Spring 2014

A bioinformatics approach to revealing the genetic basis for bacteriophage host specificity

Hayley Anne Norian

James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/honors201019

Recommended Citation
https://commons.libjmu.edu/honors201019/456

This Thesis is brought to you for free and open access by the Honors College at JMU Scholarly Commons. It has been accepted for inclusion in Senior Honors Projects, 2010-current by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
A Bioinformatics Approach to Revealing the Genetic Basis for Bacteriophage Host Specificity

A Project Presented to
the Faculty of the Undergraduate
College of Science and Mathematics
James Madison University

in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science

by Hayley Anne Norian
May 2014

Accepted by the faculty of the Department of Biology, James Madison University, in partial fulfillment of the requirements for the Degree of Bachelor of Science.

FACULTY COMMITTEE:  HONORS PROGRAM APPROVAL:

Project Advisor:  Steven G. Cresawn, Ph. D.,
Associate Professor, Biology

Barry Falk, Ph.D.,
Director, Honors Program

Reader:  Stephanie Stockwell, Ph. D.,
Assistant Professor, Integrated Science and Technology

Reader:  Crystal Scott Croshaw, M.S.,
Wet Labs Coordinator and CISE Safety Liaison,
Integrated Science and Technology
# Table of Contents

List of Figures 3

Acknowledgments 4

Abstract 5

Introduction 6

Methods 16

Results 18

Discussion 26

References 28
List of Figures

1. Venn diagram of conserved/non-conserved families in infecting phages pg 13
2. Genome map of subcluster A2 phages highlighting pham 982 pg 20
3. Clustal Omega amino acid alignment of gp 22 in cluster G phages pg 22
4. Genome map of subcluster A2 phages highlighting pham 7269 pg 23
5. Genome map of subcluster A2 phages highlighting pham 3838 pg 24
6. Genome map of subcluster L1 phages highlighting pham 3747 pg 25
Acknowledgements

The completion of this Honors Thesis would not have been possible without the help of my wonderful committee. Dr. Steven G. Cresawn served as both my mentor and guide during this research project and Dr. Stephanie Stockwell and Crystal Scott served as committee members and readers. I would like to thank each of them for their guidance, support and time throughout this entire demanding yet extremely rewarding process.
Abstract

Mycobacteriophages are bacteriophages that infect the genus *Mycobacterium*, including pathogens such as *Mycobacterium tuberculosis* and *Mycobacterium ulcerans*. Full genome sequences of 654 mycobacteriophages are currently available. A mere 20.25% of the 69,581 genes encoded by these phages have at least one known homologue in NCBI, leaving roughly 80% of known mycobacteriophages genes without a predicted function. The host range of 204 mycobacteriophages, initially isolated on *Mycobacterium smegmatis* strain mc²155, was recently determined on *M. tuberculosis* and *M. smegmatis* strains Jucho and MKD8. Three different levels of infectivity were observed: phages that were incapable of infecting the host, phages that were capable of infecting the host, and some that were at a plating efficiency less than one relative to mc²155. The phages that are capable of infecting a host are of particular interest. With so many uncharacterized genes encoded by these phages, we will take a computational approach by performing an association study using the Phamerator software to study the relationship between the complement of protein phamilies in specific genomes and host range of the corresponding phages.
Introduction

The genus *Mycobacterium* includes numerous human pathogens ranging from the ancient scourge of leprosy to the world’s leading infectious cause of death, tuberculosis (Huygen *et al*., 1996). However, a newly emerging pathogen is taking hold in sub-Saharan Africa, Central America, and Australia. This organism, *Mycobacterium ulcerans*, is the cause of a disfiguring, debilitating disease known as Buruli ulcer. The slow-growing *M. ulcerans* is the cause of the third most common mycobacterial infection in immunocompromised patients worldwide (Weir, 2002). This disease is difficult to diagnose and treat, and research aimed at improving this situation is receiving little funding or recognition. Current antibiotic treatment of the topical lesions caused by Buruli ulcer is ineffective if treatment is delayed; in most cases, the only effective means of removing the disease from the body if antibiotic treatment is not taken early is expensive and invasive surgical removal of the necrotic skin followed by skin grafting (Etuaful *et al*., 2005). However, an opportunity exists to combat this pathogen with another class of microbe, viruses that infect bacterial cells.

