* phage phiMAM1	2.00E-60	58
75[a]	-	51806	52885	DNA primase, gp61	*Serratia* phage phiMAM1	7.01E-82	46
77	-	53382	54119	Conserved hypothetical	*Salmonella* phage SKML-39	7.00E-34	36
87	-	57710	58288	Conserved hypothetical	*Vibrio sp. ER1A*	1.00E-25	40
89	-	58669	60567	Conserved hypothetical	*Salmonella* phage ViI	3.00E-16	28
96	-	62674	64422	Conserved hypothetical	*Escherichia* phage ECML-4	2.00E-31	41
100	-	65531	66265	Conserved hypothetical	*Proteobacteria bacterium JGI 0000113-L05*	3.00E-07	27
102	-	66537	68630	Conserved hypothetical	*Klebsiella* phage 0507-KN2-1	2.00E-23	38

Table 2-1: Open reading frames in the YC phage genome, continued.

ORF	*Strand*	*Coordinates*		*Significant hit*	*Organism*	*e-value*	*% id (shared aa)*
103[b]	-	68747	70957	Neck appendage protein	*Vibrio cholerae*	7.00E-13	29
109	-	74355	75989	Chaperone GroEL	*Methylophaga aminisulfidivorans*	1.00E-142	44
114	-	77872	78522	Exonuclease	*Serratia* phage phiMAM1	3.00E-48	44
117	-	79783	80520	Conserved hypothetical	*Salmonella* phage PVP-SE1	8.00E-22	36
129	-	86547	86972	Conserved hypothetical	*Salmonella* phage vB_SalM_SJ2	3.00E-05	28
133	-	88489	89013	NTP pyrophosphohydrolase	*Erwinia* phage phiEa2809	1.00E-40	44
134	-	89015	89287	Chaperone GroES	*Cardinium cBtQ1*	2.00E-04	34
135	-	89296	89727	Translation repressor regA	*Synechococcus* phage Syn19	4.47E-07	29
136	-	89724	90176	DNA polymerase clamp subunit loader, gp62	*Erwinia* phage phiEa2809	8.25E-11	34
137	-	90184	91185	DNA polymerase sliding clamp holder, gp44	*Klebsiella* phage 0507-KN2-1	6.10E-89	50
138	-	91257	91946	DNA polymerase sliding clamp, gp45	*Dickeya* phage Limestone	3.00E-31	34
139	-	92224	93828	UvsW	*Serratia* phage phiMAM1	9.00E-153	45
140	-	93871	94614	Exonuclease	*Serratia* phage phiMAM1	4.00E-59	47
141	-	94614	95102	UvsY	*Dickeya* phage Limestone	1.00E-26	38
142[b]	-	95095	95577	Tail tube, gp19	*Escherichia* phage PhaxI	6.62E-05	36
143	+	95672	96196	Conserved hypothetical	*Dickeya* phage Limestone	3.00E-06	25
144	-	96193	96528	Conserved hypothetical	*Erwinia* phage Ea9-2	5.00E-19	46
145	-	96515	98029	Homing endonuclease F-LimVII	*Dickeya* phage Limestone	4.00E-49	30
146	-	98143	99486	Major capsid, gp23	*Dickeya* phage Limestone	1.13E-91	65
147	-	99571	100377	Prohead core scaffold, gp22	*Serratia* phage phiMAM1	3.03E-32	38
148	-	100426	101094	Prohead core protease, gp21	*Erwinia* phage phiEa2809	2.00E-77	56
149	-	101097	101378	Prohead core, gp68	*Acinetobacter* phage 133	1.07E-06	37
151	-	101573	103273	Portal vertex, gp20	*Salmonella* phage Marshall	3.71E-133	47

Table 2-1: Open reading frames in the YC phage genome, continued.

ORF	*Strand*	*Coordinates*		*Significant hit*	*Organism*	*e-value*	*% id (shared aa)*
152[a]	-	103372	103902	Tail tube, gp19	*Salmonella* phage SKML-39	2.89E-18	47
153[a]	-	103915	106185	Tail sheath, gp18	*Klebsiella* phage 0507-KN2-1	2.27E-75	56
154[b]	-	106220	107971	Large terminase, gp17	*Erwinia* phage phiEa2809	N/A	56
155	-	107975	108571	Small terminase, gp16	*Klebsiella* phage 0507-KN2-1	6.00E-12	32
156	-	108564	109220	Tail sheath stabilization, gp15	*Enterobacteria* phage vB_EcoM-VR7	6.00E-12	33
157[a]	-	109217	109882	Neck, gp14	*Klebsiella* phage 0507-KN2-1	3.60E-32	33
158[a]	-	109967	110719	Neck, gp13	*Erwinia* phage phiEa2809	5.00E-33	43
163	-	120416	121438	Conserved hypothetical	*Fusobacterium varium*	7.00E-07	42
164	-	121470	122180	Baseplate subunit	*Serratia* phage phiMAM1	3.00E-31	34
165[a]	-	122164	123891	Baseplate, gp6	*Klebsiella* phage 0507-KN2-1	7.38E-99	34
168	+	125832	126335	Conserved hypothetical	*Psychromonas ossibalaenae*	4.00E-29	38
169	+	126874	127149	Integration Host Factor	*Novispirillum itersonii*	6.00E-13	41
170	+	127276	128634	rpoD	*Caldicellulosiruptor owensensis*	8.00E-25	29
171	-	128672	129124	Conserved hypothetical	*Gilvimarinus chinensis*	4.00E-06	30
172[a]	+	129184	132195	DNA PolB	*Escherichia* phage Cba120	2.02E-44	46
173[b]	+	132236	132868	5'-3' exonuclease	*Erwinia* phage phiEa2809	1.00E-10	32
176	+	133580	134383	Conserved hypothetical	*Serratia* phage phiMAM1	1.00E-53	41
181	+	135871	136488	Conserved hypothetical	*Erwinia* phage Ea35-70	1.00E-09	35
182	+	136741	137601	Conserved hypothetical	*Pseudomonas fluorescens*	2.00E-19	29
185	+	138734	139516	PhoH	*Salmonella* phage ViI	1.00E-70	46
188	-	140276	141217	Conserved hypothetical	*Pseudomonas* phage PhiPA3	1.00E-09	29
191[a]	+	142124	145435	Ribnoucleotide reductase subunit I	*Dickeya* phage Limestone	4.21E-45	51

Table 2-1: Open reading frames in the YC phage genome, continued.

ORF	*Strand*	*Coordinates*		*Significant hit*	*Organism*	*e-value*	*% id (shared aa)*
192[a]	+	145445	146569	Ribonucleotide reductase subunit II	*Serratia* phage phiMAM1	9.71E-59	52
193	+	146632	146868	Glutaredoxin	*Synechococcus* phage S-PM2	2.00E-09	48
194[a]	+	146868	147323	Glutaredoxin	*Dyella jiangningensis*	2.40E-05	31

Table 2-2: Metagenomes containing reads with hits to the YC genome. Metagenomes were used as queries against a database of all phage genomes. Shown are reads that align uniquely to the YC phage genome with at least 85 % nucleotide identity over >50bp.

MG-RAST metagenome ID	*query start*	*query end*	*Length (nt)*	*e-value*	*bit score*	*% identical*	*biome*	*sample type*
4541641.3	83	139	57	2.00E-14	78.7	91.23	grassland biome	soil
4519757.3	21	79	59	4.00E-14	76.8	89.83	terrestrial biome	soil
4573678.3	69	136	69	6.00E-12	71.3	85.51	Temperate grasslands	soil
4441592.3	690	749	60	8.00E-12	73.1	88.33	marine habitat	water
4472770.3	1	59	59	1.00E-10	65.8	86.44	terrestrial biome	humanfeces
4539063.3	15	68	54	2.00E-09	62.1	87.04	grassland biome	soil
4507669.3	315	368	54	5.00E-07	56.5	85.19	freshwater biome	wastewater

Table 2-3: Bacterial strains, phages and plasmids used in the study.

Strain, phage, or plasmid	*Genotype*	*Source or reference*
Vibrio coralliilyticus P1 (LMG23696)	Wild type	Cohen et al. (2013)
E. coli strains		
B	*sup*^{0}(Su$^-$) (ExoV+)	Lab strain
SupD	*sup*^{1}(Su$^+$) *serU*	Lab strain
BL21(DE3)	F- *ompT hsdS* (r_b^- m_b^-) *gal dcm* (DE3)	A. Segall
BL21(DE3) pEMB31 T4 gp4 WT	-	This study
BL21(DE3) pEMB31 T4 gp4 E29G	-	This study
BL21(DE3) pLysS	F- *ompT hsdS* (r_b^- m_b^-) *gal dcm* (DE3) pLysS (Cam^R)	A. Segall
BL21(DE3) pLysS pEMB31 T4 gp4 WT	-	This study
BL21(DE3) pLysS pEMB31 T4 gp4 E29G	-	This study
BL21(DE3) pLysS pEMB31 YC gp4 WT	-	This study
BL21(DE3) pLysS pEMB31 YC gp4 E32G	-	This study
MG1655 *recB*	MG1655 *recB268::*tn10 Tet CG	A. Segall
MG1655 *recB* pEMB31	-	This study
MG1655 *recB* pEMB31 T4 gp4 WT	-	This study
MG1655 *recB* pEMB31 T4 gp4 E29G	-	This study
MG1655 *recB* pEMB31 YC gp4 WT	-	This study
Phage		
T4 WT	Wild type	Lab Strain
T4 2$^-$amN51 4$^-$amN112	Amber mutation in gene 2 and 4	Dominique Belin
YC	Wild type	Cohen et al. (2013)
Plasmid		
pEMB31	pET derived, Kan^R	A. Segall
pEMB31 T4 gp4 WT	Wild type T4 gp4	This study
pEMB31 T4 gp4 E29G	E29G substitution	This study
pEMB31 YC gp4 WT	Wild type YC gp4	This study
pEMB31 YC gp4 E32G	E32G substitution	This study

Table 2-4: Primers used in the study (written 5’ to 3’). Lower case letters represent spacers used to assist in restriction enzyme binding.

Primer name	**Sequence**
T4 gp4 forward	5’tcaaccttgatcgttatCCATGGCATATTCTGGAAAATGGGTTCCTAAAAATATATCAAAG
T4 gp4 forward, histidine tag	5’tcaaccttgatcgttatGGATCCGCATATTCTGGAAAATGGGTTCCTAAAAATATATCAAAG
T4 gp4 reverse	5’tgcctatgccttaagcgcAAGCTTATGCCCCCTTAAAGCCAAGAGCTC
YC gp4 forward	5’tcaaccttgatcgttatCCATGGATGACACACAACTACAAGCAGGGATTGTACAAGGTAC
YC gp4 forward, histidine tag	5’ccttgatcgttatGGATCCACACACAACTACAAGCAGGGATTGTACAAGGTACAGAACCC
YC gp4 reverse	5’gcctatgccttaagcgcAAGCTTCTAACCGCCGCTCTTTTTCATTGCTTTCTTACGCGCC
16S (27F)	5’AGAGTTTGATCTGGCTCAG
16S (1492R)	5’ GGTTACCTTGTTACGACTT

References

Akhter, T., Zhao, L., Kohda, A., Mio, K., Kanamaru, S., Arisaka, F., 2007. The neck of bacteriophage T4 is a ring-like structure formed by a hetero-oligomer of gp13 and gp14. Biochim. Biophys. Acta - Proteins Proteomics 1774, 1036–1043. https://doi.org/10.1016/j.bbapap.2007.05.011

Appasani, K., Thaler, D.S., Goldberg, E.B., 1999. Bacteriophage T4 gp2 interferes with cell viability and with bacteriophage lambda red recombination. J. Bacteriol. 181, 1352–1355.

Arias, C.R., Garay, E., Aznar, R., 1995. Nested PCR method for rapid and sensitive detection of Vibrio vulnificus in fish, sediments, and water. Appl. Environ. Microbiol. 61, 3476–3478.

Bainton, R., Gamas, P., Craig, N.L., 1991. Tn7 transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA. Cell 65, 805–816. https://doi.org/10.1016/0092-8674(91)90388-F

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., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M. a, Pevzner, P. a, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–77. https://doi.org/10.1089/cmb.2012.0021

Ben-Haim, Y., Rosenberg, E., 2002. A novel Vibrio sp. pathogen of the coral Pocillopora damicornis. Mar. Biol. 141, 47–55. https://doi.org/10.1007/s00227-002-0797-6

Ben-Haim, Y., Thompson, F.L., Thompson, C.C., Cnockaert, M.C., Hoste, B., Swings, J., Rosenberg, E., 2003. Vibrio coralliilyticus sp. nov., a temperature-dependent pathogen of the coral Pocillopora damicornis. Int. J. Syst. Evol. Microbiol. 309–315. https://doi.org/10.1099/ijs.0.02402-0

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2015. GenBank. Nucleic Acids Res. 43, D30–D35. https://doi.org/10.1093/nar/gku1216

Besemer, J., Lomsadze, A., Borodovsky, M., 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29, 2607–2618. https://doi.org/11410670

Black, L.W., 2015. Old, new, and widely true: The bacteriophage T4 DNA packaging mechanism. Virology 480, 650–656. https://doi.org/10.1016/j.virol.2015.01.015

Black, L.W., Silverman, D.J., 1978. Model for DNA packaging into bacteriophage T4 heads. J. Virol. 28, 643–655.

Bonilla, N., Rojas, M.I., Netto Flores Cruz, G., Hung, S., Rohwer, F., Barr, J.J., 2016. Phage on tap–a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. PeerJ 4, e2261. https://doi.org/10.7717/peerj.2261

Casjens, S.R., 2011. The DNA-packaging nanomotor of tailed bacteriophages. Nat. Rev. Microbiol. 9, 647–657. https://doi.org/10.1038/nrmicro2632

Cohen, Y., Joseph Pollock, F., Rosenberg, E., Bourne, D.G., 2013. Phage therapy treatment of the coral pathogen Vibrio coralliilyticus. Microbiologyopen 2, 64–74. https://doi.org/10.1002/mbo3.52

Crooks, G., Hon, G., Chandonia, J., Brenner, S., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004.1

Davidson, A.R., Gold, M., 1992. Mutations abolishing the endonuclease activity of bacteriophage lambda terminase lie in two distinct regions of the A gene, one of which may encode a “leucine zipper” DNA-binding domain. Virology 189, 21–30. https://doi.org/10.1016/0042-6822(92)90677-H

Dixit, A.B., Ray, K., Thomas, J. a., Black, L.W., 2013. The C-terminal domain of the bacteriophage T4 terminase docks on the prohead portal clip region during DNA packaging. Virology 446, 293–302. https://doi.org/10.1016/j.virol.2013.07.011

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., Felts, B., Dinsdale, E. a., Mokili, J.L., Edwards, R. a., 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. 5, 1–11. https://doi.org/10.1038/ncomms5498

Edgar, R.S., Wood, W.B., 1966. Morphogenesis of bacteriophage T4 in extracts of mutant-infected cells. Proc. Natl. Acad. Sci. 55, 498–505. https://doi.org/10.1073/pnas.55.3.498

Efrony, R., Loya, Y., Bacharach, E., Rosenberg, E., 2006. Phage therapy of coral disease. Coral Reefs 26, 7–13. https://doi.org/10.1007/s00338-006-0170-1

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 39, 29–37. https://doi.org/10.1093/nar/gkr367

Fokine, A., Rossmann, M.G., 2014. Molecular architecture of tailed double-stranded DNA phages. Bacteriophage 4, e28281. https://doi.org/10.4161/bact.28281

Fokine, A., Zhang, Z., Kanamaru, S., Bowman, V.D., Aksyuk, A. a., Arisaka, F., Rao, V.B., Rossmann, M.G., 2013. The molecular architecture of the bacteriophage T4 neck. J. Mol. Biol. 425, 1731–1744. https://doi.org/10.1016/j.jmb.2013.02.012

Garren, M., Son, K., Raina, J.-B., Rusconi, R., Menolascina, F., Shapiro, O.H., Tout, J., Bourne, D.G., Seymour, J.R., Stocker, R., 2014. A bacterial pathogen uses dimethylsulfoniopropionate as a cue to target heat-stressed corals. ISME J. 8, 999–1007.

https://doi.org/10.1038/ismej.2013.210

Ghosh-Kumar, M., Alam, T.I., Draper, B., Stack, J.D., Rao, V.B., 2011. Regulation by interdomain communication of a headful packaging nuclease from bacteriophage T4. Nucleic Acids Res. 39, 2742–2755. https://doi.org/10.1093/nar/gkq1191

Granboulan, P., Sechaud, J., Kellenberger, E., 1971. On the fragility of phage T4-related particles. Virology 46, 407–425. https://doi.org/10.1016/0042-6822(71)90042-0

Hamilton, D.L., Luftig, R.B., 1972. Bacteriophage T4 Head Morphogenesis. J. Virol. 9, 1047–1056.

Heidelberg, J.F., Heidelberg, K.B., Colwell, R.R., 2002. Bacteria of the gamma-subclass Proteobacteria associated with zooplankton in Chesapeake Bay. Appl. Environ. Microbiol. 68, 5498–5507. https://doi.org/10.1128/AEM.68.11.5498-5507.2002

Hendrix, R., Smith, M., Burns, R., Ford, M., Hatfull, G., 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. 96, 2192–2197. https://doi.org/10.1073/pnas.96.5.2192

Hickman, A., Li, Y., Mathew, S., May, E., Craig, N., Dyda, F., 2000. Unexpected structural diversity in DNA recombination: the restriction endonuclease connection. Mol. Cell 5, 1025–1034.

Kelly, L.A., Mezulis, S., Yates, C., Wass, M., Sternberg, M., 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. Nat. Protoc. 10, 845–858. https://doi.org/10.1038/nprot.2015-053

Kovall, R.A., Matthews, B.W., 1999. Type II restriction endonucleases: Structural, functional and evolutionary relationships. Curr. Opin. Chem. Biol. 3, 578–583. https://doi.org/10.1016/S1367-5931(99)00012-5

Kuebler, D., Rao, V.B., 1998. Functional analysis of the DNA-packaging/terminase protein gp17 from bacteriophage T4. J. Mol. Biol. 281, 803–814. https://doi.org/10.1006/jmbi.1998.1952

Kushmaro, A., Rosenberg, E., Fine, M., Loya, Y., 1997. Bleaching of the coral Oculina patagonica by Vibrio AK-1. Mar. Ecol. Prog. Ser. 147, 159–165.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Bryant, S.H., 2013. CDD: Conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41, 348–352. https://doi.org/10.1093/nar/gks1243

Meyer, F., Paarmann, D., D'Souza, M., Etal., 2008. The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386. https://doi.org/10.1186/1471-2105-9-386

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shawand, P.D., Marshall, D., 2013. Using tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14, 193–202. https://doi.org/10.1093/bib/bbs012

Murialdo, H., 1988. Lethal effect of lambda DNA terminase in recombination deficient Escherichia coli. Mol. Gen. Genet. 213, 42–49.

Oram, M., Sabanayagam, C., Black, L.W., 2008. Modulation of the packaging reaction of bacteriophage T4 terminase by DNA structure. J. Mol. Biol. 381, 61–72. https://doi.org/10.1016/j.jmb.2008.05.074

Prasad Bhattacharyya, S., Basaveswara Rao, V., 1993. A novel terminase activity associated with the DNA packaging protein gp17 of bacteriophage T4. Virology. https://doi.org/10.1006/viro.1993.1452

Rao, V.B., Feiss, M., 2015. Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses. Annu. Rev. Virol. 2, 351–378. https://doi.org/10.1007/978-1-4614-5915-6

Rao, V.B., Feiss, M., 2008. The bacteriophage DNA packaging motor. Annu. Rev. Genet. 42, 647–681. https://doi.org/10.1146/annurev.genet.42.110807.091545

Rohwer, F., Edwards, R., 2002. The phage proteomic tree: A genome-based taxonomy for phage. J. Bacteriol. 184, 4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., Azam, F., 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. Limnol. Oceanogr. 45, 408–418. https://doi.org/10.4319/lo.2000.45.2.0408

Ruby, E.G., Asato, L.M., 1993. Growth and flagellation of Vibrio fischeri during initiation of the sepiolid squid light organ symbiosis. Arch. Microbiol. 159, 160–167. https://doi.org/10.1007/BF00250277

Sarnovsky, R.J., May, E.W., Craig, N.L., 1996. The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. EMBO J. 15, 6348–61.

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100.

Seguritan, V., Alves, N., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A.B., Salamon, P., Segall, A.M., 2012. Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput. Biol. 8. https://doi.org/10.1371/journal.pcbi.1002657

Snustad, D.P., 1968. Dominance interactions in Escherichia coli cells mixedly infected with bacteriophage T4D wild-type and amber mutants and their possible implications as to type of gene-product function: catalytic vs. stoichiometric. Virology 35, 550–563. https://doi.org/10.1016/0042-6822(68)90285-7

Stothard, P., Wishart, D.S., 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21, 537–539. https://doi.org/10.1093/bioinformatics/bti054

Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G., Rao, V.B., 2008. The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. Cell 135, 1251–1262. https://doi.org/10.1016/j.cell.2008.11.015

Thompson, F.L., Iida, T., Swings, J., 2004. Biodiversity of Vibrios. Microbiol. Mol. Biol. Rev. 68, 403–423. https://doi.org/10.1128/MMBR.68.3.403

Thurber, R. V, Haynes, M., Breitbart, M., Wegley, L., Rohwer, F., 2009. Laboratory procedures to generate viral metagenomes. Nat. Protoc. 4, 470–83. https://doi.org/10.1038/nprot.2009.10

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Weynberg, K.D., Voolstra, C.R., Neave, M.J., Buerger, P., van Oppen, M.J.H., 2015. From cholera to corals: Viruses as drivers of virulence in a major coral bacterial pathogen. Sci. Rep. 5, 17889. https://doi.org/10.1038/srep17889

Chapter 3 : A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage

Abstract

Diversity-generating retroelements (DGRs) are genetic cassettes that create hypervariable proteins. This work surveyed bacteriophage genomes for novel DGRs and discovered 92 DGRs that were only found in phages exhibiting a temperate lifestyle. The majority of phage-encoded DGRs were identified as prophages in bacterial hosts from the phyla Bacteroidetes, Proteobacteria and Firmicutes. Sequence reads from these previously unidentified prophages were present in viral metagenomes (viromes), indicating these prophages can produce functional viruses. Five phages possessed hypervariable proteins homologous to the tail fiber of BPP-1, whereas the functions of the remaining DGR target proteins were unknown. A novel temperate phage that harbors a DGR cassette targeting a protein of unknown function was isolated from *Bacteroides dorei*. This phage, here named *Bacteroides dorei* ΦHankyphage, lysogenizes at least 13 different *Bacteroides* spp. and was present in 34% and 21% of whole-community metagenomes and human-associated viromes, respectively.

Introduction

Phages encode genes selectively mutagenized by diversity-generating retroelements (DGRs) (Hannigan et al., 2017; Minot et al., 2012, 2013). DGRs are genetic cassettes that introduce DNA sequence variation through a unique targeted mutagenesis mechanism (Liu et al., 2002). The mechanism of DGR function is best understood in *Bordetella* phage BPP-1. BPP-1 possesses a 134 bp variable repeat (VR) within the gene encoding the phage tail fiber. Downstream of VR is a second, invariant copy designated the template repeat (TR). Nearby the TR/VR pair are genes encoding an ‘accessory variability determinant’ and a reverse

transcriptase (RT). These two proteins generate an error-prone cDNA from the TR, followed by stable incorporation of the mutagenized cDNA into the phage tail fiber gene (Liu et al., 2002; Naorem et al., 2017, Handa et al., 2018). This process, termed 'mutagenic retrohoming', yields a VR that is distinct from the TR exclusively at adenine bases. As a result, the BPP-1 tail fiber that mediates adsorption to bacterial host receptors is hypervariable, enabling tropism switching on host *Bordetella* species (Doulatov et al., 2004; Liu et al., 2002; Miller et al., 2008).

Since the initial discovery in BPP-1, the majority of DGRs identified to date are considered bacterial and archaeal in origin (Doulatov et al., 2004; Paul et al., 2015, 2017; Schillinger et al., 2012; Simon and Zimmerly, 2008). A recent survey found that 40% of DGRs are flanked by genes with sequence similarity to phages (Wu et al., 2017). However, it is currently an open question whether these cassettes are encoded by functional prophages or inactive remnants.

Here, a survey of phage genomes found that DGRs are only encoded by phages exhibiting a temperate lifestyle. Investigation of bacterial genomes found that DGRs previously considered to be bacterial in origin are encoded by functional prophages. These phage DGRs diversify proteins with unknown functions, though most possessed a C-type lectin domain. A single temperate phage, Hankyphage, was found to harbor a DGR cassette and lysogenize at least 13 different species of the genus *Bacteroides*. Hankyphage was present in whole-community metagenomes and viromes generated from microbial samples collected around the globe.

Methods

Identification of DGRs

Three databases were searched for DGR cassettes. The first database contained bacteriophage genomes available on the RefSeq database dereplicated at 98% identity using CD-HIT (n = 1881, accessed 06/2016) (Brister et al., 2015; Fu et al., 2012). A second database of Actinobacteriophage genomes (n = 2322, accessed 12/2017) was downloaded from the PhagesDB website (Russell and Hatfull, 2017). A third database of 31,946 predicted prophages was generated from 11,278 Bacterial and Archaeal genomes using the program PhiSpy (Akhter et al., 2012). All predicted prophage sequences were previously published (Kang, 2017). Sequences in each database were screened for reverse transcriptases (RTs) (Pfam model PF00078) using HMMer (Finn et al., 2011). Sequences with RTs were subsequently scanned for template repeats (TRs) and variable repeats (VRs) using the program DiGRef with default settings (minimum 10 adenines in TR, minimum 7 adenine substitutions in VR) modified to accept local inputs (Schillinger et al., 2012). Template repeat and variable repeat alignments were made using Clustal Omega and visualized with JalView (Sievers et al., 2011; Waterhouse et al., 2009). Life history assignments for phages in the databases (Refseq, PhagesDB and the DGR-containing predicted prophage regions) were accomplished using PHACTS (McNair et al., 2012). A lifestyle assignment was considered confident if the average score of the trees minus the standard deviation was > 0.5. Otherwise, no life history assignment was made.

Viral read mapping to predicted prophage-containing regions

A collection of human-associated viromes available from the NCBI sequence read archive (SRA) was previously curated (Cobián Güemes et al., 2016). Additional human viromes were added to the database and all accessions are provided. All of these viromes were

downloaded from the SRA using the fastq-dump utility, quality filtered and dereplicated with Prinseq employing the following flags: derep 12345 -lc_method entropy -lc_threshold 50 -trim_qual_left 15 -trim_qual_right 15 -trim_qual_type mean -trim_qual_rule lt -trim_qual_window 2 -trim_tail_left 5 -trim_tail_right 5 -min_len 60 -min_qual_mean 15 -ns_max_p 1 (Schmieder and Edwards, 2011). Reads from each virome were concatenated into a single file and mapped to predicted prophage-containing regions harboring a DGR (n = 170) using Bowtie2 with default parameters (Langmead and Salzberg, 2012). Coverage was calculated with Samtools and visualized using Anvio (Eren et al., 2015; Li et al., 2009). To calculate the fractional abundance of *Bacteroides dorei* Hankyphage in each sample, reads from each virome were mapped to the Hankyphage genome at 97% identity using SMALT (Ponstingl, 2010). The fractional abundance of Hankyphage in each virome was calculated as previously described (Cobián Güemes et al., 2016).

