ISOLATION AND GENOMIC CHARACTERIZATION OF 45 NOVEL BACTERIOPHAGES INFECTING THE SOIL BACTERIUM *Streptomyces griseus*

Richard H. Hale

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2018

APPROVED:

Lee E. Hughes, Major Professor
Michael S. Allen, Committee Member
Rajeev K. Azad, Committee Member
Robert C. Benjamin, Committee Member
Douglas D. Root, Committee Member
Jyoti Shah, Chair of the Department of
Biological Sciences
Su Gao, Dean of the College of Science
Victor Prybutok, Dean of the Toulouse
Graduate School

Hale, Richard H. *Isolation and Genomic Characterization of 45 Novel Bacteriophages Infecting the Soil Bacterium* Streptomyces griseus. Doctor of Philosophy (Biochemistry and Molecular Biology), December 2018, 144 pp., 10 tables, 49 figures, 65 numbered references.

Bacteriophages, or simply "phages," are the most abundant biological entities on the planet and are thought to be the largest untapped reservoir of available genetic information. They are also important contributors to both soil health and nutrient recycling and have significantly influenced our current understanding of molecular biology. Bacteria in the genus *Streptomyces* are also known to be important contributors to soil health, as well as producing a number of useful antibiotics. The genetic diversity of large (> 30) groups of other actinobacteriophages, i.e. phages infecting a few close relatives of the Streptomycetes, has been explored, but this is the first formal effort for *Streptomyces*-infecting phages.

Described here are a group of 45 phages, isolated from soil using a single Streptomycete host, *Streptomyces griseus* ATCC 10137. All 45 phages are tailed phages with double-stranded DNA. Siphoviruses predominate, six of the phages are podoviruses, and no myoviruses were observed. Notably present are seven phages with prolate icosahedral capsids. Genome lengths and genome termini vary considerably, and the distributions of each are in line with findings among other groups of studied actinobacteriophages. Interestingly, the average G+C among the 45 phages is around 11% lower than that of the isolation host, a larger disparity than reported for other groups of actinobacteriophages. Eighteen of the phages carry between 17 and 45 tRNAs and 12 of those carry a single tmRNA.

Forty-three phages were grouped into seven clusters and two subclusters based on dot plot analysis, average nucleotide identities, and gene content similarities. Two phages were not clustered with other phages in this dataset. A total of 5250 predicted genes were sorted into 1300 gene “phamilies,” with about 8% of the total phamilies having only a single member. Analysis of gene content among the 45 phages indicates first that most clusters presented here appear to be relatively isolated from one another, with phages in any one cluster generally sharing < 10% of their genes with phages in other clusters described here. Secondly, most of the phages here are more than twice as likely to share genes with phages isolated on bacteria outside of the genus *Streptomyces* than they are other phages isolated using a Streptomycete as host. These observations suggest that (1) the phage clusters here have a distinct extended host range, (2) those host ranges share overlap, and (3) *Streptomyces griseus* is likely not the preferred natural host for all phages described.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Phages and the Evolution of Molecular Biology

Scientists in the early 1900s who embraced an anti-vitalist approach to exploring biological processes, and the efforts of the Rockefeller Foundation program have both been credited with the beginnings of early molecular biology. The term, coined by Warren Weaver at the Rockefeller Institute in 1938, came to describe a research methodology that sought to describe complex biological phenomenon through chemistry and physics [1]. In addition to the above-mentioned groups, early work was also conducted by microbiologists and geneticists like Salvador Luria and Max Delbrück, who were interested in exploring the nature of “genes” and in developing a greater understanding of both microbes and the mechanisms of heredity [2]. Beginning in the 1930’s, two distinct research paths emerged within the loosely defined discipline of molecular biology. The first was the pursuit of the macromolecular chemistry of proteins, which employed methods such as x-ray diffraction analysis; the path forged by such notable scientists as Svedberg and Pauling. The second path approached biological problems by seeking “simple systems” that could be used as archetypes for the study of all biology. This approach also treated the entire biological system as a type of “black box,” complete with its intrinsic complexities, assuming that observable behaviors of the system were representative of some of the knowable processes that occurred within. It was this approach to the study of life that eventually led bacteriophages (and their bacterial hosts) to the forefront of study in molecular biology [2]. While the roots of the field were remarkably reductionistic, modern molecular biology encompasses the properties, functions, and interactions between nucleic acids and proteins, as well as the organizational principles of entire cells and, in some cases, whole organisms.