Bacteriophages are viruses that are parasitic on bacteria, and they are the most abundant biological entities on the planet. Recent estimates of their population size suggest that there are $10^{31}$ particles at any given time (Sabouri and Mohammadi, 2012). Furthermore, these particles are thought to collectively cause $10^{25}$ infections of bacterial cells per second worldwide (Lima-Mendez *et al*., 2007). Thus phages play an important environmental role through their constant cycling of carbon and other nutrients. At this
time, the complete genome sequences of 983 phages have been determined and made publicly available through GenBank. Since this represents only a small fraction of the total population, it is potentially misleading to attempt to generalize. However, a comparison of the available genome sequences shows that they are a highly genetically diverse population.

Although phages are extremely diverse in nature, most tend to be highly host-specific. An understanding of the infection process of phages can explain the high specificity of these viruses to particular hosts. As obligate intracellular parasites, phages must successfully penetrate the bacterial cell membrane and manipulate its cellular mechanisms in order to replicate and release mature virions (Rakhuba et al., 2010). In order to initiate an infection, phages must bind to specific receptors on the surface of bacteria (Goldberg et al., 1994). When binding is possible, most phages adsorb to the bacterial cell wall, although some are able to adsorb to extracellular components of the bacterium such as flagella or pili. Adsorption in many cases is mediated by tail fibers, or long projections originating at the base plate of tailed bacteriophages (Aksyuk et al., 2009). Tail fibers initiate adsorption by reversible attachment to specific receptors on the membrane of the bacterial cell, followed by irreversible attachment (Rakhuba et al., 2010). If the phage is lacking the unique component to bind to the host cell or is otherwise unable to bind its receptor, it is incapable of infecting the cell and subsequently replicating within the host.

Following adsorption via tail fibers, a conformational change occurs such that the sheath portion of the tail contracts, allowing the bacteriophage genome to penetrate the cell membrane of the bacterial host (Kostyuchenko et al., 2005). Once inside the bacterial
cell, bacteriophages that undergo successful infection typically reproduce by one of two methods: the lysogenic cycle or the lytic cycle. Lysogeny involves the integration of the phage genome into the host’s genome, where it remains dormant and is replicated as part of the bacterial chromosome (Wittebole et al., 2014). The integrated phage genome is then known as a prophage. Once an event triggers excision of the integrated prophage, viral replication may then proceed via the lytic cycle. The lytic cycle is characterized by the eventual lysis of the bacterial host upon extensive proliferation of the bacteriophage. Lytic phages manipulate the bacterial synthetic machinery to produce viral nucleic acids and proteins rather than their bacterial counterparts such that DNA replication, transcription, translation, and phage assembly and packing may occur (Labrie et al., 2010).