Prophage genome annotation

Open reading frames (ORFs) in predicted prophage-containing regions were called by Prodigal and annotated by BLASTp searches against the NCBI COG database, where hits are considered significant if the e-value is $< 10^{-5}$ (Galperin et al., 2015; Hyatt et al., 2010). Additional annotations were determined by HMM searches against the Pfam-A database with a bitscore cutoff of > 25 (Finn et al., 2016; Marchler-Bauer et al., 2013). ORFs containing a variable repeat were analyzed using Phyre2 and Phobius (Käll et al., 2007; Kelly et al., 2015). To identify the host range of Hankyphage, the phage genome was used as a query against all subject *Bacteroides* genomes available on RefSeq using BLASTn (accessed 06/2017). Query-subject alignments were visualized using Easyfig (Sullivan et al., 2011).

Metagenome read mapping to Hankyphage

A collection of whole-community metagenomes available from the NCBI SRA was previously curated (Torres et al., 2017). An untargeted selection of 37,583 metagenomes was downloaded using the fastq-dump utility of the SRA toolkit and mapped to the Hankyphage genome using Bowtie2 with default parameters. Binary Alignment Map (BAM) files were sorted and indexed using Samtools version 1.7 (Li et al., 2009). Heatmaps were generated from the BAM files using publicly available scripts (Levi et al., 2017).

Phage induction, PCR and transmission electron microscopy

A culture of *Bacteroides dorei* strain CL02T12C06 (HM-719) was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project. *B. dorei* was maintained on supplemented brain heart infusion media (BHIS) [38 g/L BHI, 5 g/L yeast extract, 1.5 g/L L-cysteine, 1 mM $CaCl_2$, 10 mM $MgSO_4$, 1 μg/L Vitamin K3, 0.24 μg/L histidine-hematin]. BHIS broth was inoculated with a glycerol stock of *B. dorei* and cultured anaerobically at 37 °C. Early-log cultures were induced for prophage with 8 μg/mL carbadox and harvested after 8-10 hours of growth. The cultures were clarified of bacterial cells through chloroform treatment and syringe-driven 0.22 μm filtration. The clarified supernatant was stored at 4 °C over 0.1 volumes chloroform and used for downstream assays. For PCR, the phage lysate was treated with 3 U of DNase I and incubated for 1 hour at 37 °C, followed by heat inactivation for 10 minutes at 99 °C. For transmission electron microscopy, 30 μl of the purified phage lysate was added to glow-discharged 300-mesh copper grids coated with 10 nm formvar and 1 nm carbon (Electron Microscopy Sciences, PA) for 3 minutes. To remove the salts in buffer, the grids were rinsed 3 times by drops of distilled water (20 μl). The grids were then negatively stained with uranyl acetate (0.5%) for 15 seconds, dried and examined with a FEI Tecnai T12 TEM (FEI, Hillsboro, OR) operating at 120 kV at the SDSU Electron

Microscopy Facility. Micrographs were taken by an AMT HX41 side mounted digital camera (Advanced Microscopy Technique, Woburn, MA).

Results

DGRs found in three temperate phages

A comprehensive survey of phage genomes was conducted to identify novel phage-encoded DGRs. Searching publicly available genomes from isolated phages (n = 4203, NCBI Refseq and PhagesDB databases) identified six phages that possessed a reverse-transcriptase (RT). Three of these six possessed features characteristic of a DGR cassette: an adjacent variable repeat (VR)/template repeat (TR) pair containing $\geq$ 7 adenine mismatches (Table 3-1). The paradigmatic DGR-harboring phage BPP-1 was identified, as well as two marine *Vibrio* phages. The VR of the BPP-1 DGR lied within the phage tail fiber gene, while the VRs of the two marine *Vibrio* phages were encoded within genes of unknown function. All three genes harbored a C-type lectin domain, consistent with previous structural analyses of DGR target proteins (Le Coq and Ghosh, 2011).

DGRs encoded by novel temperate phages

To more exhaustively search for DGRs in temperate phages, the survey was expanded to include temperate phages integrated into genomes as prophages. A database of 31,946 predicted prophage-containing regions generated by PhiSpy from 11,278 Bacterial genomes was scanned for DGRs (Kang et al., 2017). PhiSpy utilizes both sequence homology-dependent and independent approaches to predict regions harboring an integrated phage (Akhter et al., 2012). In total, 170 regions harbored a DGR cassette (Table 3-2). Previous surveys identified 25 of the 170 DGRs and assigned them as bacterial in origin, while 57 were reported as associated with a prophage or inactive phage remnant (Table 3-2) (Cornuault et al., 2018; Doulatov et al., 2004; Paul et al., 2017; Schillinger et al., 2012; Simon and

Zimmerly, 2008; Wu et al., 2017). This analysis identified 74 novel DGRs and expands their observed association with temperate phages across diverse Phyla.

Virions produced by prophages can be captured during viral-particle enrichment steps as part of viral metagenome preparation. Thus, predictions of the PhiSpy algorithm were tested by determining whether viral metagenome reads could be recruited to sequences of all 170 of the predicted prophage regions. Because 104 of the 170 DGR-containing lysogens belonged to the Bacteroidetes or Firmicutes phyla (Table 3-2), whose members dominate the human gastrointestinal tract (Huttenhower et al., 2012), viral metagenomes from human skin and fecal samples were collected (n = 1,366). Approximately three billion quality-filtered sequence reads were concatenated into a single file and aligned to each predicted prophage region (Figure 3-1).

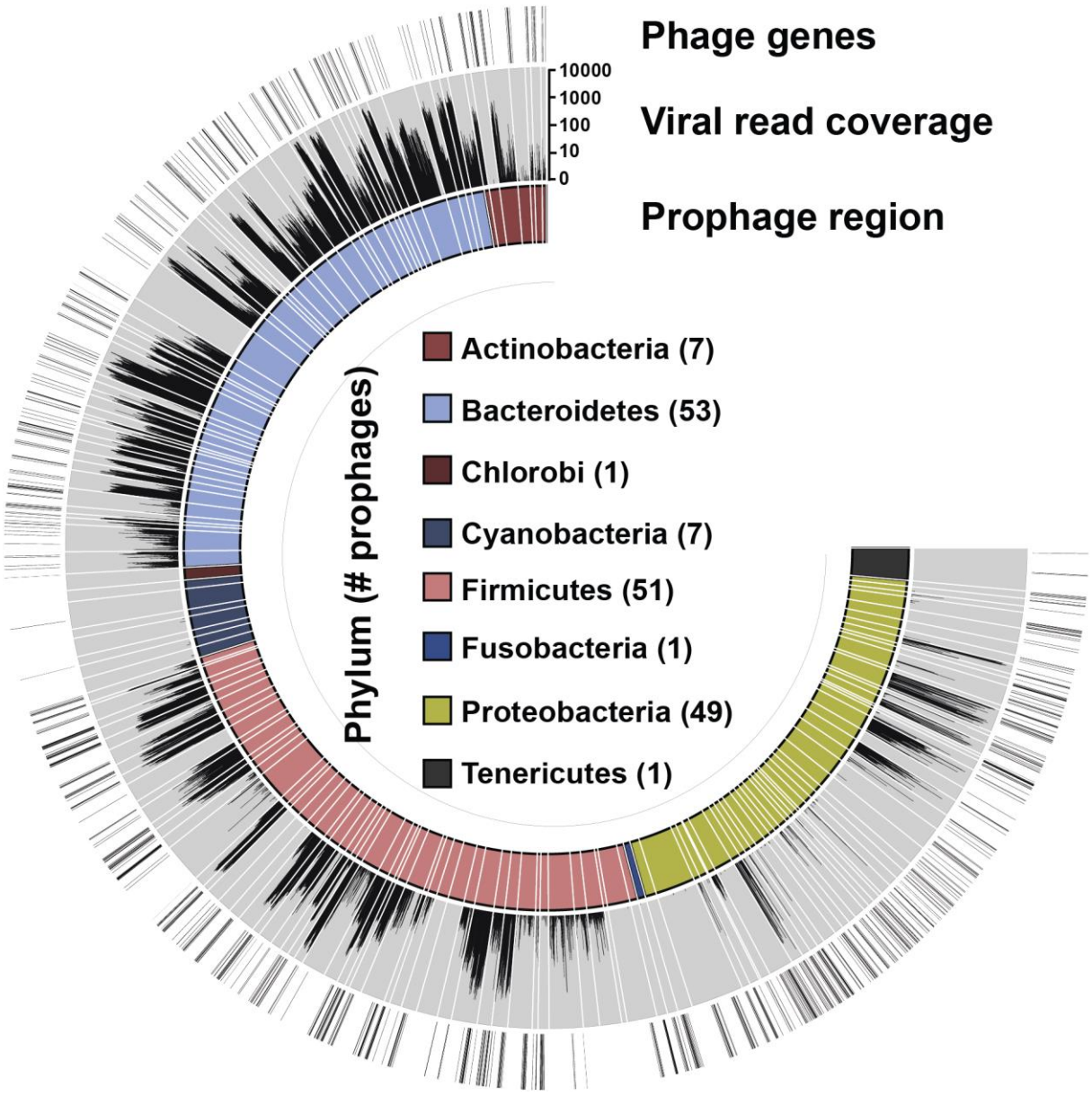

Figure 3-1: Predicted prophage-containing regions with a DGR (n = 170) are depicted on the inner ring and colored according to host phylum. Each predicted prophage-containing region was used as a reference to align sequencing reads from human viromes. The coverage plot of reads for each region is shown on the middle ring. ORFs whose annotations contained the keyword 'phage' are indicated with a hash mark on the outer ring. Sequences of all predicted prophage-containing regions and their annotations are available in additional files 12 and 5, respectively.

Ninety-two of the 170 regions recruited viral reads at >10x mean fold coverage, demonstrating the presence of an integrated phage capable of producing virions (Table 3-2). Predicted prophage regions contained identifiable phage genes, such as capsids, tails and integrases, validating the recruitment of reads to a temperate phage. Through the bioinformatic approach presented here, 92 DGRs were classified as viral in origin.

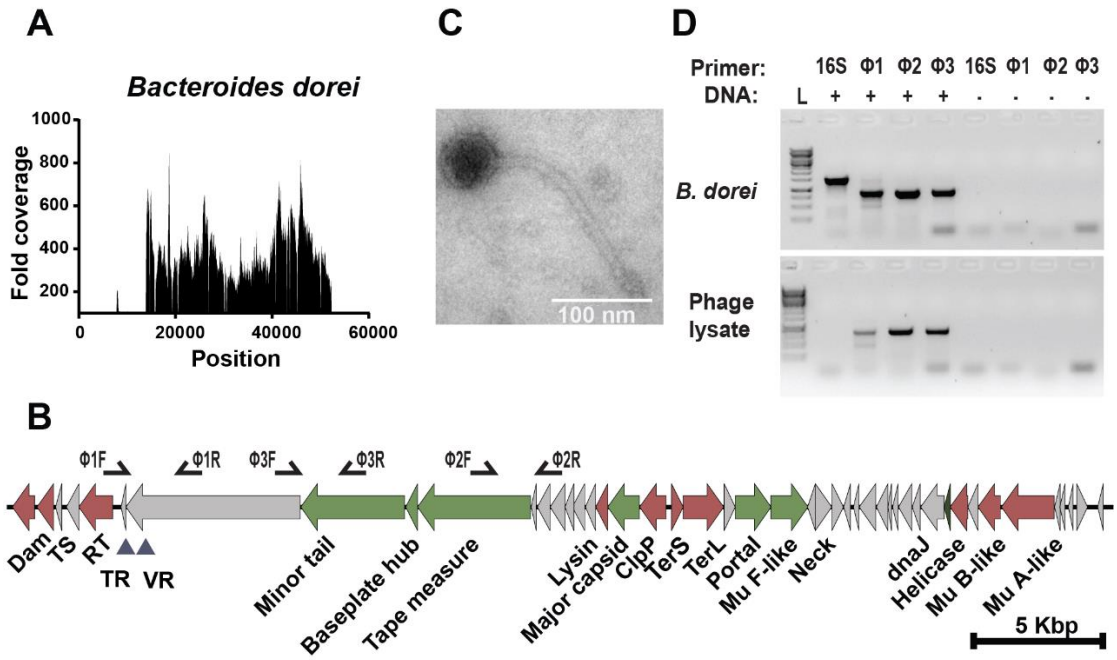

Figure 3-2: *Bacteroides dorei* is lysogenized by a DGR-containing temperate phage.(A) Coverage plot of viral reads aligning to a 66kb region in *Bacteroides dorei* predicted to contain a prophage. (B) ORFs encoded by the discrete 43kb region with high read coverage are represented by arrows and colored according to function. Red, enzymatic; green, structural; grey, hypothetical. The DGR components are indicated as RT, reverse transcriptase; TR, template repeat; VR, variable repeat. (C) Transmission electron microscopy image of a phage lysate prepared from an induced culture of *B. dorei* (scale bar = 100 nm). (D) Agarose gel showing PCR products using primers indicated on panel (B) and universal 16S rRNA gene primers. The upper panel depicts PCR products using *B. dorei* DNA as a template and the lower panel depicts products from the purified phage lysate. All primer sequences are listed in Table 3-6. For both panels the four lanes on the right are no-template controls.

All of the phages possessing a DGR, including BPP-1 and the two *Vibrio* phages, exhibit a temperate lifestyle (Liu et al., 2002; Oakey et al., 2002; Zabala et al., 2009). To determine if this observation was an artefact of temperate phages being overrepresented in the Refseq, PhagesDB and prophage collection, all phage life histories were classified using PHACTS (McNair et al., 2012). The majority of phages in these databases could not have a lifestyle assigned (n = 2,911). Where a lifestyle could be confidently assigned, 1,181 phages were predicted to be temperate and 332 were predicted to be lytic. However, despite the database bias towards temperate phages, the observed number of DGRs in temperate phages was still higher than expected (χ^2, $^{**}P = 0.0012$). Therefore, DGRs are preferentially encoded by temperate phages.

Bacteroides dorei is lysogenized by a DGR-containing temperate phage

A 66kb region in *Bacteroides dorei* predicted to contain a DGR-encoding prophage was further examined. This region recruited reads from 13 different human viral studies with a mean fold coverage of 229x (Figure 3-2A). The reads aligned to a discrete 43kb locus encoding 45 ORFs. Seven ORFs with significant homology to known phage structural proteins could be annotated, including two tail proteins, a baseplate, portal, neck and two capsid genes (Figure 3-2B and Table 3-4). Genes required for phage genome packaging and integration were detected, including a large and small terminase and phage Mu-like transposase. The DGR cassette was composed of an RT with two adjacent 117bp repeats that were mismatched at 14 positions corresponding to adenine bases. Proximal to the RT gene was a small ORF whose amino acid sequence had a similar isoelectric point and molecular weight as the accessory variability determinant of BPP-1 (Liu et al., 2002). The VR was located at the C terminus (residues 2188-2225) of a large ORF possessing a C-type lectin

domain. Collectively, these features indicated the 43kb region was a prophage genome that harbored a DGR cassette.

Phage induction assays on a culture of *B. dorei* were performed to validate mature virion production *in vitro*. A culture of *B. dorei* was obtained and cultured anaerobically at 37°C. The culture was treated with the antibiotic carbadox, a known prophage inducing agent (Bearson et al., 2014; Johnson et al., 2017; Kohler et al., 2000). After 10-12 hours of anaerobic growth, bacterial cells were removed. Transmission electron microscopy identified a virus with a 200 nm flexible tail typical of the *Siphoviridae* (Figure 3-2C). Lastly, we confirmed the observed phage contained the predicted prophage genotype using PCR (Figure 3-2D). This phage was named *Bacteroides dorei* Hankyphage and exemplifies the power of the bioinformatic pipeline to identify inducible prophages, including those that encode DGRs.

Hankyphage exhibits broad-host range and global distribution

Plaque assays combining a cell-free phage lysate and susceptible bacterial hosts can determine the minimal host range of a phage (Adams, 1959). A phage lysate was prepared from an induced culture of *B. dorei* and combined with a naïve strain of *Bacteroides dorei* in top agar under anaerobic conditions. Lawns of the naïve strain or additional *Bacteroides* spp. did not yield observable plaques (data not shown). As an alternative, the NCBI Refseq database of bacterial genomes (accessed 07/2017) was queried for the Hankyphage genome. The complete, 43kb Hankyphage genome was present in 13 unique *Bacteroides* species at $\geq$ 95% nucleotide identity (Figure 3-3A). Consistent with Mu-like transposition, the site of integration was random and flanked by 5 bp direct repeats (Allet, 1979) (Figure 3-3A). Thus, Hankyphage exhibits broad host-range and each tropic variant was assigned a unique indication (Hankyphage p00 – Hankyphage p12).

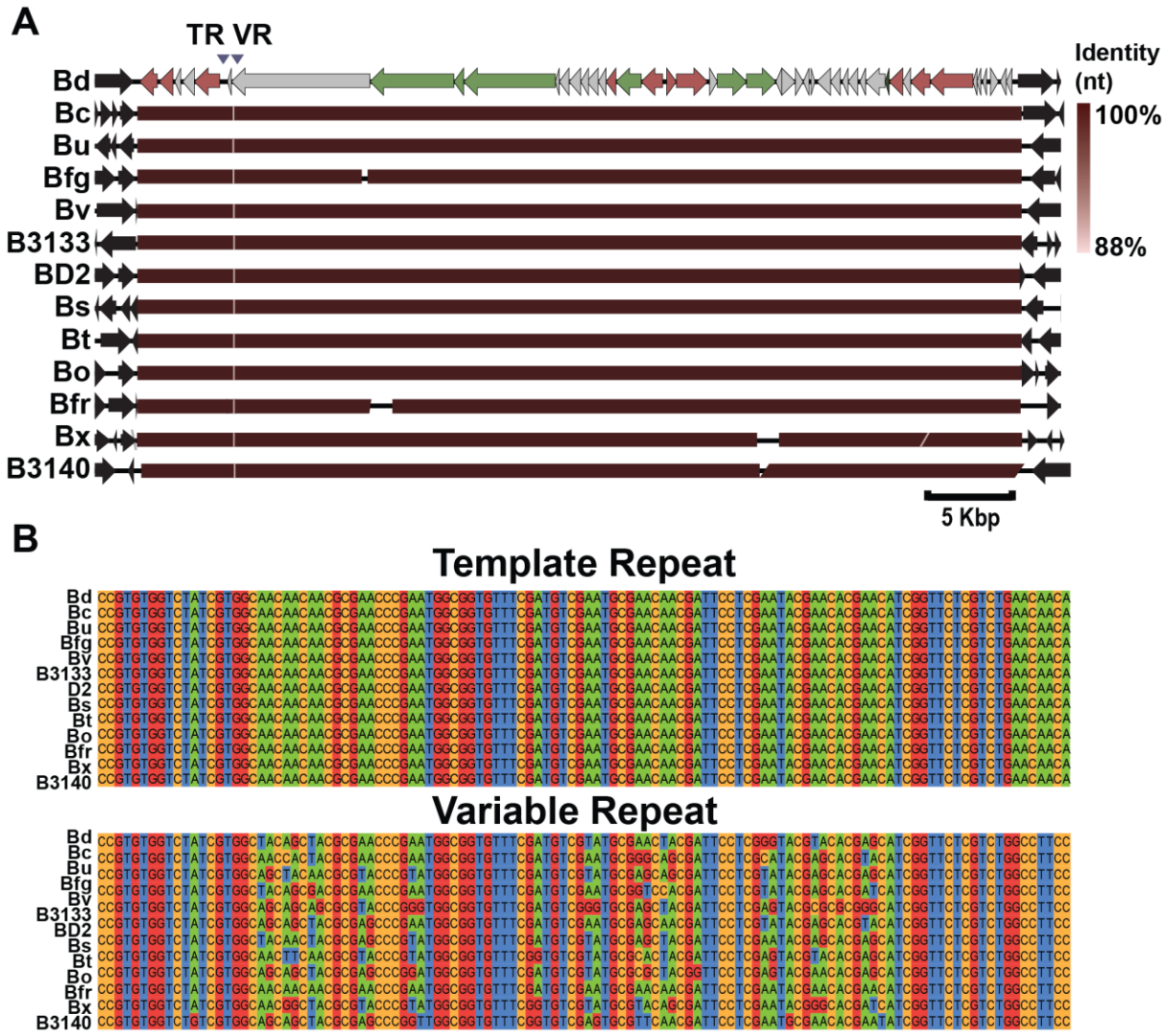

Figure 3-3: Hankyphage lysogenizes 13 *Bacteroides* species. (A) The Hankyphage genome (top) is color coded as before and was used to query the NCBI Refseq database of bacterial genomes. The resulting alignments are color coded according to the pairwise nucleotide identity. Regions flanking Hankyphage (2 kb) in each subject are depicted as black arrows and produced no significant alignments to each other. The subject sequences are Bc, *B. cellulosilyticus*; Bu, *B. uniformis*; Bfg, *B. finegoldii*; Bv, *B. vulgatus*; B3133, *B. sp. 3_1_33FAA*; BD2, *B. sp. D2*; Bs, *B. salyersiae*; Bt, *B. thetaiotaomicron*; Bo, *B. ovatus*; Bfr, *B. fragilis*; Bx, *B. xylanisolvens*; and B3140, *B. sp. 3_1_40A* (B) Alignments of the TRs and VRs from Hankyphage in each *Bacteroides* species are shown. Each nucleotide base is color coded for visualization of mismatches in the variable repeat.

To test if the DGR cassette diversifies the variable repeat in different *Bacteroides* hosts, the TRs and cognate VRs of each tropic variant were extracted for multiple alignment (Clustal Omega). The aligned TRs contained no substitutions while the VRs exhibited adenine-specific substitutions in asparagine codons, a distinguishing attribute of DGR-mediated variation (Liu et al., 2002) (Figure 3-3B). None of the tropic variants shared the same VR sequence. Interestingly, two tropic variants (*B. dorei* and *B. cellulosilyticus*) only differed in the sequence of their VR along the entirety of their aligned genomes. The VR contained 21 variable adenines, enabling Hankyphage to explore a sequence space of potentially 10^{12} unique variants. Despite the similar capacity for sequence variation as the BPP-1 tail fiber, Hankyphage's DGR target protein shared no sequence similarity with any characterized protein (Table 3-4).

To estimate the ubiquity of Hankyphage in the environment, total-community metagenomes and viromes available on the NCBI SRA were searched for reads matching the Hankyphage genome. An untargeted selection of 37,583 total-community metagenomes from the NCBI SRA (accessed 03/2018) spanning diverse microbial environments were curated previously (Torres et al., 2017). In total, 12,774 metagenomes contained $\geq$ 1 read aligning to the Hankyphage genome (12,774 / 37,583, 34%) (Figure 3-4). The same analysis was repeated using the previous database of 1,366 human-associated viromes sampled from geographically separated individuals. Sequencing reads recruited at $\geq$ 97% nucleotide identity could be recovered from 221 of 1,366 viromes. All 221 viromes were derived from human fecal samples (221/1038) while none of the human skin virome reads were recruited (0/329). For a given virome sample, Hankyphage was present at $4x10^{-8}$ - $4x10^{-2}$ fractional abundance. Using these data to estimate the number of Hankyphage virions, approximately 3.3×10^{17}

exist globally (Table 3-7). In summary, Hankyphage is present in microbial environments, namely the human gastrointestinal tract, from locations sampled around the world.

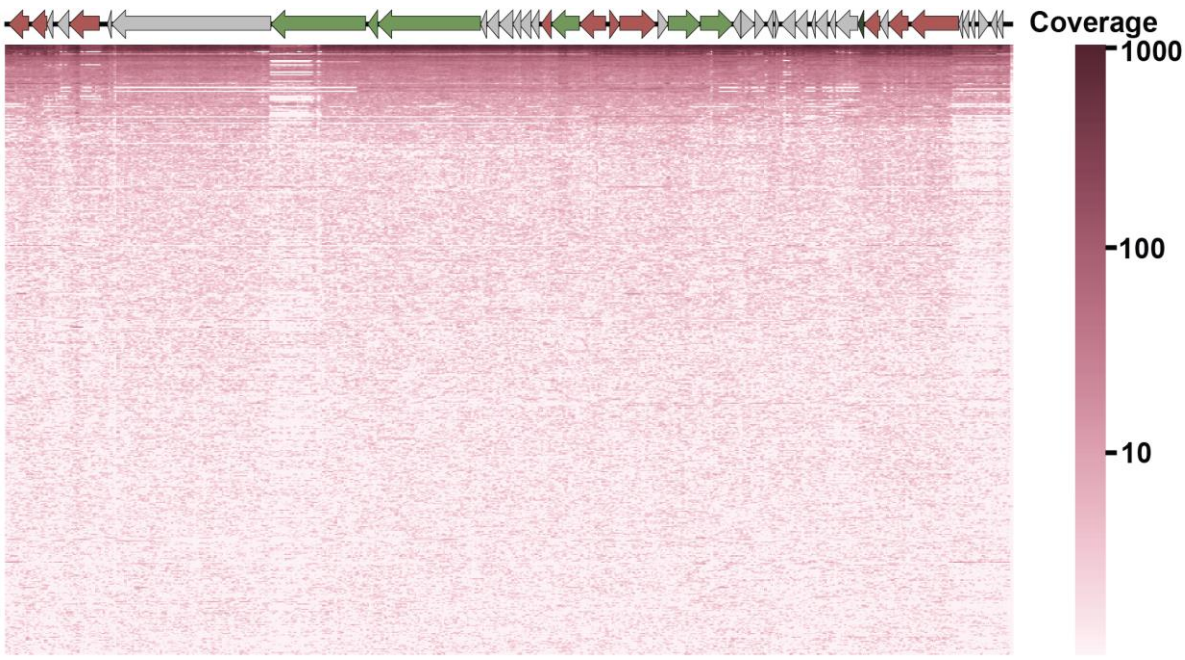

Figure 3-4: Coverage plot of the Hankyphage genome in metagenomes. Each row is a metagenome (n = 500). The x axis of the heat map displays the 42,831 bp length of the Hankyphage genome. The color bar is scaled proportionally to the coverage of each nucleotide in the Hankyphage genome.

Function of hypervariable proteins

To assess the functional role of all the DGR cassettes identified, structures of the variable repeat-containing ORFs were predicted using Phyre2 (Kelly et al., 2015). Five of the ORFs exhibited structural homology to the major tropism determinant of phage BPP-1, signifying these phages may use a similar strategy as BPP-1 to mutagenize their tail fibers and broaden host range (Table 3-5) (Liu et al., 2002). None of the other hypervariable proteins were homologous to known phage proteins; however, 57 hypervariable genes were homologous to a DGR target protein from *Thermus aquaticus* and 29 were homologous to a pilin tip from *Bacteroides ovatus*. The pilin structure was recently shown to contain a C-type

lectin domain, supporting previous observations that this fold is a conserved target of DGRs (Le Coq and Ghosh, 2011; Xu et al., 2016). Overall, the functional roles of these hypervariable proteins remain to be determined.