It is reasonably assumed that bacteriophages have been encountered by scientists since the earliest of microbiologists. However, the first recorded encounter occurred in 1915 when pathologist Frederick William Twort described the “glassy transformation” of *Micrococcus* colonies by a, then, unknown agent [3]. It was two years later that the term bacteriophage was coined by Felix d’Hérelle, after discovering that the microbial agent was capable of lysing bacterial cultures in liquid media as well as killing bacteria growing on agar plates in discreet patches that he termed “plaques” [4]. d’Hérelle was able to correctly deduce that the serially transmissible agent was viral in nature, multiplying at the host’s expense, that the plaque count provided a way to enumerate titers of the bacteriophages, and that the phage multiplied in “waves” that represented cycles of infection, reproduction, release, and reinfection [4]. The assertion that phages were viruses of bacteria was not one that was easily accepted in the scientific community of the time. In fact, until the early 1940’s, many researchers believed that the lysis occurring with bacteriophage infection was due to a type of autocatalytic enzyme that was endogenous to the host and not the phage. d’Hérelle was responsible for founding two important research foci within phage biology. The first was phage therapy to treat infectious disease in a pre-antibiotic world, and the second was the investigations into the biological nature of the bacteriophages themselves. His findings shed light on many aspects of the phage-host relationship and he was the first to suggest that, based on experimental findings regarding changes in host characteristics, there were actually many different “races” of bacteriophage [5].

By the 1930's, unraveling the mechanisms of heredity had become a pressing issue in the scientific community. Studies involving model organisms suggested that "the gene" had a dual nature, capable of being both highly stable in order to reliably confer traits from one generation to the next, and capable of rare mutations that were remarkably stable as well. It was the general thinking of the time that genes were not physical entities at all, but rather simple "factors" or "determinants." The concept of the gene as a physical unit of hereditary information was not conventionally accepted until the late 1950's. From early after their discovery, bacteriophages were seen as valuable tools in the search for the definitive nature of genes. Because of their ability to produce progeny that were similar to the parental phage, their seemingly small size, and the presumption that the phage had a simple structure, phages were proposed to be the simplest form of life and, in some cases, "naked genes" [6].

In 1938, German physicist Max Delbrück came to work in the United States at the California Institute of Technology with the initial goal of studying the elusive nature of the gene in fruit flies. It was here that he met Emory Ellis, who was studying bacteriophages as a model for oncogenic viruses. Impressed with the work, Delbrück saw an opportunity to study the bacteriophage as a "black box" system for heredity. He would begin to recruit other scientists and organize communications and collaborations between the likes of Salvador Luria, Jacques Bronfenbrenner, and Alfred Hershey in the advancement of phage biology. It was because of Delbrück that research groups began to primarily focus research on an "approved" group of bacteriophages so that information could be compared across laboratories. This group of *Escherichia coli* phages would become the go-to model organisms for lytic phage research and come to be known as the “T-phages” [7].

Arising from investigations into the nature of gene mutation, a pressing question came to light: were mutations the result of exposure to selective growth conditions or did they occur at random, only becoming evident when the organism was exposed to the proper selective conditions? In the 1940’s and early 1950’s, two particularly influential studies involving bacterial resistance to phage infection provided evidence that gene mutation occurred randomly and prior to exposure to the selective conditions. Both studies had sizable implications in the study of evolutionary biology and were paramount in establishing the gene as a “physical” entity. The first, conducted by Luria and Delbrück, was a statistical approach that suggested that phage-resistant bacterial mutants must exist within the population prior to exposure to the infectious agent [8]. This method was, however, indirect and met with some skepticism. In 1952, Lederberg and Lederberg published a study that provided direct evidence that mutations were occurring prior to the application of the selective agent [9]. Bacterial strains were tested against both bacteriophage and streptomycin in what would become the first documented instance of using velvet cloth as a transfer tool for replica plating.

The phenomenon known as lysogeny was observed in phage research almost from the beginning, however by the end of the 1940’s the mechanisms of the phenomenon were far from clear. It was the work of André Lwoff and Antoinette Gutmann in 1950 that began to illuminate the nature of lysogeny and introduced the term “prophage.” Lwoff and Gutmann were able to observe phage induction and release from single cells through the use of both microscopic observation and a micromanipulator [10].

The major groundwork for the operon concept of gene regulation was provided by the work of François Jacob while studying the mechanisms of lysogeny in *Pseudomonas pyocanea* [10]. It was with great fortune that the 1950's would see the lysogenic counterpart to the T-phages with the discovery of bacteriophage λ by Esther Lederberg in 1951 [11]. Although lysogeny had been studied for several years leading up to the discovery of bacteriophage λ, this particular phage provided the scientific community with a prototypic lysogenic phage through which it would eventually develop a greater understanding of both regulation of gene expression and the underlying mechanisms of lysogeny.

