Fall 2015

The evolutionary selective pressures exerted on A3 actinobacteriophages

Cheyenne Weeks-Galindo
James Madison University

The evolutionary selective pressures exerted on A3 actinobacteriophages

Cheyenne Weeks-Galindo

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Science

Biology

December 2015

FACULTY COMMITTEE:

Committee Chair: Steven Cresawn

Committee Members/Readers:

Reid Harris

James Herrick

Acknowledgments

Initial thanks go to my adviser Dr. Steve Cresawn for welcoming me into his research lab. I am appreciative for his patience, knowledge, optimism, and unyielding support. I am also grateful to Dr. James Herrick and Dr. Reid Harris for being members of my committee. I will always remember my early conversations with them about microbial ecology.

Giving thanks to the department for allowing me the chance to earn a master's degree and the professors I had the pleasure of learning from, I leave with a greater depth of knowledge in biology, research, and teaching; and with ambitious goals for my future. I declare a special thanks to Dr. Terrie Rife for not only her inspiring teaching style but also the countless times she helped me when questions arose while conducting research. Apart from the academic support I received from the department, Dr. Patrice Ludwig and Dr. Sharon Babcock provided a space for me to express my non-academic concerns about my time in the program and responded with encouragement and compassion. My friends and family have provided endless support both academically and emotionally. I am especially grateful to my partner Lucas Painter for staying by my side through the most difficult of times and always believing in my abilities and potential.

I lastly would like to acknowledge Dr. Graham Hatfull and the SEA-PHAGES program for contributing the phages, Dan Russell and Charlie Bowman for the sequencing, Dr. Welkin Pope for the annotation oversight, and the HHMI for funding the work.

Table of Contents

List of Tables

List of Figures

Abstract

This study identified evolutionary selective pressures within subcluster A3 actinobacteriophages. These phages are able to infect the clinically important genus *Mycobacterium*. Understanding the selective pressures on genes in these phage genomes is a step toward understanding the adaptations that result from short-term and long-term associations of phages and bacteria that have been co-evolving for perhaps billions of years. In this study 149 phamilies (phage protein families) of homologous gene sequences were analyzed using Datamonkey. Complete data were obtained for 57 phamilies. Of these, eleven phamilies were affected by recombination, three showed evidence of predominantly diversifying selection, and twenty-four have a function. In the near future, a study investigating the protein structure of qualified phamilies (those with ≥10 sequences in the analysis) would provide further insight into the selection identified in this study. As more actinobacteriophage genomes are sequenced and annotated, analysis with Datamonkey should be repeated with larger and more diverse alignments.

I. INTRODUCTION

This study investigates evolution at the genomic level in a group of bacteriophages. Bacteriophages, phages for short, are viruses that infect bacteria; virus indicates an obligate intracellular parasite. There are an estimated 10^{31} phage particles in the biosphere (Wommack and Colwell, 2000), making them the most abundant of all biological entities. Bacteriophages are highly dynamic with an estimated 10^{23} phage infections occurring globally every second (Suttle, 2007). A 2011 study suggests that the global phage population has been evolving for two to four billion years (Pope et al. 2011), resulting in a genetically diverse population that may represent the greatest genetic reservoir in the biosphere (Hatfull, 2008).

Bacteriophages are key biotic drivers of microbial diversification (Rodriguez-Valera et al. 2009 and Clokie et al. 2011). Whether the study is on the microbial communities of soil (Gómez et al. 2011, Griffiths et al. 2011), leaves (Koskella et al. 2011, Lindow and Brandl 2003), the ocean (Marston et al. 2012, Pommier et al. 2012), or the human body (Smillie et al. 2011, Hooper, et al. 2012), each reveals the importance of the local ecology in driving microbial diversification (Koskella and Meaden, 2013).

As components of microenvironments, phages exert important pressures on bacteria and consequently on the larger environment. Phages have been shown to alter competition among bacteria (Bohannan and Lenksi, 2006, Joo et al., 2006, and Koskella et al., 2012), maintain bacterial diversity (Buckling and Rainey, 2002, and Rodriguez-Valera et al., 2009), and mediate horizontal gene transfer among bacteria (Kidambi et al., 1994, Canchaya et al., 2003). Phages carry out these roles by utilizing two different

lifecycles, lytic and lysogenic, classifying them into two main categories, lytic phages and temperate phages.

The lytic growth cycle consists of a series of sequential steps beginning with adsorption of the phage to a host cell receptor and progressing through injection of the linear viral genomic DNA into the cytoplasm, viral genome circularization, viral genome replication and protein synthesis, assembly of new virus particles, and packaging of viral DNA. The pathway culminates with lysis of the infected cell and release of the newly synthesized viral particles. The temperate cycle includes each of these steps, but also includes a period of dormancy following genome circularization wherein few viral genes are thought to be expressed. The dormant period is of indeterminate length and is typically characterized by site-specific integration of the viral genomic DNA into the host chromosome.

The lifecycles begin in similar fashion and then diverge after circularization of the viral genome. The bacteriophage infection cycles begin with the adsorption of the phage to a bacterial cell, followed by injection of the viral genome into the cytoplasm of the host bacteria. Once the phage genome has entered the host cell it circularizes to avoid degradation by bacterial exonucleases. After circularization of the viral genome the virus genome will either enter a lytic or lysogenic cycle. The lytic cycle will proceed by replicating its DNA and transcribing messenger RNA for the synthesis of viral proteins, shortly thereafter killing the host. The lysogenic cycle will initiate integration of the viral genome into the host genome or the viral genome will exist separately, without causing lysis of the bacterium (figure 1).

Lytic and lysogenic lifecycles have differing effects on the bacterial community. Lytic phages can cause the rapid decline of a specific population of bacteria; in so doing, bacterial diversity may be maintained and/or competition between bacteria may be influenced. Infection with a temperate phage forms a lysogen. A lysogen is a bacterial cell that harbors a prophage. Lysogens form ostensibly stable relationships between the host and the phage (Levin and Lensky, 1983) and through this relationship a phage is able to confer genetic material that benefits its host. This phenomenon is known as lysogenic conversion. Lysogenic conversion can confer a variety of phenotypes to the host, including increased fitness of the lysogen, improved pathogenicity, or resistance to antibiotics (Waldor and Mekalanos, 1996; Brüssow et al., 2004).

The dynamic interaction between phage and bacterium suggests an important process of co-evolution between the organisms. Co-evolution is defined as the process of reciprocal adaptation and counter-adaptation between interacting species (Janzen, 1980). Bacteria can evolve to resist phage infection by *de novo* mutation and have evolved other mechanisms to avoid infection (Koskella and Brockhurst, 2014). Mechanisms include restriction-modification (Hyman and Abedon, 2010), which results in the death of the phage and the survival of the bacterium (Hyman and Abedon, 2010) and resistance to adsorption of the phage to the bacterium, resulting in reduced interaction between the phage and the bacterium (Hyman and Abedon, 2010). Despite the evolution of bacteria resistant to phage infection, a 2013 study demonstrated that phage lineages persist over time (2.5 years), noting high substitution rates in lytic phage genomes (Minot et al., 2013). This type of interaction can be summarized as phage-bacteria systems

experiencing fluctuating associations over short time spans (days) and stabilized relationships over long time periods (years) (Needham et al., 2013).

The bacteriophage-bacteria system is dependent on the ability of a phage to successfully infect a bacterium. Ways in which infection is avoided may be by preventing adsorption of the phage to the bacterium or by preventing the phage from taking over host machinery (e.g. modification of bacterial ribosomes). Adsorption to the host is dependent on bacteriophage tail fibers. Tail fibers are fibrous proteins located distally from the head of the bacteriophage that form non-covalent bonds with receptors on the surface of the bacterial host (figure 2). A 2012 paper investigating bacteriophage host preference indicated that mutations in the tail fiber genes enabled phages to overcome host barriers to infection (Jacobs-Sera et al., 2012). In 2009, Vos et al. concluded that bacteriophages undergo adaptations in response to the local microbial environment. Bacteria and bacteriophages were isolated from soil (Vos et al., 2009). The bacteria were identified and colonized to yield clones and the clones were infected with locally found bacteriophages or with bacteriophages found in an unrelated soil sample (Vos et al., 2009). The foreign bacteriophages were 9% less effective in infecting the bacteria than the local phages (Vos et al., 2009).

The bacteriophages in this study come predominantly from one bacterial host. *Mycobacterium smegmatis*, a member of the phylum *Actinobacteria*. All bacteriophages in this study can therefore be considered actinobacteriophages (phages infecting bacteria in the phylum *Actinobacteria*) or mycobacteriophages (phages infecting the genus *Mycobacterium*). *Actinobacteria* can be readily found in both terrestrial and aquatic environments, with a notable presence in soil for their nutrient recycling (Stach and Bull,

2005; Goodfellow and Williams, 1983). This phylum is considered to be comprised of one of the largest taxa among the 18 major lineages recognized within the domain *Bacteria* (Stackebrandt et al., 1997). Bacteria in this phylum tend to have high GC content in their genomes, they exhibit a variety of morphologies, including coccoid, rod-coccoid, hyphal forms, and branched mycelium (Ventura et al., 2007). Some members of the phylum produce antibiotics in the form of secondary metabolites (Lechevalier and Lechevalier, 1967). The phylum also includes various pathogens, notably the causative agents of the human diseases tuberculosis, leprosy, and Buruli ulcer.