Discussion

By combining a prophage prediction algorithm with mapping reads from viromes, 92 phage-encoded DGRs were identified. Many of the proteins targeted for diversification contain a subtype of the C-type lectin fold, suggesting a role in binding interactions with diverse ligands. Phage DGR proteins lack detectible transmembrane regions or signal peptides found on surface-displayed bacterial DGR proteins (Arambula et al., 2013; Le Coq and Ghosh, 2011) (Table 3-5). While this supports their phage origin, discerning whether the DGR target proteins are components of functional prophages or defective remnants is difficult. Here, a generalized bioinformatic pipeline was employed to identify functional prophages where viral read data were available.

In one case, an integrated prophage from *B. dorei* was induced, thereby verifying the bioinformatic prediction. This temperate bacteriophage, Hankyphage, is the fifth phage to be isolated that possesses a DGR. Hankyphage was detected as a complete prophage in 13 different *Bacteroides* species at $\geq$ 95% nucleotide identity. Furthermore, Hankyphage's complete genome was recovered in whole-community and viral metagenomes. These observations suggest that Hankyphage can infect a wide-range of *Bacteroides* in diverse environments despite not infecting naïve *Bacteroides* hosts *in vitro*. Additional factors may be required for a productive infection outside the gastrointestinal tract. Previous analyses of intestinal viral communities demonstrate few genotypes are shared between individuals (Reyes et al., 2010). However, analysis of metagenomes generated from human fecal samples

collected around the globe disclose that Hankyphage resides in the gastrointestinal tract of approximately 50% of the human population (Table 3-7).

Despite there being hundreds of lytic phage genomes in public databases, DGRs were only identified in temperate phages. Given that the global phage diversity far exceeds that of cultured isolates, identification of a DGR-containing lytic phage may be forthcoming (Cobián Güemes et al., 2016). Nevertheless, DGR-diversification may provide increased functionality for temperate phages. Targeted mutagenesis of phage genes would offset accurate replication by bacterial DNA polymerases, poising temperate phages for dynamic responses to changing environmental conditions. That the majority of DGR-containing phages lysogenize human gut commensals suggests that DGRs are advantageous in host-associated environments. Similarly, DGRs are enriched in host-associated bacteria and archaea with reduced genome sizes, highlighting the parallel emergence of hypervariability cassettes in host-associated environments (Paul et al., 2017).

The variable repeats of known DGRs lie within C-type lectin folds and Ig-like domains (Minot et al., 2012; Wu et al., 2017). Both C-type lectins and Ig-like domains are encoded by phages, including Hankyphage and two other *Bacteroides* phages (Dutilh et al., 2014; Fraser et al., 2006; Reyes et al., 2013). In phage T4, the Ig-like domain-containing protein Hoc mediates phage adherence to mucus, thereby increasing the rate of T4 adsorption onto *E. coli* host cells (Barr et al., 2015). Consistent with this Bacteriophage Adherence to Mucus (BAM) model, Hankyphage may employ a similar strategy. By utilizing a C-type lectin to adhere to mucus, Hankyphage may increase the rate of encounter with *Bacteroides*

hosts in mucus. Targeted hypervariation in the C-type lectin could modulate phage diffusion in mucus, enabling optimal search strategies in different mucosal environments.

In this study, we present evidence that DGR cassettes are encoded by temperate bacteriophages integrated in the chromosomes of bacteria as prophages. Many of these DGRs were previously considered to be bacterial in origin. While the functions of these hypervariable phage proteins remain enigmatic, this research will enable future dissection of their roles in phage life cycles. Moreover, isolation of a single phage shared amongst an estimated one-half of the human population is an exception to our current understanding that the human viral community is largely composed of unique viruses.

Acknowledgements

Chapter 3, in full, has been submitted for publication. Sean Benler, Ana Cobian-Guemes, Katelyn McNair, Shr Hau Hung, Kyle Levi, Robert Edwards, and Forest Rohwer. 2018. The dissertation author was the primary investigator and author of this paper.

Appendix

Table 3-1: DGRs identified in isolated phages from the Refseq database. For each DGR, annotation of the VR-encoding ORF was made using Phyre2. A domain analysis was conducted using HMMer against the Pfam-A database. Lifestyle assignment was extracted from the literature.

Phage (Accession)	***Hypervariable ORF annotation (Phyre 2)***	***Confidence***	***PDB code***	***Domains (PFAM accession)***	***Lifestyle***	***Reference***
Vibrio phage VHML (NC_004456)	*Thermus aquaticus* variable protein	100	5VF4	FGE-sulfatase (PF03781)	Temperate	Oakey et al. (2002)
Vibrio phage VP58.5 (NC_027981)	*Thermus aquaticus* variable protein	100	5VF4	FGE-sulfatase (PF03781)	Temperate	Zabala et al. (2009)
Bordetella phage BPP-1 (NC_005357)	Major tropism determinant	100	2IOU	FGE-sulfatase (PF03781)	Temperate	Liu et al. (2002)

Table 3-2: Taxonomic details of the 170 predicted prophage-containing regions that harbor a DGR cassette. For each region, the unique identifier and lowest level taxonomic assignment of the organism is provided. The average coverage of viral sequencing reads, as well as length, of each region is reported.

Prophage ID	***Organism***	***Mean viral read coverage***	***length (nt)***
g.345.pp.1	*Bifidobacterium longum* DJO10A	217.2	39864
g.31087.pp.3	*Bifidobacterium breve* HPH0326	83.4	117203
g.1381.pp.1	*Collinsella aerofaciens* ATCC 25986	5.8	33756
g.31354.pp.6	*Enterorhabdus caecimuris* B7	0.0	60491
g.1527.pp.1	*Collinsella stercoris* DSM 13279	2.4	53572
g.1947.pp.2	*Collinsella intestinalis* DSM 13280	0.2	97506
g.1738.pp.2	*Eggerthella lenta* DSM 2243	5.2	40121
g.1648.pp.1	*Bacteroides* sp. 1 1 14	201.9	49814
g.1648.pp.4	*Bacteroides* sp. 1 1 14	89.0	59111
g.1599.pp.6	*Bacteroides* sp. 1 1 30	568.1	94623
g.1649.pp.4	*Bacteroides* sp. 1 1 6	17.3	74713
g.2867.pp.3	*Bacteroides* sp. 2 1 56FAA	123.8	55511
g.1653.pp.13	*Bacteroides* sp. 2 2 4	3198.8	76735
g.1616.pp.11	*Bacteroides* sp. 3 1 12	486.5	34734
g.1616.pp.10	*Bacteroides* sp. 3 1 12	9.3	51061
g.1601.pp.2	*Bacteroides* sp. 3 1 23	195.4	192015
g.1602.pp.4	*Bacteroides* sp. 3 1 33FAA	205.4	79195
g.1656.pp.9	*Bacteroides* sp. 3 1 40A	306.8	51263
g.3439.pp.4	*Bacteroides fragilis* 638R	12.0	137840
g.1606.pp.6	*Bacteroides* sp. 9 1 42FAA	3215.9	59172
g.1552.pp.2	*Bacteroides stercoris* ATCC 43183	982.0	65038
g.1368.pp.5	*Bacteroides ovatus* ATCC 8483	27.8	85783
g.26941.pp.2	*Prevotella* sp. C561	0.4	58530
g.28909.pp.6	*Bacteroides dorei* CL02T00C15	229.2	65252
g.28919.pp.8	*Bacteroides ovatus* CL02T12C04	101.0	75463
g.28919.pp.10	*Bacteroides ovatus* CL02T12C04	2.4	49694
g.28910.pp.10	*Bacteroides dorei* CL02T12C06	246.1	52339

Table 3-2: Taxonomic details of the 170 predicted prophage-containing regions**,** continued.

Prophage ID	***Organism***	***Mean viral read coverage***	***length (nt)***
g.28916.pp.7	*Bacteroides fragilis* CL07T00C01	1529.3	124403
g.28917.pp.2	*Bacteroides fragilis* CL07T12C05	2545.0	69841
g.28925.pp.2	*Bacteroides vulgatus* CL09T03C04	214.3	72871
g.28922.pp.9	*Bacteroides thetaiotaomicron* CL09T03C10	865.1	37378
g.28922.pp.3	*Bacteroides thetaiotaomicron* CL09T03C10	41.4	59064
g.2289.pp.1	*Bacteroides* sp. D1	26.0	79661
g.2290.pp.4	*Bacteroides* sp. D2	691.2	67154
g.2290.pp.2	*Bacteroides* sp. D2	25.2	127504
g.2542.pp.7	*Bacteroides sp*. D22	494.5	121461
g.27099.pp.2	*Parabacteroides* sp. D25	56.3	118432
g.31350.pp.6	*Bacteroides uniformis* dnLKV2	269.6	32818
g.31350.pp.5	*Bacteroides uniformis* dnLKV2	16.9	44469
g.31351.pp.4	*Bacteroides sartorii* dnLKV3	63.1	37534
g.31348.pp.2	*Bacteroides thetaiotaomicron* dnLKV9	2.5	97392
g.1522.pp.3	*Alistipes putredinis* DSM 17216	1574.2	54617
g.27343.pp.3	*Alistipes finegoldii* DSM 17242	995.5	49352
g.1750.pp.2	*Bacteroides finegoldii* DSM 17565	2942.4	62520
g.1750.pp.6	*Bacteroides finegoldii* DSM 17565	113.2	135741
g.3070.pp.2	*Odoribacter splanchnicus* DSM 20712	260.9	54224
g.3043.pp.2	*Bacteroides vulgatus* PC510	284.9	55314
g.436.pp.7	*Bacteroides thetaiotaomicron* VPI-5482	14.4	100086
g.29562.pp.7	*Bacteroides salyersiae* WAL 10018	184.9	60689
g.3099.pp.1	*Alistipes shahii* WAL 8301	159.2	99074
g.2820.pp.2	*Bacteroides xylanisolvens* XB1A	16.9	80227
g.27474.pp.2	*Bacteroides oleiciplenus* YIT 12058	1520.6	118993
g.27983.pp.1	*Bernardetia litoralis* DSM 6794	0.0	151752

Table 3-2: Taxonomic details of the 170 predicted prophage-containing regions**,** continued.

Prophage ID	***Organism***	***Mean viral read coverage***	***length (nt)***
g.3007.pp.3	*Arthrospira platensis* NIES-39	0.0	46498
g.3007.pp.10	*Arthrospira platensis* NIES-39	0.0	63859
g.3007.pp.8	*Arthrospira platensis* NIES-39	0.0	45913
g.3007.pp.12	*Arthrospira platensis* NIES-39	0.0	35127
g.26931.pp.6	*Oscillatoria acuminata* PCC 6304	0.0	40574
g.2640.pp.3	*Paenibacillus* sp. oral taxon 786 D14	0.0	59976
g.32274.pp.1	*Anoxybacillus* sp. DT3-1	0.1	51189
g.72.pp.2	*Paenibacillus mucilaginosus* KNP414	0.0	116160
g.2060.pp.4	*Bacillus cereus* R309803	1.4	125567
g.3088.pp.2	*Paenibacillus vortex* V453	0.0	72363
g.31101.pp.1	*Butyricicoccus pullicaecorum* 1.2	17.3	97645
g.31357.pp.5	*Oscillibacter* sp. 42738	5.8	30897
g.32467.pp.6	*Hungatella hathewayi* 12489931	3.0	83872
g.2844.pp.4	*Lachnospiraceae* 1 4 56FAA	40.1	50370
g.1615.pp.3	*Clostridiales* 1 7 47FAA	366.4	55255
g.27478.pp.3	*Clostridium clostridioforme* 2 1 49FAA	22.7	90760
g.27137.pp.4	*Subdoligranulum* sp. 4 3 54A2FAA	2126.4	50046
g.1613.pp.5	*Ruminococcus* sp. 5 1 39BFAA	565.6	72808
g.2839.pp.2	*Lachnospiraceae* 5 1 63FAA	9.1	41846
g.27096.pp.15	*Lachnospiraceae* 7 1 58FAA	490.8	54175
g.27096.pp.1	*Lachnospiraceae* 7 1 58FAA	301.8	89432
g.27096.pp.7	*Lachnospiraceae* 7 1 58FAA	66.2	75423
g.27096.pp.3	*Lachnospiraceae* 8 1 58FAA	54.4	74371
g.27128.pp.10	*Clostridium* 7 3 54FAA	167.6	35589
g.27128.pp.1	*Clostridium* 7 3 54FAA	74.5	42669
g.32458.pp.11	*Clostridium clostridioforme* 90A1	15.1	124462
g.32460.pp.7	*Clostridium clostridioforme* 90A4	4.6	60100
g.1372.pp.5	*Faecalibacterium prausnitzii* A2-165	229.5	67965

Table 3-2: Taxonomic details of the 170 predicted prophage-containing regions**,** continued.

Prophage ID	***Organism***	***Mean viral read coverage***	***length (nt)***
g.2234.pp.4	Ruminococcaceae D16	996.5	35444
g.1910.pp.5	*Clostridium asparagiforme* DSM 15981	42.5	60199
g.29595.pp.3	*Desulfurispora thermophila* DSM 16022	0.0	76968
g.2827.pp.9	*Eubacterium rectale* DSM 17629	304.3	38931
g.2827.pp.4	*Eubacterium rectale* DSM 17629	235.5	83454
g.2418.pp.1	*Halanaerobium praevalens* DSM 2228	0.0	52523
g.3257.pp.2	*Desulfotomaculum kuznetsovii* DSM 6115	0.0	17026
g.1763.pp.12	*Desulfotomaculum acetoxidans* DSM 771	0.0	57982
g.3102.pp.5	*Coprococcus catus* GD/7	7.4	66866
g.2141.pp.5	*Roseburia intestinalis* L1-82	670.8	51428
g.3105.pp.8	*Faecalibacterium prausnitzii* L2-6	249.8	128425
g.1375.pp.2	*Clostridium* M62/1	322.0	68880
g.996.pp.5	*Desulfotomaculum reducens* MI-1	0.0	55509
g.26274.pp.7	*Alkaliphilus metalliredigens* QYMF	0.0	80619
g.2830.pp.9	*Faecalibacterium prausnitzii* SL3/3	36.5	29788
g.2831.pp.4	*Ruminococcus* SR1/5	77.5	70467
g.3125.pp.4	*Clostridium symbiosum* WAL-14163	391.8	96577
g.3126.pp.4	*Clostridium symbiosum* WAL-14673	1.4	119259
g.28311.pp.1	Megasphaera	6.6	115414
g.1980.pp.1	*Anaerococcus lactolyticus* ATCC 51172	0.3	58674
g.26742.pp.5	*Fusobacterium ulcerans* 12 1B	0.0	47606
g.963.pp.4	*Magnetospirillum magneticum* AMB-1	0.0	36223
g.963.pp.2	*Magnetospirillum magneticum* AMB-1	0.0	54361

Table 3-2: Taxonomic details of the 170 predicted prophage-containing regions**,** continued.

Prophage ID	***Organism***	***Mean viral read coverage***	***length (nt)***
g.2482.pp.4	*Sideroxydans lithotrophicus* ES-1	0.2	55744
g.23643.pp.2	*Burkholderia cepacia* H111	13.4	47047
g.27505.pp.2	Pseudogulbenkiania NH8B	9.7	77560
g.24302.pp.1	*Janthinobacterium lividum* PAMC 25724	0.0	52239
g.31758.pp.2	Pandoraea SD6-2	0.0	23801
g.1257.pp.4	*Leptothrix cholodnii* SP-6	0.0	58690
g.27080.pp.7	*Desulfobacula* Tol2	0.0	104610
g.27080.pp.5	*Desulfobacula* Tol2	0.0	75984
g.27016.pp.1	Desulfovibrio U5L	0.0	60322
g.1144.pp.4	*Stigmatella aurantiaca* DW4/3-1	0.0	286337
g.3399.pp.3	Gammaproteobacteria HdN1	0.0	37777
g.24569.pp.1	*Vibrio cyclitrophicus* 1F273	0.0	33812
g.31903.pp.2	*Pseudomonas fluorescens* 1YdBTEX2	1024.0	52322
g.23533.pp.2	*Pseudomonas aeruginosa* 213BR	11.1	46816
g.530.pp.1	*Psychrobacter arcticus* 273-4	0.0	74819
g.23534.pp.3	*Pseudomonas aeruginosa* 9BR	11.5	46816
g.23381.pp.3	*Pseudomonas fluorescens* A506	409.2	59385
g.31624.pp.2	*Thioalkalivibrio* sp. ALMg9	0.0	56806
g.27954.pp.1	*Escherichia coli* B41	253.1	82479
g.30632.pp.1	*Pseudomonas* sp. CBZ-4	592.4	57553
g.27535.pp.5	*Escherichia coli* G58-1	204.0	67770
g.24636.pp.3	*Pseudomonas* sp. GM18	1.7	41848
g.24647.pp.2	*Pseudomonas* sp. GM67	2.6	56068
g.28684.pp.1	*Vibrio cholerae* HE-09	0.0	28223
g.86.pp.1	*Pseudomonas fluorescens* HK44	0.6	26902
g.25008.pp.2	*Rhodanobacter thiooxydans* LCS2	0.0	30708
g.28509.pp.2	*Pseudomonas* sp. M1	36.4	64455
g.25178.pp.2	*Pseudomonas* sp. M47T1	22.5	58574
g.31227.pp.3	*Acinetobacter* sp. NIPH 899	117.1	47100
g.1207.pp.8	*Stenotrophomonas* sp. SKA14	13.6	44217
g.1343.pp.3	*Xenorhabdus bovienii* SS-2004	0.0	135748
g.1293.pp.2	*Legionella pneumophila* Corby	0.0	21830

Table 3-3: Nucleotide coordinates of the DGR cassettes. For each predicted prophage-containing region, the location of the reverse transcriptase, template repeat and variable repeat is listed. Only the first TR/VR pair is reported.

Prophage ID	***Reverse transcriptase***	***Template repeat***	***Variable repeat***	***Reference***
g.1032.pp.1	21289-20015	21275-21167	21819-21711	Simon & Zimmerly (2008)
g.1144.pp.4	122927-121281	85098-85002	85980-85884	
g.1207.pp.8	36002-37321	36070-36124	32221-32275	Wu et al. (2017)
g.1257.pp.4	25140-24121	25401-25286	25993-25878	Wu et al. (2017)
g.1293.pp.2	2304-3503	2090-2229	1370-1509	Wu et al. (2017)
g.1302.pp.2	45827-46765	45542-45640	44884-44982	
g.1343.pp.3	102647-104017	75177-75235	130948-131006	
g.1354.pp.3	37266-38405	36942-37056	36267-36381	Wu et al. (2017)
g.1363.pp.1	19748-21046	19497-19607	19155-19265	Wu et al. (2017)
g.1363.pp.4	38932-38003	39349-39226	39980-39857	Wu et al. (2017)
g.1365.pp.3	41623-42768	41309-41418	41040-41149	Wu et al. (2017)
g.1368.pp.5	78709-80130	12089-12011	78030-78108	
g.1372.pp.5	57163-58308	56768-56881	56185-56298	Wu et al. (2017)
g.1375.pp.2	3952-2729	4244-4149	4507-4412	Schillinger et al. (2012)
g.1380.pp.3	45866-44856	39546-39670	46837-46713	Wu et al. (2017)
g.1381.pp.1	31581-32816	31241-31367	30503-30629	Wu et al. (2017)
g.1522.pp.3	4686-4102	5393-5280	5938-5825	Wu et al. (2017)
g.1527.pp.1	20877-19549	20997-20870	21661-21534	Wu et al. (2017)
g.1552.pp.2	11439-10252	11662-11584	54710-54640	
g.1599.pp.6	90872-92059	90631-90728	90430-90527	Schillinger et al. (2012)
g.1601.pp.2	131838-133259	131944-132058	131173-131287	
g.1602.pp.4	25780-24476	15452-15337	26474-26359	Schillinger et al. (2012)
g.1606.pp.6	24947-24531	35308-35252	25972-25916	
g.1613.pp.5	13201-11924	13524-13412	13792-13680	Wu et al. (2017)
g.1615.pp.3	39861-40871	39545-39670	38891-39016	
g.1616.pp.10	47614-48624	47355-47432	9402-9479	Wu et al. (2017)
g.1616.pp.11	2530-1454	3070-2952	3427-3309	Schillinger et al. (2012)
g.1648.pp.1	44019-45512	44171-44259	43442-43530	Wu et al. (2017)
g.1648.pp.4	25487-26977	25304-25379	24295-24370	Wu et al. (2017)
g.1649.pp.4	60828-62249	12089-11999	60149-60239	

Table 3-3: Nucleotide coordinates of the DGR cassettes, continued.

Prophage ID	*Reverse transcriptase*	*Template repeat*	*Variable repeat*	*Reference*
g.1653.pp.13	34969-34154	35361-35251	35648-35538	
g.1656.pp.9	2926-1622	15453-15361	3621-3529	Wu et al. (2017)
g.1738.pp.2	28698-29840	28353-28481	27623-27751	Schillinger et al. (2012)
g.1750.pp.2	47366-48667	47051-47166	46448-46563	Schillinger et al. (2012)
g.1750.pp.6	14993-13824	15454-15337	15823-15706	
g.1763.pp.12	23665-22418	17771-17721	51406-51356	Wu et al. (2017)
g.1894.pp.4	30207-28951	30494-30343	30758-30607	Schillinger et al. (2012)
g.1894.pp.6	20059-18665	47290-46284	20062-19056	
g.1895.pp.2	3479-2229	3882-3801	4131-4050	Wu et al. (2017)
g.1910.pp.5	8281-7154	8598-8479	8847-8728	
g.1947.pp.2	86766-87827	86327-86457	85599-85729	Wu et al. (2017)
g.1980.pp.1	15379-14261	15631-15538	13615-13522	Wu et al. (2017)
g.2060.pp.4	113063-114961	10778-10827	49480-49529	
g.2141.pp.5	40597-41430	39885-39958	39238-39311	Wu et al. (2017)
g.2234.pp.4	5884-4661	6172-6080	4098-4006	Schillinger et al. (2012)
g.226.pp.8	41002-41535	41651-41765	39880-39994	Wu et al. (2017)
g.2289.pp.1	64048-65469	64126-64216	63369-63459	
g.2290.pp.2	12167-10746	12089-11999	12846-12756	
g.2290.pp.4	39718-41013	39391-39508	39105-39222	
g.23381.pp.3	9413-8418	9617-9518	11032-10933	Wu et al. (2017)
g.23533.pp.2	12688-11618	12838-12740	13545-13447	Wu et al. (2017)
g.23534.pp.3	12688-11618	12801-12739	14194-14132	
g.23643.pp.2	39090-40085	38910-38962	37768-37857	Wu et al. (2017)
g.2412.pp.2	41539-42840	47411-48403	79949-80941	
g.2418.pp.1	23902-21791	33264-33200	26358-26294	
g.24302.pp.1	8816-7800	9004-8904	10956-10856	Wu et al. (2017)
g.24569.pp.1	3332-2289	3578-3466	4119-4007	Paul et al. (2017)
g.24636.pp.3	33269-34270	30143-30243	32373-32473	Wu et al. (2017)
g.24647.pp.2	41938-42993	30143-30243	40361-40463	Wu et al. (2017)
g.24656.pp.5	20358-21356	20040-20161	19430-19551	Paul et al. (2017)
g.2482.pp.2	26911-25817	27052-26956	28123-28027	Wu et al. (2017)
g.2482.pp.4	12673-11738	12986-12870	14003-13921	Wu et al. (2017)

Table 3-3: Nucleotide coordinates of the DGR cassettes, continued.

Prophage ID	*Reverse transcriptase*	*Template repeat*	*Variable repeat*	*Reference*
g.27080.pp.7	56127-57437	75997-76060	91263-91326	
g.27096.pp.1	51577-50264	51873-51742	52528-52397	Wu et al. (2017)
g.27096.pp.15	27049-28221	26626-26748	25580-25702	Wu et al. (2017)
g.27096.pp.3	30637-29456	30994-30872	31758-31636	Wu et al. (2017)
g.27096.pp.7	32098-33273	31736-31876	31410-31550	Wu et al. (2017)
g.27099.pp.2	53178-54434	52812-52930	52469-52587	
g.27128.pp.1	8100-6898	8379-8293	8637-8551	Wu et al. (2017)
g.27128.pp.10	21323-22522	21058-21153	20256-20351	Wu et al. (2017)
g.27137.pp.4	47452-48612	47185-47293	46970-47078	
g.27343.pp.3	9751-10686	5389-5280	8854-8963	Wu et al. (2017)
g.27474.pp.2	94268-95083	35361-35252	93589-93698	
g.27478.pp.3	43302-44339	2036-2111	22959-23034	Wu et al. (2017)
g.27505.pp.2	41742-40636	42829-42727	43525-43423	Wu et al. (2017)
g.27535.pp.5	18442-18876	7498-7556	39052-39110	
g.2764.pp.3	49705-48608	46934-46816	47249-47131	Wu et al. (2017)
g.27954.pp.1	18445-18879	7498-7556	39057-39115	
g.27983.pp.1	62022-63527	75876-76417	60038-60399	
g.2820.pp.2	8071-6650	12089-11999	8750-8660	Wu et al. (2017)
g.2827.pp.4	60560-59520	39546-39671	61499-61374	Guo et al. (2011)
g.2827.pp.9	33560-34810	3883-3757	32907-33033	
g.2830.pp.9	25301-26485	24991-25043	24736-24788	Wu et al. (2017)
g.2831.pp.4	53888-55045	53553-53646	52958-53051	Wu et al. (2017)
g.28311.pp.1	97171-98145	96774-96846	96211-96283	Wu et al. (2017)
g.2839.pp.2	3916-3044	4471-4364	4735-4628	Wu et al. (2017)
g.2844.pp.4	26446-25415	26617-26550	27746-27679	Wu et al. (2017)
g.28509.pp.2	26918-25752	26959-26860	27642-27543	Wu et al. (2017)
g.2867.pp.3	15709-14486	16104-15985	16470-16351	Schillinger et al. (2012)
g.28684.pp.1	23505-24506	23246-23365	22701-22820	Wu et al. (2017)
g.28908.pp.3	2926-1622	15452-15337	3620-3505	
g.28909.pp.6	17974-16670	15452-15337	18668-18553	
g.28910.pp.10	17974-16670	15452-15337	18668-18553	Wu et al. (2017)
g.28916.pp.7	78271-77456	35361-35252	41843-41952	Wu et al. (2017)
g.28917.pp.2	42522-43337	35361-35252	41843-41952	

Table 3-3: Nucleotide coordinates of the DGR cassettes, continued.