Francis Crick was the first to coin the so-called central dogma of molecular biology in 1958 [12]. He hypothesized that information flowed from both DNA to DNA and from DNA to RNA, and ultimately to proteins. Taking into consideration the existence of RNA viruses, he additionally speculated that it would be possible for information to flow from RNA to DNA and from RNA to RNA [13]. Prior to this, the experimental evidence of the role of RNA as an intermediate in the proposed flow of information was gained in 1956 by Elliot Volkin of Oak Ridge National Laboratory. According to Volkin, *Escherichia coli* B cells that had been infected with bacteriophage T2 showed no net gain in the number of RNA molecules following phage infection. However, the turnover of RNA molecules was remarkably extensive during the first minutes following infection, as shown by the incorporation of radiolabeled phosphorus and carbon. Through comparison of RNA species that existed in both uninfected and infected cells, it was determined that a new type of RNA appeared just after infection with the bacteriophage and that the nucleotide composition in the new species resembled more closely the bacteriophage rather than the infected host. Additionally, the cells were treated with chloramphenicol and deprived of essential amino acids in order to prohibit protein synthesis. It was found that the unique RNA species still appeared after infection, even in the absence of protein synthesis, suggesting that the informational flow in phage infected cells traveled from DNA to RNA to protein [14].

1.2 Abundance, Morphology, and Taxonomy

Bacteriophages are believed to be the most abundant genetic entities on the planet, with a global estimate of ≥ 10^{31} particles [15]. Phages are known to be practically ubiquitous, with terrestrial and aquatic phage isolates having been recovered from every continent (including Antarctica), and from coastal waters and other marine environments, including deep-sea hydrothermal vents [16]. In marine phages, it has been shown that the production and distribution of bacteriophage particles is determined primarily by the productivity and density of the host bacterial populations. This phage-to-host ratio has been frequently expressed at around 10:1 [16]. Many attempts have been made to directly determine phage counts in soil without the aid of enrichment and several of them have produced viable counts. However, these results have been widely variable. Notably in 2002, an attempt at direct counts of phages in soil was conducted using TEM. By direct observation of both phage particles and VLPs, the researchers have estimated an average of 1.5×10^7 particles per gram of soil. This number, according to the authors, may actually underestimate the actual counts by up to 40-fold due to the destruction of phages during sampling, secondary to the "abrasive nature" of soil [17].

As of the 2013 Taxonomic Release, the International Committee on Taxonomy of Viruses (ICTV) recognizes 253 total species of bacteriophages, which are organized into 64 genera, 6 subfamilies, 18 families, and two orders (ictvonline.org/virusTaxonomy.asp). The current taxonomical structure of bacteriophages is summarized in Table 1.1. As is common in virology, families are principally defined according to the nature of their nucleic acid composition and particle morphology. At this time, no universal criteria exist for genus and species demarcation.

In 1991, the ICTV adopted the following definition for the viral species [18]: *"A virus species is a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche."* A polythetic class is a group of entities, wherein all of the members of the group share a list of properties in common; however, no single property is possessed by every member in the group. Almost since its inception, this species definition has caused controversy among the rank-and-file within the ICTV community. Many members disagreed with the use of the term "polythetic," as they believed it to simply be a synonym for the word "variable." There have been several proposals throughout the years to replace the existing definition. In 2008, one such proposal suggested that nucleotide motifs be considered a property of a virus that, if present in all members of a species, would define the species as a monothetic class. This proposal was countered with the reasoning that a part of a thing is a thing and is not a property.

Table 1.1: Current taxonomy of bacteriophages as determined by the International Committee on the Taxonomy of Viruses (ICTV), 2013 Viral Taxonomy release. Features listed are those which are important in the assignment of bacteriophages to a particular family.

Orders	Families	Subfamilies	Genera	Species	Features
Caudovirales	*Myoviridae*	*Peduovirinae* *Hpunalikevirus* *P2likevirus*	6 within subfamilies, 12 not in subfamilies	83 within genera, 3 unassigned	Contractile tails
	Podoviridae	*Autographivirinae* *Picovirinae*	5 within subfamilies, 6 not in subfamilies	36 within genera, 8 unassigned	Short, noncontractile tail
	Siphoviridae	NONE	10 genera	31 within genera	Long, noncontractile tail
Ligamenvirales	*Lipothrixviridae*	NONE	4 genera	9 within genera	Envelope, lipids
	Rudiviridae	NONE	1 genus	3 within genus	Resembles TMV
UNASSIGNED	*Ampullaviridae*	NONE	1 genus	1 species	Bottle-shaped, enveloped
	Bicaudaviridae	NONE	1 genus	1 species	Lemon-shaped with 2, bipolar tails
	Clavaviridae	NONE	1 genus	1 species	Bacilliform
	Corticoviridae	NONE	1 genus	1 species	Complex capsid, lipids
	Cystoviridae	NONE	1 genus	1 species	Envelope, lipids
	Fuselloviridae	NONE	2 genera	9 within genera	Spindle-shaped, no capsid
	Globuloviridae	NONE	1 genus	2 species	Helical, enveloped
	Guttaviridae	NONE	2 genera	2 species within genera	Ovoid, enveloped, fibers at one end
	Inoviridae	NONE	2 genera	43 within genera	Filaments or rods
	Leviviridae	NONE	2 genera	4 within genera	NONE
	Microviridae	*Gokushovirinae*	3 within subfamily, 1 not in subfamily	12 within genera	NONE
	Plasmaviridae	NONE	1 genus	1 species	Envelope, lipids, no capsid
	Tectiviridae	NONE	1 genus	1 within genus, 1 unassigned	Internal lipoprotein vesicle
TOTALS: 2	18	6	64	253	N/A

It was instead proposed that a nucleotide motif would sufficiently serve as a diagnostic marker for determining which species a virus belonged to after the species itself had been established [18]. Largely, attempts at redefining viral species as monothetic classes have failed because of the mosaic nature of bacteriophage genomes, discussed in some detail in a later section. This tendency toward recombination and reassortment of genomic segments that is prevalent among many groups of phages produces, in single phage types, a polyphyletic genome. It is argued by many that it is practically impossible to represent this multidimensional phylogeny in a monophyletic scheme [18].