Mycobacterium tuberculosis, *Mycobacterium leprae*, and *Mycobacterium ulcerans* cause tuberculosis disease, leprosy or Hansen's disease, and Buruli ulcer disease, respectively. Mycobacteria are aerobic, non-motile rods characterized by their complex lipid-rich cell walls which stain acid-fast due to their ability to resist de-staining (Pfyffer, 2007). In addition to the well-known human pathogens, this genus contains many environmental species and nontuberculoid pathogenic species (September et al., 2004).

In order to investigate the counter adaptation made by bacteriophages in response to the evolution of bacteria resistant to phage infection, this study determines sites of diversifying and purifying selection in the genomes of select mycobacteriophages. Diversifying or purifying selection can be distinguished from neutral mutation by constructing a sequence alignment and calculating the ratio of non-synonymous (dN) changes per non-synonymous site to synonymous changes per synonymous site (dS). When dN differs significantly from dS the selection is non-neutral. The results of this study may provide important insights both for research investigating the clinical applications of mycobacteriophages and research investigating microbial ecology.

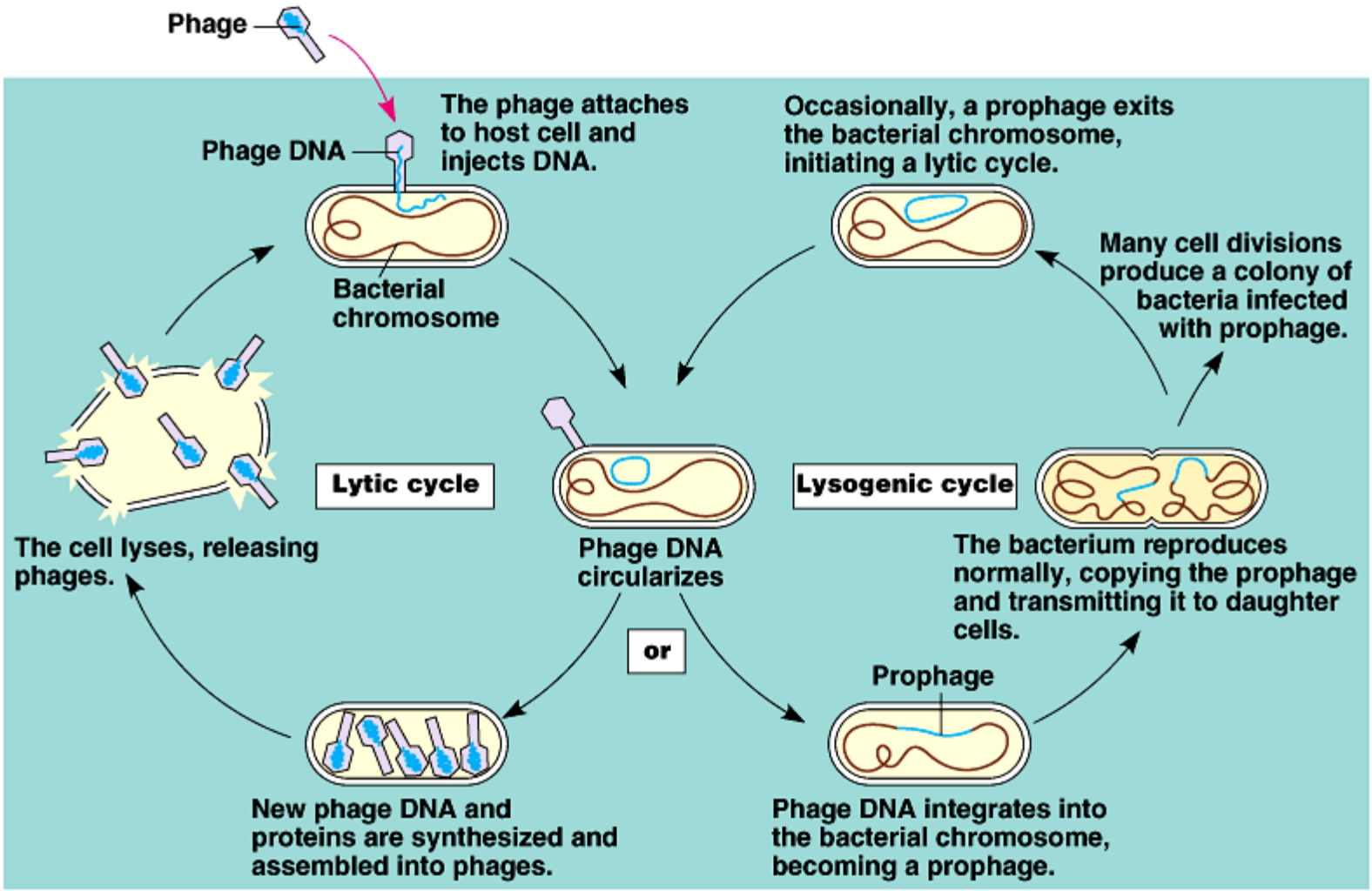

Figure 1. The lytic and lysogenic lifecycles of bacteriophages are demonstrated.

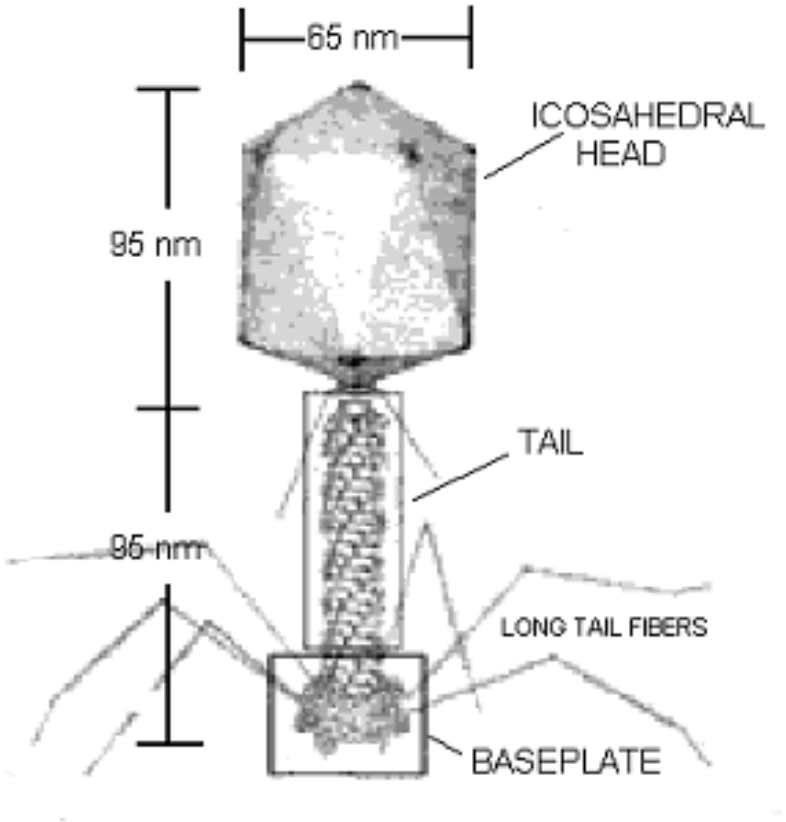

Figure 2. A bacteriophage schematic based on the T4 phage which infects *Escherichia coli*. Importantly, the tail structure of the phages in this study differ from the tail of the T4 bacteriophage. Source: http://www.science.marshall.edu/straitho/sem_projects/spring_01/Aoune%20s%20folder/structural_studies_of_t4.htm

II. BACKGROUND

The isolation of bacteriophages capable of infecting of *M. smegmatis* dates back to the 1940's (Hatfull, 2010). By 1954 a paper was published on a phage active against *M. tuberculosis* (Froman et al., 1954). In 1987, the phage TM4 was isolated from *Mycobacterium avium* and used to construct shuttle phasmids. TM4 replicates as a plasmid in the bacterium *Escherichia coli* and as a phage in mycobacteria (Jacobs et al., 1987). Through the efforts of various educational platforms ranging from the high school to the university level, 876 mycobacteriophage genomes have been completely sequenced (Hanauer et al., 2006; Hatfull et al., 2006; Hatfull, 2012; Hatfull, 2013; and phagesdb.org). The majority of the sequenced genomes came from phages that were isolated from environmental soil samples using the host *M. smegmatis* mc^2155 (Hatfull, 2014), however some phages were isolated from stool samples of tuberculosis patients (Carter and Redmond, 1963).

Current literature offering comparative data on mycobacteriophage genomes reveals vast genetic diversity. Mycobacteriophage genomes can be classified into clusters or as singletons based on genome sequence similarity. Clusters are formed when the phage genomes share at least 50% nucleotide identity across the entire length of their genomes, singletons are phages that do not fit into any existing cluster (Cresawn et al., 2011). Clusters may be regrouped into subclusters when phages within a cluster have similar gene content and/or genome organization (Jacobs-Sera et al. 2012). For example, cluster A actinobacteriophages are regrouped into 17 subclusters. Currently, there are 30 distinct types of mycobacteriophage genomes formed by clusters or singletons which share little or no nucleotide identity to each other (Hatfull, 2014).