Successful adsorption, penetration, and injection of the bacteriophage genetic material into the bacterial host cell does not guarantee productive assembly and release of the virus during a lytic infection. Many host defense mechanisms exist that allow the bacterial cell to recognize foreign genetic material and degrade it or otherwise halt the phage replication cycle, often at the cost of the infected cell. These defense mechanisms further restrict the already narrow host range of bacteriophages. Restriction-modification systems, involving restriction endonucleases, are one well-known mechanism of host defense against bacteriophages. Restriction endonucleases are enzymes that cleave DNA at or near specific recognition sequences as a defense mechanism against viruses (Loenen et al., 2014). When unmethylated phage DNA enters the host cell, it can be methylated by a bacterial methylase to avoid restriction modification and thus proceed to subsequent steps of the lytic infection life cycle, including DNA replication, transcription,
translation, and phage assembly (Labrie et al., 2010). However, in most cases, the unmethylated viral DNA is first recognized by restriction enzymes of the host’s restriction-modification system. While the host DNA is preserved, the foreign DNA of the virus is cut at the restriction sites and rapidly degraded. Restriction enzymes have proven to be of monumental importance and have led to many developments in the fields of genetic research and molecular biology. These endonucleases, with the assistance of DNA ligase, allow for the insertion of genes into plasmids during gene cloning and protein expression (Roberts, 1976). They are also useful in DNA fingerprinting as an alternative to gene sequencing. Restriction enzymes also play a critical role in producing recombinant DNA by cleaving a DNA molecule such that its “sticky ends” could be joined to a similarly cut plasmid by DNA ligase (Loenen et al., 2014). This development enabled DNA to be cloned into Escherichia coli.

Upon infection by lytic phages, only a small portion of phage-infected bacterial cells (approximately 1 in $10^6$) survive (Labrie et al., 2010). These phage-resistant bacterial cells may then utilize another mechanism of host defense involving clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (cas) genes that is referred to collectively as the CRISPR/cas System. CRISPRs are short repeat sequences of about 23 to 47 base pairs (bp) separated by hypervariable sequences of about 21 to 72 bp called spacers (Horvath and Barrangou, 2010). These sequences are present in many prokaryotic genomes and seem to provide a DNA-encoded form of acquired immunity against invading nucleic acid, such as viruses and plasmids (Bhaya et al., 2011). Spacer sequences between repeats share 100% nucleotide identity with components of viral genome sequence. For this reason, they are considered to be
segments of viral DNA that have been incorporated into the host genome after previous exposure to a virus and subsequent survival of the infected bacterial cell (Labrie et al., 2010). CRISPR loci are frequently located adjacent to cas genes, which are hypothesized to provide the CRISPR-encoded immunity; inactivation of cas genes impairs host integration of novel repeat-spacer sequences following phage exposure (Horvath and Barrangou, 2010). Upon transcription of the repeat-spacer sequence, the CRISPR transcript that is generated is then processed into small RNA molecules that correspond to a spacer surrounded at either end by two partial repeats. The small CRISPR RNAs (crRNAs) act as guides for the cas interference machinery, bringing them towards foreign nucleic acid molecules with a sequence matching the spacer sequences of the CRISPR (Bhaya et al., 2011). Ultimately, the cas interference machinery degrades the foreign genetic material and the infection ceases. The role of cas genes in the CRISPR/cas system was demonstrated when inactivation of the CRISPR1-associated csn1-like gene resulted in loss of phage resistance, although matching spacers were detected (Horvath and Barrangou, 2010).

Abortive infection (Abi) systems offer another mechanism that can influence the capability of a bacterial strain to defend itself from productive infection by a particular phage. Abi systems provide protection for populations of bacterial cells via altruistic death of the infected cell (Labrie et al., 2010). A toxin-antitoxin (TA) system with an Abi phenotype termed ToxIN has been characterized (Fineran et al., 2009). TA systems rely on the dual activity of a toxin and an antagonistic antitoxin. When expressed in a balance, the antitoxin is able to neutralize the toxin and avoid its damaging effects on the cell. However, when the balance is altered, the toxin is released and may act on important
cellular processes or result in bacterial cell death (Labrie et al., 2010). In such a way, TA systems are capable of inhibiting phage genome replication, transcription, and translation. In the case of the ToxIN system of Gram-negative Erwinia carotovora subspecies atroseptica (Eca), the ToxI protein acts to suppress the toxic effects of the ToxN protein (Fineran et al., 2009).