It is largely difficult to discuss phage morphology and phage taxonomy exclusively, as the taxonomy of bacteriophages relies heavily upon morphological characteristics. It is therefore prudent for the following discourse to present both topics simultaneously and discuss the current understandings of bacteriophage group morphologies as they relate to the taxonomy of the collective phages within that particular group. A representation of gross morphologies displayed by bacteriophage families is illustrated in Figure 1.1.

Prior to 1998, the ICTV had established two orders of viruses, both exclusively comprised of members that infected animal hosts. Bacteriophages made their debut in this arena with the order *Caudovirales*, consisting of the families *Myoviridae*, *Podoviridae*, and *Siphoviridae*. These families, the so-called “tailed phages” are marked by having either sheathed contractile, short non-contractile, or long non-contractile tails, respectively. Although tail-like structures appear elsewhere in virology, the tailed phages are unique in the consistent and regular presence of this head-tail morphology [19]. This particular attribute, as well as other physiological properties, suggests that these phages constitute a monophyletic evolutionary group.

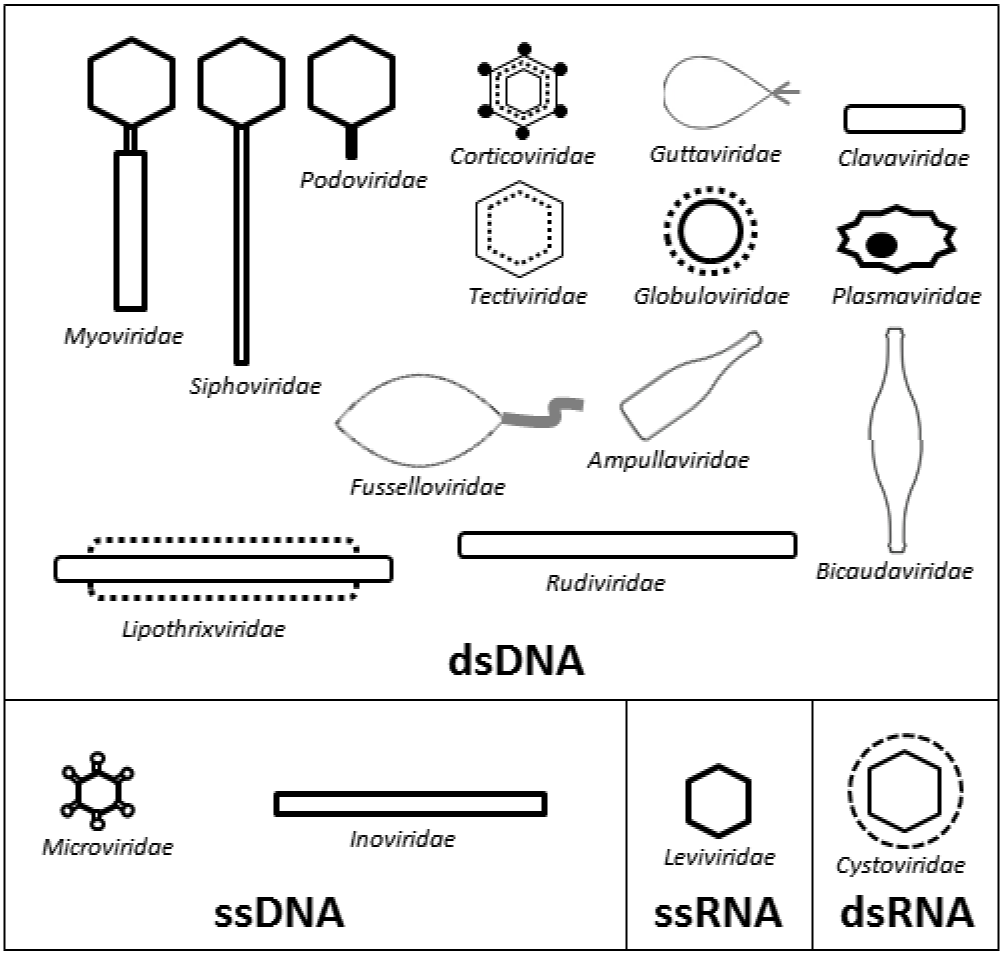


Figure 1.1: Representative morphologies of bacteriophage families. Solid lines indicate proteinaceous components and dotted lines represent lipids.