The formation of a subcluster is partly a function of the hallmark mosaic architecture of the bacteriophage genome, in that phage genomes can be thought of as being comprised of specific groupings of interchangeable modules, where each module may be present in one or more genomes (Pedulla et al., 2003; Hendrix et al., 1999). The mosaic patterns can be understood by various evolutionary processes, point mutations or nucleotide substitutions, homologous recombination aided by conserved boundary sequences (Susskind and Botstein 1978; Clark et al., 2001), and illegitimate or non-homologous recombination (Pedulla et al., 2003; Hendrix et al., 1999). In this last process the recombination observed is a result of selection for gene function (Hendrix, 2003). Therefore, genomes within a subcluster share patterns of mosaicism.

Recombination may be driven by homologous gene sequences, though it can also occur at non-homologous sites. Bacteriophage mosaicism is most evident in amino acid sequences. For example, when comparing two genomes, genes that code for homologous proteins may have little nucleotide conservation (figure 3). Recombination, particularly when non-homologous, often leads to deleterious mutations resulting in loss of function. In bacteriophages illegitimate recombination is commonly seen.

Some generalizations can be made based on current comparative genomics of bacteriophages. Typically, the structural genes are encoded in the left end of the genome and small genes of unknown function are encoded in the right end of the genome. The middle of the genome is generally a mix of enzymes and small genes of unknown function. Genome length is determined in part based on physical packaging constraints imposed by the interior volume of the viral capsid. The presence of some genes within

the genome may simply be based on these constraints. If so, these genes will exist under vastly different selective pressures than essential structural genes.

This study hones in on the mycobacteriophage genomes that make up subcluster A3. Cluster A the largest and one of the most diverse genome clusters, and some of the mycobacteriophages within it have been previously characterized and used in host range studies (table 1.) The study compared the plating efficiency, or infectivity, of various A3 phages on *M. smegmatis* strains Jucho and MKD8 which are distinct from stain mc^2155 and a strain of *M. tuberculosis* (Jacobs-Sera et al., 2012). Plating efficiency is relative to that on *M. smegmatis* mc^2155, on which the viruses were propagated (Jacobs-Sera et al., 2012). This host range data was particularly interesting because of the variability of the plating efficiency between the phages and the ability of some phages to infect *M. tuberculosis* mc^27000. At the time of analysis, there were 17 actinobacteriophages in subcluster A3, currently there are 57. This is a testament to the growing actinobacteriophage genome sequence dataset. A3 was chosen because of the existing host range data, the manageable size of the subcluster, and the size and diversity of cluster A.

Table 1. Plating efficiencies of subcluster A3 bacteriophages on *M. smegmatis* strains MKD8 and Jucho and *M. tuberculosis* $mc^2$7000. When the plating efficiency is one, it is equal to that on *M. smegmatis* $mc^2$155, greater than one means greater infectivity and less than one means diminished infectivity. When a less than sign is present the plating efficiency was under the level of detection. Adapted from Jacobs-Sera et al., 2012.

Phage	Cluster	*M. smegmatis* strain Jucho	*M. smegmatis* strain MKD8	*M. tuberculosis* strain $mc^2$7000
Bxz2	A3	1.0	$< 10^{-6}$	1.5
HelDan	A3	7.7×10^{-2}	$< 10^{-7}$	$< 10^{-7}$
JHC117	A3	3.3×10^{-6}	$< 10^{-9}$	$< 10^{-9}$
Microwolf	A3	1.0	$< 10^{-4}$	2.2
Rockstar	A3	1.0	1.0	5.0
Vix	A3	1.0	$< 10^{-6}$	4.5×10^{-1}

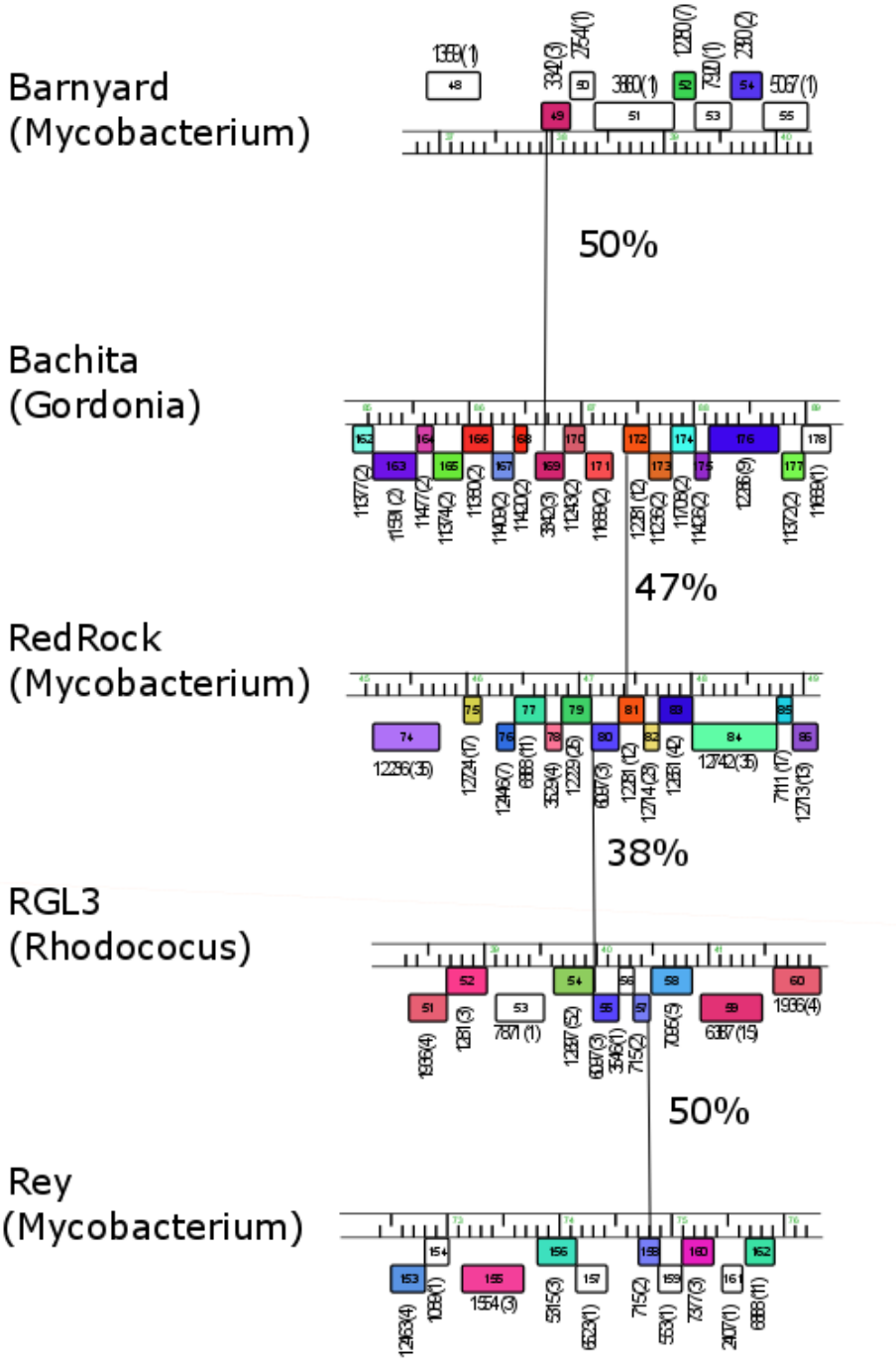

Figure 3. Genome segments of five actinobacteriophages isolated from different genera, genus in parentheses. Neighboring genome segments share a protein phamily in common, indicated by vertical lines. The total amino acid identity shared between proteins is shown as a percentage. There is no nucleotide conservation between the genomes indicating a case of illegitimate recombination. Adapted form Hatfull 2015.

The estimation of synonymous (dS) and non-synonymous (dN) nucleotide substitution rates is accepted as a standard tool to study evolution at the molecular level (Pond et al., 2006). The ratio of dN/dS is accepted as a measure of selective pressure (Nielsen and Yang, 1998 and Yang et al. 2000). Datamonkey determines the difference (dN-dS). Point mutations and recombination are important drivers of genetic variation and adaptation. dN substitutions change the primary amino acid sequence and can affect protein function. dS substitutions, conversely, do not change the amino acid sequence and are considered to be of minimal consequence to the organism. Datamonkey (http://www.datamonkey.org) (Pond and Frost, 2005), a web-based resource, was used to compute rates of synonymous and non-synonymous substitutions and infer purifying and diversifying selection in this study.

Selection can be neutral or non-neutral. When the rate of dN differs significantly from dS there is convincing evidence for non-neutral evolution (Pond et al., 2006). When the rate of dS is larger than dN there is evidence for purifying selection and diversifying selection is inferred when the rate of dN is greater than that of dS. Datamonkey can estimate substitution rates in the presence of recombination (Poon et al. 2009). This capability is essential for analyzing bacteriophage genes due to the mosaic architecture of their genomes.