The ability of phages to specifically infect their target bacteria has proven extremely useful in diagnostics. Methods exploiting this specificity for bacterial identification are typically quick and cost-effective. For example, phage typing is a method used to determine the source of an infection that involves inoculating the bacteria with different phages of known host specificity in order to differentiate the particular strain causing the infection (Schofield et al., 2012). Plaques, or circular clearings in the agar, are visible when the phage successfully infects and lyses the bacterial host cells. More recently, a method involving the detection of light produced by luciferase or fluorescent reporter phages has been used to observe infections (Rybniker et al., 2006). Specifically, a TM4 mycobacteriophage was engineered to contain a gene encoding enhanced green fluorescent protein (EGFP) in order to detect drug-resistant strains of Mycobacterium tuberculosis (Rondon et al., 2011). This method is both a rapid and economical means of detecting drug-resistant tuberculosis and may be clinically useful.

As public health concerns over antibiotic resistance in pathogenic organisms escalate, a need for novel mechanisms of treating bacterial infection arises. The specificity of phages to their hosts, among other factors, makes them potentially useful as therapeutic agents to combat bacterial infection. Using phages to treat pathogenic bacterial infections is termed phage therapy. The ultimate lysis of the bacterial cell is
common between both phage infections and antibiotics, however many differences
between them exist. Some antibiotics, including β-lactam antibiotics such as penicillins,
cause bacterial cell death by inhibiting cell wall peptidoglycan synthesis (Holten and
Onusko, 2000). As the structural integrity of the cell wall is lost, the bacterial cell
becomes susceptible to osmotic pressure and eventually lyses. Phages, on the other hand,
bind and infect a particular host, utilize the machinery of that host cell to replicate, and
eventually lyse bacteria from within the cell with the help of enzymes such as lysozyme,
holin, and hydrolase (Young, 1992). Lysis serves to release the replicated phages and is
therefore essential to the spread of viral infections.

Exposing bacteria to antibiotics naturally selects for organisms containing
antibiotic resistance genes. Although the extent of diversity among the bacteriophage
population has yet to be determined, it is evident that phages are much more diverse than
the very limited number of antibiotics currently in use (Wittebole et al., 2014). Although
utilizing phage therapy will inevitably result in some resistance, considering the diversity
and number of phages in existence, it is almost a non-issue. Phages are also much more
specific for their host than antibiotics, which can act on a much broader spectrum
(Kutateladze and Adamia, 2010). Since phages can only infect particular hosts, the
likelihood that the normal flora within the body will be affected is greatly reduced.
Leaving the human microbiota undisturbed can potentially reduce the risk of
opportunistic infection during treatment (Wittebole et al., 2014). The use of phages as
therapeutic agents to treat pathogenic bacterial infections in humans has not been
approved in the United States, however phages are currently being used here as a means
of controlling growth of bacterial pathogens and spoilage organisms in food and the food-
processing environment (Brovko et al., 2012). While phage therapy lost favor in the United States after the pre-antibiotic era ended and antibiotics were introduced, phage therapy is still commonly practiced in the country of Georgia and other parts of Eastern Europe (Wittebole et al., 2014).

The diverse population of bacteriophages that exist may be organized by which bacterium they infect. Mycobacteriophages, viruses that infect the genus Mycobacterium, are of particular interest. The complete genome sequences of 654 mycobacteriophages have been determined. Comparative genomic analysis at both the nucleotide and gene content levels shows that while all 654 mycobacteriophages infect Mycobacterium smegmatis strain mc²155, the host on which they were isolated, they represent a diverse population as a whole. These phages have been grouped into 62 distinct groups, termed clusters, subclusters, and singletons, based on average nucleotide identity and protein family composition. Clusters are closely related genomes based on these parameters, while subclusters represent genomes within the same cluster that may be further divided based on differences in degrees of nucleotide similarity between members of the same cluster. Singletons are single genomes with little to no sequence similarity with any another sequenced genome at this time. These classifications vary dramatically in size: singletons, the smallest of these distinct groups, are each the only representative of their genome architecture. Meanwhile, there are 72 subcluster A1 genomes representing that architecture. However, it is important to note that there is likely some sampling bias as the mycobacteriophages represented here have been collected over a number of years, from a relatively small number of sites, and mostly in early fall. Therefore, the current
distribution probably does not accurately depict the actual diversity of the population at any given time.