Particles consist of a head with cubic symmetry and a helical tail, an arrangement said to have “binary symmetry.” Virions generally have no envelope, are largely lacking in lipids, and icosahedral heads (or their elongated, prolate derivatives) predominate. The virion head is a protein shell with 5-fold symmetry that contains a single, linear dsDNA molecule. The helical, protein tail has 6-fold symmetry and is joined to the head by a connector. A few seemingly conserved minor dimensions, such as tail width, suggest that there are selective advantages for the conservation of these dimensions [19]. Order

Caudovirales is additionally characterized by the mode of infection, wherein the proteinaceous virion shell remains outside of the host as the DNA enters. DNA replication is achieved through the formation of concatemers, which are then cleaved to produce progeny with appropriate unit-length DNA. Virion assembly is sequential and begins with the formation of a prohead containing a protein shell with an internal scaffolding protein and a portal protein. Then, proteolytic cleavage of the capsid subunit occurs, and DNA is packaged into the prohead as head maturation occurs. Finally, the tail is attached to the mature, DNA-filled head and virion assembly is complete [19].

The families within the order *Caudovirales*, while sharing many gross morphological similarities, are highly different in the finer structural aspects of the tail. These structural differences alone lend to additional considerations. There is a substantial disparity in the numbers and functions of genes required by each of the families according to tail morphology. Additionally, modes of infection vary in the mechanism by which the DNA is injected into the host and virion assembly varies as well. In the assembly pathways, it has been shown that the longer tails are assembled independently of the capsid and then attached after completion while the shorter tails are assembled directly to the capsid [19].

The order *Ligamenvirales* consists of two families, *Lipothrixviridae* and *Rudiviridae*, both of which are known to infect within the domain Archaea. The two families, like the tailed phages before them, share a common gross morphology and nucleic acid composition. These phages are filamentous in shape and contain linear, dsDNA as the genetic material [20]. Family *Lipothrixiviridae* is currently comprised of four genera, collectively comprised of nine species. Family *Rudiviridae* is currently comprised of a single genus which is comprised of three species. Two major structural differences demarcate the two families within *Ligamenvirales*, the first being the rigidity of the virion particle itself. Lipothrix viruses are flexible and capable of bending, while the Rudiviruses are rigid and show a consistent linear shape when observed with electron microscopy [21]. The second major morphological difference is the presence of a phospholipid envelope surrounding the Lipothrix viruses, where the Rudiviruses lack this feature altogether. When the nature of this envelope was investigated in the bacteriophage SIFV, it was found that the phospholipid components of the envelope were able to be shed with a detergent treatment and then were capable of self-reassembly. Additionally, the thin layer chromatograph of these phospholipids was compared to that of the host's own lipids after a similar preparation. It was found that the two were highly similar, suggesting that the phospholipid envelope of the Lipothrix viruses is derived from the host's own lipid pool [20].

Families which are currently unassigned to orders comprise the bulk of the total number of phage families, while comprising a remarkably small percentage of the total number of phage species currently described and characterized in the literature. As of 2013, bacteriophages outside of the order *Caudovirales* represented only about 3% of the characterized phage isolates. As such, these families will not be discussed in detail. It is important to note, however, that these families constitute a wide variety of morphotypes and employ a wide range of infection strategies [22- 27].

1.3 Infection Cycles and Phage-Host Interactions

The molecular characterization of a bacteriophage often lends only minimal contemplation to the ecology of phage production, i.e. the interactions between the phage and its natural environment. Current estimates place phage to bacteria ratios around 10:1 [28] and suggest that around 10^{23} infections/second occur on a global scale [29]. Phages are an integral part of the microbial community and phage-host interactions have been shown to be a huge driving force in the shaping of microbial communities.

The common phage lifecycle, while highly variable in detail, generally involves mechanisms of adsorption, infection, and release of viral particles. In addition to the aforementioned stages, the phenomenon of phage decay exists, whereby the phage particle simply becomes inactive. Adsorption begins with a diffusion-mediated extracellular search, followed by bacteriophage collision with a host cell, subsequent attachment between phage and bacterium, and ultimately, nucleic acid uptake into the host cytoplasm. Adsorption is followed by infection, which consists of the eclipse period and the maturation period of phage progeny. The eclipse period is comprised of the period preceding phage-progeny maturation, whether that period is limited before the maturation, or prolonged through either lysogeny or pseudolysogeny. Release of viral particles occurs via one of several mechanisms, including host lysis, extrusion, or budding.