Datamonkey was developed to detect and quantify the evolutionary pressures that contribute to genetic variation (Yang et al., 2002 and Pond and Frost, 2004). Prior to Datamonkey, the two predominant methods used to quantify evolutionary pressures were a likelihood-based approach (Nielsen and Yang, 1998) and a parsimony-based counting method (Suzuki and Gojobori, 1999). Datamonkey is unique in that it offers algorithms

that integrate these approaches (Kosakovsky Pond and Frost, 2005) and was designed by a group of researchers studying viral evolution.

Sequence alignments that are affected by recombination present obstacles when analyzed with traditional algorithms to detect dS and dN substitution rates. Recombination can interfere with the construction of a phylogenetic lineage (Posada and Crandall, 2001) and distort inferred substitution rates (Schierup and Hein 2000). Datamonkey manages this limitation with genetic algorithm recombination detection (GARD) (Pond et al. 2006a). The GARD model searches an alignment for putative points of recombination and quantifies the level of significance at each break (Pond et al., 2006a). A breakpoint serves as an indicator to construct a new phylogenetic tree and GARD produces multiple phylogenies that model the evolution of non-recombinant fragments (Pond et al. 2006a).

Datamonkey offers three basic models that carry out site-specific inferences, single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and random effects likelihood (REL) (Pond et al., 2006b). While these three models use different statistical methods to quantify diversifying or purifying selective pressures, they are considered to be comparable and not significantly different (Pond et al., 2006b). However, SLAC is considered to be more conservative than REL or FEL (Poon et al., 2009). When one of these models infers that there is positive selection, it assumes that the rate of substitution is constant over evolutionary time (Pond et al., 2006). Alignment size can also affect inferred substitution rates.

Positive selection is more readily identified in smaller alignments than in larger ones (Murrell et al., 2012; Yokoyama et al., 2008; Chen and Sun, 2011). When additional

sequences are added to an alignment and positive selection disappears, due to a site reverting later in the phylogeny, the sequences experiencing positive selection are lost in SLAC, FEL, or REL models. Datamonkey offers another model that can be used to identify episodes of positive selection in specific branches of a phylogeny. This model is the mixed effects model of evolution (MEME) (Murrell et al., 2012).

MEME allows for the substitution rate to vary throughout evolutionary history (Murrell et al., 2012). Similarly, internal fixed effects likelihood (IFEL) (Pond et al., 2006) captures selective pressures that occur in internal branches as opposed to external branches, in other words, positive selection that may disappear in a larger alignment (Pond et al., 2006; Poon et al., 2009). IFEL was developed to identify positive selection that increased the fitness of HIV-1 in a specific host population, but was lost in a larger more diverse HIV-1 sample (Pond et al., 2006). Therefore, IFEL may be useful in identifying host range determinants in A3 phages. Positive selection that is observed in episodic events (MEME) or only in internal branches (IFEL) may help explain how plating efficiencies vary between related mycobacteriophages.

Based on what is known about bacteriophage genomes there are some predictions that can be made regarding recombination and selective pressures on specific genes. The evolutionary events that can affect DNA sequence can be thought as a spectrum from minimal change to drastic change, where purifying selection is minimal, recombination leads to drastic change, and diversifying selection is in the middle. It is predicted that necessary genes like genes that code for structural proteins or enzymes involved with virus particle assembly would be at the minimal change end of the spectrum and be under purifying selection. Genes that are considered to be tolerant to change code for tail fiber

proteins, enzymes that overcome host restriction, and enzymes for lysogeny and may therefore be under diversifying selection. At the far end of the spectrum genes that may be affected by recombination are the small genes of unknown function that are present at high density at the right end of the genomes. This region of the genomes varies considerably among otherwise closely related phages, implying that there is minimal selection against drastic changes in these genes.

II. Methods

Sequences of homologous genes, known as phamilies, were obtained from Phamerator (Cresawn et al., 2011) and exported in FastA format. Phamerator serves as a database and a tool for comparative bacteriophage genomics. The Phamerator database used to obtain sequence data was Actinobacteriophage_554. Phamilies used in the study can be seen in table 2. Before uploading phamilies to Datamonkey, the sequences were first aligned using webPRANK (Löytynoja and Goldman, 2010, http://www.ebi.ac.uk/goldman-srv/webprank/). webPRANK is an online server that allows a user to align sequences with a phylogeny-aware alignment algorithm (Löytynoja and Goldman, 2010). A phylogeny-aware alignment was necessary to avoid errors in the downstream Datamonkey analysis. Common alignment programs do not consider phylogeny in their placement of gaps and fail to distinguish insertion and deletion events (Wong et al., 2008; Löytynoja and Goldman, 2008). Therefore, the webPRANK alignment accommodated gaps, insertions, and deletions better than other alignment algorithms. Phage phamily sequences were uploaded to webPRANK and the "align translated codons" option was chosen. All other options were unchanged from the default.

Datamonkey (Delport et al., 2010; Pond and Frost, 2005, http://www.datamonkey.org) was used to determine the dN and dS substitution rates (Poon et al. 2009). Datamonkey computes user uploaded sequences through models, using the HyPhy package (Pond et al. 2005) as its computational engine. GARD was always the first model used. GARD determined the presence of recombination. When recombination was present GARD phylogenetic trees were used in downstream models,

when no recombination was found a neighbor-joining phylogenetic tree was used. After selecting an appropriate tree, the following models were used in no particular order, SLAC, FEL, IFEL, REL, and MEME. Datamonkey offered various data formats including html summaries, .csv files, and plots.

IV. RESULTS

A total of 149 phamilies were exported from Phamerator and contained at least one bacteriophage member from subcluster A3. After data was processed, the phamilies were organized into seven distinct groups, no recombination (46 phamilies), significant recombination (11), insignificant recombination (53), no alignment (14), incomplete analysis (21), empty FastA file (2), and alignment error (2) (table 2). The groups are color coded with five colors (figure 4). Only two groups have complete data from Datamonkey, no recombination and significant recombination, for a total of 57 phamilies. For all other groups, limited data was obtained (92 phamilies).

Recombination without significance was the primary reason that a phamily was omitted from downstream analyses. Without significance, the exact point of recombination was uncertain and therefore GARD phylogenies could be inaccurate. There were 11 phamilies with significant breakpoints (table 2). Complete data that infers selection includes five of the Datamonkey models, SLAC, FEL, IFEL, REL, and MEME. Results could have been affected by alignment size, both in length of sequence and number of sequences. After data were uploaded to Datamonkey, duplicates were determined and all but one were removed from the analysis (Poon et al., 2009). According to the Datamonkey tutorial (Poon et al., 2009) a minimum of ten sequences should be used for SLAC, FEL, IFEL, and REL models. Due to the removal of duplicates some phamilies with ten or more sequences in the alignment had less than ten in the analysis (table 3).

The majority of phamilies showed evidence of both purifying and diversifying selection, phamily 2191 had no selection, and phamily 2981 showed only positive selection (table 4). Eighty-eight

Table 2. Phamilies categorized into seven groups based on alignment and the presence of recombination. In cases of insignificant recombination the GARD model identified recombination but it was not supported statistically. A phamily was categorized with no alignment when there was only one sequence and into incomplete analysis when there were only two sequences. The phamilies in empty FastA file had no sequences and the gene sequences in alignment were incomplete.

No recombination	Insignificant Recombination	Significant recombination	No Alignment	Incomplete analysis	Empty FastA file	Alignment error
41	115	565	355	315	12492	12490
115	340	1838	874	721	12548	12744
217	700	4481	1852	901		
387	1493	5065	2781	1594		
1613	2838	5218	4185	2465		
1661	2847	5238	5964	3014		
1670	3045	6333	6950	3243		
1706	3146	12269	9337	3425		
1807	3580	12412	10901	3826		
1827	4151	12479	12224	4800		
2018	4559	12742	12379	4831		
2153	4899		12426	6676		
2160	5594		12656	6692		
2191	6289			6825		
2589	6912			7537		
2981	7209			8899		
4045	7531			12264		
4048	7622			12267		
4075	7638			12484		

4832
7347
7832
9297
11075
11241
11785
11976
12148
12229
12304
12305
12323
12340
12353
12356
12364
12396
12495
12526
12580
12581
12660
12669
12670
12690
12692

7655
7714
8536
9056
12109
12210
12233
12235
12245
12249
12286
12362
12363
12381
12402
12469
12473
12520
12533
12544
12550
12565
12602
12603
12652
12655
12659
12662

12722
12725

12675
12691
12714
12741
12761
12771

Table 3. Number of sequences used in the Datamonkey analysis. Datamonkey identified duplicate sequences and removed all but one from the analysis. All phamilies are in the order of which they appear in the genome of Bxz2, save the last ten phamilies which are not found in Bxz2. In the fourth column, we can verify if the alignment size reaches ten, the minimum number

Phamily	Number of Sequences in Alignment	Duplicate Sequences Identified by Datamonkey	Number of Sequences in Analysis	Number of Codon Sites in Analysis
11075	12	7	6	303
12412	16	8	9	792
4048	16	8	9	321
4481	16	4	13	1710
5218	16	7	10	318
12396	16	6	11	999
12580	12	8	5	525
12323	12	10	3	186
12495	16	8	9	303
1827	16	9	8	381
12356	16	9	8	483
11241	16	10	6	342
12269	14	8	7	330
2981	4	1	3	1569
2018	15	5	11	618
4832	12	7	6	144
12305	16	6	11	465
7347	16	8	9	261
1706	16	7	10	201