Clusters and subclusters are determined by Phamerator, a bioinformatic tool capable of both comparative genome analysis and representations of said genomes. Phamerator organizes genomes and proteins into related groups based on nucleotide and amino acid identity. In order to sort proteins into families (phams) of related amino acid sequences, Phamerator performs pairwise amino acid sequence comparisons between predicted protein products of a set of phage genomes (Cresawn et al., 2011). These phams are organized in such a way as to allow the relationships between different phages to be analyzed using genome maps, which, in turn, illustrate the mosaic nature and potentially the evolutionary history of phage genomes.

While all of the sequenced mycobacteriophages infect *M. smegmatis* strain mc$^2$155, a subset also infect other related mycobacterial hosts. These include human pathogens such as *Mycobacterium leprae*, *M. tuberculosis*, and *M. ulcerans*. The host range of 204 mycobacteriophages, initially isolated on *Mycobacterium smegmatis* strain mc$^2$155, was recently determined on *M. tuberculosis* and *M. smegmatis* strains Jucho and MKD8. (Jacobs-Sera et al., 2012). The quantification of host range phenotypes was described as an efficiency of plating relative to mc$^2$155. Phages with an efficiency of plating of one for a particular host have the same titer that they have on mc$^2$155. Those with an efficiency of plating of zero for a particular host do not detectably infect that host. Efficiency values between zero and one can indicate a reduced replication rate or the emergence of viral mutants that can infect a host that is typically non-permissive for
that virus. All three types of efficiency were observed, and efficiencies generally correlated with the phage genome clusters.

Two large sets of data currently exist without any efficient means of correlating the information contained within them: mycobacteriophage genome sequence data and mycobacteriophage host range data. While 654 mycobacteriophages genomes are currently sequenced and in the efforts of a single paper, the efficiencies of plating of over 204 mycobacteriophages were determined, there is no obvious way to draw conclusions or inferences about the relationship between genome composition and host specificity. The majority of bacteriophage genomes consist of genes of unknown function with no known homologues. The few exceptions seem to be comprised of highly conserved structural genes. Of the 69,581 genes encoded by the sequenced mycobacteriophages, a mere 20.25% of the genes have at least one known homologue in NCBI. This leaves roughly 80% of all sequenced mycobacteriophage genes without even a predicted function. With so many uncharacterized genes encoded by both the phages and their hosts, we will take a computational approach by performing an association study to identify individual genes or combinations of genes that are linked to host susceptibility or resistance to each phage. A more thorough understanding of the molecular basis for host range will contribute to the utility of phages as therapeutic and diagnostic tools.
Methods

Phamerator, the bioinformatic software program used for comparative genomic analysis and representation of phage genomes, is written entirely in Python computer programming language. This software organizes related gene products into phamilies based on amino acid sequence similarity utilizing both BLASTP and CLUSTALW to perform pairwise amino acid sequence comparisons. A CLUSTALW threshold of 32.5% identity and a BLASTP e-value cut off of $10^{-50}$ served as optimal parameters for building phamilies; these values allowed phamilies of homologous proteins to be built without false phamily assembly of only closely related domains within the proteins themselves. Phamerator also performs automated searches of GenBank and NCBI to indicate previously identified proteins and conserved domains.