Prophage ID	*Reverse transcriptase*	*Template repeat*	*Variable repeat*	*Reference*
g.28919.pp.10	45169-46809	44872-44993	44490-44611	Wu et al. (2017)
g.28919.pp.8	45314-46804	25266-25379	44082-44195	Wu et al. (2017)
g.28920.pp.6	40612-41631	40010-40126	39725-39841	Wu et al. (2017)
g.28920.pp.7	51326-52816	25266-25358	50092-50184	Wu et al. (2017)
g.28922.pp.3	3780-2527	4160-4050	4512-4402	
g.28922.pp.9	2986-1691	39392-39508	3598-3482	
g.28923.pp.1	15261-16301	14872-14985	12394-12507	
g.28925.pp.2	11019-9979	11213-11137	50254-50178	
g.28927.pp.3	9803-8739	10026-9961	10609-10544	Wu et al. (2017)
g.28933.pp.1	3976-2723	4317-4225	4659-4567	Wu et al. (2017)
g.29562.pp.7	2209-905	15452-15337	2903-2788	
g.29595.pp.3	39706-40779	39410-39526	38841-38957	
g.3007.pp.10	16363-15812	20929-20870	15584-15525	
g.3007.pp.12	24447-25922	26156-26212	25329-25385	
g.3007.pp.3	15459-15432	9649-9854	9259-9464	
g.3007.pp.8	15184-15480	29781-29831	22410-22460	
g.3043.pp.2	37984-39288	15454-15337	37301-37418	Schillinger et al. (2012)
g.30632.pp.1	14644-13499	14752-14650	16155-16053	
g.3070.pp.2	15427-13910	15529-15408	15886-15765	Schillinger et al. (2012)
g.3088.pp.2	32355-33416	27059-27130	21656-21727	
g.3099.pp.1	48800-47790	48986-48937	85029-84980	Schillinger et al. (2012)
g.3102.pp.5	45058-43802	45351-45263	45607-45519	Wu et al. (2017)
g.3105.pp.8	107629-108858	107337-107458	109366-109487	Wu et al. (2017)
g.31087.pp.3	113416-114603	113050-113155	112813-112918	
g.31101.pp.1	33922-33053	34538-34402	34797-34661	
g.31227.pp.3	14359-13265	14628-14518	15190-15080	
g.3125.pp.4	63219-61954	63389-63324	63617-63552	Wu et al. (2017)
g.3126.pp.4	66439-65237	67026-66931	68511-68416	Wu et al. (2017)
g.31348.pp.2	7794-6451	7996-7947	8368-8319	
g.31350.pp.5	3178-1538	3481-3356	3892-3767	

Table 3-3: Nucleotide coordinates of the DGR cassettes, continued.

Prophage ID	***Reverse transcriptase***	***Template repeat***	***Variable repeat***	***Reference***
g.31350.pp.6	29700-30734	29472-29582	28927-29037	
g.31351.pp.4	36494-37534	36266-36374	36067-36175	
g.31354.pp.6	13124-11910	13470-13343	14204-14077	
g.31357.pp.5	24727-25845	24392-24485	26508-26601	
g.31410.pp.3	21526-22653	21442-21524	20690-20772	
g.3151.pp.2	90551-91546	9617-9517	89652-89752	Schillinger et al. (2012)
g.31515.pp.2	2030-1044	2540-2424	4596-4480	
g.31540.pp.1	14528-13455	14648-14552	15795-15699	
g.31624.pp.2	4763-3732	5048-4939	5579-5470	
g.31758.pp.2	14582-15919	14738-14831	14070-14163	
g.31903.pp.2	30182-31342	30143-30243	28753-28853	
g.32274.pp.1	11991-10924	12219-12127	10377-10285	
g.32364.pp.2	6764-5628	6841-6748	7962-7871	Schillinger et al. (2012)
g.32458.pp.11	64699-63662	2029-2110	65489-65408	
g.32460.pp.7	2231-3268	2029-2110	1441-1522	
g.32467.pp.6	54059-55363	44686-44747	496-557	
g.3257.pp.2	12180-11854	13135-13061	5731-5657	
g.3399.pp.3	14820-16076	29742-29816	29355-29429	
g.3439.pp.4	111672-112886	74035-74118	69826-69909	Schillinger et al. (2012)
g.345.pp.1	1857-613	2166-2061	2405-2300	Wu et al. (2017)
g.436.pp.7	27452-25794	42995-42929	44405-44339	Doulatov et al. (2004)
g.488.pp.2	39358-39861	52801-52936	60731-60866	
g.530.pp.1	10857-9790	11109-10994	11677-11562	Simon & Zimmerly (2008)
g.680.pp.4	30069-28975	21991-21927	51224-51160	Wu et al. (2017)
g.72.pp.2	102916-103296	5919-5997	35865-35943	
g.86.pp.1	5928-5125	6074-6024	13373-13307	
g.963.pp.2	42579-43694	42490-42545	41778-41833	Wu et al. (2017)
g.963.pp.4	7772-6645	7860-7803	9176-9091	Wu et al. (2017)
g.996.pp.5	44021-45049	43726-43842	43140-43256	Paul et al. (2017)

Table 3-4: Open reading frames (ORFs) encoded in the Hankyphage genome. Functional annotations for each ORF were determined by comparisons to the Pfam-A, Phyre2, or conserved domain database. The database and accession to the significant hit of each ORF is listed with the corresponding e-value or confidence score.

ORF	***Start***	***Stop***	***Strand***	***Annotation***	***Database (accession)***	***Score or E-value***
1	219	1073	-	DNA methyltransferase	Pfam-A (PF02086)	2.70E-09
2	1169	1807	-	thymidylate synthase	Pfam-A (PF00303)	2.70E-17
3	1886	2128	-	hypothetical		
4	2308	2793	-	hypothetical		
5	2790	4094	-	reverse transcriptase	Pfam-A (PF00078)	
6	4211	4414	-	hypothetical		
7	4611	11348	-	hypothetical		
8	11375	15391	-	tail-associated	Phyre2 (3GS9)	98.8
9	15438	15902	-	baseplate hub	Phyre2 (2X8K)	98
10	15904	20295	-	tape measure	CDD (cI28567)	1.41E-05
11	20296	20496	-	hypothetical		
12	20535	21026	-	hypothetical		
13	21030	21605	-	hypothetical		
14	21602	21925	-	hypothetical		
15	21891	22370	-	hypothetical		
16	22398	22784	-	hypothetical		
17	22797	23261	-	lysin	Pfam-A (PF01510)	4.10E-08
18	23261	24484	-	major capsid protein	Phyre2 (2FT1)	97.3
19	24496	25533	-	clp-protease	Pfam-A (PF00574)	1.50E-30

Table 3-4: Open reading frames (ORFs) encoded in the Hankyphage genome, continued.

ORF	***Start***	***Stop***	***Strand***	***Annotation***	***Database (accession)***	***Score or E-value***
20	25730	26176	+	TerS	Pfam-A (PF06056)	3.00E-04
21	26178	27746	+	TerL	Phyre2 (4BIJ)	100
22	27765	28190	+	hypothetical	Pfam-A (DUF1320)	2.70E-10
23	28192	29580	+	portal	Phyre2 (4ZJN)	
24	29574	31013	+	Mu F head completion	Pfam-A (04233)	1.00E-08
25	30996	31211	-	hypothetical		
26	31298	31930	+	hypothetical		
27	31935	32381	+	neck	Phyre2 (2GJV)	97.6
28	32391	32660	-	hypothetical		
29	32768	32983	-	antiterminator	Phyre2 (4MO1)	93.6
30	32986	33636	-	hypothetical		
31	33658	34101	-	hypothetical		
32	34104	34244	-	hypothetical		
33	34285	34473	-	hypothetical		
34	34515	35027	-	hypothetical	Phyre2 (3IWF)	90.7
35	35048	35353	-	hypothetical		
36	35373	36278	-	hypothetical		
37	36282	36518	-	chaperone DnaJ	Phyre2 (2CTT)	95.4
38	36515	37192	-	helicase	Phyre2 (5LKM)	100
39	37195	37593	-	hypothetical		
40	37590	38459	-	MuB	CDD (cI26673)	4.18E-06
41	38495	40534	-	MuA	Phyre2 (4FCY)	100
42	40550	40756	-	hypothetical		
43	40761	40961	-	lacI-like regulator	Phyre2 (3H5F)	93.1
44	41081	41272	-	hypothetical		
45	41501	41827	+	hypothetical		

Table 3-5: Annotation of the VR-containing ORFs in DGRs. All ORFs with a VR were annotated using Phyre2. Annotations were only assigned if the confidence is greater than 97; otherwise, the annotation is listed as N/A. The presence or absence of transmembrane-spanning regions (Tm) and signal peptides were identified using Phobius and indicated as 'yes', 'no' and 'signal peptide', respectively.

Prophage ID	***VR ORF annotation***	***PDB code***	***Confidence***	***Tm***
g.1032.pp.1	TaqVP	5vf4A	100	No
g.1144.pp.4	Superkiller protein 3	4bujF	100	No
g.1207.pp.8	Pilin tip	4epsA	98.5	No
g.1257.pp.4	TaqVP	5vf4A	100	No
g.1293.pp.2	Pilin tip	4epsA	99.9	Signal peptide
g.1302.pp.2	Pilin tip	4epsA	99.8	No
g.1343.pp.3	N/A	N/A	N/A	No
g.1354.pp.3	TaqVP	5vf4A	100	No
g.1363.pp.1	TaqVP	5vf4A	100	No
g.1363.pp.4	TaqVP	5vf4A	100	No
g.1365.pp.3	TaqVP	5vf4A	100	No
g.1368.pp.5	Pilin tip	4epsA	99.9	No
g.1372.pp.5	TaqVP	5vf4A	100	No
g.1375.pp.2	TaqVP	5vf4A	100	No
g.1380.pp.3	MTD-P1	2iouC	100	No
g.1381.pp.1	N/A	N/A	N/A	No
g.1522.pp.3	TaqVP	5vf4A	100	No
g.1527.pp.1	N/A	N/A	N/A	No
g.1552.pp.2	TaqVP	5vf4A	100	No
g.1599.pp.6	TaqVP	5vf4A	99.9	No
g.1601.pp.2	Spore coat protein H	5jd9A	99.5	No
g.1602.pp.4	Spore coat protein H	5jd9A	99.8	No
g.1606.pp.6	TaqVP	5vf4A	100	No
g.1613.pp.5	Uncharacterized protein	4e40A	99.8	No
g.1615.pp.3	MTD-P1	2iouC	100	No
g.1616.pp.10	TaqVP	5vf4A	100	No
g.1616.pp.11	TaqVP	5vf4A	100	No
g.1648.pp.1	Spore coat protein H	5jd9A	99.4	No
g.1648.pp.4	Spore coat protein H	5jd9A	99.4	No
g.1649.pp.4	Pilin tip	4epsA	99.9	No

Table 3-5: Annotation of the VR-containing ORFs in DGRs, continued.

Prophage ID	*VR ORF annotation*	*PDB code*	*Confidence*	*Tm*
g.1656.pp.9	Spore coat protein H	5jd9A	99.8	No
g.1738.pp.2	N/A	N/A	N/A	No
g.1750.pp.2	Spore coat protein H	5jd9A	99.8	No
g.1750.pp.6	Spore coat protein H	5jd9A	99.8	No
g.1763.pp.12	N/A	N/A	N/A	No
g.1894.pp.4	TaqVP	5vf4A	98.4	No
g.1894.pp.6	TaqVP	5vf4A	98.4	No
g.1895.pp.2	TaqVP	5vf4A	100	No
g.1910.pp.5	TaqVP	5vf4A	100	No
g.1947.pp.2	N/A	N/A	N/A	No
g.1980.pp.1	Receptor-type Tyrosine-protein phosphatase Mu	2v5yA	99.7	No
g.2060.pp.4	N/A	N/A	N/A	No
g.2141.pp.5	N/A	N/A	N/A	No
g.2234.pp.4	Anosmin-1	1zlgA	99.8	No
g.226.pp.8	Chromophore maturation protein PvdO	5hhaB	100	No
g.2289.pp.1	Pilin tip	4epsA	99.9	No
g.2290.pp.2	Spore coat protein H	5jd9A	99.7	No
g.2290.pp.4	Spore coat protein H	5jd9A	99.7	No
g.23381.pp.3	Pilin tip	4epsA	99.8	No
g.23533.pp.2	Pilin tip	4epsA	99.7	No
g.23534.pp.3	Pilin tip	4epsA	99.8	No
g.23643.pp.2	Pilin tip	4epsA	99	No
g.2412.pp.2	transposase	1mus	99.4	No
g.2418.pp.1	N/A	N/A	N/A	No
g.24302.pp.1	N/A	N/A	N/A	No
g.24569.pp.1	TaqVP	5vf4A	100	No
g.24636.pp.3	Pilin tip	4epsA	99.9	No
g.24647.pp.2	Pilin tip	4epsA	99.8	No
g.24656.pp.5	MTD-P1	2iouC	100	No
g.2482.pp.2	N/A	N/A	N/A	No
g.2482.pp.4	N/A	N/A	N/A	No
g.25008.pp.2	Pilin tip	4epsA	98	No
g.25178.pp.2	Pilin tip	4epsA	99.8	Signal peptide
g.2542.pp.7	TaqVP	5vf4A	100	No

Table 3-5: Annotation of the VR-containing ORFs in DGRs, continued.

Prophage ID	*VR ORF annotation*	*PDB code*	*Confidence*	*Tm*
g.2615.pp.2	Adenine-N6-DNA-Methyltransferase	1aqjB	99.6	No
g.26274.pp.7	N/A	N/A	N/A	No
g.2640.pp.3	RTX-adhesin	5irbB	99.8	No
g.26473.pp.3	N/A	N/A	N/A	No
g.2650.pp.3	Pilin tip	4epsA	99.1	No
g.26528.pp.6	N/A	N/A	N/A	No
g.26742.pp.5	Flavocytochrome C fumarate reductase	1d4c	100	No
g.26919.pp.1	N/A	N/A	N/A	No
g.26931.pp.6	N/A	N/A	N/A	No
g.26941.pp.2	Spore coat protein H	5jd9A	99.4	No
g.27016.pp.1	TaqVP	5vf4A	100	No
g.27080.pp.5	Protease Do-like 5	4ic5	98.7	No
g.27080.pp.7	Protease Do-like 5	4ic5	98.7	No
g.27096.pp.1	TaqVP	5vf4A	100	No
g.27096.pp.15	TaqVP	5vf4A	100	No
g.27096.pp.3	TaqVP	5vf4A	100	No
g.27096.pp.7	TaqVP	5vf4A	100	No
g.27099.pp.2	TaqVP	5vf4A	99.5	No
g.27128.pp.1	TaqVP	5vf4A	99.9	No
g.27128.pp.10	TaqVP	5vf4A	99.9	No
g.27137.pp.4	TaqVP	5vf4A	100	No
g.27343.pp.3	TaqVP	5vf4A	97.8	No
g.27474.pp.2	TaqVP	5vf4A	100	No
g.27478.pp.3	N/A	N/A	N/A	No
g.27505.pp.2	Pilin tip	4epsA	99.9	No
g.27535.pp.5	N/A	N/A	N/A	No
g.2764.pp.3	Chromophore maturation protein PvdO	5hhaB	100	No
g.27954.pp.1	N/A	N/A	N/A	No
g.27983.pp.1	N/A	N/A	N/A	No
g.2820.pp.2	Pilin tip	4epsA	99.9	No
g.2827.pp.4	TaqVP	5vf4A	100	No
g.2827.pp.9	TaqVP	5vf4A	100	No
g.2830.pp.9	TaqVP	5vf4A	98.5	No
g.2831.pp.4	TaqVP	5vf4A	100	No

Table 3-5: Annotation of the VR-containing ORFs in DGRs, continued.

Prophage ID	*VR ORF annotation*	*PDB code*	*Confidence*	*Tm*
g.2839.pp.2	TaqVP	5vf4A	100	Signal peptide
g.2844.pp.4	N/A	N/A	N/A	No
g.28509.pp.2	Pilin tip	4epsA	99.9	No
g.2867.pp.3	TaqVP	5vf4A	100	No
g.28684.pp.1	MTD-P1	2iouC	100	No
g.28908.pp.3	N/A	N/A	N/A	No
g.28909.pp.6	Spore coat protein H	5jd9A	99.7	No
g.28910.pp.10	Spore coat protein H	5jd9A	99.8	No
g.28916.pp.7	N/A	N/A	N/A	No
g.28917.pp.2	N/A	N/A	N/A	No
g.28919.pp.10	TaqVP	5vf4A	100	No
g.28919.pp.8	TaqVP	5vf4A	100	No
g.28920.pp.6	Spore coat protein H	5jd9A	99.7	No
g.28920.pp.7	Spore coat protein H	5jd9A	99.7	No
g.28922.pp.3	Spore coat protein H	5jd9A	99.7	No
g.28922.pp.9	Spore coat protein H	5jd9A	99.7	No
g.28923.pp.1	Pilin tip	4epsA	100	Yes
g.28925.pp.2	TaqVP	5vf4A	100	No
g.28927.pp.3	TaqVP	5vf4A	100	Yes
g.28933.pp.1	Spore coat protein H	5jd9A	99.4	No
g.29562.pp.7	Spore coat protein H	5jd9A	99.8	No
g.29595.pp.3	TaqVP	5vf4A	100	No
g.3007.pp.10	Cas9	5vgbA	99.6	No
g.3007.pp.12	Cas9	5vgbA	99.6	No
g.3007.pp.3	Cas9	5vgbA	99.6	No
g.3007.pp.8	Cas9	5vgbA	99.6	No
g.3043.pp.2	Spore coat protein H	5jd9A	99.8	No
g.30632.pp.1	Pilin tip	4epsA	99.9	No
g.3070.pp.2	TaqVP	5vf4A	100	No
g.3088.pp.2	N/A	N/A	N/A	No
g.3099.pp.1	TaqVP	5vf4A	100	No
g.3102.pp.5	TaqVP	5vf4A	98.3	No
g.3105.pp.8	Fibronectin(III)-like module	3pe9A	96.6	No
g.31087.pp.3	TaqVP	5vf4A	99.7	No
g.31101.pp.1	TaqVP	5vf4A	100	No
g.31227.pp.3	TaqVP	5vf4A	100	No

Table 3-5: Annotation of the VR-containing ORFs in DGRs, continued.

Prophage ID	***VR ORF annotation***	***PDB code***	***Confidence***	***Tm***
g.3126.pp.4	N/A	N/A	N/A	No
g.31348.pp.2	TaqVP	5vf4A	99.9	No
g.31350.pp.5	TaqVP	5vf4A	100	No
g.31350.pp.6	TaqVP	5vf4A	100	No
g.31351.pp.4	TaqVP	5vf4A	100	No
g.31354.pp.6	N/A	N/A	N/A	No
g.31357.pp.5	Receptor-type Tyrosine-protein phosphatase Mu	2v5yA	99.8	No
g.31410.pp.3	Pilin tip	4epsA	99.7	No
g.3151.pp.2	Pilin tip	4epsA	99.8	No
g.31515.pp.2	Sulfoxide synthase EgtB	4x8bA	100	No
g.31540.pp.1	Pilin tip	4epsA	99	No
g.31624.pp.2	MTD-P1	2iouC	100	No
g.31758.pp.2	Pilin tip	4epsA	98.9	No
g.31903.pp.2	Pilin tip	4epsA	99.8	No
g.32274.pp.1	Adhesin	4m00A	99.5	No
g.32364.pp.2	Pilin tip	4epsA	99.1	No
g.32458.pp.11	N/A	N/A	N/A	No
g.32460.pp.7	N/A	N/A	N/A	No
g.32467.pp.6	Phosphoadenosine phosphosulfate reductase	2oq2B	100	No
g.3257.pp.2	Pyrophosphokinase	1cbka	100	No
g.3399.pp.3	N/A	N/A	N/A	No
g.3439.pp.4	Fimbrial protein	4qb7A	100	No
g.345.pp.1	N/A	N/A	N/A	No
g.436.pp.7	N/A	N/A	N/A	No
g.488.pp.2	surfactin synthetase	2vsqA	100	No
g.530.pp.1	TaqVP	5vf4A	100	No
g.680.pp.4	Rep-associated tyrosine transposase	4er8A	99.8	No
g.72.pp.2	N/A	N/A	N/A	No
g.86.pp.1	Pilin tip	4epsA	99.8	No
g.963.pp.2	Pilin tip	4epsA	99.9	No
g.963.pp.4	Pilin tip	4epsA	99.9	No
g.996.pp.5	TaqVP	5vf4A	100	No

Table 3-6: Primers used in the study (written 5’ to 3’).

Primer name	***Sequence***
Φ1F	5’GCCTCCTCAATCAACCCTAATG
Φ1R	5’ACCGATGTTCGTGTTCGTATTC
Φ2F	5’ ATTCACCGAAGACAGTACCC
Φ2R	5’ GAAGGAATACGGAAAGATTACGC
Φ3F	5’ GGAGGAAAGGACATAATAACCG
Φ3R	5’CAACTACCTTTACTGGTACGG
16s (27F)	5’AGAGTTTGATCMTGGCTCAG
16s (1492R)	5’ GGTTACCTTGTTACGACTT

Equation S1: Calculation of the global number of Hankyphage virions.

$$Virions = \phi_{Hanky} \times H \times p \times n \qquad (1)$$

Where ϕ_{Hanky} is the average fractional abundance of Hankyphage in the human viral community, H equals the number of human associated VLPs, p is the fraction of individuals with Hankyphage and n is the number humans globally.

The average fractional abundance of Hankyphage (ϕ_{Hanky}) was calculated from supplemental data 1. The number of human-associated VLPs (H) is calculated in Cobian-Guemes et al. (Cobián Güemes *et al.*, 2016). To calculate the fraction of individuals with Hankyphage (p), 1811 metagenomes from 11 SRA projects were analyzed (supplemental data 2) and are listed below. If a metagenome contains reads mapping to $\geq$ 10% of the Hankyphage genome, Hankyphage is considered present in the metagenome.

Table 3-7: Metagenomes surveyed for Hankyphage.

Project	***Metagenomes***	***Location***	***Fraction of metagenomes with Hankyphage***
SRP075633	57	USA	0.63
SRP000319	6	USA	0.67
SRP072561	17	Luxembourg	0.59
ERP013092	24	Finland	0.59
ERP005989	44	Denmark	0.55
ERP007090	30	Sweden	0.53
SRP067755	75	USA	0.51
ERP014167	8	China	0.50
SRP002163	1427	USA	0.50
ERP015450	118	China	0.42
Average			0.53

Solving for each variable, this leads to

$$Virions = 2.2 \times 10^{-5} \cdot 3.19 \times 10^{12} \cdot 0.53 \cdot 7.4 \times 10^{9} = 3.3 \times 10^{17} \qquad (2)$$

References

Adams, M.H. (1959). Bacteriophages (New York: Interscience Publishers, Inc.).

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. *40*, 1–13.

Allet, B. (1979). Mu insertion duplicates a 5 base pair sequence at the host inserted site. Cell *16*, 123–129.

Arambula, D., Wong, W., Medhekar, B. a, Guo, H., Gingery, M., Czornyj, E., Liu, M., Dey, S., Ghosh, P., and Miller, J.F. (2013). Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. Proc. Natl. Acad. Sci. U. S. A. *110*, 8212–8217.

Barr, J.J., Auro, R., Sam-Soon, N., Kassegne, S., Peters, G., Bonilla, N., Hatay, M., Mourtada, S., Bailey, B., Youle, M., et al. (2015). Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. Proc. Natl. Acad. Sci. *112*, 13675–13680.

Bearson, B.L., Allen, H.K., Brunelle, B.W., Lee, I.S., Casjens, S.R., and Stanton, T.B. (2014). The agricultural antibiotic carbadox induces phage-mediated gene transfer in Salmonella. Front. Microbiol. *5*, 1–8.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral Genomes resource. Nucleic Acids Res. *43*, D571–D577.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. Annu. Rev. Virol. *3*, 197–214.

Le Coq, J., and Ghosh, P. (2011). Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. Proc. Natl. Acad. Sci. *108*, 14649–14653.

Cornuault, J.K., Petit, M.-A., Mariadassou, M., Benevides, L., Moncaut, E., Langella, P., Sokol, H., and De Paepe, M. (2018). Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes. Microbiome *6*, 65.

Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R., Zimmerly, S., and Miller, J. (2004). Tropism Switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature *431*, 476–481.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 1–11.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont,

T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. *39*, 29–37.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E. V (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. *43*, D261–D269.

Hannigan, G.D., Zheng, Q., Meisel, J.S., Minot, S.S., Bushman, F.D., and Grice, E.A. (2017). Evolutionary and functional implications of hypervariable loci within the skin virome. PeerJ *5*, e2959.

Huttenhower, C., Fah Sathirapongsasuti, J., Segata, N., Gevers, D., Earl, A.M., Fitzgerald, M.G., Young, S.K., Zeng, Q., Alm, E.J., Alvarado, L., et al. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Hyatt, D., Chen, G., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal : prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*.

Johnson, T.A., Looft, T., Severin, A.J., Bayles, D.O., Nasko, D.J., Wommack, K.E., Howe, A., and Allen, H.K. (2017). The In-Feed Antibiotic Carbadox Induces Phage Gene Transcription in the Swine Gut Microbiome. MBio *8*, e00709-17.

Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. *35*, 429–432.

Kang, H.S. (2017). Phispy prophage predictions.

Kang, H.S., McNair, K., Cuevas, D., Bailey, B., Segall, A., and Edwards, R.A. (2017). Prophage genomics reveals patterns in phage genome organization and replication. BioRxiv 1–28.

Kelly, L.A., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015). The Phyre2 web

portal for protein modelling, prediction, and analysis. Nat. Protoc. *10*, 845–858.

Kohler, B., Karch, H., and Schmidt, H. (2000). Antibacterials that are used as growth promoters in animal husbandry can affect the release of Shiga-toxin-2 converting bacteriophages and Shiga toxin 2 from Escherichia coli strains. Microbiology *146*, 1085–1090.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Levi, K. BAM scripts.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, M., Deora, R., Doulatov, S.R., Gingery, M., Eiserling, F.A., Preston, A., Maskell, D.J., Simons, R.W., Cotter, P.A., Parkhill, J., et al. (2002). Reverse Transcriptase – Mediated Tropism Switching in Bordetella. Science (80-.). *295*, 2091–2095.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., et al. (2013). CDD: Conserved domains and protein three-dimensional structure. Nucleic Acids Res. *41*, 348–352.

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

Miller, J.L., Le Coq, J., Hodes, A., Barbalat, R., Miller, J.F., and Ghosh, P. (2008). Selective ligand recognition by a diversity-generating retroelement variable protein. PLoS Biol. *6*, 1195–1207.

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2012). Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. *109*, 3962–3966.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. *110*, 12450–12455.

Naorem, S.S., Han, J., Wang, S., Lee, W.R., Heng, X., Miller, J.F., and Guo, H. (2017). DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. Proc. Natl. Acad. Sci. *114*, E10187–E10195.

Oakey, H.J., Cullen, B.R., and Owens, L. (2002). The complete nucleotide sequence of the Vibrio harveyi bacteriophage VHML. J. Appl. Microbiol. *93*, 1089–1098.