115	14	10	5	183
2160	13	8	6	189
12479	16	8	9	336
41	11	7	5	189
565	16	5	12	2283
12353	13	9	5	207
7832	12	8	5	183
9297	16	4	13	768
1661	12	6	7	165
12581	16	7	10	516
1807	13	10	4	105
12690	12	7	6	825
11976	16	5	12	444
11785	15	9	7	258
5065	16	8	9	810
2191	8	5	3	75
12692	12	6	7	309
2589	16	6	11	246
12670	16	11	6	270
5238	16	5	12	885
1670	12	6	7	456
12229	13	6	8	258
12364	16	8	9	930
12742	23	10	15	1512
217	11	6	6	216
12304	10	7	4	159
387	11	9	3	117
6333	16	9	8	162

1838	11	5	7	193
4045	3	0	3	366
4075	3	0	3	108
12148	3	0	3	162
12340	16	7	10	273
12526	3	0	3	129
12660	3	0	3	363
12669	12	7	6	180
1613	10	8	3	291
2153	4	0	4	186

phamilies are modeled on A3 phage Bxz2, they are color coded according to data group, and have a plus or minus sign above the gene (plus for diversifying or minus for purifying selection) when one selection type was predominant (figure 4). Of the 57 phamilies for which Datamonkey analysis was completed, 24 have a putative function (table 4).

Nine of the phamilies for which complete data was obtained (115, 217, 1661, 1706, 1807, 2153, 2160, 2191, and 4832) contain only proteins from subcluster A3 phages, however not all A3 phages encode proteins that are members of these phamilies. Twenty–four analyzed phamilies each contain proteins from all 17 A3 bacteriophages that were included in this study (table 4). Phamily 1706 is of particular interest because it includes a member from every A3 phage, and only A3 phages. In addition to phamily 1706, three other phamilies are discussed in more detail. Phamily 4481 contains all members of A3 and had the most sites undergoing purifying selection, pham 2981 was found to be under only diversifying selection, and pham 12396 was chosen because there were more than 10 sequences analyzed, it has a known function, and it demonstrated both diversifying and purifying selection.

Phamily 1706 is not affected by recombination and had ten sequences in the analysis (after removal of duplicates), 201 codon sites, and no known function. Gene 42 from phage Bxz2 is a member of phamily 1706 and is located in the middle of the genome (figure 4). There were two codon sites identified by MEME that are under diversifying selection, sites 10 and 28. Codon site 28 was identified by FEL and MEME. Tables that demonstrate the codon and amino acid substitutions of sites 10 and 28 can be

seen in tables 5 and 6, respectively. The amino acids of these sites can be seen in figures 5 and 6. There were ten sites found by the FEL model that

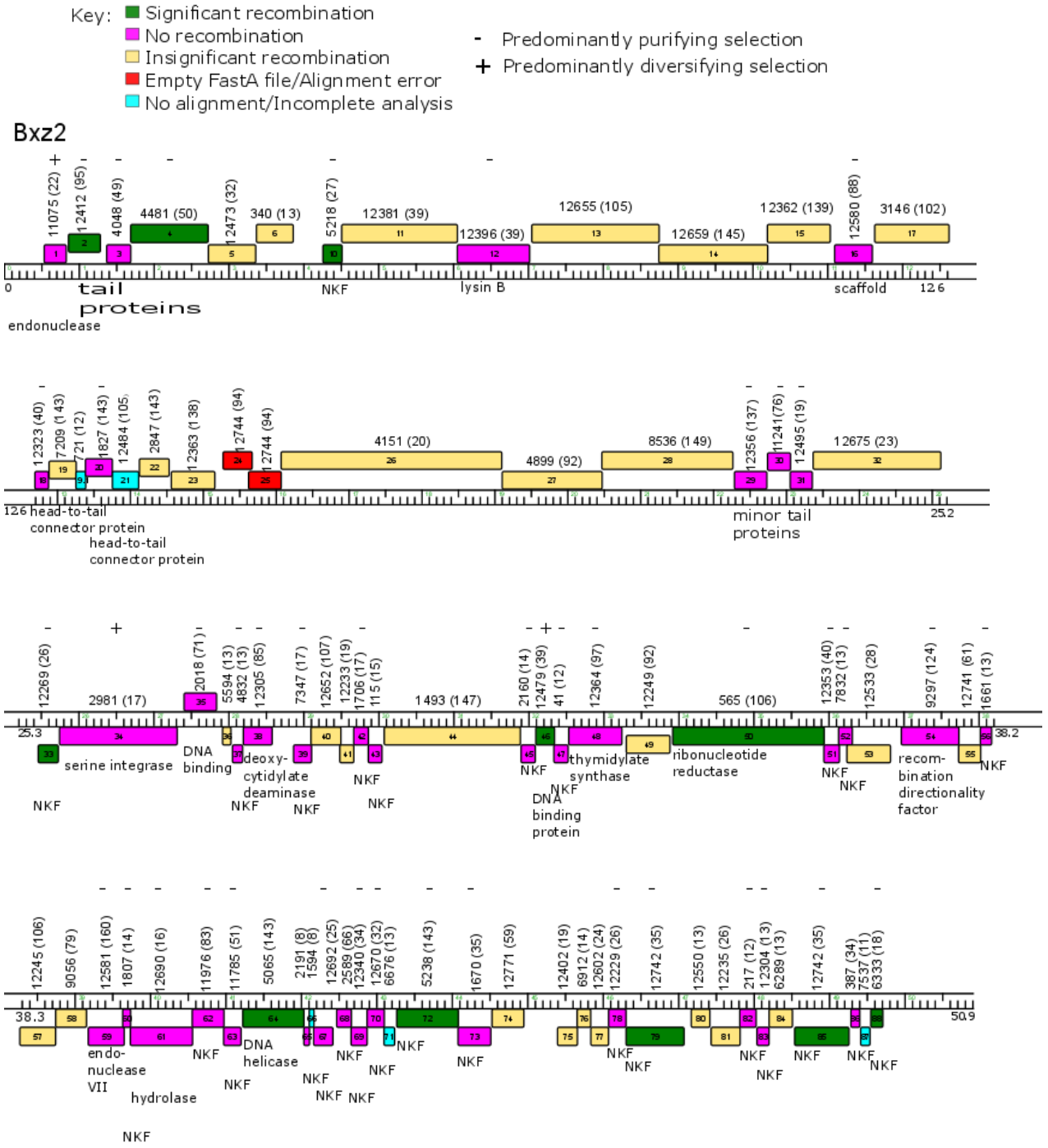

Figure 4. The genome of mycobacteriophage Bxz2 is presented in four pieces. The base pairs in kilo-bases are indicated at the ends of the genome segments. Boxes above or below the ruler represent genes coded on the top or bottom strand of DNA. The colors represent the group the phamily fell into in the initial findings. Bxz2 includes 88 of the 149 phamilies in the analysis. The predominant type of selection was determined by the greatest number of sites under either type of selection within a gene.

Table 4. Recombination, selection, and function of the 53 phamilies with complete data. The phamilies are ordered as in table 3. Putative function is acquired from Phamerator and was double checked using HHpred. Recombination and selection data was determined via Datamonkey. Positive and Negative Selection columns were obtained from the consensus report which did not include IFEL selection, thereby demonstrated in separate columns. The Overall column is determined by the ratio of the number of positive to negative sites of selection, when there were more positive sites the overall was considered positive, less positive was considered negative, and an equal ratio was considered equal.

Phamily	Function	Recombination	Sites under positive selection	Sites under negative selection	IFEL Positive	IFEL Negative	Overall
11075	endonuclease	NO	3	2	0	1	+
12412	tail protein	YES	3	47	0	33	-
4048	tail protein	NO	1	30	0	20	-
4481	collagen-like tail protein	YES	20	255	6	106	-
5218	NKF	YES	1	25	0	17	-
12396	lysin B	NO	2	73	0	63	-
12580	scaffold	NO	0	1	0	0	-
12323	head to tail connector protein	NO	0	1	0	0	-
12495	minor tail protein	NO	0	28	0	19	-
1827	head to tail connector protein	NO	0	40	0	29	-
12356	minor tail protein	NO	1	58	0	29	-
11241	minor tail protein	NO	1	48	0	18	-
12269	NKF	YES	3	22	1	12	-
2981	Ser integrase	NO	31	0	0	0	+
2018	DNA binding	NO	3	72	1	30	-
4832	NKF	NO	1	1	0	1	=
12305	dCMP deaminase	NO	2	47	0	37	-
7347	NKF	NO	2	21	0	20	-
1706	NKF	NO	2	10	0	6	-
115	NKF	NO	0	15	0	11	-