A host range database was established utilizing the current database of bacteriophage genome data within the program Phamerator. Phamerator utilizes MySQL database software to populate phage and gene tables with information in GenBank records. In addition to these tables, another table containing the bacterial hosts used in the host range study was created. A second table comprised of the efficiencies of plating of 204 phages on the different bacterial hosts relative to their infection of *M. smegmatis* strain mc²155 was generated. A genome-wide association study (GWAS) concerning host range was performed. Software was written to find conserved protein phamilies that exist in any phage that infects a particular host to determine whether or not these conserved proteins correlate with the ability of the phages to infect the host in question. The
information from the GWAS was then depicted on the genomic maps generated within the program by selecting the “show host range data” option from the pull-down menu, at which point the program colored the conserved protein families on the genomic maps according to the host(s) that the phage containing the conserved proteins could infect.
Results

After the host range data was added to the MySQL table, a GWAS was performed to determine the number of conserved and non-conserved protein families that existed between mycobacteriophages with an efficiency of plating on hosts *M. tuberculosis*, Jucho, and MKD8 within one order of magnitude of mc²155. This information was initially organized into a Venn diagram to depict the distribution of these families among phages that could infect one or more of the hosts (Figure 1). Somewhat surprisingly, the mycobacteriophages that were capable of infecting *M. smegmatis* strain Jucho were the most distinct; 55% of the protein families found in phages that infect Jucho are not found in phages that infect MKD8 or *M. tuberculosis*. Meanwhile, 21.4% and 0.6% of protein families found in phages infecting MKD8 and *M. tuberculosis* respectively are unique to those groups. This occurrence may be partially explained by the number of mycobacteriophages observed to infect each host: 14 phages infected *M. tuberculosis*, 23 infected MKD8, and 99 infected Jucho. Of these mycobacteriophages capable of successful infection, one was observed to infect both *M. tuberculosis* and MKD8, nine were observed to infect both Jucho and MKD8, 11 were observed to infect both *M. tuberculosis* and Jucho, and one phage was capable of infecting all three hosts. A total of 88 phages were incapable of infecting any of the three hosts other than *M. smegmatis* strain mc²155, the host on which they were originally isolated.
Figure 1. The number of protein families in phages infecting each of three hosts is shown as a Venn diagram. Two families are found only in genomes that infect the human pathogen *M. tuberculosis*.

The considerably higher rate at which the sample of tested mycobacteriophages were able to infect Jucho compared to MKD8 and *M. tuberculosis* may be a result of significant diversity within the host *Mycobacterium* species that were utilized. While a sequenced genome is currently unavailable for *M. smegmatis* strain Jucho, *M. smegmatis* strain MKD8 has been sequenced and studied. MKD8 appears to be significantly structurally different from *M. smegmatis* mc²155: it lacks a 55.2kb genome duplication present in mc²155 and roughly 1.6% of the genome consists of single-nucleotide polymorphisms, or SNPs (Gray *et al.*, 2013). A total of 649 insertions and deletions greater than 19 bp in length have also been observed. Subsets of these insertions and deletions as well as the SNPs present in the genome may be responsible for the phenotypic differences observed within these strains of *M. smegmatis*.

Upon analysis of the genome sequences of those 204 phages included in the host range study, pham 982 was identified as a protein phamily of particular interest (Figure 2). Pham 982 is conserved in only five of the 654 mycobacteriophages genomes that have
been sequenced to date, and each of the phages containing this protein phamily belong to subcluster A2. Only two of the five phages have had their host range on *M. tuberculosis*, Jucho, and MKD8 tested, but both phages are able to successfully infect Jucho and *M. tuberculosis*. These two phages, D29 and L5, represent two of the three total A2 phages known to infect *M. tuberculosis*. The third phage, Turbido, lacks pham 982. From this initial analysis, it seems that pham 982 may correlate with the ability of phages L5 and D29 to infect *M. tuberculosis*. HHpred analysis returns a restriction endonuclease as the closest match to the amino acid sequence of pham 982, but with an e-value of 6.6, it remains unclear what the exact function of this protein is.