Phage-host interactions are complex, extensive, and are thought to widely drive the evolution of both entities in a rapid and dynamic fashion [29]. As proposed by Leigh Van Halen in 1973, the so-called “Red-Queen Hypothesis” is currently the most widely accepted hypothesis concerning the nature of the relationship between parasite and host. In “Through the Looking Glass,” the Red Queen tells Alice, “Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!” This hypothesis, also referred to as the “evolutionary arms-race hypothesis,” concludes that the tight co-evolutionary interaction displayed by parasite and host is an uneasy balance between changes on one side versus changes on the other. Any drastic changes (e.g. running faster) on one side without a corresponding change on the other could lead to the near-extinction of the static member of the pair. The result is the need for both species to constantly evolve in order to maintain the same level of fitness [30]. Given the nature of this relationship, it is prudent to discuss the mechanisms by which phage and host maintain this balance by presenting the current understanding of these mechanisms in a move/countermove fashion and in the successive steps of phage infection.

Phage infections may be foiled at a number of different steps. Adsorption, narrowly defined as the recognition and attachment of phage virions to appropriate receptor molecules on the surface of the bacterial cell, may be blocked by either preventing the phage-host encounter altogether or by a loss of the receptor molecule a phage uses to recognize its host. The most common way that the phage-host encounter may be prevented is through the “masking” of receptors by barriers such as extracellular polymers [31]. Many such barriers are shown to be plasmid-encoded and include such modifications as (a) increasing the lipid level in the lipoteichoic acid on a cell’s surface and (b) producing a galactose and/or rhamnose layer in order to shield the receptor from phage-encounter [31]. Another strategy employed by hosts to prevent adsorption is the loss of the functional phage receptor. Receptors are bacteria-encoded and may include lipopolysaccharides, teichoic acids, capsules, proteins, or other molecules that are exposed on the host surface. These molecules often serve in an essential capacity for the metabolic or other functions of the host. As such, a total loss or drastic down-regulation of these molecules may be very costly to host fitness. However, mutation in these receptors is widely found to be a mechanism of phage resistance in bacteria [32]. This is likely due to the fact that a discreet mutation in the receptor may render it unusable by the phage as a receptor molecule without deleteriously altering the functionality of the molecule for its intended purpose in the bacteria [32]. Phages have many mechanisms to overcome blocks in the availability of phage receptors. Random mutation and/or recombination are thought to be the most employed of these mechanisms by which the phage may alter its anti-receptor. These alterations on part of the phage may lead to recognition of the modified receptor, recognition of a new receptor, or a relaxation of recognition stringencies. The latter two strategies may, and often do, result in an expansion of the phage's host-range [31].

If phage receptor binding cannot be averted, the host may be protected from productive phage infection by preventing the takeover of host cell machinery, through mechanisms collectively referred to as "restriction." It is important to note that these systems result in destruction of the phage while allowing the bacterium to survive the encounter. In some instances, uptake blocks are employed, preventing the phage nucleic acid molecule from reaching the host cytoplasm. The most obvious of these mechanisms is resident-prophage encoded through a mechanism dubbed "superinfection exclusion." Superinfection exclusion is distinctly different from homoimmunity in that the nucleic acid molecule never reaches the cytoplasm [31]. Restriction-modification systems are regularly employed to further prevent exogenous DNA from affecting normal cell functions. These broad-reaching mechanisms act on nucleic acid molecules that are able to reach the host cytoplasm, both phage and plasmid alike. Restriction enzymes cleave DNA molecules indiscriminately at specific nucleotide recognition sequences. These mechanisms are usually paired with a modification system to protect host DNA from cleavage. Most often, phages are shown to evade this system through two mechanisms. The first is through random mutations that remove recognition sequences from the phage DNA, and the second is to carry modification enzymes encoded within the phage genome, effectively modifying the phage DNA and therefore evading destruction [30]. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) systems are a direct, targeted system through which a bacterium may gain "acquired immunity" from specific phages containing specific target sequences. It is believed that these sequences, which reside in the bacterial genome and are comprised of phage genomic regions with interspaced, short, palindromic regions of DNA, are acquired through a previous phage encounter in which the phage genome was degraded after infection and the host cell survived the encounter. The system is additionally thought to operate much like an RNA interference system, wherein the phage DNA is recognized by the CRISPR strand and host restriction enzymes used to subsequently destroy the hybridized, double-stranded DNA product [33].

1.4 Genomics and Evolution of the Tailed Phages

Full-genome DNA sequencing began in 1977 with a team of scientists led by Fred Sanger, using the *Escherichia coli*-infecting bacteriophage φX174 [34]. The sequencing of this ssDNA virus was quickly followed by the complete sequencing of the dsDNA, and fellow coliphages, λ and T7 in 1982 and 1983, respectively [35]. Phages, because of their relatively small size and ease of isolation, were a natural target for these first full-genome sequencing attempts. As of this writing, the NCBI phage genome database (https://www.ncbi.nlm.nih.gov/genome) contains 1384 full-genome sequences for bacteriophages. Of these, 1298 belong to the order *Caudovirales*, representing about 94% of reported sequences; comprised of 348 phages in the Family *Myoviridae*, 237 in *Podoviridae*, and 692 in *Siphoviridae*. There are currently 21 tailed phages that are unclassified with respect to taxonomical family. Reported genome sizes range from the 11,660nt podovirus, *Mycoplasma* phage P1, to the 497,513nt myovirus, *Bacillus* phage G.