2160	NKF	NO	0	10	0	7	-
12479	HTH DNA binding	YES	0	30	0	16	-
41	NKF	NO	3	2	0	0	+
565	ribonucleotide reductase	YES	12	232	5	145	-
12353	NKF	NO	1	9	0	2	-
7832	NKF	NO	0	8	0	2	-
9297	recombination directionality factor	NO	3	71	0	51	-
1661	NKF	NO	0	5	0	0	-
12581	endonuclease VII	NO	0	43	1	35	-
1807	NKF	NO	1	4	0	0	-
12690	esterase/lipase	NO	0	27	0	14	-
11976	NKF	NO	3	42	1	32	-
11785	NKF	NO	2	17	1	13	-
5065	DNAB-like helicase	YES	0	101	1	50	-
2191	NKF	NO	0	0	0	0	0
12692	NKF	NO	0	4	0	1	-
2589	NKF	NO	3	14	0	12	-
12670	NKF	NO	2	20	0	9	-
5238	RecB	YES	2	74	0	58	-
1670	NKF	NO	5	9	0	0	-
12229	NKF	NO	1	16	0	1	-
12364	thymidylate synthase	NO	7	48	0	38	-
12742	NKF	YES	0	20	0	13	-
217	NKF	NO	0	19	0	2	-
12304	NKF	NO	0	2	0	0	-
387	NKF	NO	0	1	0	0	-
6333	NKF	YES	1	22	0	13	-
1613	NKF	NO	0	18	0	0	-
1838	DNA binding	YES	0	10	0	9	-
2153	NKF	NO	0	7	0	0	-
4045	NKF	NO	0	32	0	0	-
4075	NKF	NO	2	2	0	0	=

12148	NKF	NO	0	5	0	0	-
12340	NKF	NO	2	17	0	3	-
12526	NKF	NO	0	12	0	0	-
12660	recombination directionality factor	NO	0	8	0	0	-
12669	NKF	NO	2	18	1	8	-

Table 5. Codon site 10 in pham 1706. Under the column branch, a node number indicates an internal branch and a phage name indicates an external branch. The pink shading represents only non-synonymous substitutions. This table is composed of data from Datamonkey.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Node 4	GTG	ACG	Valine	Threonine	0	2
Phantastic	ACG	GTG	Threonine	Valine	0	2

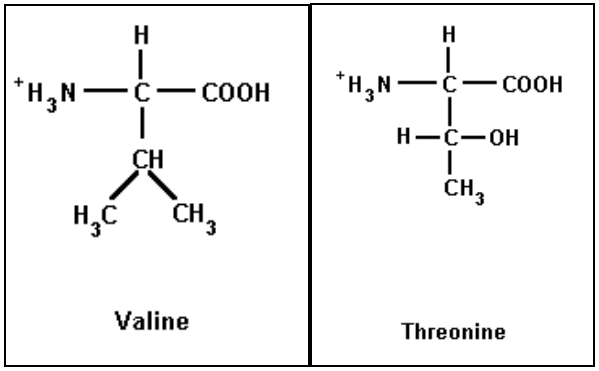

Figure 5. The amino acids valine and threonine were replace with each other after codon substitutions.

Table 6. Codon site 28 in pham 1706. Green shading represents only synonymous substitutions and orange shading represents mixed substitutions. All branch points are internal.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Node 3	GGA	GGG	Glycine	Glycine	1	0
Node 4	GGG	GGC	Glycine	Glycine	1	0
Node 13	GGA	CAG	Glycine	Glutamine	1	2

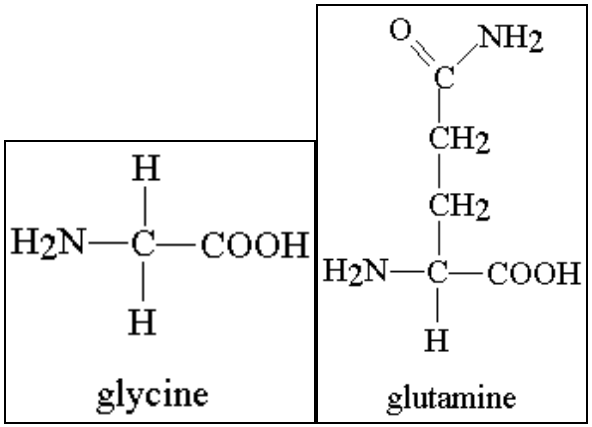

Figure 6. The amino acids glycine and glutamine that are found at codon site 28 in pham 1706.

Table 7. Codon site 44 in pham 1706, two branch sites had synonymous substitutions (shaded in green) and one has non-synonymous substitutions.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Node 4	GAC	GAT	Aspartic acid	Aspartic acid	1	0
Rockstar	GAT	GAC	Aspartic acid	Aspartic acid	1	0
QuinnKiro	GAC	GAG	Aspartic acid	Glutamic acid	0	1

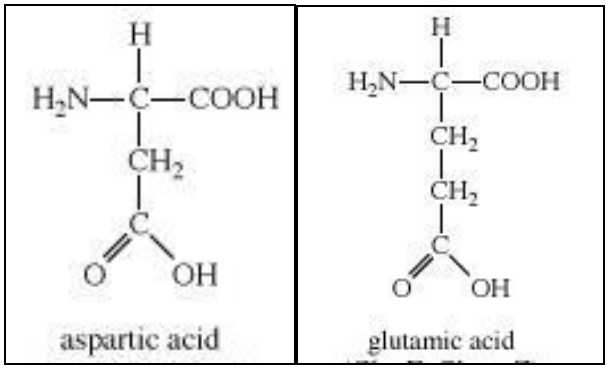

Figure 7. The amino acids aspartic acid and glutamic acid. At site 44 of pham 1706, QuinnKiro has a change from aspartic acid to glutamic acid.

Table 8. Codon site 509 in pham 4481. Pink shaded rows indicate only non-synonymous substitutions and green indicates only synonymous.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Node 11	GCC	TCC	Alanine	Serine	0	1
Heldan	GCC	TCC	Alanine	Serine	0	1
Phantastic	GCC	GAC	Alanine	Aspartic acid	0	1
QuinnKiro	GCC	GCG	Alanine	Alanine	1	0
Tiffany	TCC	AAC	Serine	Asparagine	0	2

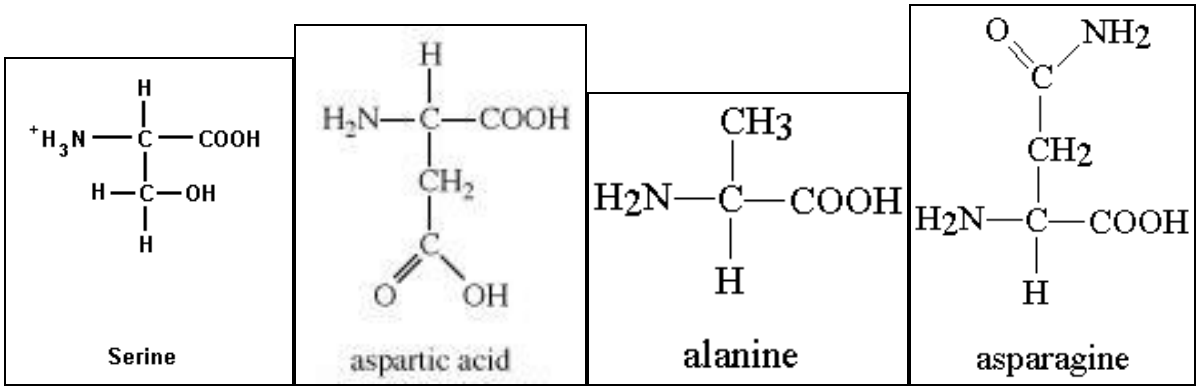

Figure 8. The amino acids serine, aspartic acid, alanine, and asparagine. These amino acids are either retained or replaced at codon site 509 in pham 4481.

undergo purifying selection, SLAC and IFEL agreed with codon site 44, IFEL identified six sites, including site 44 (table 7), amino acids are seen in figure 7.

Phamily 4481, gene number 4 on Bxz2 (figure 4) is affected by recombination, its putative function is a tail protein (part of the long tail connecting the capsid to the baseplate and tail fibers in figure 2), and 1710 sites were analyzed. There were 20 sites identified by MEME that are under diversifying selection, sites 509 (table 8 and figure 8) and 557 were also identified by FEL. Sites 85, 215, 239, and 263 were identified also by IFEL, which identified a total of 6 sites under diversifying selection. There were 255 sites identified by either REL, FEL or SLAC that are under purifying selective pressure; IFEL identified 106 sites. The following sites were identified by three models, 28 (table 9), 82, 90, 119, 125, 121, 159, 183, 196, 232, 249, 250, 252, 258, 260, 273, 306, 321, 355, 356, 382, 383, 388, 391, 406, 461, 480, 483, 495, 496, 502, 508, 515, 516, 519, 520, 521, 525, 525, 529, and 533, all under purifying selection.

Phamily 2981 (Bxz2 gene 34) encodes the serine integrase required for lysogeny in some phages (figure 4). In total, 1569 codons were analyzed in this phamily, and it was not identified as being affected by recombination. There were also no sites found to be under purifying selective pressure by any model, though 31 sites were identified by REL to be under diversifying selective pressure. Phamily 12396, gene 12 in Bxz2 (figure 4) was not identified to be affected by recombination, has a putative function as a lysin B, and 999 sites were analyzed. MEME identified two sites under diversifying selection, sites 151 (table 10 and figure 9) and 288. Additionally, 73 sites under purifying selection were identified by SLAC or FEL and 63 were identified by IFEL. The following sites are

under consensus by all three models for purifying selection, 30, 34, 44, 57, 95 (table 11), 153, 166, and 171.