Paul, B.G., Bagby, S.C., Czornyj, E., Arambula, D., Handa, S., Sczyrba, A., Ghosh, P., Miller, J.F., and Valentine, D.L. (2015). Targeted diversity generation by intraterrestrial archaea and archaeal viruses. Nat. Commun. *6*, 6585.

Paul, B.G., Burstein, D., Castelle, C.J., Handa, S., Arambula, D., Czornyj, E., Thomas, B.C., Ghosh, P., Miller, J.F., Banfield, J.F., et al. (2017). Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat. Microbiol. *2*, 17045.

Ponstingl, H. (2010). SMALT.

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

Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc. Natl. Acad. Sci. *110*, 20236–20241.

Russell, D.A., and Hatfull, G.F. (2017). PhagesDB: The actinobacteriophage database. Bioinformatics *33*, 784–786.

Schillinger, T., Lisfi, M., Chi, J., Cullum, J., and Zingler, N. (2012). Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. BMC Genomics *13*, 430.

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

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*.

Simon, D.M., and Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. Nucleic Acids Res. *36*, 7219–7229.

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

Torres, P.J., Edwards, R.A., and McNair, K.A. (2017). PARTIE: A partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. Bioinformatics *33*, 2389–2391.

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

Wu, L., Gingery, M., Abebe, M., Arambula, D., Czornyj, E., Handa, S., Khan, H., Liu, M., Pohlschroder, M., Shaw, K.L., et al. (2017). Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res. 1–14.

Xu, Q., Shoji, M., Shibata, S., Naito, M., Sato, K., Elsliger, M.A., Grant, J.C., Axelrod, H.L., Chiu, H.J., Farr, C.L., et al. (2016). A Distinct Type of Pilus from the Human Microbiome. Cell *165*, 690–703.

Zabala, B., Hammerl, J.A., Espejo, R.T., and Hertwig, S. (2009). The Linear Plasmid Prophage Vp58.5 of Vibrio parahaemolyticus Is Closely Related to the Integrating Phage VHML and Constitutes a New Incompatibility Group of Telomere Phages. J. Virol. *83*, 9313–9320.

Chapter 4 Experimental approaches to measure bacteriophage adherence to mucus: challenges and prospects

The Bacteriophage Adherence to Mucus (BAM) model proposes that adherent lytic phages provide a dynamic defense of their metazoan hosts (Barr et al., 2013). To better define the role of bacteriophage in mucosal immunity, three specific aims were undertaken: (1) determine the effect of Ig-like domains on phage hunting strategies in a mucosal environment, (2) develop novel methods to measure phage adherence to mucus and (3) test diverse phages for a mucus-adherent phenotype. The chapter presented here synthesizes the methods and preliminary results obtained from addressing each specific aim. Specific emphasis is given on the technical challenges associated with each methodological approach. Experiments are proposed to elucidate how bacteriophages adapt to the metazoan mucosal environment. Finally, a model of temperate phages utilizing DGRs to generate mucosal immunity is proposed.

Specific Aim 1: Determine the effect of Ig-like domains on phage hunting strategies in a mucosal environment.

This aim used two experimental approaches to compare the fitness of T4 versus T4 *hoc*-. The first approach employed a continuous culture of *E. coli* and T4 with to observe differences in infection dynamics in the presence or absence of mucus. A second approach explored simultaneous addition of both T4 and T4 *hoc*- to a standard culture of *E. coli*. The population size of each phage genotype was distinguished by plating on suppressor and non-suppressor hosts.

Protocols

Measurement of T4 and E. coli dynamics in chemostats. Continuous cultures of *E. coli* BL21 marked with kanamycin resistance were maintained in LB supplemented with kanamycin to a final concentration of 50 μg mL^{-1}. Chemostats were set up as previously described (Sanchez et al., 2015). Chemostats were inoculated with a 1:100 dilution of overnight bacteria and allowed to equilibrate for 24 hours. Wild-type T4 was introduced to a final concentration of 10^5 pfu/mL. Aliquots of cultures were sampled over time and the optical density at 600 nm recorded using a spectrophotometer.

Direct competition of T4 and T4 hoc- using amber mutants: A mutant of T4 was obtained that harbors amber mutations in gene 43 and 44. Gene 44 encodes a DNA polymerase clamp loader and gene 43 encodes a DNA polymerase (Miller et al., 2003). High titer stocks of T4 43- 44- were maintained on *Escherichia coli* supE. For the competition assay, *E. coli* supE was diluted 1:100 from an overnight culture into fresh LB media (Sigma). Next, 10^7 pfu/mL of both T4 43- 44- (*hoc*+) and T4 *hoc*- were added to the culture and incubated at 37 °C. Aliquots of the culture were sampled over time and clarified using chloroform followed by 0.22 μm syringe-driven filtration. Phage particles were enumerated by the top agar method using both *E. coli* supE and *E. coli* B hosts. The population size of T4 43- 44- phage was determined by subtracting the total number of plaques on the *E. coli* B host from the total number of plaques on the *E. coli* supE host.

Results

Measurement of T4 and E. coli dynamics in chemostats. Infection of a continuous culture of *E. coli* by T4 caused clarification of the culture and a concomitant drop in optical density compared to the uninfected control (Figure 4-1).

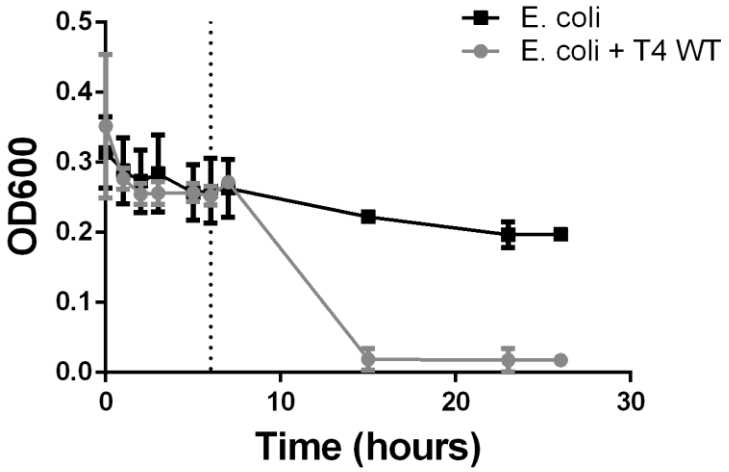

Figure 4-1: Addition of T4 to continuous cultures of *E. coli*. Chemostats were inoculated with *E. coli* and allowed to equilibrate for 24 hours. Cultures were infected (dashed line) with T4 at a MOI of 0.01 (n = 5) or left uninfected (n = 4). Aliquots of the culture were removed at indicated times for optical density measurements.

The control cultures exhibited a modest decrease over time but stabilized at an optical density of 0.2. Attempts with T4 *hoc*- or media supplemented with mucus were not performed.

Direct competition of T4 and T4 hoc- using amber mutants. The population sizes of T4 43- 44- and T4 *hoc*- in the same mixture can be determined by plating the phages on permissive and non-permissive hosts. Subtraction of the number of plaques obtained on the non-permissive host (corresponding to the number of T4 *hoc*- phages in the mixture) from the number of plaques on the *E. coli* supE host (corresponding to the total number of phages in the mixture) yields the number of T4 43- 44- phages. In this assay, T4 43- 44- is observed to reach a higher titer than T4 *hoc*- when both phages are simultaneously added to the permissive host (Figure 4-2).

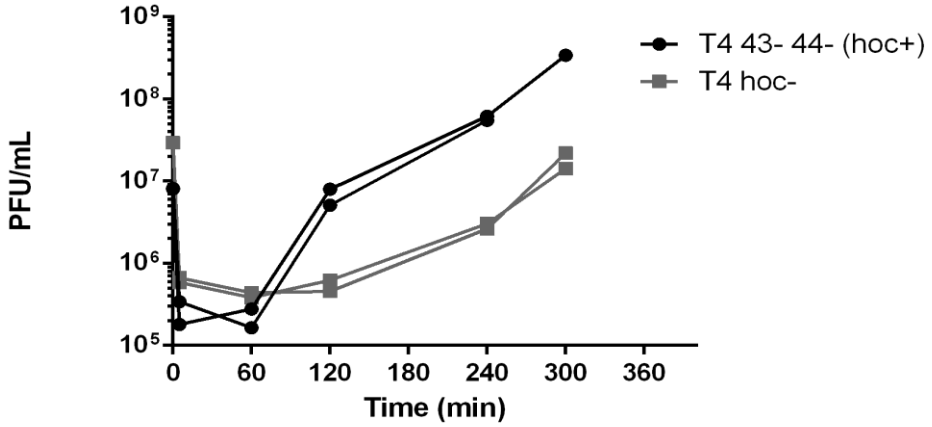

Figure 4-2: Coinfection of *E. coli* supE with 10^7 phage·mL^{-1} T4 43- 44- and T4 hoc-. Shown is the titer of each phage population as determined by deductive plating (n = 2).

However, the estimates of each phage population size are unreliable (discussed below).

Discussion

Infection of *E. coli* by T4 caused clarification of the culture, whereas an uninfected control stabilized in optical density over the course of 48 hours (Figure 4-1). Future experiments comparing T4 and T4 *hoc*- should be performed to observe any differences in the rate of clarification. In a microfluidic environment with mucus-producing epithelial cells, T4 decreased the titer of *E. coli* by three orders of magnitude whereas T4 *hoc*- exhibited no significant antimicrobial effect (Barr et al., 2015). Supplementation of the media in chemostats with mucus may demonstrate a similar advantage of T4 compared to T4 *hoc*-. However, this experiment faces several significant technical challenges. Due to the volume requirements of the chemostat setup (2 – 4 L of media consumed per day), several grams of mucin are required to maintain the media at a final concentration of 1% w/v. Settling of the material over time leads to a non-uniform distribution in the liquid and could cause large technical variation. Additionally, bacterial or fungal contamination is difficult to manage.

These technical challenges need to be considered for experiments exploring phage-mucus interactions within a continuous culture system.

Co-addition of amber mutant and T4 *hoc*- to a standard culture of bacteria demonstrated that T4 43- 44- exhibited a competitive advantage over T4 *hoc*- (Figure 4-2). However, the relative proportion of each phage genotype as determined by deductive plating is inaccurate. Coinfection of *E. coli* by two T4 phages can result in recombination between the two genotypes (Benzer, 1955). Consequently, in this assay the *hoc*+ containing phages may or may not contain amber mutations in genes 43 and 44. Therefore, using amber mutants and deductive plating cannot be used to distinguish replicating phage populations.

Future competition experiments could employ coaddition of T4 43- 44- and T4 *hoc*- and quantify the populations deductive plating as above, so long as the phage are not replicating. For example, T4 43- 44- and T4 *hoc*- could be added simultaneously to a selective surface (e.g., a glycan-displaying bead), washed and quantified by deductive plating. In this experiment, recombination between the two genotypes does not occur as no single cell is coinfected with two different phage genotypes.

Specific aim 2: Development of methods to measure binding kinetics of phage with mucus.

T4 is hypothesized to adhere to mucus via hoc-glycan interactions. To measure the binding affinity of hoc with glycans, two complementary approaches were explored. The first method is a bead-based screen to identify candidate glycan structures that could serve as ligands for *hoc*. The second method explored capillary electrophoresis as a high throughput method to identify mucus-adhering phages.

Method A: Identification of glycan structures that interact with T4. A novel platform to display mucin-like polymers was recently developed (Cohen et al., 2016). The assay employs magnetic beads conjugated with 1 of 43 different polymers of mammalian glycan structures. In this aim, the adsorption of T4 and T4 *hoc*- to the library of 43 glycan structures was examined.

Method B: Measure binding kinetics of T4 with mucus. The goal of this aim is to develop a high-throughput assay to detect phage adherence to mucus. Previous methods, such as the classical adsorption assay, are time consuming and technically challenging. To circumvent this problem, this aim employed capillary electrophoresis as a tool to measure phage-mucus interactions. Measurement of binding events, such as lectin-sugar interactions, can be quantitated by shifts in electrophoretic mobility (Kuhn et al., 1994). In this aim, T4 was electrophoresed through a capillary and mobility shifts in the presence or absence of mucin determined.

Protocols

Glycan screen using mucin-mimetic beads. Working stocks of bacteriophage were prepared as previously described. A 96-well plate was blocked with 0.5% BSA for 30 minutes, followed by a triple rinse with PBS or SM buffer. 10 μL of glycan-conjugated magnetic beads were added to duplicate wells and triple rinsed with buffer. The residual supernatant was removed using a magnetic stand and 100 μL of phage stock added to each well. The plate was sealed with microseal foil, vortexed and incubated for 1 hour at room temperature on an orbital shaker. Next, the plate was spun for 2 minutes at 2000 rpm and placed on a magnetic stand. The supernatant fraction was collected into a new plate and stored at 4 °C. The beads were washed by adding 100 μL of buffer to the wells. The washing was

repeated for a total of three washes and all wash fractions stored at 4 °C. The beads were resuspended in 60 μL of buffer and stored at 4 °C. The titer of phage in all fractions was determined through top-agar plating.

Capillary electrophoresis of T4. T4 and T4 *hoc*- were stained with SYBR Gold (Invitrogen, CA) at a final concentration between 100x - 500x overnight. Excess stain was removed by dialysis against SM buffer using a 100 kD pore size Amicon (Millepore) ultracentrifuge filter. Stained stocks of T4 and T4 *hoc*- at 10^{10} pfu/mL were stored at 4 °C in SM buffer protected from light. 20 μL aliquots of stained phage was electrophoresed in uncoated silica, polyvinyl alcohol (SciEx, part no. 477601), neutral (SciEx, part no. 477441), or 1,2-Dilauroyl-sn-phosphocholine (DLPC) coated capillaries. The best results were obtained with Polyvinyl alcohol (PVA) or neutral capillaries. To assay mucin-phage interactions, a 1% w/v mucin solution was prepared in SM buffer. Next, a 10 μL of stained T4 phage was resuspended in 90 μL of 1% mucin and electrophoresed. Alternatively, a 1% mucin solution was injected into the capillary using 0.5 psi of pressure for 5 seconds prior to electrophoresis with stained phage. Capillaries were trimmed to 31 cm and electrophoresed with 10 mM pH 6.9 phosphate running buffer at 20 kV reverse polarity.

Results

Glycan screen using mucin-mimetic beads. T4 and T4 *hoc*- (10^7 pfu/mL) were applied to mucin-mimetic beads. After a one-hour incubation, the supernatant fraction and the beads were titered using top-agar plating. Both phages exhibited equivalent titers to an unconjugated bead control in either the supernatant or resuspended bead fraction (Figure 4-3).

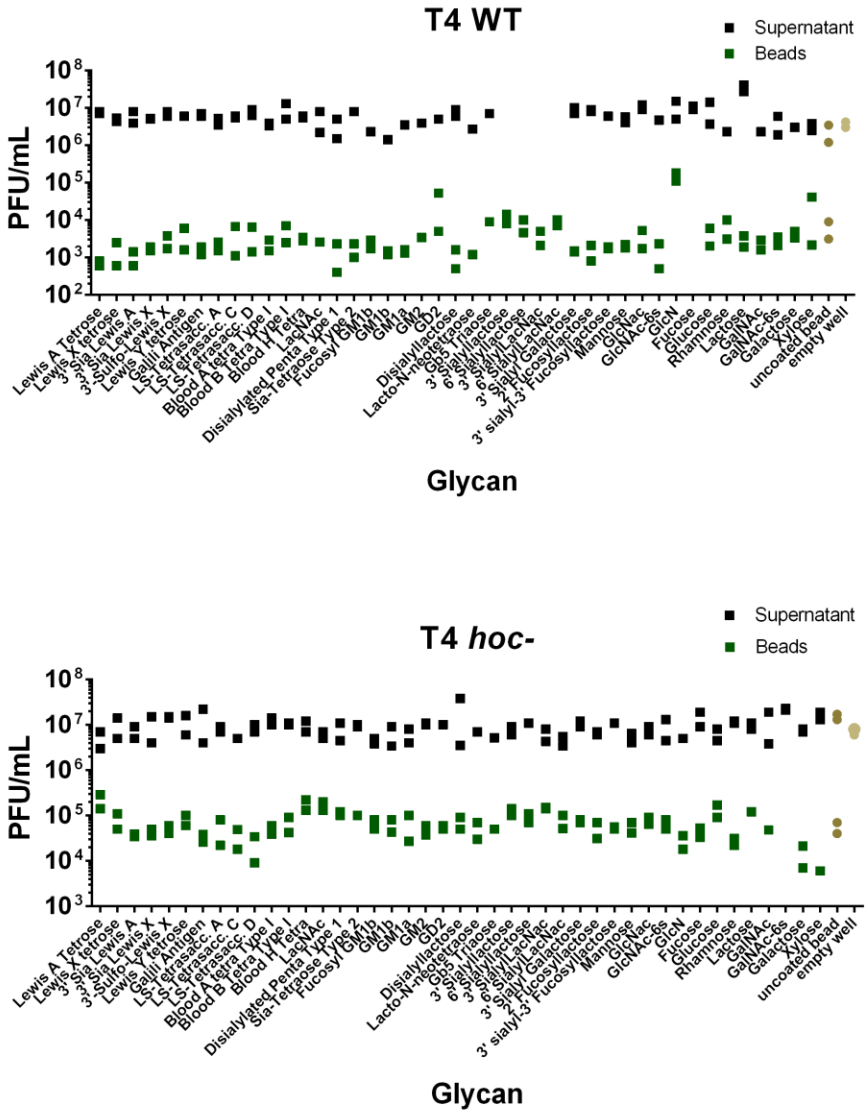

Figure 4-3: Adsorption of T4 to a library of 43 mucin mimetic beads. Each glycan-displaying bead was incubated with T4 or T4 *hoc*- phage and the titer determined for the supernatant fraction or the bead fraction after triple wash and resuspension with PBS. A magnetic bead without a displayed glycan ('uncoated bead') and an empty well are included as controls. Each technical plating replicate is shown as a separate symbol (n = 1).

Across all beads, the mean titer of T4 was 7.5 x 10^3 and T4 *hoc*- was 7.0 x 10^4. The glycan structures Lewis antigen (Lewis A tetrose), N-acetyl lactosamine (LacNAc), Disialylated pentasaccharide type I and GD2 caused variable retention of T4 and T4 *hoc*-. To further investigate the binding kinetics to these glycans, the assay was repeated and the supernatant, wash and bead fractions titered using top-agar plating. The Lewis antigen retained more T4 phage particles in the bead fraction than T4 *hoc*- (Figure 4-4).

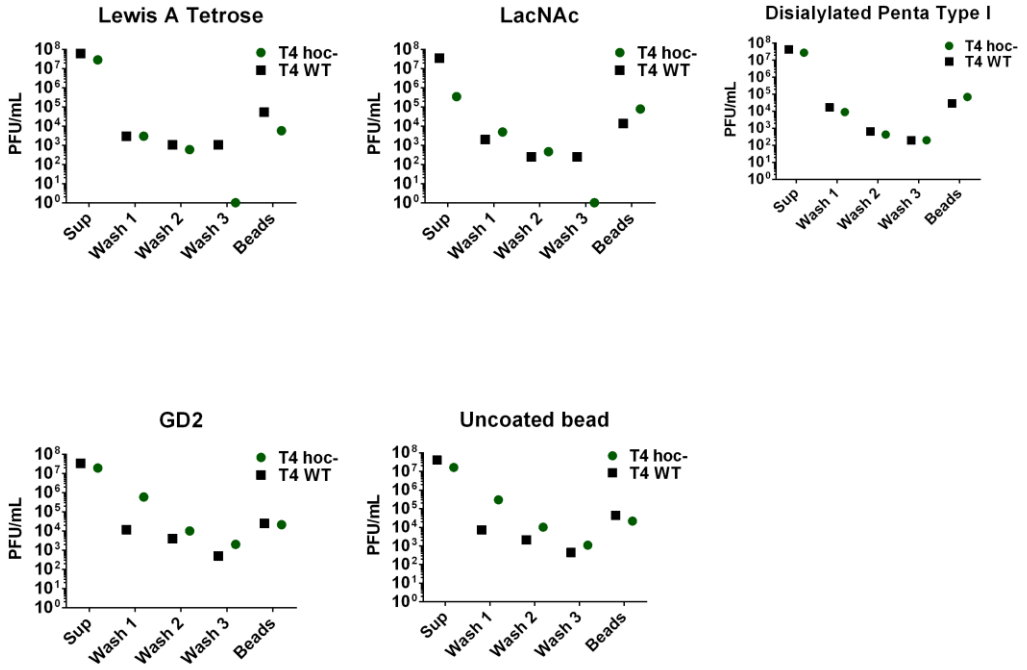

Figure 4-4: Adsorption kinetics of T4 and T4 *hoc*- to mucin mimetic beads. T4 or T4 *hoc*- were added to glycan-displaying beads and the titer determined for the supernatant, wash and resuspended bead fractions in SM buffer. A magnetic bead without a displayed glycan ('uncoated bead') is used as a control. The average of three technical replicates is shown as a single symbol (n = 1).

Conversely, more T4 *hoc*- than T4 phage were recovered in the bead fraction of LacNAc and Disialylated pentasaccharide type I. The glycans GD2 and the unconjugated control bead retained equivalent titers of T4 and T4 *hoc*-. The most significant difference between the two

phages was observed in the supernatant fraction of the LacNAc bead, where 3.5 x 10^7 pfu/mL of T4 was recovered compared to 3.5 x 10^5 pfu/mL of T4 *hoc*-.

Capillary electrophoresis of T4. Electrophoresis of SYBR-gold stained phage was achieved using a PVA-coated capillary device coupled with a laser-intensity fluorescence detector. T4 was migration was observed as a single peak at 1.33 minutes, whereas T4 *hoc*- migration appeared as multiple peaks between 1.5 – 2 minutes (Figure 4-6A, B). Electrophoresis of SYBR gold without phage did not result in any peaks (Figure 4-6C). Similarly, electrophoresis of unstained phage did not result in any peaks and top-agar plating of the eluent confirmed that viable phage particles migrated to the destination reservoir (data not shown).

Two approaches were attempted to measure the binding interactions of T4 with mucus. The first method resuspended T4 in a 1% mucin solution and the mixture electrophoresed. Peak fluorescence was observed between 1.5 – 2 minutes (Figure 4-5).

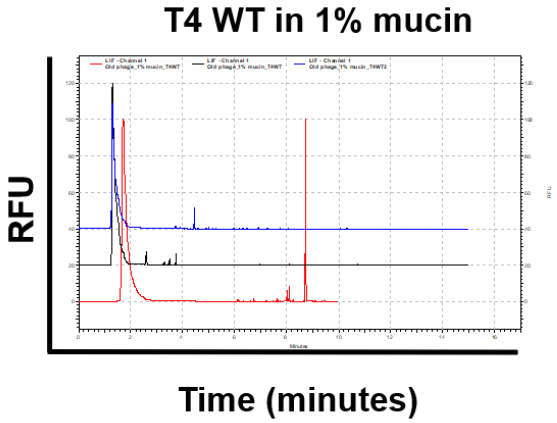

Figure 4-5: Electropherogram of SYBR gold stained T4 complexed with mucin. T4 was added to a 1% w/v solution of mucin prior to electrophoresis. The migration time is between 1.5 – 2 minutes. Depicted are three technical replicates that are arbitrarily offset for clarity. Electrophoresis of T4 *hoc*- was not attempted.

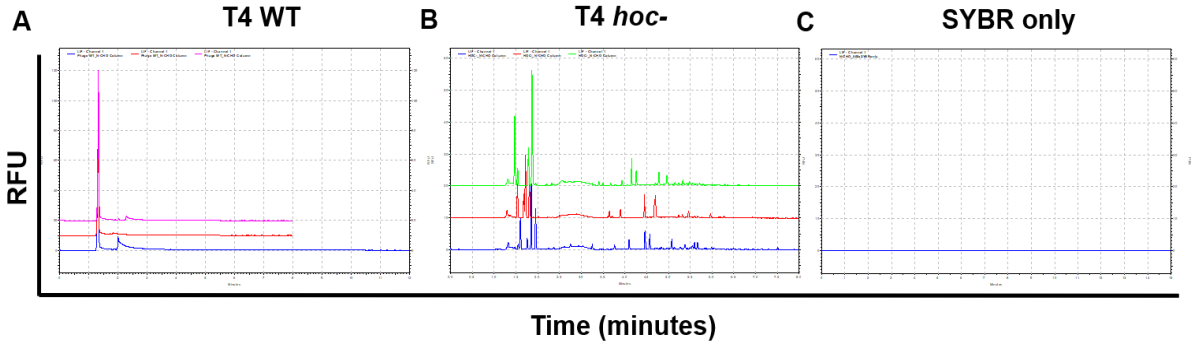

Figure 4-6: Electropherograms of SYBR gold stained T4 and T4 *hoc*- through a PVA-coated capillary. Fluorescence peaks appear at 1.33 minutes for wild-type T4 (A) and between 1.5 – 2 minutes for T4 *hoc*- (B). Depicted are three technical replicates that are arbitrarily offset for clarity. A solution of SYBR gold (500x) does not produce a detectable signal (C).

Results were not obtained for T4 *hoc*-. In the second method, the capillary was injected with a 1% mucin solution prior to electrophoresis. Pre-injection of mucin caused T4 to electrophorese more diffusely, from 1.33 minutes – 3 minutes (Figure 4-7). Similarly, T4 *hoc*- exhibited the same diffuse electrophoresis and did not qualitatively differ from wild-type T4.

Discussion

Identification of glycan binding partners of T4. In the initial screen of glycan structures for binding activity, we observed that T4 *hoc*- was bound to the beads more than T4, irrespective of the glycan structure (Fig B). T4 *hoc*- phage particles aggregate in the absence of magnesium ions (Yamaguchi and Yanagida, 1980), which was absent in the PBS buffer. Aggregates of T4 *hoc*- could bind the beads more strongly than non-aggregated T4 particles and increase the observed titer in the bead fraction. Using SM buffer on subsequent assays resulted in more equivalent titers of T4 and T4 *hoc*- in the bead fraction (Figure 4-4).

A significant technical issue encountered was the variation between technical replicates in the number of plaques obtained for each fraction. To circumvent this issue,

addition of T4 and T4 *hoc*- were added to the same well and quantification performed by qPCR. The results of this assay are reported elsewhere (Shr-Hau Hung, unpublished) and indicate that T4 may be approximately 5-fold enriched on LacNAc beads compared to T4 *hoc*-. A future experiment could employ an amber mutant of T4 that is added simultaneously with T4 *hoc*- and each population quantified with deductive plating.