![Genomic map representation of subcluster A2 mycobacteriophages](image)

**Figure 2.** Genomic map representation of subcluster A2 mycobacteriophages whose efficiencies of plating on *M. tuberculosis*, Jucho, and MKD8 were tested. Pham 982 is conserved in only 5 phages, including both L5 and D29, which successfully infect Jucho and *M. tuberculosis* with an efficiency of plating within one order of magnitude relative to *M. smegmatis* strain mc²155.
A recently sequenced and annotated mycobacteriophage, Rover14, shares considerable sequence homology with Cluster G phage Angel. Angel has been observed to infect both Jucho and *M. tuberculosis* at efficiencies of plating comparable to mc2155 and is the only Cluster G phage known to infect *M. tuberculosis*. Of the three hosts that were tested, closely related Cluster G phage Halo is only observed to infect Jucho, forming plaques at an efficiency of plating of 1.7 relative to mc2155. When Halo is plated on *M. tuberculosis*, it forms plaques at an efficiency 6.0x10^{-4} lower than on mc2155. However, if plaques are picked from these *M. tuberculosis* plates, harvested and subsequently plated onto mc2155 and *M. tuberculosis*, equivalent titers are observed on both mc2155 and *M. tuberculosis*. The entire genome of the Halo expanded host range mutant was sequenced, and a single non-silent mutation in putative minor tail protein gp22 was identified. This mutation substitutes an alanine residue at position 604 with a glutamic acid residue (Jacobs-Sera et al., 2012). Angel, which infects *M. tuberculosis* at high efficiency, has an alanine at position 604, suggesting that the glutamic acid residue at this position is not an absolute requirements for infecting *M. tuberculosis* (Figure 3). Interestingly, the homologous gene in Rover14 contains a glutamic acid residue at position 604, just as the mutant Halo phage with enhanced host range does. This observation suggests that like the Halo mutant, Rover14 will infect *M. tuberculosis* at high efficiency.
Figure 3. Clustal Omega amino acid alignment of gp 22 in cluster G mycobacteriophages Rover14, Angel, and wild-type Halo. Rover14 contains a glutamic acid residue at position 604, while Angel and wild-type Halo contain an alanine residue.

Utilizing the view by host range data function in Phamerator, comparisons of protein phamily composition between phages belonging to the same cluster or subcluster may be advanced. Upon generation of a genomic map of subcluster A2 mycobacteriophages, an interesting variation in these phages with high nucleotide similarity is observed. While many of the subcluster A2 phages contain the pham 7269, there seem to be two distinct locations at which this protein phamily is found (Figure 4). Certain mycobacteriophages, such as Trixie, EagleEye, Pukovnik, and RedRock, encode this gene product close to the left end of the genome. Other mycobacteriophages, like Odin, L5, Che12, and D29 encode pham 7269 toward the center of their genomes. Turbido is the only subcluster A2 mycobacteriophage on this map lacking this protein pham; this phage is also the only tested subcluster A2 mycobacteriophage found not to infect the bacterial host Jucho at an efficiency of plating within one order of magnitude of mc²155. As a result, each instance of pham 7269 is color-coded in green to represent its conservation in phages within the same cluster or subcluster capable of infecting Jucho and lack of conservation in those phages that were not found to infect this host. It is unclear what effect, if any, the positioning of this gene may have on its function, but it is
interesting to note that of the four mycobacteriophages with pham 7269 located towards the middle of their genomes, two of these phages are capable of infecting *M. tuberculosis*.

**Figure 4.** Genomic map representation of subcluster A2 mycobacteriophages. Pham 7269 is conserved in eight of the nine phages represented here, with some phages coding for this protein phamily toward the left end of the genome and others toward the middle.

Within this same subcluster, it seems significant to note that only two of the nine mycobacteriophages representing subcluster A2 here contain pham 3838: Trixie and Turbidio. Of the 204 mycobacteriophages tested in the host range study on hosts *M.*
tuberculosis, MKD8, and Juho, Trixie and Turbido were the only two phages belonging to subcluster A2 determined to infect *M. smegmatis* strain MKD8. When the host range data is displayed on the genome maps of these bacteriophages, pham 3838 is color-coded in dark blue to represent the conservation of this gene product in subcluster A2 phages capable of infecting MKD8 and non-conservation in those incapable of infection (Figure 5). This phamily is represented in only 18 mycobacteriophage genomes to date.