Table 9. Codon site 28 in pham 4481. Only synonymous substitutions were made as indicated by green shading.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Node 7	ACG	ACA	Threonine	Threonine	1	0
Node 8	ACA	ACT	Threonine	Threonine	1	0
Rockstar	ACT	ACG	Threonine	Threonine	1	0

Table 10. Codon site 151 in pham 12396. Only non-synonymous changes were observed as indicated by pink shading.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Heldan	ACG	GAG	Threonine	Glutamic acid	0	2

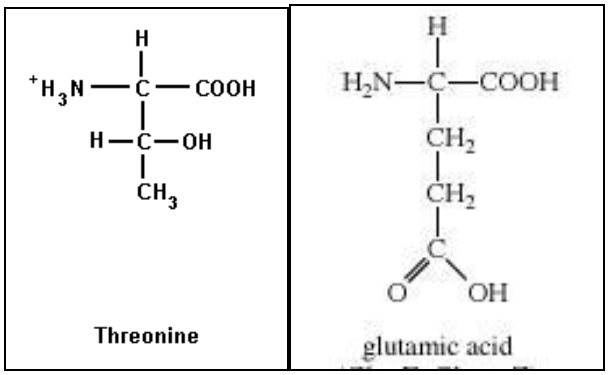

Figure 9. The amino acids threonine and glutamic acid. Codon site 151 in pham 12396 amino acid replacement.

Table 11. Codon site 95 in pham 12396. Only synonymous changes were observed as indicated by green shading.

Branch	From Codon	To Codon	From AA	To AA	Synonymous Substitutions	Non-synonymous Substitutions
Rockstar	GAC	GAT	Asp	Asp	1	0
QuinnKiro	GAC	GAT	Asp	Asp	1	0

V. DISCUSSION

This study aimed to identify evidence of evolutionary selective pressures in the genomes of bacteriophages. Specifically, diversifying and purifying selective pressure were investigated in protein-coding sequences. Identifying the presence of recombination was imperative to accurately infer selection. This study utilized sequence files available on Phamerator (Cresawn et al., 2011). The program webPRANK (http://www.ebi.ac.uk/goldman-srv/webprank/, Löytynoja and Goldman, 2010) was used to create alignments, and Datamonkey (http://www.datamonkey.org, Delport et al., 2010; Pond and Frost, 2005) was used to infer the presence of diversifying or purifying selection within gene phamilies.

The sequences in the analysis were groups of homologous gene sequences called phamilies. Each gene in a bacteriophage genome is assigned to a phamily or is considered an orpham (no homologous sequences) (figure 4). A total of 149 phamilies were of interest in this study. Of these phamilies, 92 were not included in the final analysis for various reasons (table 2). It is noteworthy that the two phamilies in table 2 under alignment error encode a protein that is produced by a programmed ribosomal frameshift. The frameshifted gene was not represented in the sequence file. Rather, the sequences were the incomplete gene fragments upstream and downstream of the ribosomal slippage site. Of the 57 phamilies for which complete data were obtained, 11 of them were affected by recombination (table 2 and figure 4). Insignificant recombination was responsible for the majority of phamilies being omitted from the final analysis (table 2). This may be explained by the particular dataset used in this study. This study used incomplete phamilies, meaning any gene

not present in a phage genome from subcluster A3 was discarded from the initial FastA file of gene sequences. Therefore, the recombination had to be significant between A3 genomes and could not account for recombination that is evident between phage of different clusters. This is to say, that if the alignment had included other cluster phages, recombination may have been significant. This was unavoidable, however, as larger datasets would have required weeks to months of computational time for a single analysis.

Referring to the spectrum of evolutionary mutation where minimal change is represented by purifying selection and drastic change is represented by recombination, with diversifying selection in the middle, we now have an idea of what methods of change are selected for in the A3 genomes. Minimal change was seen in the majority of phamilies including ones that code for a scaffold protein, three tail fibers proteins, and two head-to-tail connector proteins, table 4 and figure 4. The genes coding for tail fiber proteins were expected to be tolerant to change. There are six genes coding for tail fiber proteins in A3 phage genomes. The three other phamilies with tail fiber genes fell into the category of insignificant recombination (table 2), it may be the case that the tail fiber genes for which data was obtained are not actively involved in adsorption to the host. Overall this suggests that there are various genes that are of use for these viruses including many genes that currently have no known function, table 4 and figure 4.

In the middle of the spectrum where diversifying selection lies, three genes were identified, table 4 and figure 4. They code for an endonuclease, a serine integrase, and a gene of unknown function, table 4 and figure 4. All of these phamilies contained

less than 10 sequences in the alignment for analysis by Datamonkey (table 3). Murrell et al. (2012), Yokoyama et al. (2008), and Chen and Sun (2011) suggest that positive selection is more readily identified in smaller alignments. These three phamilies found to be under diversifying selection should be viewed with some skepticism. A larger alignment could clarify this inference.

At the far end of the spectrum of change is recombination. Eleven phamilies were affected by recombination, tables 2 and 4 and figure 4. In addition to being affected by recombination these were all found to be under purifying selection. The genes coded for tail sheath proteins, DNA associated enzymes, a ribonucleotide reductase, and genes of unknown function, table 4 and figure 4. Initially, it was thought that the function of the non-structural proteins could be carried out by proteins in the host cell. This would make the virally encoded ones expendable. However, these phamilies are present in all or most A3 phages suggesting that recombination plays a role in the adaptation of these genes.

Phamily 1706, tables 4, 5, 6 and 7; figures 6, 7, and 8, is the only phamily that does not include any phage genome outside of the A3 subcluster even before discarding other clusters. This may signify a gene that is essential for subcluster A3 phages but not other phages. With just 201 codon sites, this relatively short gene was not found to be subject to recombination. Two sites under diversifying selection and 10 sites under purifying selection were detected. These observations, when taken together with its conservation in all A3 phage genomes suggests the existence of a functional role for this gene phamily. The amino acid substitutions are highlighted in figures 5, 6, and 7. Codon site 10 replaces valine with threonine at one branch and

threonine with valine at another (table 5), suggesting they are interchangeable. Threonine is more polar than valine with a hydroxyl in the R-group. However, as this is only one codon site of many under selection not much can be said about how it affects function.

Phamily 4481 encodes a putative collagen-tail-like protein involved in forming the phage tail that connects the capsid to the baseplate and tail fibers (figure 2). The phamily included all A3 phages and is affected by recombination. Of the 1,710 total sites in the analysis, 20 sites were identified as under diversifying selection and 255 were identified as under purifying selection. All this suggests that 4481 is a structural protein that is being selected to remain the same in various parts of the gene with diversifying selection occurring at specific sites within the gene. Figure 8 shows the amino acids that can be found at site 509 (table 8). Three of the four amino acids are polar. The recombination identified in tandem with many sites under purifying selection suggests that tail protein adaptation is commonly handled by bulk substitution of large segments of the gene rather than by tweaking the existing sequence through accumulation of point mutations.

Phamily 12396, gene 12 in Bxz2 (figure 4), encodes lysin B proteins. Lysin B aids in lysis of the cell wall of mycobacteria. This phamily is not affected by recombination, has two sites under diversifying selection and 73 under purifying selection. These data support the hypothesis that this is a gene required for successful infection and recombination could be lethal. One site under diversifying selection, site 151 (table 10 and figure 9) is apparently under selection for larger and more polar amino acids R-groups.

Overall, the results of this study manage to infer preliminary data on the purifying and diversifying selective pressure in bacteriophage genomes of subcluster A3. A limitation experienced in this study was the time limit on analysis determined by Datamonkey. Initially, non-A3 bacteriophages were included in the alignments, however, the majority of Datamonkey models did not complete the analysis due to exceeding the allotted time limit. It was after this that non-A3 phages were discarded from the alignments. Phamilies that have complete data and had ten or more sequences in the alignment provide the most reliable data about the evolutionary selective pressures occurring within the A3 subcluster. The Bzx2 genome map (figure 4) is a visualization of these the overall findings.

The preliminary data suggest that recombination is used in surprising ways to diversify gene sequences in A3 phage genomes. It should be noted that while the majority of phamilies with complete data were under mostly purifying selection there were often various sites under diversifying selection as well (table 4). It would be exciting to see a study that modeled the proteins encoded by the genes from this study and identify where the amino acids under selection are found. As the number of sequenced actinobacteriophage genomes increases, these analyses should be repeated. Future analyses should include not only A3 phage sequences, but also sequences from all other actinobacteriophages. Of course, the time limit issue would need to be ameliorated. A repeat is especially necessary for phamilies that had less than ten sequences in the analysis.

In this study the bacterial hosts are members of the ubiquitous phylum *Actinobacteria* and the clinically important genus *Mycobacterium*. As such, our

findings provide preliminary data on how bacteriophages counter-adapt to changes in their bacterial hosts and may be the first steps down a path that leads to novel insights about evolutionary pressures on human pathogens. Clinical and ecological studies may converge as some of the co-evolutionary patterns that are responsible for the short-term and long-term associations between phages and bacteria are elucidated. For example, if bacteriophage therapy is to be considered as a viable alternative to antibiotics, knowing what genomic changes are responsible for short-term fluctuating associations and stabilized long-term associations (Needham et al., 2010) will be useful in developing effective treatment regimens. Effective and safe treatment regimens for bacteriophage therapy may be dependent on understanding how phages administered for therapy will adapt to the commensal bacteria of the patient.