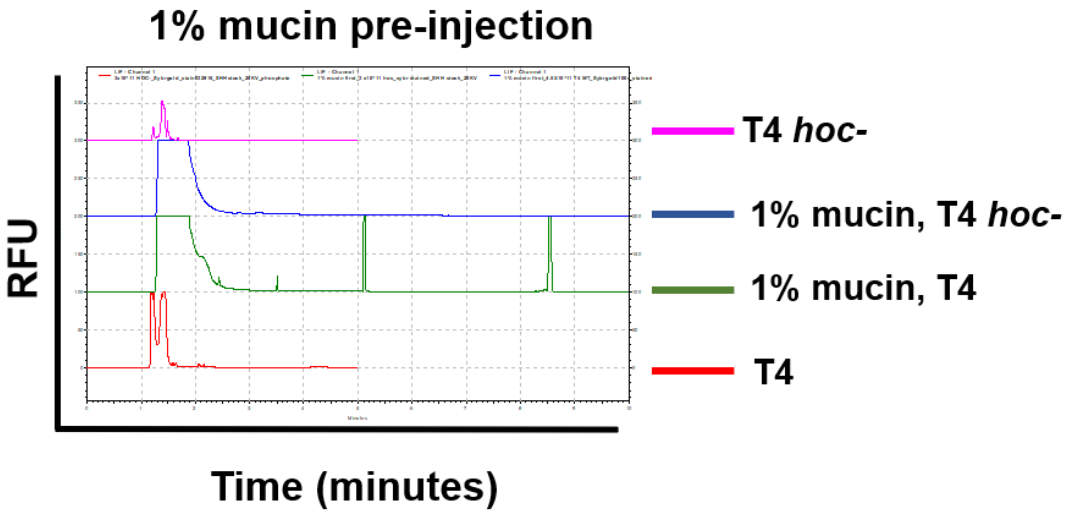

Figure 4-7: Electropherogram of SYBR gold-stained T4 and T4 *hoc*- through a capillary pre-injected with mucin. A 1% w/v solution of mucin was injected using 0.5 psi of pressure for 5 seconds prior to electrophoresis. The migration time is between 1.5 – 2 minutes. Each experimental condition is arbitrarily offset for clarity.

While the technical aspects of the glycan-bead assay remain to be fully optimized, such as modification of the incubation time, conjugation of the polymers to the beads, number of washes, etc., the diversity of glycan structures may be too vast for a medium-throughput assay. Additionally, glycan structure, linkage, modification, topology and conformation influences their interactions with proteins (Cohen, 2015). Exploration of virome-glycome interactions will require novel methods to identify candidate interactions. A previous screening of 610 mammalian glycan targets yielded structures that enhanced T4 adsorption by

an estimated 10-fold compared to T4 *hoc*- (Barr et al., 2013). Ideally, these structures could be incorporated into a format similar to the one attempted here for a more quantitative measurement.

Measurements of T4-mucus binding kinetics. Using capillary electrophoresis, the migration of T4 and T4 *hoc*- was most successful when using PVA or neutral-coated capillaries. Capillary coating prevents wall adsorption of biological particles and has been used for the electrophoresis of viruses (Kremser et al., 2004). Usage of a DLPC coated capillary gave variable results and may be attributable to the in-house coating procedure or unanticipated adsorption of phage to the coating.

Despite the successful electrophoresis of stained T4 with PVA or neutral-coated capillaries, we did not obtain conclusive evidence for T4 interaction with mucus. Mixture of T4 with a 1% w/v mucin solution yields a migration time comparable to T4 alone, but additional replicates and measurements with T4 *hoc*- are needed. Pre-injection of 1% mucus altered the migration time of both T4 and T4 *hoc*-, but no discernable difference between the two phage was observed. Furthermore, pre-injection of mucus and electrophoresis of a buffer-only control resulted in a peak similar to that observed with stained phage (data not shown). One explanation is that residual SYBR-gold from the phage preparation stains DNA present in the mucus.

Future experiments should optimize the staining procedure to prevent excess background signal. To our knowledge, electrophoresis of mucin macromolecules has not been reported and may pose a significant technical challenge for this assay. Candidate mono- or

disaccharides could be used as an alternative to mucus and have been successfully used in the study of lectins with capillary electrophoresis (Kuhn et al., 1994).

Specific aim 3: Identify diverse mucus-adherent phages

Bacteriophage T4 and T3 exhibit a mucus-adherent phenotype mediated by an Ig-like domain-containing protein (Barr et al., 2013). The goal of this aim was to identify Ig-like and lectin domains in diverse phages and measure their adherence for mucus. The first approach measured the mucus-adherent phenotype of a phage previously isolated from the water above diseased corals, bacteriophage YC. YC is a lytic phage that infects *Vibrio coralliilyticus,* a mucus-adherent member of the *Pocillopora damicornis* coral holobiont (Cohen et al., 2013; Garren et al., 2014). The second approach performed a bioinformatic survey of isolated phages for Ig-like and lectin domains to generate target phages for subsequent mucus-adherence assays.

Method A: Test the hypothesis that bacteriophage YC is mucus-adherent. Using previously established techniques, YC was examined for a mucus-adherent phenotype. Attempts to purify coral-mucus to use for adherence assays were undertaken.

Method B: Identify Ig-like and lectin-like domains in bacteriophage genomes. A comprehensive database of Ig-like and lectin Hidden Markov Models (HMMs) was constructed to find potential mucus-adherent phages. Emphasis was placed on identifying temperate phages for future exploration of lytic-lysogeny decisions in mucosal environments.

Protocols

Qualitative mucus adherence assay of YC. Marine-broth tryptone (MBT) agar plates were coated with 1 mL of 1% w/v mucin (type III porcine stomach mucin, Sigma) and allowed to dry for 90 minutes. An uncoated MBT agar plate was included as a control. Stocks

of YC were serially diluted 10^{-2} into 5 mL of MBT and added to the plates. The plates were incubated at 37 °C for 30 minutes, liquid decanted and allowed to dry for 45 minutes. The plates were overlaid with 1 mL of overnight *V. coralliilyticus* P1 and 3 mL of MBT top agar. The plates were incubated at 30 °C overnight and plaques quantified the following day.

Quantitative mucus adherence assay of YC: The adsorption rate of YC to *V. coralliilyticus* was determined as described previously (Adams, 1959). Briefly, an overnight culture of *V. coralliilyticus* was subcultured to a final concentration of 10^6 cfu/mL in 5 mL of MBT or MBT + 1 % w/v mucus (type III porcine mucus, Sigma). A 50 µL aliquot of YC at 10^5 pfu/mL was added to the bacterial culture and 500 µL samples taken over time. Each sample was centrifuged at 10,000 x g for 5 minutes to pellet bacterial cells and adsorbed phage. The supernatant of each sample was removed and treated with 0.1 volumes of chloroform. The titer of unadsorbed YC at each timepoint was determined using top-agar plating as described above.

Coral mucus isolation: Coral mucus was milked by aerial exposure to promote release of mucus. The volume of mucus was measured and mixed 1:1 with PBS buffer at pH 7.3 containing proteolysis inhibitors (10 mM EDTA, 100 mM aminocaproic acid, 10 mM *N*-ethylmaleimide, 1 mM iodoacetamide, 5 mM benzamide HCl, 1mM PMSF) (Jatkar et al., 2010). The solution was kept at < -80 °C prior to extraction. The solution was centrifuged at 13,000 x g at 10 °C for 30 minutes to pellet debris. Next, the samples were desalted using a 100kD pore-size Amicon filter (Millepore) and the retentate lyophilized for 24 hours.

Bioinformatic survey of phage genomes for Ig-like and lectin domains. To identify Ig-like and lectin domains in bacteriophage genomes, HMM models from diverse superfamilies

were extracted from the Pfam-A HMM database using the hmmfetch tool from the HMMER package (Finn et al., 2011). These included HMM models from the carbohydrate binding domain (CL0203), carbohydrate binding module (CBM), C-type lectin (CL0056), concanavalin (CL0004), E-set (CL0159), galactose-binding (CL0202) and Ig-like (CL0011) clans. In total 189 HMM models were concatenated into a single database using the hmmpress function of the HMMER package. The database of HMM models was used to query bacteriophage genomes available on RefSeq (*n* = 1881, accessed 06/2016). Bacteriophage genomes were de-replicated at 98% identity using CD-Hit and translated into 6 reading frames using the EMBOSS package (Fu et al., 2012). Each genome was queried against the database using hmmsearch with a bitscore cutoff of 25. The results were parsed to identify the single best hit and the protein sequence extracted. The sequences were compared to the NCBI non-redundant database using BLASTp to recover the functional annotation of the proteins. Host-assignments of each bacteriophage were extracted from the virus-host DB http://www.genome.jp/virushostdb/) and morphology determined from the Refseq viral database (accessed 06/2016). Phage lifestyle was determined by searching individual genomes for phage integrase HMM model PF00589. Genomes containing hits to integrase are classified as a temperate, whereas phage lacking an integrase hit are classified as lytic.

Results

YC does not exhibit adherence to mucus. A modified top-agar assay was performed to determine if YC adheres to mucin macromolecules. Plain agar plates and agar plates coated with 1% w/v mucin were prepared. YC phage suspensions were incubated on the plates for 30 minutes and subsequently decanted to remove unbound phage particles. Plates were overlaid with top-agar containing *V. coralliilyticus* and the number of phage plaques enumerated the

following day. No difference was observed between uncoated and mucin-coated agar plates (Figure 4-8A, $t = 0.16$, $df = 18$, $P = 0.87$).

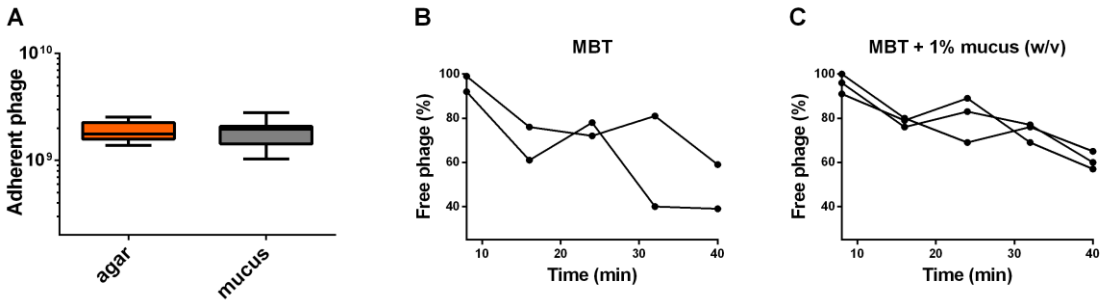

Figure 4-8: YC does not exhibit a mucus-adherence phenotype. Phage adherence to an uncoated agar plate or agar plate coated with mucin ($n = 1$, $t = 0.16$, $df = 18$, $P = 0.87$) (A). Adsorption assay measuring the percentage of free phage remaining during a 40-minute period in marine broth tryptone (B, $n = 2$) and marine broth tryptone supplemented with 1% mucin (C, $n = 3$).

A classical adsorption assay was employed to measure the rate of YC adsorption to *V. coralliilyticus* cells (Adams, 1959). YC was added to control (0%) or 1% mucin solutions containing *V. coralliilyticus* to give a final concentration of 3×10^5 phage particles·mL^{-1} and 4×10^6 bacteria·mL^{-1}. The solutions were sampled during the eclipse period of YC infection (Cohen et al., 2013). Centrifugation of the sample and treatment of the supernatant with chloroform halts phage adsorption. The number of free phage particles in each sample were enumerated and used to calculate the adsorption constant (k) of YC in both control and 1% mucin solutions. The adsorption rate of YC in the control solution ($k = 5.27 \times 10^{-9}$ mL/min) was not significantly different than in the 1% mucin solution ($k = 3.1 \times 10^{-9}$ mL/min, $t = 2.045$, $df = 3$, $P = 0.13$) (Figure 4-8B, C).

To explore whether the source of mucin may influence YC adsorption to *V.* coralliilyticus, the mucus from corals was explored for use in downstream adsorption assays. Coral mucus was harvested from three coral species by inversion and aerial exposure. The exuded liquid was used for purification of the component mucin glycoproteins (Coddeville et al., 2011). Milligram amounts of material were obtained from *Colpophylia natans* and no material was obtained from *Diploria strigosa* or *Acropora*

Table 4-1: Purification of coral mucus. The volume of exuded mucus recovered after milking (initial volume) and the amount recovered after dialysis (final volume) is listed. The dialyzed material was lyophilized for 24 hours and weighed.

Coral species	Initial volume	Final volume	Final yield (dry weight)
C. natans	115 mL	10 mL	0.2 g
D. strigosa	120 mL	0.3 mL	0.0
A. palmata	110 mL	3.5 mL	0.0

palmata (Table 4-1). Currently, adherence assays require grams of material and thus no attempt was made to quantitate YC adsorption to cells in the presence of coral mucus.

B) Bioinformatic survey of phage genomes for Ig-like and lectin domains. To ascertain the presence of Ig-like and lectin domains in bacteriophages, an exhaustive survey of completely sequenced phage genomes was performed. Hidden Markov Models of 189 families belonging to 6 Ig-like and lectin superfamilies was used to query 1,881 sequenced genomes. DNA sequences of the genomes were translated into all reading frames to capture domains appended to ORFs through frameshifting. In total, 389 phages from all three families of the *Caudovirales* contain at least one domain in the database (Appendix, Table 4-2). More Ig-like domains (67%) were detected than lectin domains (33%). The functions of the proteins are unknown in most cases (‘hypothetical’ annotation, 53%), but in some phage the domains are encoded by tail (30%) or capsid proteins (12%). To estimate the lifestyle of all phages in

this survey, each genome was queried for the presence of an integrase and classified as temperate if a significant match was found. By this approximation, most of the phage with an Ig-like or lectin domain exhibit a lytic lifestyle (74%).

Discussion

The mucus-adherence assays performed here did not demonstrate an adherent phenotype for bacteriophage YC. Additionally, no detectable Ig-like or lectin domain-containing proteins were present in the sequenced genome of YC (this thesis, reported elsewhere). Together, these preliminary results lead to a rejection of the hypothesis that YC adheres to mucus.

While attempts were made to isolate and purify coral mucin, assays were not completed using this material. Coral mucins exhibit unique glycosylation patterns compared to other animals (Coddeville et al., 2011). Glycosylation patterns can dictate the specificity of binding interactions with viruses and bacteria (Cohen et al., 2016; Comstock, 2009). YC or other phages of a coral holobiont may use the glycan moieties displayed on coral mucins as ligands for adherence. As the concentration of protein in milked coral mucus samples is low (0.1 mg/mL), collecting enough material for adherence assays to test this hypothesis could be problematic (Jatkar et al., 2010).

The survey of bacteriophage genomes presented here provides a comprehensive list of bacteriophages that may exhibit a mucus-adherent phenotype mediated by an Ig-like or lectin domain. Ig-like and lectin domains are commonly found on phage structural proteins and bind diverse carbohydrates during the infection process (Fraser et al., 2006, 2007; Simpson et al., 2015). Ig-like domains can be dispensable, such as in the Hoc protein of T4, or required for efficient assembly of the virion, such as in P22 (Fokine et al., 2004; Parent et al., 2010).

Lambda possesses and Ig-like fold on the tail sheath protein gpV that is semi-dispensable: deletion incurs a temperature-sensitive phenotype and complementation with a truncated gpV lacking the Ig-like domain results in a 1-log decrease in titer compared to the full length protein (Katsura, 1981; Xu et al., 2014). Discerning between Ig-like domains that play essential structural roles and those involved in binding events is difficult to determine bioinformatically. Phage proteins harboring Ig-like or lectin domains could be pleiotropic and serve both structural and binding roles. The phage identified here can be used in future experiments testing the role of Ig-like and lectin domains in mediating phage adherence to mucus.

Prospectus

Examining the role of Ig-like domains in phage diffusion through mucus. Ig-like domains encoded by phage remain promiscuous and deceitful. While the function of some have been probed experimentally, most remain uncharacterized and await examination (Table 4-2). Moreover, it is still unclear whether Ig-like domains specifically mediate the adherence of phage to mucus. Experiments with T4 demonstrate that the entire hoc protein is required for adherence, but experiments using deletion constructs lacking Ig-domains have not yet been performed due to the difficulty of modifying the T4 genome. It is thus unknown whether T4 diffusion through mucus is altered by Ig-like domains binding to glycans (or other ligands), or if the entire hoc protein inhibits diffusion through the mucus mesh network due to steric hindrance. To discern between these two mechanisms, experiments using bacteriophages T5 and phi29 are proposed.

T5 could be assayed for a mucus adherence phenotype using previously generated deletion constructs of the decoration protein pb10. Pb10 possesses a C-terminal Ig-like

domain and the protein is dispensable for growth under laboratory conditions (Vernhes et al., 2017). Moreover, complementation of T5 *pb10*- with full length pb10 or a C-terminal deletion construct is robust *in vitro* (Vernhes et al., 2017). Complementation of T5 *pb10*- with the Ig deletion construct and subsequent mucus adherence assays would implicate the immunoglobulin domain in mediating adherence to mucus.

To test if steric hindrance by structural proteins alters phage diffusion, *Bacillus subtilis* phage phi29 may offer a tractable phage for experimentation. Phi29 possesses a large auxiliary head fiber that is structurally similar to the fibritin protein of T4 and does not contain an Ig-like domain (Morais et al., 2005). The head fiber is non-essential for growth and treatment with DMSO removes the fiber without reducing phage infectivity (Reilly et al., 1977; Salas et al., 1972). Quantification of phi29 with and without head fibers through a network of polymers, such as mucus, would provide experimental evidence addressing steric interactions during phage diffusion.

Examining the effect of mucus on phage diffusion. An outstanding question to be addressed is if phage diffusion is affected by the source of mucus (pig, human, coral, etc.). The source of mucus can drastically impact its rheological properties, which could range from a highly viscous solution to a true gel (Bansil et al., 1995). Preliminary attempts to purify coral mucus were unsuccessful and are challenging to supply continually for experimentation. Hagfish slime was considered as an alternative, renewable source of mucus. Hagfish slime is composed of mucus and intermediate filament proteins at concentrations of approximately 0.016 mg/mL and 0.022 mg/mL dry weight, respectively (Ewoldt et al., 2011). Purifying sufficient quantities of mucin protein (> 1 gram) would then require at least 62 liters of

hagfish slime. A second alternative source of mucin protein could be human salivary mucus. Mucins in human salivary secretions are at concentrations of approximately 2 mg/mL and purification methods are relatively facile (M. Cohen, personal communication). Measurement of virion diffusion through mucus prepared from different animals could identify critical parameters of phage predation of bacteria.

Developing methods to assay phage adherence to mucus. Several methods were explored to measure the binding kinetics of phage to mucins or glycans. Optimization of the capillary electrophoresis approach attempted here may yield a powerful platform to study phage-mucin interactions. By coating the capillary wall with mucins or mucin mimetics, shifts in phage electrophoretic mobility can be used to quantify the affinity of interactions. Attachment of mucins to glass slides has been achieved previously and may be feasible to scale up to silica capillary tubing (Gunning et al., 2013). Synthetic mucin mimetics have been reported and could also be adapted to a capillary format (Cohen et al., 2016; Kramer et al., 2015; Xiao et al., 2018).

Investigating the role of DGRs in BAM immunity. Diversity-generating retroelements (DGRs) are encoded by ubiquitous temperate phages within the human gastrointestinal tract. Based on this finding, as well as the literature, I hypothesize hypervariable bacteriophages are an element of metazoan mucosal immunity (Figure 4-9). Virion release from commensal bacteria via spontaneous induction yields a mucus layer replete with phage. For example, spontaneous induction in the mouse gastrointestinal tract is estimated to occur in 1.6 % of the bacterial population (De Paepe et al., 2016). During induction and viral morphogenesis, targeted mutations introduced by DGRs result in virions with unique C-type lectins or Ig-like

domains (Liu et al., 2002). Diversified lectins could then exhibit varying dissociation constants to mucosal glycans. Positive selection would favor phage with an optimal mucin-binding phenotype that increases host encounter rate while avoiding wash-out into the intestinal lumen (Barr et al., 2015). In this model, mucus-adherent phage lyse closely related naïve bacteria. The spatial structuring of phage and bacteria in the mucosal environment may favor lytic behavior by temperate phage (Silveira and Rohwer, 2016). Commensals, on the other hand, are protected from lysis via superinfection immunity (Susskind and Botstein, 1978). In doing so, temperate phages benefit the metazoan host by contributing to the selection of commensal mucosal bacteria.

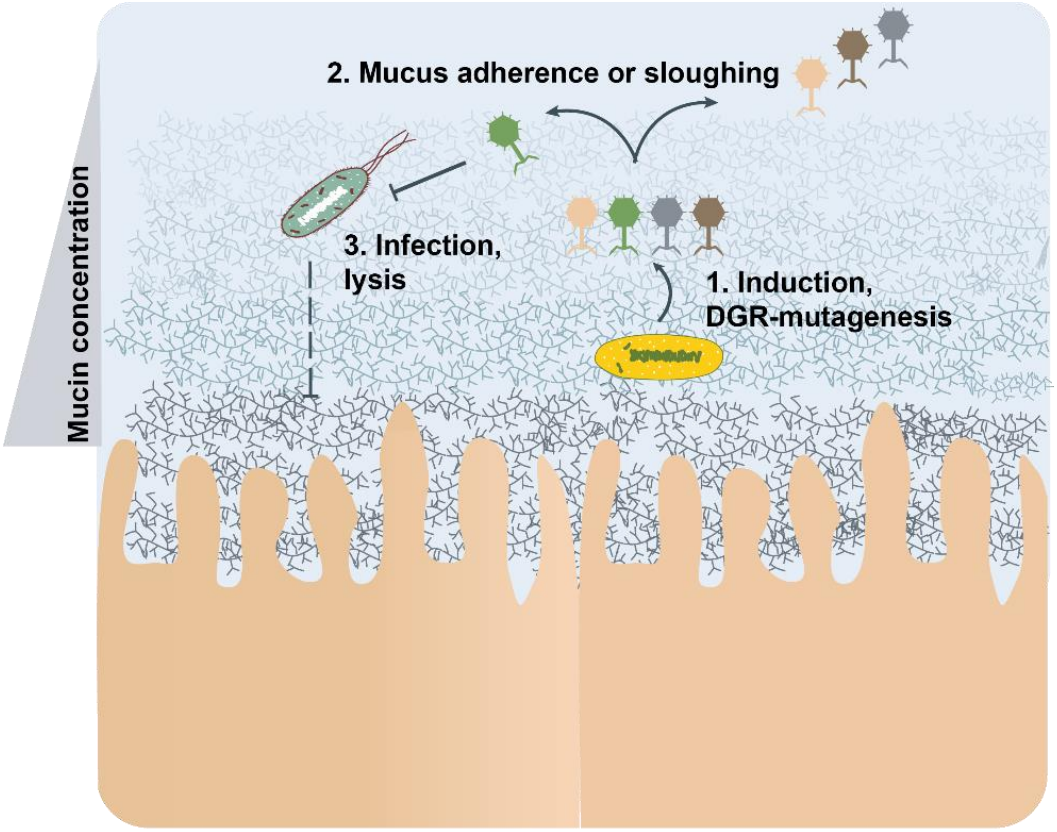

Figure 4-9: Model of hypervariable bacteriophage-mediated mucosal immunity. Prophages residing within commensal bacteria (yellow) deter colonization by incoming bacteria (teal). During induction, DGRs create hypervariable Ig-like or C-type lectin domains. These domains enable effective search strategies in a rapidly changing mucosal environment. Broad host-range phage infect non-lysogens, while commensals are protected via superinfection immunity.

Appendix

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database. Listed are phage that possess an Ig-like or lectin domain, sorted alphabetically by domain annotation. Lifestyle assignments were made by searching for the presence of an integrase (methods). The number of Ig-like or lectin domains is listed (#) along with the Refseq ORF annotation.