A hallmark of tailed bacteriophage genomes is the presence of genetic mosaicism, wherein genetic modules are regularly traded, lost, or acquired through non-homologous recombination events. This recombination produces unique combinations of these modules which result in viable daughter phages that may or may not resemble either of the parents. While the modules of mosaicism are most often thought of as individual genes, it has been observed that phages may trade entire regions of their genomes or simply trade single protein domains with some success. Although the observed host range of individual phage isolates may suggest otherwise, the extent to which phylogenetically related modules can be found in phages that infect non-related hosts suggests that all tailed phages have access to a common gene pool. Recombination appears to occur at distinct nodes which most often align with the boundaries of open reading frames (ORFs), as suggested by nucleotide sequence similarity analysis between phage DNA regions that appear to share an evolutionary history. While this suggests that these recombination events are directed, this particular conclusion may be premature and there currently exists two hypotheses regarding the nature of phage recombination and the resulting mosaicism. The first involves the presence of short, conserved boundary sequences which are targeted by phage- or host-encoded recombinases, although this model would not seem to account for the majority of exchanges [35]. The second model suggests that the recombination events are random in nature and that the vast majority of these produce non-viable offspring. Those that survive the recombination event would necessarily need to maintain an appropriate genome length and contain necessary functional gene products to produce viable particles, therefore giving the appearance that these recombination events occur only at the boundaries of ORFs [35].

In an additional layer of mosaicism, phages that appear to share no significant nucleotide identity along the length of genomic sequence have been shown to contain genes that share an evolutionary history when analyzed at the amino acid level. This observation reinforces, to some degree, the concept of the gene as the functional module of genetic mosaicism. It has been hypothesized, however, that illegitimate recombination may occur at the boundary of functional domains within proteins, producing new combinations of these functional domains and, in essence, new proteins altogether. While this may occur, it is at such low frequencies that it is difficult to observe using classical genomic methodology [36].

Genome architecture, and often size, among the tailed phages appear to largely follow certain themes, which are particularly related to phage morphotype. This phenomenon is most readily observed within the *Siphoviridae*, where a clear synteny regularly exists among the virion structural and assembly genes [37]. Located on what is described as the “left arm” of the genome, this cluster of genes has a typically canonical arrangement. Even in isolates where a direct syntenous relationship between genes does not exist, these very genes are found grouped into segments according to function. Genome size also appears to be correlated to phage morphotype within the order *Caudovirales*. Increasing complexity of the tail structure requires an additional number of genes to be present in order to assemble a final, functional tail product. Therefore, it is without much surprise that podoviruses tend to have smaller genomes than the siphoviruses, who in turn are generally smaller than the myoviruses. It is currently unclear why the upper limits of genome size exist for each of the phage families.

With respect to genomics, the most comprehensively studied of the tailed phages are those that infect bacteria belonging to the genus *Mycobacterium*. In addition to the vast number of sequences available (discussed below), this group of bacteriophages provides particularly interesting insights into phage genomics in that the isolates span two of the three families within the order *Caudovirales*; and all are capable of infecting and producing viable phage particles in a single host, i.e. *Mycobacterium smegmatis* mc^{2}155.

In addition to insights on genome evolution and genetic exchange mechanisms between phages, bacteriophage genomics has also provided a framework for the further classification of bacteriophage isolates, particularly those for which full-genome sequencing data is available. While not formally recognized by the ICTV (for reasons discussed below), bacteriophages can be grouped into “cluster” and “subcluster” assignments, based on overall similarity as determined through a holistic approach using (1) dot plot analysis, (2) average nucleotide identity (ANI), and (3) gene content similarity (GCS). As recently as 2016, the definition of such clusters was generally accepted as being “≥ 50% of the nucleotide sequence of the genome is recognizably similar and syntenic (at word length = 10) to other members of the cluster" [38]. However, as phage researchers grappled with phages that straddled the line with respect to nucleotide identity but shared a sizeable number of genes that were related at the amino acid level, the focus became more centered around GCS than contiguous nucleotide sequence similarity alone. It should be noted that the use of cluster and subcluster assignments is meant supply a convenient lens through which phages may be viewed and compared and is not meant to imply a taxonomic relationship.

The term "subcluster" is used less formally to demarcate those phages *within a cluster* that are more closely related to each other than to other phages within the same cluster. There currently exist no stringent guidelines for subcluster assignment; and as that may be, the placement of phages into this taxonomical structure remains truly more art than science. This informal taxonomy, as it exists, is not without concern; particularly when considering the mosaic nature of bacteriophage genomes and the inevitability that phages exist that may be easily clustered into more than one of the existing clusters. When taking this into consideration, along with the discovery of evolutionarily-related modules in bacteriophages that infect hosts of different genera, a new clustering scheme has been implemented, wherein all bacteriophages of the order *Actinomycetales* are clustered into a single clustering scheme. Building on the clustering scheme established and first described by the Pittsburgh Bacteriophage Institute, a group of scientists working with phages of the *Enterobacteriaceae* have begun to group phages into so-called "superclusters" based on the syntenic arrangement of genes according to the genes' functions in phage development and/or infection. According to the developers of this particular schematic arrangement, this represents a level of evolutionary conservation that pre-dates contiguous sequence similarity at both the level of nucleotide and amino acid identity [38].