**Figure 5.** Subcluster A2 mycobacteriophages Trixie (top) and Turbido (bottom) are the only two tested mycobacteriophages known to infect *M. smegmatis* strain MKD8. These two mycobacteriophages share pham 3838 in common while the other A2 phages represented lack it. As a result, this phamily is color-coded dark blue.

Of the three subcluster L1 mycobacteriophages tested in the host range study, only two were capable of infecting *M. smegmatis* strain MKD8 at a plating efficiency comparable to mc2155. When analyzing the genomes of these three mycobacteriophages, JoeDirt, LeBron, and UPIE, pham 3747 was of particular interest (Figure 6). While the phages capable of infecting MKD8, JoeDirt and LeBron, both contain this phamily, this
gene product is deleted in the otherwise closely related genome of UPIE, the phage incapable of infecting MKD8. Interestingly, JoeDirt and LeBron are the only two sequenced mycobacteriophages that contain pham 3747. This protein phamily, because of its conservation in the only subcluster L1 mycobacteriophages known to infect MKD8, should be investigated and studied further to determine if it plays a role in the specificity of these phages for this host.

Figure 6. Subcluster L1 mycobacteriophages with known host range on *M. tuberculosis* and *M. smegmatis* strains MKD8 and Jucho. Both JoeDirt (top) and LeBron (middle) are capable of infecting MKD8, while UPIE (bottom) lacks this capability. Pham 3747 is only known to exist in subcluster L1 mycobacteriophages JoeDirt and LeBron and is therefore color-coded dark blue.
Discussion

A number of factors can influence the susceptibility of a given bacterial strain to infection with a particular phage. Well characterized examples include restriction endonucleases which can degrade the genomic DNA of phages upon its injection into the cell, and toxin/anti-toxin systems. The more recently discovered CRISPR/cas system has been shown to function as a form of adaptive immunity, providing yet another barrier that phages must surmount in order to carry out a productive infection. However, it is likely that other systems also exist to protect bacterial cells from infection, and likewise that pathways that circumvent these systems remain undiscovered in phage genomes. Thus, it is critical to explore the correlation of protein phamilies in the sequenced mycobacteriophages with the host range of those phages.

While all currently sequenced and annotated mycobacteriophages, as members of the order Caudovirales, share structural similarity in the form of a flexible tail and dsDNA genome contained within an icosahedral head, much diversity exists between these phages (Fokine and Rossmann, 2014). The mosaic genomic organization of these phages, as well as the potential evolutionary history, may be observed utilizing Phamerator. Previously, an efficient means of displaying host range data simultaneously with genomic structure did not exist. This software serves to combine host range data compiled in the laboratory with genome arrangement in order to determine a correlation between the two. Color-coding conserved protein phamilies between related phages capable of infection of a particular host allows for the investigation of those specific
proteins and the potential role that they play in phage infection and host specificity. Considering the relatively high percentage of mycobacteriophage gene products with no known function, potential elucidation of key factors involved in host range determination is an exciting prospect.

A better understanding of this large and diverse population has real world implications. Although bacteriophages may not currently be an integral part of medical care for humans, they have already been implemented in several other fields. Phages are utilized as disinfectants in meat-packaging plants and have even been introduced into veterinary medicine (Brovko et al., 2012). Mycobacteriophages in particular provide opportunities for advancement and acceleration of diagnostics for the typically slow-growing genus of bacteria that they infect; diagnosis of mycobacterial infections and determination of antibiotic-resistant strains of pathogenic bacteria is greatly expedited using mycobacteriophages tagged with EGFP (Rondon et al., 2011).


7. Fokine, A., Rossmann, M.G. 2014. Molecular architecture of tailed double-