References

Bohannan, B. J. M., & Lenski, R. E. (2000). The relative importance of competition and predation varies with productivity in a model community. *American Naturalist, 156*(4), 329-340.

Brüssow, H., Canchaya, C., & Hardt, W. (2004). Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews, 68*(3), 560.

Buckling, A., & Rainey, P. B. (2002). The role of parasites in sympatric and allopatric host diversification *Nature, 420*(6915), 496-499.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. , & Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Current Opinion in Microbiology, 6*(4), 417-424.

Cater, J. C., & Redmond, W. B. (1963). Mycobacterial phages isolated from stool specimens of patients with pulmonary disease. *The American Review of Respiratory Disease, 87,* 726-729.

Chen, J., & Sun, Y. (2011). Variation in the analysis of positively selected sites using nonsynonymous/synonymous rate ratios: An example using influenza virus. *PloS one, 6*(5), e19996.

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

Cresawn, S. G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R. W., & Hatfull, G. F. (2011). Phamerator: A bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics, 12(1), 395.*

Delport, W., Poon, A. F. Y., Frost, S. D. W., & Kosakovsky Pond, S. L. (2010). Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics, 26*(19), 2455-2457.

Pope. W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A., …& Filliger, L.Z. (2011). Expanding the diversity of mycobacteriophages: Insights into genome architecture and evolution. *PloS one 6*(1), e16329.

Froman, S., Will, D.W., & Bogen, E. (1954). Bacteriophage active against virulent Mycobacterium tuberculosis. *American Journal of Public Health and the Nations Health, 44*(10), 1326-1333.

Gómez, P., & Buckling, A. (2011). Bacteria-phage antagonistic coevolution in soil. *Science, 332*(6025), 106-109

Goodfellow, M., & Williams, S. T. (1983). Ecology of actinomycetes. *Annual Review of Microbiology, 37*, 189-216.

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., & Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environmental Microbiology, 13*(6), 1642-1654.

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., & Hatfull, G. F. (2006). Teaching scientific inquiry. *Science, 314*(5807), 1880-1881.

Hatfull, G. F. (2008). Bacteriophage genomics. *Current Opinion in Microbiology, 11*(5), 447-453.

Hatfull, G. F. (2012). Complete genome sequences of 138 mycobacteriophages. *Journal of Virology, 86*(4), 2382-2384.

Hatfull, G.F. (2013). Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science [SEA-PHAGES] Program, KwaZulu-Natal Research Institute for Tuberculosis and HIV [K-RITH] Mycobacterial Genomics Course, University of California—Los Angeles Research Immersion Laboratory in Virology, Phage Hunters Integrating Research and Education [PHIRE] Program. Complete Genome Sequences of 63 Mycobacteriophages. Genome Announc 1: e00847-13.

Hatfull, G. F. (2014). Mycobacteriophages: Windows into tuberculosis. *PLoS Pathogens, 10*(3), e1003953.

Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E., . . . & Hendrix, R. W. (2006). Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genetics, 2*(6), e92.

Hatfull, G. F. (2010). Mycobacteriophages: Genes and genomes. *Annual Review of Microbiology, 64*(1), 331-356.

Hooper, L. V., Littman, D. R., & Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science, 336*(6086), 1268-1273.

Jacobs, W. R., Tuckman, M., & Bloom, B. R. (1987). Introduction of foreign DNA into mycobacteria using a shuttle phasmid. *Nature, 327*(6122), 532.

Jacobs-Sera, D., Marinelli, L. J., Bowman, C., Broussard, G. W., Guerrero Bustamante, C., Boyle, M. M., . . . & Hatfull, G. F. (2012). On the nature of mycobacteriophage diversity and host preference. *Virology, 434*(2), 187-201.

Joo, J., Gunny, M., Cases, M., Hudson, P., Albert, R., & Harvill, E. (2006). Bacteriophage-mediated competition in Bordetella bacteria. *Proceedings of the Royal Society B: Biological Sciences, 273*(1595), 1843-1848.

Kidambi, S. P., Ripp, S., & Miller, R. V. (1994). Evidence for phage-mediated gene transfer among pseudomonas aeruginosa strains on the phylloplane. *Applied and Environmental Microbiology, 60*(2), 496-500.

Konrad Scheffler, & Cathal Seoighe. (2006). Robust inference of positive selection from recombining coding sequences. *Bioinformatics, 22*(20), 2493.

Kosakovsky Pond, S. L., Frost, S. D. W., Grossman, Z., Gravenor, M. B., Richman, D. D., & Leigh Brown, A. J. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Computational Biology, 2*(6), 0530-0538.

Koskella, B., Lin, D. M., Buckling, A., & Thompson, J. N. (2012). The costs of evolving resistance in heterogeneous parasite environments. *Proceedings of the Royal Society B: Biological Sciences, 279*(1735), 1896-1903. doi:10.1098/rspb.2011.2259

Koskella, B., & Meaden, S. (2013). Understanding bacteriophage specificity in natural microbial communities. *Viruses, 5*(3), 806-823.

Koskella, B., Thompson, J. N., Preston, G. M., & Buckling, A. (2011). Local biotic environment shapes the spatial scale of bacteriophage adaptation to bacteria. *American Naturalist, 177*(4), 440-451.

Lechevalier, H. A., & Lechevalier, M. P. (1967). Biology of actinomycetes. *Annual Review of Microbiology, 21,* 71-100.

Lindow, S. E., & Brandl, M. T. (2003). Microbiology of the phyllosphere. *Applied and Environmental Microbiology, 69*(4), 1875-1883.

Löytynoja, A., & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science, 320*(5883), 1632-1635.

Löytynoja, A., & Goldman, N. (2010). webPRANK: A phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics, 11(*1), 579.

Marston, M. F., Pierciey Jr., F. J., Shepard, A., Gearin, G., Qi, J., Yandava, C., . . . & Martiny, J. B. H. (2012). Rapid diversification of coevolving marine synechococcus

and a virus. *Proceedings of the National Academy of Sciences of the United States of America, 109*(12), 4544-4549.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics, 8*(7), e1002764.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics, 148*(3), 929-936.

Pfyffer, G. E. (2007). Mycobacterium: General Characteristics, laboratory detection, and staining procedures. In *Manual of Clinical Microbiology* (pp. 163-183). , Washington D.C.: ASM Press.

Pfyffer, G. E., & Palicova, F. (2011). Mycobacterium: General characteristics, laboratory detection, and staining procedures. In *Manual of Clinical Microbiology* (pp. 472-502). Washington D.C.: ASM Press.

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

Pommier, T., Douzery, E. J. P., & Mouillot, D. (2012). Environment drives high phylogenetic turnover among oceanic bacterial communities. *Biology Letters, 8*(4), 562.

Pond, S. L. K., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution, 23*(10), 1891-1901.

Pond, S. L. K., & Frost, S. D. W. (2005). Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England), 21*(10), 2531-2533.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics (Oxford, England), 21*(5), 676-679.

Poon, A. F. Y., Frost, S. D. W., & Pond, S. L. K. (2009). Detecting signatures of selection from DNA sequences using Datamonkey. *Bioinformatics for DNA Sequence Analysis (pp. 163-183).* Humana Press.

Posada, D., & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences of the United States of America, 98*(24), 13757-13762.

Rodriguez-Valera, F., Martin-Cuadrado, A., Rodriguez-Brito, B., Pašic, L., Thingstad, T. F., Rohwer, F., & Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology, 7*(11), 828-836.

Schierup, M. H., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics, 156*(2), 879-891.

Smillie, C. S., Smith, M.B., Friedman, J., Cordero, O. X., David, L. A., Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature, 480*(7376), 241-244.

Stach, J. E. M., & Bull, A. T. (2005). Estimating and comparing the diversity of marine Actinobacteria. *Antonie Van Leeuwenhoek, 87*(1), 3-9.

Stackebrandt, E., Sproer, C., Rainey, F. A., Burghardt, J., Päuker, O., & Hippe, H. (1997). Phylogenetic analysis of the genus Desulfotomaculum: evidence for the misclassification of Desulfotomaculum guttoideum and description of Desulfotomaculum orientis as Desulfosporosinus orientis gen. nov., comb. nov. *International Journal of Systematic Bacteriology, 47*(4), 1134-1139.

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology, 5*(10), 801-812.

Suzuki, Y., & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution, 16*(10), 1315-1328.

Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Evolution, 59*(1), 11.

Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., & Van Sinderen, D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiology & Molecular Biology Reviews, 71*(3), 495-548.

Waldor, M. K., & Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science, 272*(5270), 1910-1914.

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiology & Molecular Biology Reviews, 64*(1), 69-114.

Wong, K. M., Suchard, M. A., & Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science, 319*(5862), 473-476.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A.M.K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics, 155*(1), 431-449.

Yokoyama, S., Tada, T., Zhang, H., & Britt, L. (2008). Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences, 105*(36), 13480-13485.