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_019524	Enterobacter phage Enc34	*Enterobacter cancerogenus*	lytic	Siphoviridae	Big_1	2	phage structural protein
NC_018832	Providencia phage Redjac	*Providencia stuartii MRSN 2154*	lytic	unclassified	Big_1	1	phage structural protein
NC_029045	Salmonella phage 37	*Salmonella*	lytic	Siphoviridae	Big_1	1	hypothetical protein
NC_025442	Salmonella phage Chi	*Salmonella enterica*	lytic	Siphoviridae	Big_1	2	hypothetical protein chi_062
NC_021779	Salmonella phage FSLSP030	*Salmonella enterica subsp. enterica serovar Dublin*	lytic	Siphoviridae	Big_1	2	hypothetical protein SP030_00110
NC_021780	Salmonella phage FSLSP088	*Salmonella enterica subsp. enterica serovar Typhimurium*	lytic	Siphoviridae	Big_1	1	hypothetical protein SP088_00245
NC_021783	Salmonella phage iEPS5	*Salmonella enterica subsp. enterica serovar Typhimurium*	lytic	Siphoviridae	Big_1	2	bacterial Ig-like domain-containing protein
NC_019417	Salmonella phage SPN19	*Salmonella sp.*	lytic	Siphoviridae	Big_1	1	hypothetical protein SP124_00250
NC_027991	Staphylococcus phage SA1	*Staphylococcus aureus*	lytic	Myoviridae	Big_1	1	hypothetical protein
NC_023556	Achromobacter phage JWAlpha	*Achromobacter xylosoxidans*	lytic	Podoviridae	Big_2	4	hypothetical protein JJJB_0019
NC_028908	Achromobacter phage phiAxp-3	*Achromobacter xylosoxidans*	lytic	Podoviridae	Big_2	2	hypothetical protein HQ29_06765, partial
NC_023581	Acinetobacter phage Presley	*Acinetobacter nosocomialis M2*	lytic	Podoviridae	Big_2	1	decoration protein
NC_004165	Bacillus phage B103	*Bacillus subtilis*	lytic	Podoviridae	Big_2	1	major head protein
NC_011048	Bacillus phage phi29	*Bacillus subtilis*	lytic	Podoviridae	Big_2	1	major head protein
NC_015253	Brochothrix phage A9	*Brochothrix thermosphacta*	lytic	Myoviridae	Big_2	2	gp154.1
NC_015252	Brochothrix phage NF5	*Brochothrix thermosphacta*	temperate	Siphoviridae	Big_2	1	gp6.1
NC_028672	Cronobacter phage PBES 02	*Cronobacter*	lytic	Myoviridae	Big_2	2	putative tail fiber protein 1

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_023717	Cronobacter phage CR9	*Cronobacter sakazakii*	lytic	Myoviridae	Big_2	3	putative structural protein
NC_019400	Cronobacter phage vB_CsaM_GAP31	*Cronobacter sakazakii*	lytic	Myoviridae	Big_2	2	hypothetical protein GAP31_232
NC_019401	Cronobacter phage vB_CsaM_GAP32	*Cronobacter sakazakii*	lytic	Myoviridae	Big_2	2	Ig domain-containing protein
NC_017974	Cronobacter phage CR3	*Cronobacter sakazakii*	lytic	Myoviridae	Big_2	2	putative tail fiber protein 1
NC_024354	Cronobacter phage CR8	*Cronobacter sakazakii*	lytic	Myoviridae	Big_2	2	putative tail fiber protein 1
NC_018454	Cronobacter phage phiES15	*Cronobacter sakazakii ES15*	temperate	Siphoviridae	Big_2	1	hypothetical protein
NC_018454	Cronobacter phage phiES15	*Cronobacter sakazakii ES15*	temperate	Siphoviridae	Big_2	1	tail protein
NC_028795	Enterobacter phage E-3	*Enterobacter cloacae*	lytic	Podoviridae	Big_2	1	minor capsid protein 10B
NC_027335	Enterococcus phage ECP3	*Enterococcus faecalis*	lytic	Myoviridae	Big_2	1	putative Ig-like protein
NC_029009	Enterococcus phage EFDG1	*Enterococcus faecalis*	lytic	Myoviridae	Big_2	3	putative Ig-like protein
NC_029026	Enterococcus phage EFLK1	*Enterococcus faecalis*	lytic	Myoviridae	Big_2	2	putative Ig-like protein
NC_009904	Enterococcus phage phiEF24C	*Enterococcus faecalis*	lytic	Myoviridae	Big_2	2	putative Ig-like protein
NC_024212	Enterococcus phage VD13	*Enterococcus faecalis*	lytic	Siphoviridae	Big_2	2	fibronectin type III domain protein
NC_018086	Enterococcus phage BC611	*Enterococcus faecalis*	lytic	Siphoviridae	Big_2	2	major tail protein
NC_029016	Enterococcus phage vB_IME198	*Enterococcus faecalis*	lytic	Siphoviridae	Big_2	2	major tail protein
NC_011811	Erwinia phage phiEa21-4	*Erwinia amylovora*	lytic	Myoviridae	Big_2	1	conserved structural protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_028840	Escherichia phage slur09	*Escherichia*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_003298	Enterobacteria phage T3	*Escherichia coli*	lytic	Podoviridae	Big_2	1	minor capsid protein 10B
NC_018835	Enterobacteria phage NJ01	*Escherichia coli*	lytic	Podoviridae	Big_2	4	bacterial surface protein
NC_023693	Enterobacteria phage phi92	*Escherichia coli*	lytic	Myoviridae	Big_2	2	Phi92_gp152
NC_010324	Enterobacteria phage Phieco32	*Escherichia coli*	lytic	Podoviridae	Big_2	4	bacterial surface protein
NC_012223	Enterobacteria phage SSL2009a	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	hypothetical protein EpSSL_gp63
NC_005833	Enterobacteria phage T1	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	hypothetical protein T1p48
NC_019517	Enterobacteria phage vB_EcoM-FV3	*Escherichia coli*	lytic	Myoviridae	Big_2	1	hypothetical protein FV3_00027
NC_028903	Escherichia phage 172-1	*Escherichia coli*	lytic	Podoviridae	Big_2	4	bacterial surface protein
NC_019725	Escherichia phage ADB-2	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	hypothetical protein B508_00205
NC_018859	Escherichia phage ECBP2	*Escherichia coli*	lytic	Podoviridae	Big_2	6	bacterial Ig-like domain
NC_019724	Escherichia phage HK578	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	hypothetical protein HK578_058
NC_020079	Escherichia phage phAPEC8	*Escherichia coli*	lytic	Myoviridae	Big_2	3	hypothetical protein phAPEC8_00173
NC_028248	Escherichia phage slur16	*Escherichia coli*	lytic	Myoviridae	Big_2	1	hypothetical protein Ec2_0026
NC_027395	Phage vB_EcoP_SU10	*Escherichia coli*	lytic	Podoviridae	Big_2	4	bacterial surface protein
NC_027356	Enterobacteria phage DT57C	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	putative major tail protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_001901	Enterobacteria phage N15	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	tail protein
NC_017969	Escherichia phage Akfv33	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_010583	Escherichia phage Eps7	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_023593	Escherichia phage KBNP1711	*Escherichia coli*	lytic	Podoviridae	Big_2	4	putative tail fiber protein
NC_005859	Escherichia phage T5	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_015269	Salmonella phage Spc35	*Escherichia coli*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_009514	Enterobacteria phage cdtI	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_019717	Enterobacteria phage HK225	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	phage tail protein
NC_019711	Enterobacteria phage HK629	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	Tail Tube Protein
NC_019723	Enterobacteria phage HK630	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	Tail Tube Protein
NC_001416	Enterobacteria phage lambda	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	Tail Tube Protein
NC_019706	Enterobacteria phage mEp043 c-1	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	tail fiber protein
NC_019704	Enterobacteria phage mEp237	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	tail protein
NC_019716	Enterobacteria phage mEp460	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_021190	Enterobacteria phage phi80	*Escherichia coli*	temperate	Siphoviridae	Big_2	2	minor tail protein
NC_019720	Enterobacterial phage mEp213	*Escherichia coli*	temperate	Siphoviridae	Big_2	1	tail fiber protein
NC_022968	Enterobacteria phage 4MG	*Escherichia coli K-12*	lytic	Myoviridae	Big_2	2	Ig domain-containing protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_011041	Escherichia phage rv5	*Escherichia coli O157*	lytic	Myoviridae	Big_2	1	hypothetical protein BY33_26155
NC_022323	Escherichia phage 2 JES-2013	*Escherichia coli O157:H7*	lytic	Myoviridae	Big_2	1	hypothetical protein BY33_26155
NC_024134	Escherichia phage vB_EcoM_FFH2	*Escherichia coli O157:H7*	lytic	Myoviridae	Big_2	1	hypothetical protein
NC_024139	Escherichia phage vB_EcoS_FFH1	*Escherichia coli O157:H7*	lytic	Siphoviridae	Big_2	1	putative major tail protein
NC_027399	Klebsiella phage K64-1	*Klebsiella pneumoniae*	lytic	Myoviridae	Big_2	1	putative structural protein
NC_019526	Enterobacteria phage vB_KleM-RaK2	*Klebsiella sp.*	lytic	Myoviridae	Big_2	1	putative structural protein
NC_004112	Lactobacillus phage A2	*Lactobacillus casei*	temperate	Siphoviridae	Big_2	1	hypothetical protein
NC_028888	Lactobacillus phage CL1	*Lactobacillus paracasei*	lytic	Siphoviridae	Big_2	1	group 2 bacterial Ig family protein
NC_028835	Lactobacillus phage CL2	*Lactobacillus paracasei*	lytic	Siphoviridae	Big_2	1	group 2 bacterial Ig family protein
NC_028911	Lactobacillus phage iLp1308	*Lactobacillus paracasei*	lytic	Siphoviridae	Big_2	1	group 2 bacterial Ig family protein
NC_012530	Lactobacillus phage Lb338-1	*Lactobacillus paracasei*	lytic	Myoviridae	Big_2	1	hypothetical protein lb338_phage_157
NC_019916	Lactobacillus phage ATCC8014	*Lactobacillus plantarum*	lytic	Siphoviridae	Big_2	1	minor capsid protein
NC_006936	Lactobacillus phage phijl1	*Lactobacillus plantarum*	lytic	Siphoviridae	Big_2	1	hypothetical protein phiJL1_ORF64b
NC_011104	Lactobacillus phage Lrm1	*Lactobacillus rhamnosus*	temperate	Siphoviridae	Big_2	1	hypothetical protein
NC_027120	Lactococcus phage 1358	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	3	putative phage structural protein
NC_009817	Lactococcus phage KSY1	*Lactococcus lactis*	lytic	Podoviridae	Big_2	1	gp055

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_024215	Lactococcus phage P078	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	2	hypothetical protein P092_0011
NC_012663	Lactococcus phage P087	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	1	structural protein
NC_024203	Lactococcus phage P092	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	2	hypothetical protein P092_0011
NC_024208	Lactococcus phage P118	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	2	hypothetical protein P118_0011
NC_024214	Lactococcus phage P162	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	2	hypothetical protein P162_0011
NC_021861	Lactococcus phage BM13	*Lactococcus lactis*	lytic	Siphoviridae	Big_2	1	major tail protein
NC_004066	Lactococcus phage ul36	*Lactococcus lactis*	temperate	Siphoviridae	Big_2	1	hypothetical protein
NC_003216	Listeria phage A118	*Listeria monocytogenes*	lytic	Siphoviridae	Big_2	1	gp13
NC_009811	Listeria virus A511	*Listeria monocytogenes*	lytic	Myoviridae	Big_2	1	gp145
NC_009810	Listeria phage A500	*Listeria monocytogenes*	temperate	Siphoviridae	Big_2	1	Ig-like virion protein
NC_028871	Listeria phage vB_LmoS_188	*Listeria monocytogenes*	temperate	Siphoviridae	Big_2	1	hypothetical protein
NC_028929	Listeria phage vB_LmoS_293	*Listeria monocytogenes*	temperate	Siphoviridae	Big_2	1	hypothetical protein EL89_07770
NC_009815	Listeria phage A006	*Listeria monocytogenes*	temperate	Siphoviridae	Big_2	1	Tsh
NC_024384	Listeria phage LP-030-3	*Listeria monocytogenes serotype 4b str. F2365*	temperate	Siphoviridae	Big_2	1	hypothetical protein EL89_07770
NC_023859	Microbacterium phage vB_MoxS-ISF9	*Microbacterium oxydans*	lytic	Siphoviridae	Big_2	1	putative capsid protein
NC_008206	Mycobacterium phage Wildcat	*Mycobacterium*	lytic	Siphoviridae	Big_2	2	gp30
NC_022071	Mycobacterium phage Crossroads	*Mycobacterium*	temperate	Siphoviridae	Big_2	1	unnamed protein product
NC_022052	Mycobacterium phage Whirlwind	*Mycobacterium*	temperate	Siphoviridae	Big_2	1	major tail subunit

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_014461	Mycobacterium phage bron	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	gp8
NC_028878	Mycobacterium phage Archie	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	2	unnamed protein product
NC_014461	Mycobacterium phage bron	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	gp13
NC_015584	Mycobacterium phage Faith1	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	unnamed protein product
NC_028843	Mycobacterium phage Lolly9	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	unnamed protein product
NC_023732	Mycobacterium phage Rumpelstiltskin	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	unnamed protein product
NC_028778	Mycobacterium phage Snenia	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Big_2	1	unnamed protein product
NC_020201	Pectobacterium phage phiTE	*Pectobacterium atrosepticum SCRI1043*	lytic	Myoviridae	Big_2	2	hypothetical protein phiTE_236
NC_019522	Pectobacterium phage ZF40	*Pectobacterium carotovorum*	lytic	Myoviridae	Big_2	2	putative tail-fiber protein
NC_016161	Pediococcus phage clP1	*Pediococcus damnosus*	lytic	Siphoviridae	Big_2	1	hypothetical protein clP1_004
NC_016762	Pseudomonas phage phi297	*Pseudomonas aeruginosa*	temperate	Siphoviridae	Big_2	2	Ig domain-containing protein
NC_027292	Pseudomonas phage Pf-10	*Pseudomonas fluorescens*	lytic	Podoviridae	Big_2	2	minor capsid protein 10B
NC_015264	Pseudomonas phage phiIBB-PF7A	*Pseudomonas fluorescens*	lytic	Podoviridae	Big_2	2	minor capsid protein 10B
NC_021062	Pseudomonas phage Phi-S1	*Pseudomonas fluorescens*	lytic	Podoviridae	Big_2	1	minor capsid protein
NC_023005	Pseudomonas phage PPpW-4	*Pseudomonas plecoglossicida*	lytic	Podoviridae	Big_2	1	putative Ig domain-containing protein
NC_019490	Riemerella phage RAP44	*Riemerella anatipestifer*	temperate	Siphoviridae	Big_2	1	hypothetical protein M949_1800

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_027351	Salmonella phage SSE-121	*Salmonella*	lytic	Myoviridae	Big_2	1	40 gene product
NC_016071	Salmonella phage PVP-SE1	*Salmonella enterica subsp. enterica serovar Enteritidis*	lytic	Myoviridae	Big_2	1	40 gene product
NC_015938	Salmonella phage 7-11	*Salmonella enterica subsp. enterica serovar Newport*	lytic	Podoviridae	Big_2	3	major head protein
NC_028754	Salmonella phage Shivani	*Salmonella enterica subsp. enterica serovar Typhimurium*	lytic	Siphoviridae	Big_2	1	putative major tail protein
NC_018843	Salmonella phage SSU5	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	lytic	Siphoviridae	Big_2	4	MULTISPECIES: hypothetical protein
NC_020083	Serratia phage phiMAM1	*Serratia sp.*	lytic	Myoviridae	Big_2	2	putative phage tail fiber
NC_026010	Shigella phage pSf-2	*Shigella flexneri*	lytic	Siphoviridae	Big_2	1	hypothetical protein B508_00205
NC_007066	Staphylococcus phage G1	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	ORF080
NC_013195	Staphylococcus phage P954	*Staphylococcus aureus*	lytic	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_023009	Staphylococcus phage Sb-1	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	ORF080
NC_007053	Staphylococcus phage 3A	*Staphylococcus aureus*	lytic	Siphoviridae	Big_2	1	tail protein
NC_019448	Staphylococcus phage GH15	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	putative major tail protein
NC_019726	Staphylococcus phage JD007	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_025416	Staphylococcus phage MCE-2014	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	putative major tail protein
NC_025426	Staphylococcus phage P108	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	putative major tail protein
NC_023573	Staphylococcus phage phiSA012	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_022920	Staphylococcus phage S25-3	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_022918	Staphylococcus phage S25-4	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_007021	Staphylococcus phage Twort	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	putative major tail protein
NC_022090	Staphylococcus phage vB_SauM_Remus	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_020877	Staphylococcus phage vB_SauM_Romulus	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_005880	Staphylococcus virus K	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	major tail protein
NC_019511	Staphylococcus virus SA11	*Staphylococcus aureus*	lytic	Myoviridae	Big_2	1	tail protein
NC_028775	Staphylococcus phage 23MRA	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_005356	Staphylococcus phage 77	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_019921	Staphylococcus phage phi5967PVL	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_008617	Staphylococcus phage phiNM3	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_025460	Staphylococcus phage phiSa119	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_002321	Staphylococcus phage PVL	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein PVL_12
NC_023499	Staphylococcus phage StauST398-4	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_007047	Staphylococcus phage 187	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	ORF031
NC_007051	Staphylococcus phage 2638A	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	ORF014

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_007052	Staphylococcus phage 42E	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_007054	Staphylococcus phage 47	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_011612	Staphylococcus phage Ipla35	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_021773	Staphylococcus phage JS01	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	phage tail protein
NC_011344	Staphylococcus phage phi2958PVL	*Staphylococcus aureus*	temperate	unclassified	Big_2	1	tail protein
NC_020199	Staphylococcus phage phi7401PVL	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_012784	Staphylococcus phage phiPVL-CN125	*Staphylococcus aureus*	temperate	unclassified	Big_2	1	phage tail protein
NC_008689	Staphylococcus phage Pvl108	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	phage tail protein
NC_002661	Staphylococcus phage Slt	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein
NC_019513	Staphylococcus phage SMSAP5	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_021323	Staphylococcus phage StauST398-2	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_009761	Staphylococcus phage tp310-1	*Staphylococcus aureus*	temperate	unclassified	Big_2	1	phage tail protein
NC_009762	Staphylococcus phage tp310-2	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_009763	Staphylococcus phage tp310-3	*Staphylococcus aureus*	temperate	unclassified	Big_2	1	phage tail protein
NC_028862	Staphylococcus phage vB_SauS_phi2	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_022758	Staphylococcus phage YMC/09/04/R1988	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	tail protein, partial

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_002486	Staphylococcus prophage phiPV83	*Staphylococcus aureus*	temperate	Siphoviridae	Big_2	1	phage tail protein
NC_026016	Staphylococcus phage phiBU01	*Staphylococcus aureus HIP07256*	temperate	Siphoviridae	Big_2	1	hypothetical protein, partial
NC_004617	Staphylococcus phage 13	*Staphylococcus aureus subsp. aureus NCTC 8325*	temperate	Siphoviridae	Big_2	1	phage tail protein
NC_004616	Staphylococcus phage Phi12	*Staphylococcus aureus subsp. aureus NCTC 8325*	temperate	Siphoviridae	Big_2	1	tail protein, partial
NC_028765	Staphylococcus phage phiIPLA-RODI	*Staphylococcus sp.*	lytic	Myoviridae	Big_2	1	major tail protein
NC_025417	Staphylococcus phage vB_SauM_Team1	*Staphylococcus xylosus*	lytic	Myoviridae	Big_2	1	ORF080
NC_021868	Streptococcus phage SPQS1	*Streptococcus pneumoniae D39*	lytic	Siphoviridae	Big_2	2	major tail protein
NC_018270	Weissella phage phiYS61	*Weissella cibaria*	lytic	Podoviridae	Big_2	1	major capsid protein with bacterial immunoglobulin-like domain
NC_019919	Yersinia phage phiR201	*Yersinia enterocolitica*	lytic	Siphoviridae	Big_2	2	phage neck whiskers
NC_001271	Yersinia phage phiYeO3-12	*Yersinia enterocolitica (type O:3)*	lytic	Podoviridae	Big_2	1	minor capsid protein 10B
NC_009760	Bacillus phage 0305phi8-36	*Bacillus thuringiensis*	lytic	Myoviridae	Big_5	2	virion structural protein
NC_021325	Clostridium phage vB_CpeS-CP51	*Clostridium perfringens*	temperate	Siphoviridae	CBM_4_9	6	hypothetical protein phiCP51_0022
NC_005857	Klebsiella phage phiKO2	*Klebsiella oxytoca*	lytic	Siphoviridae	CBM_4_9	7	Gp21
NC_012530	Lactobacillus phage Lb338-1	*Lactobacillus paracasei*	lytic	Myoviridae	CBM_4_9	4	putative tail fibre protein
NC_007066	Staphylococcus phage G1	*Staphylococcus aureus*	lytic	Myoviridae	CBM_4_9	1	hypothetical protein phi_Fi200W_ORF094

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_019726	Staphylococcus phage JD007	*Staphylococcus aureus*	lytic	Myoviridae	CBM_4_9	1	hypothetical protein
NC_005880	Staphylococcus virus K	*Staphylococcus aureus*	lytic	Myoviridae	CBM_4_9	1	hypothetical protein phi_Fi200W_ORF094
NC_025417	Staphylococcus phage vB_SauM_Team1	*Staphylococcus xylosus*	lytic	Myoviridae	CBM_4_9	1	hypothetical protein phi_Fi200W_ORF094
NC_004827	Haemophilus phage Aaphi23	*Aggregatibacter actinomycetemcomitans*	temperate	Myoviridae	CBM_5_12	1	putative tail fiber protein
NC_028657	Pseudomonas phage YMC11/02/R656	*Pseudomonas aeruginosa*	temperate	Siphoviridae	CBM_5_12	1	hypothetical protein, partial
NC_007902	Sodalis phage phiSG1	*Sodalis glossinidius str. 'morsitans'*	temperate	unclassified	CBM_5_12	1	hypothetical protein SGPHI_0001
NC_020837	Synechococcus phage S-CAM1	*Synechococcus sp. WH 7803*	lytic	Myoviridae	CBM_5_12	4	hypothetical protein SXBG_00229
NC_001447	Acholeplasma phage L2	*Acholeplasma laidlawii*	temperate	Plasmaviridae	DUF5011	4	envelope protein
NC_023719	Bacillus phage G	*Bacillus megaterium*	lytic	Myoviridae	F5_F8_type_C	5	gp364
NC_028749	Brevibacillus phage Sundance	*Brevibacillus laterosporus*	temperate	Siphoviridae	F5_F8_type_C	5	Hypothetical protein
NC_019406	Caulobacter phage CcrColossus	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	3	putative tail protein
NC_019410	Caulobacter phage karma	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	2	putative tail protein
NC_019407	Caulobacter phage magneto	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	2	putative tail protein
NC_019405	Caulobacter phage phiCbK	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	2	putative tail protein
NC_019408	Caulobacter phage rogue	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	3	putative tail protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_019411	Caulobacter phage swift	*Caulobacter vibrioides*	temperate	Siphoviridae	F5_F8_type_C	2	putative tail protein
NC_024360	Listeria phage LMSP-25	*Listeria monocytogenes*	lytic	Myoviridae	F5_F8_type_C	3	hypothetical protein
NC_024383	Listeria phage LP-083-2	*Listeria monocytogenes*	lytic	Myoviridae	F5_F8_type_C	3	hypothetical protein AG2_096
NC_019522	Pectobacterium phage ZF40	*Pectobacterium carotovorum*	lytic	Myoviridae	F5_F8_type_C	1	hypothetical protein ZF40_0065
NC_028980	Pseudomonas phage PAE1	*Pseudomonas aeruginosa*	lytic	Siphoviridae	F5_F8_type_C	1	hypothetical protein ORF032
NC_028809	Pseudomonas phage PaMx74	*Pseudomonas aeruginosa*	lytic	Siphoviridae	F5_F8_type_C	1	hypothetical protein
NC_028809	Pseudomonas phage PaMx74	*Pseudomonas aeruginosa*	lytic	Siphoviridae	F5_F8_type_C	1	Ig family protein
NC_018282	Pseudomonas phage MP1412	*Pseudomonas aeruginosa PAO1*	lytic	Siphoviridae	F5_F8_type_C	1	tail fiber structural protein
NC_012697	Silicibacter phage DSS3phi2	*Ruegeria pomeroyi DSS-3*	lytic	Podoviridae	F5_F8_type_C	2	hypothetical protein DSS3P2_gp36
NC_020862	Sulfitobacter phage phiCB2047-B	*Sulfitobacter sp. CB2047*	lytic	Podoviridae	F5_F8_type_C	1	hypothetical protein SUFG_00043
NC_020862	Sulfitobacter phage phiCB2047-B	*Sulfitobacter sp. CB2047*	lytic	Podoviridae	F5_F8_type_C	1	hypothetical protein SUFG_00075
NC_008562	Microcystis aeruginosa phage Ma-LMM01	*Microcystis aeruginosa*	lytic	Myoviridae	FGE-sulfatase	1	hypothetical protein MaLMM01_gp037
NC_008562	Microcystis aeruginosa phage Ma-LMM01	*Microcystis aeruginosa*	lytic	Myoviridae	FGE-sulfatase	1	hypothetical protein MaLMM01_gp062
NC_029002	Microcystis phage MaMV-DC	*Microcystis aeruginosa*	lytic	Myoviridae	FGE-sulfatase	1	hypothetical protein MaMVDC_63
NC_029002	Microcystis phage MaMV-DC	*Microcystis aeruginosa*	lytic	Myoviridae	FGE-sulfatase	1	serine/threonine protein kinase
NC_004456	Vibrio phage VHML	*Vibrio harveyi*	temperate	Myoviridae	FGE-sulfatase	1	ORF35

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_027981	Vibrio phage VP58.5	*Vibrio parahaemolyticus*	temperate	Myoviridae	FGE-sulfatase	1	lytic
NC_021561	Vibrio phage pYD38-B	*Vibrio sp. YD38*	lytic	Siphoviridae	FGE-sulfatase	1	hypothetical protein VPSG_00008
NC_004166	Bacillus phage SPP1	*Bacillus subtilis*	lytic	Siphoviridae	fn3	2	hypothetical protein SPP1p107
NC_009760	Bacillus phage 0305phi8-36	*Bacillus thuringiensis*	lytic	Myoviridae	fn3	3	virion structural protein
NC_009760	Bacillus phage 0305phi8-36	*Bacillus thuringiensis*	lytic	Myoviridae	fn3	6	virion structural protein
NC_021537	Halorubrum phage CGphi46	*Halorubrum sp. M4 25-10-8A*	temperate	unclassified	fn3	3	hypothetical protein HALG_00046
NC_027990	Lactobacillus phage LBR48	*Lactobacillus brevis*	temperate	Myoviridae	fn3	2	hypothetical
NC_008198	Mycobacterium phage PBI1	*Mycobacterium*	lytic	Siphoviridae	fn3	2	gp25
NC_008200	Mycobacterium phage PLot	*Mycobacterium*	lytic	Siphoviridae	fn3	2	gp25
NC_011290	Mycobacterium phage Gumball	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	fn3	2	gp23
NC_024209	Mycobacterium phage Hawkeye	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	fn3	2	hypothetical protein PBI_HAWKEYE_26
NC_011285	Mycobacterium phage Troll4	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	fn3	2	gp25
NC_024371	Mycobacterium phage Damien	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	fn3	1	major tail subunit
NC_023578	Mycobacterium phage Oaker	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	fn3	1	major tail subunit
NC_023738	Mycobacterium phage Thibault	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	fn3	3	major tail subunit
NC_027204	Sinorhizobium phage phiM19	*Sinorhizobium meliloti*	lytic	Myoviridae	Glyco_hydro_16	1	putative beta-glucanase

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_028676	Sinorhizobium phage phiM9	*Sinorhizobium meliloti*	lytic	Myoviridae	Glyco_hydro_16	1	putative beta-glucanase
NC_028945	Sinorhizobium phage phiN3	*Sinorhizobium meliloti*	lytic	Myoviridae	Glyco_hydro_16	1	putative beta-glucanase
NC_015720	Gordonia phage GTE2	*Nocardia otitidiscaviarum*	lytic	Siphoviridae	He_PIG	1	main tail protein
NC_027356	Enterobacteria phage DT57C	*Escherichia coli*	lytic	Siphoviridae	Ig_2	3	head protein
NC_028825	Escherichia phage vB_EcoM_AYO145A	*Escherichia coli O145:NM*	lytic	Myoviridae	Ig_2	3	hypothetical protein SP010_00592
NC_021534	Vibrio phage pYD38-A	*Vibrio sp. YD38*	lytic	Siphoviridae	Ig_2	1	major tail subunit
NC_025414	Citrobacter phage Miller	*Citrobacter*	lytic	Myoviridae	Ig_3	3	Wac whiskers
NC_029013	Citrobacter phage IME-CF2	*Citrobacter freundii*	lytic	Myoviridae	Ig_3	3	Wac whiskers
NC_019398	Cronobacter phage vB_CsaM_GAP161	*Cronobacter sakazakii*	lytic	Myoviridae	Ig_3	2	whisker protein
NC_020078	Cronobacter phage vB_CskP_GAP227	*Cronobacter sakazakii*	lytic	Podoviridae	Ig_3	2	hypothetical protein GAP227_44
NC_025452	Dickeya phage RC-2014	*Dickeya*	lytic	Myoviridae	Ig_3	1	unknown structural protein
NC_019925	Dickeya phage Limestone	*Dickeya solani*	lytic	Myoviridae	Ig_3	1	unknown structural protein
NC_007023	Enterobacteria phage RB43	*Enterobacteriaceae*	lytic	Myoviridae	Ig_3	3	Wac whiskers
NC_014467	Enterobacteria phage RB16	*Escherichia coli*	lytic	Myoviridae	Ig_3	3	Wac whisker protein
NC_029023	Escherichia phage JH2	*Escherichia coli*	lytic	Myoviridae	Ig_3	1	hypothetical protein JH2_114
NC_024146	Enterobacteria phage 9g	*Escherichia coli*	lytic	Siphoviridae	Ig_3	1	tail subunit
NC_016158	Escherichia phage HK639	*Escherichia coli*	temperate	Siphoviridae	Ig_3	2	Phage neck whiskers

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_027337	Escherichia phage vB_EcoM-VpaE1	*Escherichia coli B*	lytic	Myoviridae	Ig_3	2	putative tail protein
NC_029021	Enterobacteria phage JenK1	*Escherichia coli K-12*	lytic	Siphoviridae	Ig_3	1	tail subunit
NC_028997	Enterobacteria phage JenP2	*Escherichia coli K-12*	lytic	Siphoviridae	Ig_3	1	tail subunit
NC_027369	Escherichia phage EC6	*Escherichia coli O157:H7*	lytic	Myoviridae	Ig_3	1	hypothetical protein
NC_028872	Escherichia phage HY02	*Escherichia coli O157:H7*	lytic	Myoviridae	Ig_3	3	putative tail protein
NC_012749	Escherichia phage wV8	*Escherichia coli O157:H7*	lytic	Myoviridae	Ig_3	3	putative tail protein
NC_011356	Enterobacteria phage YYZ-2008	*Escherichia coli O157:H7*	temperate	unclassified	Ig_3	1	tail protein
NC_028808	Escherichia phage phiSUSP1	*Escherichia coli str. K-12 substr. MG1655*	lytic	Myoviridae	Ig_3	2	hypothetical protein CPT_Moogle90
NC_028935	Escherichia phage phiSUSP2	*Escherichia coli str. K-12 substr. MG1655*	lytic	Myoviridae	Ig_3	2	hypothetical protein CPT_Moogle90
NC_014036	Klebsiella phage KP15	*Klebsiella pneumoniae*	lytic	Myoviridae	Ig_3	1	whisker protein
NC_020080	Klebsiella phage KP27	*Klebsiella pneumoniae*	lytic	Myoviridae	Ig_3	2	whisker protein
NC_022343	Klebsiella phage 0507-KN2-1	*Klebsiella pneumoniae*	lytic	Myoviridae	Ig_3	1	tail fiber protein
NC_029042	Salmonella phage 38	*Salmonella*	lytic	Myoviridae	Ig_3	1	unnamed protein product
NC_005282	Salmonella phage FelixO1	*Salmonella*	lytic	Myoviridae	Ig_3	2	major tail protein
NC_019910	Salmonella phage SKML-39	*Salmonella*	lytic	Myoviridae	Ig_3	1	tail fiber protein
NC_018279	Salmonella phage vB_SosS_Oslo	*Salmonella*	temperate	Siphoviridae	Ig_3	1	hypothetical protein
NC_028698	Salmonella phage f18SE	*Salmonella enterica*	lytic	Siphoviridae	Ig_3	2	hypothetical protein
NC_027329	Salmonella phage HB-2014	*Salmonella enterica*	lytic	Myoviridae	Ig_3	1	hypothetical protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_027360	Enterobacteriaphage UAB_Phi87	*Salmonella enterica subsp. enterica*	lytic	Myoviridae	Ig_3	3	hypothetical protein Phi87_21
NC_009232	Salmonella phage SETP3	*Salmonella enterica subsp. enterica serovar Enteritidis*	lytic	Siphoviridae	Ig_3	2	putative tail protein
NC_022754	Salmonella phage SETP7	*Salmonella enterica subsp. enterica serovar Enteritidis*	lytic	Siphoviridae	Ig_3	2	putative tail protein
NC_027119	Salmonella phage Det7	*Salmonella enterica subsp. enterica serovar Typhimurium*	lytic	Myoviridae	Ig_3	1	tail fiber protein
NC_006949	Enterobacteria phage ES18	*Salmonella enterica subsp. enterica serovar Typhimurium*	temperate	Siphoviridae	Ig_3	1	hypothetical protein
NC_019545	Salmonella phage SPN3UB	*Salmonella enterica subsp. enterica serovar Typhimurium*	temperate	Siphoviridae	Ig_3	1	hypothetical protein
NC_013693	Shigella phage Ag3	*Shigella boydii*	lytic	Myoviridae	Ig_3	1	putative tail protein
NC_027991	Staphylococcus phage SA1	*Staphylococcus aureus*	lytic	Myoviridae	Ig_3	2	putative tail protein
NC_027393	Vibrio phage J2	*Vibrio cholerae*	lytic	Podoviridae	Ig_3	3	hypothetical protein VP2p14
NC_005879	Vibrio phage VP2	*Vibrio cholerae*	lytic	Podoviridae	Ig_3	3	hypothetical protein VP2p14
NC_027397	Vibrio phage QH	*Vibrio cholerae*	lytic	Podoviridae	Ig_3	3	major tail subunit
NC_005891	Vibrio phage VP5	*Vibrio cholerae*	lytic	Podoviridae	Ig_3	3	major tail subunit
NC_027118	Vibrio phage phiVC8	*Vibrio cholerae O1*	lytic	Podoviridae	Ig_3	4	tail protein
NC_019529	Vibrio phage pVp-1	*Vibrio parahaemolyticus*	lytic	Siphoviridae	Ig_3	1	major tail protein
NC_021344	Escherichia phage Lw1	*Escherichia coli BL21(DE3)*	lytic	Myoviridae	I-set	3	whisker protein
NC_006938	Phage phiJL001	*alpha proteobacterium JL001*	lytic	Siphoviridae	Laminin_G_2	5	lytic
NC_019443	Synechococcus phage metaG-MbCM1	*Synechococcus*	lytic	Myoviridae	Laminin_G_2	2	gp14 neck protein
NC_026923	Synechococcus phage ACG-2014d	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_2	4	hypothetical protein CPUG_00079
NC_006938	Phage phiJL001	*alpha proteobacterium JL001*	lytic	Siphoviridae	Laminin_G_3	5	gp70

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_023007	Bacillus phage vB_BanS-Tsamsa	*Bacillus anthracis*	temperate	Siphoviridae	Laminin_G_3	1	hypothetical protein
NC_025423	Bacillus phage CP-51	*Bacillus cereus*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein
NC_028982	Bacillus phage JL	*Bacillus cereus*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein JL_219
NC_028983	Bacillus phage Shanette	*Bacillus cereus*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein SHANETTE_219
NC_028805	Brevibacillus phage Jenst	*Brevibacillus laterosporus*	temperate	Siphoviridae	Laminin_G_3	2	MULTISPECIES: hypothetical protein
NC_028749	Brevibacillus phage Sundance	*Brevibacillus laterosporus*	temperate	Siphoviridae	Laminin_G_3	1	hypothetical protein
NC_019406	Caulobacter phage CcrColossus	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_019410	Caulobacter phage karma	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_019407	Caulobacter phage magneto	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_019405	Caulobacter phage phiCbK	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_019408	Caulobacter phage rogue	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_019411	Caulobacter phage swift	*Caulobacter vibrioides*	temperate	Siphoviridae	Laminin_G_3	1	putative lectin-like domain protein
NC_021800	Cellulophaga phage phi46:1	*Cellulophaga baltica*	temperate	Siphoviridae	Laminin_G_3	1	SGNH hydrolase
NC_007581	Clostridium phage c-st	*Clostridium botulinum*	temperate	Myoviridae	Laminin_G_3	1	non-toxic non-hemmagglutinin
NC_028672	Cronobacter phage PBES 02	*Cronobacter*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein CR8_067
NC_017974	Cronobacter phage CR3	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein CR8_067

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_024354	Cronobacter phage CR8	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CR8_197
NC_023717	Cronobacter phage CR9	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein CR9_175
NC_019400	Cronobacter phage vB_CsaM_GAP31	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein GAP31_094
NC_019400	Cronobacter phage vB_CsaM_GAP31	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein GAP31_108
NC_017974	Cronobacter phage CR3	*Cronobacter sakazakii*	lytic	Myoviridae	Laminin_G_3	3	lytic
NC_023557	Erwinia phage Ea35-70	*Erwinia amylovora*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein Ea357_113
NC_023610	Erwinia phage PhiEaH1	*Erwinia amylovora*	lytic	Siphoviridae	Laminin_G_3	1	hypothetical protein
NC_023610	Erwinia phage PhiEaH1	*Erwinia amylovora*	lytic	Siphoviridae	Laminin_G_3	2	hypothetical protein
NC_019504	Erwinia phage vB_EamM-Y2	*Erwinia amylovora*	lytic	Myoviridae	Laminin_G_3	1	gp85
NC_010106	Enterobacteria phage phiEcoM-GJ1	*Escherichia coli*	lytic	Myoviridae	Laminin_G_3	1	phage protein
NC_017969	Escherichia phage Akfv33	*Escherichia coli*	lytic	Siphoviridae	Laminin_G_3	1	putative tail fiber protein
NC_008562	Microcystis aeruginosa phage Ma-LMM01	*Microcystis aeruginosa*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein MaLMM01_gp041
NC_029002	Microcystis phage MaMV-DC	*Microcystis aeruginosa*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein MaMVDC_45
NC_011267	Mycobacterium phage Solon	*Mycobacterium smegmatis str. MC2 155*	lytic	Siphoviridae	Laminin_G_3	1	gp6
NC_028876	Mycobacterium phage Ariel	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	LamG
NC_023690	Mycobacterium phage Courthouse	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	LamG

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_028953	Mycobacterium phage MiaZeal	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	LamG
NC_026584	Mycobacterium phage Minerva	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	hypothetical protein PBI_WANDA_50
NC_022066	Mycobacterium phage Redno2	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	lysin A
NC_022067	Mycobacterium phage Wanda	*Mycobacterium smegmatis str. MC2 155*	temperate	Siphoviridae	Laminin_G_3	1	hypothetical protein PBI_WANDA_50
NC_023865	Pectobacterium phage PM1	*Pectobacterium carotovorum subsp. carotovorum*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein PM1_060
NC_020845	Prochlorococcus phage MED4-213	*Prochlorococcus*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein CPMG_00027
NC_006884	Prochlorococcus phage P-SSM4	*Prochlorococcus*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPYG_00004
NC_021071	Cyanophage P-RSM1	*Prochlorococcus marinus str. MIT 9303*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPPG_00213
NC_013021	Cyanophage PSS2	*Prochlorococcus marinus str. MIT 9313*	lytic	Siphoviridae	Laminin_G_3	4	fiber protein
NC_015290	Prochlorococcus phage P-SSM7	*Prochlorococcus marinus str. NATL1A*	lytic	Myoviridae	Laminin_G_3	3	structural protein
NC_020855	Cyanophage P-RSM6	*Prochlorococcus marinus str. NATL2A*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPXG_00128
NC_028663	Cyanophage P-TIM40	*Prochlorococcus marinus str. NATL2A*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPXG_00128
NC_021559	Prochlorococcus phage P-SSM3	*Prochlorococcus marinus str. NATL2A*	lytic	Myoviridae	Laminin_G_3	3	hypothetical protein PRAG_00216
NC_015280	Prochlorococcus phage P-HM1	*Prochlorococcus marinus subsp. pastoris str. CCMP1986*	lytic	Myoviridae	Laminin_G_3	1	baseplate wedge initiator
NC_022096	Pseudomonas phage PaBG	*Pseudomonas aeruginosa PAO1*	lytic	Myoviridae	Laminin_G_3	8	putative adhesin
NC_028950	Ralstonia phage RSL2	*Ralstonia*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_028950	Ralstonia phage RSL2	*Ralstonia*	lytic	Myoviridae	Laminin_G_3	3	laminin
NC_028899	Ralstonia phage RSF1	*Ralstonia solanacearum*	lytic	Myoviridae	Laminin_G_3	1	10 gene product
NC_028899	Ralstonia phage RSF1	*Ralstonia solanacearum*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein
NC_028899	Ralstonia phage RSF1	*Ralstonia solanacearum*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein
NC_029050	Salmonella phage 21	*Salmonella*	lytic	Myoviridae	Laminin_G_3	1	hypothetical protein CR3_066
NC_027351	Salmonella phage SSE-121	*Salmonella*	lytic	Myoviridae	Laminin_G_3	1	10 gene product
NC_016071	Salmonella phage PVP-SE1	*Salmonella enterica subsp. enterica serovar Enteritidis*	lytic	Myoviridae	Laminin_G_3	1	10 gene product
NC_006820	Synechococcus phage S-PM2	*Synechococcus*	lytic	Myoviridae	Laminin_G_3	2	Hypothetical-Protein | belonging to T4-LIKE GC: 864
NC_006820	Synechococcus phage S-PM2	*Synechococcus*	lytic	Myoviridae	Laminin_G_3	2	virion structural protein
NC_008296	Synechococcus phage syn9	*Synechococcus*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPUG_00079
NC_019443	Synechococcus phage metaG-MbCM1	*Synechococcus*	lytic	Myoviridae	Laminin_G_3	4	fiber protein
NC_015569	Synechococcus phage S-CRM01	*Synechococcus sp.*	lytic	Myoviridae	Laminin_G_3	1	hypothetical secreted protein
NC_015282	Synechococcus phage S-SM1	*Synechococcus sp. WH 6501*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein SSM1_102
NC_021072	Cyanophage Syn30	*Synechococcus sp. WH 7803*	lytic	unclassified	Laminin_G_3	2	hypothetical protein CPRG_00210, partial
NC_015285	Prochlorococcus phage Syn33	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	3	structural protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_027130	Synechococcus phage ACG-2014b	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	gp7
NC_026923	Synechococcus phage ACG-2014d	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	structural protein
NC_026928	Synechococcus phage ACG-2014e	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein
NC_026924	Synechococcus phage ACG-2014g	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	gp7
NC_023587	Synechococcus phage ACG-2014h	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	3	structural protein
NC_021530	Synechococcus phage S-CAM8	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	3	structural protein
NC_021536	Synechococcus phage S-IOM18	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	3	hypothetical protein SWYG_00065
NC_013085	Synechococcus phage S-RSM4	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	structral protein
NC_020851	Synechococcus phage S-SKS1	*Synechococcus sp. WH 7803*	lytic	Siphoviridae	Laminin_G_3	6	hypothetical protein SWZG_00005
NC_020851	Synechococcus phage S-SKS1	*Synechococcus sp. WH 7803*	lytic	Siphoviridae	Laminin_G_3	4	hypothetical protein SWZG_00008
NC_023584	Synechococcus phage S-MbCM100	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	YadA domain protein
NC_020486	Synechococcus phage S-RIM8	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	lytic
NC_019444	Synechococcus phage ACG-2014c	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	2	gp8 baseplate wedge protein
NC_026927	Synechococcus phage ACG-2014f	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	1	baseplate wedge initiator
NC_026927	Synechococcus phage ACG-2014f	*Synechococcus sp. WH 7803*	lytic	Myoviridae	Laminin_G_3	1	gp8 baseplate wedge protein
NC_020875	Synechococcus phage S-SSM4	*Synechococcus sp. WH 8018*	lytic	Myoviridae	Laminin_G_3	2	structural protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_015288	Prochlorococcus phage Syn1	*Synechococcus sp. WH 8101*	lytic	Myoviridae	Laminin_G_3	3	hypothetical protein Syn1_217
NC_015281	Synechococcus phage S-ShM2	*Synechococcus sp. WH 8102*	lytic	Myoviridae	Laminin_G_3	2	structural protein
NC_015289	Synechococcus phage S-SSM5	*Synechococcus sp. WH 8102*	lytic	Myoviridae	Laminin_G_3	4	fiber
NC_015286	Synechococcus phage Syn19	*Synechococcus sp. WH 8109*	lytic	Myoviridae	Laminin_G_3	2	hypothetical protein CPTG_00203, partial
NC_029047	Verrucomicrobia phage P8625	*Verrucomicrobia str. ICC8625*	lytic	Siphoviridae	Laminin_G_3	2	lytic
NC_028829	Vibrio phage ValKK3	*Vibrio alginolyticus NBRC 15630*	lytic	Myoviridae	Laminin_G_3	5	gp35
NC_023568	Vibrio phage VH7D	*Vibrio harveyi*	lytic	Myoviridae	Laminin_G_3	3	hypothetical protein JCM19296_3255
NC_013599	Xylella phage Xfas53	*Xylella fastidiosa*	temperate	Podoviridae	Laminin_G_3	1	gp45
NC_012638	Enterobacteria phage RB14	*Enterobacteriales*	lytic	Myoviridae	PKD	3	large head outer capsid protein
NC_019503	Enterobacteria phage ime09	*Escherichia coli*	lytic	Myoviridae	PKD	4	head outer capsid protein
NC_025448	Enterobacteria phage RB27	*Escherichia coli*	lytic	Myoviridae	PKD	4	outer capsid protein Hoc
NC_025419	Enterobacteria phage RB3	*Escherichia coli*	lytic	Myoviridae	PKD	4	large head outer capsid protein
NC_008515	Enterobacteria phage RB32	*Escherichia coli*	lytic	Myoviridae	PKD	4	large head outer capsid protein
NC_000866	Enterobacteria phage T4	*Escherichia coli*	lytic	Myoviridae	PKD	3	Hoc head outer capsid protein
NC_028448	Escherichia phage slur08	*Escherichia coli*	lytic	Myoviridae	PKD	6	putative large head outer capsid protein
NC_024794	Escherichia phage vB_EcoM_PhAPEC2	*Escherichia coli*	lytic	Myoviridae	PKD	3	head outer capsid protein

Table 4-2: Ig-like domains in isolated bacteriophage in the Refseq database, continued.

Accession	Phage	Host	Lifestyle	Morphology	Domain	#	ORF annotation (BLAST)
NC_019505	Escherichia phage wV7	*Escherichia coli*	lytic	Myoviridae	PKD	4	head outer capsid protein
NC_024124	Escherichia phage vB_EcoM_JS09	*Escherichia coli*	lytic	Myoviridae	PKD	4	ADP-ribosyltransferase
NC_027979	Escherichia phage RB68	*Escherichia coli B*	lytic	Myoviridae	PKD	3	large head outer capsid protein
NC_027983	Enterobacteria phage AR1	*Escherichia coli O157:H7*	lytic	Myoviridae	PKD	4	head outer capsid protein
NC_025449	Escherichia phage ECML-134	*Escherichia coli O157:H7*	lytic	Myoviridae	PKD	4	capsid and scaffold protein
NC_027349	Escherichia phage HY01	*Escherichia coli O157:H7*	lytic	Myoviridae	PKD	3	putative head outer capsid protein, Hoc
NC_025437	Shigella phage Shf125875	*Shigella flexneri*	lytic	Myoviridae	PKD	3	capsid decoration protein
NC_015457	Shigella phage Shfl2	*Shigella flexneri*	lytic	Myoviridae	PKD	3	putative large head outer capsid protein
NC_025829	Shigella phage pSs-1	*Shigella sonnei*	lytic	Myoviridae	PKD	4	capsid and scaffold protein
NC_027353	Yersinia phage phiD1	*Yersinia pestis*	lytic	Myoviridae	PKD	4	capsid and scaffold protein
NC_027404	Yersinia phage PST	*Yersinia pseudotuberculosis*	lytic	Myoviridae	PKD	5	capsid decoration protein
NC_023719	Bacillus phage G	*Bacillus megaterium*	lytic	Myoviridae	Pur_ac_phosph_N	2	gp352
NC_024217	Nitrincola phage 1M3-16	*Nitrincola*	lytic	unclassified	Pur_ac_phosph_N	3	putative glycosyl hydrolase
NC_009603	Microbacterium phage Min1	*Microbacterium nematophilum*	temperate	Siphoviridae	TIG	1	hypothetical protein MPMin1_gp57

References

Adams, M.H. (1959). Bacteriophages (New York: Interscience Publishers, Inc.).

Bansil, R., Stanley, E., and LaMont, J.T. (1995). Mucin biophysics. Annu. Rev. Physiol. *57*, 635–657.

Barr, J.J., Auro, R., Furlan, M., Whiteson, K.L., Erb, M.L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A.S., Doran, K.S., et al. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc. Natl. Acad. Sci. U. S. A. *110*, 10771–10776.

Barr, J.J., Auro, R., Sam-Soon, N., Kassegne, S., Peters, G., Bonilla, N., Hatay, M., Mourtada, S., Bailey, B., Youle, M., et al. (2015). Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. Proc. Natl. Acad. Sci. *112*, 13675–13680.

Benzer, S. (1955). Fine Structure of a Genetic Region in Bacteriophage. Proc. Natl. Acad. Sci. *41*, 344–354.

Coddeville, B., Maes, E., Ferrier-Pagès, C., and Guerardel, Y. (2011). Glycan profiling of gel forming mucus layer from the scleractinian symbiotic coral oculina arbuscula. Biomacromolecules *12*, 2064–2073.

Cohen, M. (2015). Notable Aspects of Glycan-Protein Interactions. Biomolecules *5*, 2056–2072.

Cohen, M., Fisher, C.J., Huang, M.L., Lindsay, L.L., Plancarte, M., Boyce, W.M., Godula, K., and Gagneux, P. (2016). Capture and characterization of influenza A virus from primary samples using glycan bead arrays. Virology *493*, 128–135.

Cohen, Y., Joseph Pollock, F., Rosenberg, E., and Bourne, D.G. (2013). Phage therapy treatment of the coral pathogen Vibrio coralliilyticus. Microbiologyopen *2*, 64–74.

Comstock, L.E. (2009). Importance of Glycans to the Host-Bacteroides Mutualism in the Mammalian Intestine. Cell Host Microbe *5*, 522–526.

Ewoldt, R.H., Winegard, T.M., and Fudge, D.S. (2011). Non-linear viscoelasticity of hagfish slime. Int. J. Non. Linear. Mech. *46*, 627–636.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. *39*, 29–37.

Fokine, A., Chipman, P.R., Leiman, P.G., Mesyanzhinov, V. V, Rao, V.B., and Rossmann, M.G. (2004). Molecular architecture of the prolate head of bacteriophage T4. Proc. Natl. Acad. Sci. U. S. A. *101*, 6003–6008.

Fraser, J.S., Yu, Z., Maxwell, K.L., and Davidson, A.R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. *359*, 496–507.

Fraser, J.S., Maxwell, K.L., and Davidson, A.R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr. Opin. Microbiol. *10*, 382–387.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Garren, M., Son, K., Raina, J.-B., Rusconi, R., Menolascina, F., Shapiro, O.H., Tout, J., Bourne, D.G., Seymour, J.R., and Stocker, R. (2014). A bacterial pathogen uses dimethylsulfoniopropionate as a cue to target heat-stressed corals. ISME J. *8*, 999–1007.

Gunning, a P., Kirby, A.R., Fuell, C., Pin, C., Tailford, L.E., and Juge, N. (2013). Mining the "glycocode"--exploring the spatial distribution of glycans in gastrointestinal mucin using force spectroscopy. FASEB J. *27*, 2342–2354.

Jatkar, A. a., Brown, B.E., Bythell, J.C., Guppy, R., Morris, N.J., and Pearson, J.P. (2010). Coral mucus: The properties of its constituent mucins. Biomacromolecules *11*, 883–888.

Katsura, I. (1981). Structure and function of the major tail protein of bacteriophage lambda. Mutants having small major tail protein molecules in their virion. J. Mol. Biol. *146*, 493–512.

Kramer, J.R., Onoa, B., Bustamante, C., and Bertozzi, C.R. (2015). Chemically tunable mucin chimeras assembled on living cells. Proc. Natl. Acad. Sci. 201516127.

Kremser, L., Blaas, D., and Kenndler, E. (2004). Capillary electrophoresis of biological particles: Viruses, bacteria, and eukaryotic cells. Electrophoresis *25*, 2282–2291.

Kuhn, R., Frei, R., and Christen, M. (1994). Use of Capillary Affinity Electrophoresis for the Determination of Lectin-Sugar Interactions. Anal. Biochem. *218*, 131–135.

Liu, M., Deora, R., Doulatov, S.R., Gingery, M., Eiserling, F.A., Preston, A., Maskell, D.J., Simons, R.W., Cotter, P.A., Parkhill, J., et al. (2002). Reverse Transcriptase – Mediated Tropism Switching in Bordetella. Science (80-.). *295*, 2091–2095.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Rüger, W. (2003). Bacteriophage T4 Genome. Microbiol. Mol. Biol. Rev. *67*, 86–156.

Morais, M.C., Choi, K.H., Koti, J.S., Chipman, P.R., Anderson, D.L., and Rossmann, M.G. (2005). Conservation of the capsid structure in tailed dsDNA bacteriophages: The pseudoatomic structure of φ29. Mol. Cell *18*, 149–159.

De Paepe, M., Tournier, L., Moncaut, E., Son, O., Langella, P., and Petit, M.A. (2016). Carriage of Lamda Latent Virus Is Costly for Its Bacterial Host due to Frequent Reactivation in Monoxenic Mouse Intestine. PLoS Genet. *12*, e1005861.

Parent, K.N., Khayat, R., Tu, L.H., Suhanovsky, M.M., Cortines, J.R., Teschke, C.M., Johnson, J.E., and Baker, T.S. (2010). P22 Coat Protein Structures Reveal a Novel Mechanism for Capsid Maturation: Stability without Auxiliary Proteins or Chemical Crosslinks. Structure *18*, 390–401.

Reilly, B.E., Nelson, R.A., and Anderson, D.L. (1977). Morphogenesis of bacteriophage phi 29 of Bacillus subtilis: mapping and functional analysis of the head fiber gene. J. Virol. *24*, 363–377.

Salas, M., Vasquez, C., Mandez, E., and Vinuela, E. (1972). Fibers of Bacteriophage phi29. Virology *50*, 180–188.

Sanchez, S.E., Cuevas, D. a., Rostron, J.E., Liang, T.Y., Pivaroff, C.G., Haynes, M.R., Nulton, J., Felts, B., Bailey, B. a., Salamon, P., et al. (2015). Phage Phenomics: Physiological approaches to characterize novel viral proteins. J. Vis. Exp. JoVE 1–18.

Silveira, C.B., and Rohwer, F.L. (2016). Piggyback-the-Winner in host-associated microbial communities. Npj Biofilms Microbiomes *2*, 16010.

Simpson, D.J., Sacher, J.C., and Szymanski, C.M. (2015). Exploring the interactions between bacteriophage-encoded glycan binding proteins and carbohydrates. Curr. Opin. Struct. Biol. *34*, 69–77.

Susskind, M.M., and Botstein, D. (1978). Molecular Genetics of Bacteriophage P22. Microbiol. Rev. *42*, 385–413.

Vernhes, E., Renouard, M., Gilquin, B., Cuniasse, P., Durand, D., England, P., Hoos, S., Huet, A., Conway, J.F., Glukhov, A., et al. (2017). High affinity anchoring of the decoration protein pb10 onto the bacteriophage T5 capsid. Sci. Rep. *7*, 1–14.

Xiao, Q., Ludwig, A.-K., Romanò, C., Buzzacchera, I., Sherman, S.E., Vetro, M., Vértesy, S., Kaltner, H., Reed, E.H., Möller, M., et al. (2018). Exploring functional pairing between surface glycoconjugates and human galectins using programmable glycodendrimersomes. Proc. Natl. Acad. Sci. 201720055.

Xu, J., Hendrix, R.W., and Duda, R.L. (2014). Chaperone-protein interactions that mediate assembly of the bacteriophage lambda tail to the correct length. J. Mol. Biol. *426*, 1004–1018.

Yamaguchi, Y., and Yanagida, M. (1980). Head shell protein hoc alters the surface charge of bacteriophage T4. Composite slab gel electrophoresis of phage T4 and related particles. J. Mol. Biol. *141*, 175–193.