1.5 Actinobacteriophages and the Actinobacteriophage Database

Actinobacteria are those bacteria belonging to the phylum Actinobacteria. They are Gram-positive, high G+C, organisms, which may be either soil-dwelling or aquatic. They are known to be sources of both (1) contributions to nutrient recycling through their fungi-like contributions to the decomposition of organic matter, as well as (2) a wide range of secondary metabolites, many having antibiotic activities. In addition to their importance in both soil (and ecosystem) health and activities, Actinobacteria colonies often grow as extensive mycelia, more closely resembling fungi in their appearance and structure than a typical bacterial colony would. The cellular complexity of bacteria belonging to this phylum ranges extensively, from some of the largest and most complex bacterial cells to some of the smallest free-living prokaryotic cells described [39].

Streptomyces has long been established as the largest genus of Actinobacteria with well over 500 species described. Additionally, around 70% of clinically useful antibiotics derive from species in the genus [40]. Streptomycetes are also unique in that several of the genomes that have been completely sequenced indicate a linear chromosome, as opposed to the circular chromosome that is largely typical of bacteria. *Streptomyces* species have also become important as hosts for heterologous protein expression, and are displacing the historical powerhouse, *E. coli*, that has been used for decades. [41 – 43].

Because of the (1) unusual growth habits of *Streptomyces* species; and (2) wide array of potential uses as genetic tools; phages that infect in the genus *Streptomyces* show particular promise and should be ripe for deeper study. Indeed, there have been extensive characterizations of a few of these phages over the past 6 decades, particularly in their ability to cross-infect across multiple species in the same genus [41, 44]. In the age of genomics, further study into the similarities and differences of these phages, using gene-content-based methodology, is sure to illuminate host-specific evolutionary adaptations that are unique and interesting due to the unusual growth habits of the Streptomycete hosts. Further, it may provide insights into the development of new genetic tools for the manipulation of the host bacteria.

The Actinobacteriophage Database (www.phagesdb.org), formerly known as the Mycobacteriophage Database, was created in an effort by Graham Hatful and the Pittsburgh Bacteriophage Institute to provide a central database for information regarding bacteriophages isolated using *M. smegmatis* mc^{2}155 as host. As early as 2012, the database began adding additional “subdatabases” to the website, providing a platform for participating organizations to deposit information regarding phages isolated on a range of hosts falling in the phylum Actinobacter. In 2015, the Mycobacteriophage Database officially became the Actinobacteriophage Database and information for all hosts were available in a single, convenient database. Alongside this transition, the program known as Phamerator [45] (a program discussed in detail in Chapter 4) expanded its database to include options to compare phage gene composition across phages isolated using a wide range of Actinomycete hosts. These host genera, included in the database as well as Phamerator, currently include: *Actinoplanes* (one species, one strain), *Arthrobacter* (four species, five strains), *Brevibacterium* (two species, two strains), *Corynebacterium* (three species, five strains), *Dietza* (one species, one strain), *Gordonia* (eight species, eight strains), *Kocuria* (one species, one strain), *Microbacterium* (seven species, eight strains), *Mycobacterium* (six species, eight strains), *Propionibacterium* (two species, eleven strains), *Rhodococcus* (five species, 13 strains), *Rothia* (one species, one strain), *Streptomyces* (22 species, 28 strains), *Tetrasphaera* (one species, one strain), and *Tsukamurella* (two species, three strains).

In addition to a wide range of Actinobacterial hosts, the database contains phage isolate information for phages that have been isolated from a large geographic area. There are currently over 200 institutions across the United States that are participating in programs that isolate phages that ultimately end up with information in the database.

Of the 15 genera included in the database, those with the highest number of reported phage isolates include *Mycobacterium* (9765 phages), *Gordonia* (1153 phages), *Microbacterium* (715 phages), *Arthrobacter* (627 phages), and *Streptomyces* (547 phages). A review of the percentage of total isolates within these genera tell a different story. Of the reported *Streptomyces* isolates, 167 (~31%) have been sequenced and are either submitted to GenBank (122 phages) or are in various stages of finishing, annotation, or QC checks (45 phages). With respect to the percentage of total isolates that have been sequenced, those that infect Streptomycete hosts are solely outpaced by those that infect in the genus *Arthrobacter* (~38%). The percentage of total isolates that have been sequenced which infect hosts in *Gordonia*, *Mycobacterium*, and *Microbacterium* are roughly 26%, 16%, and 15%, respectively.

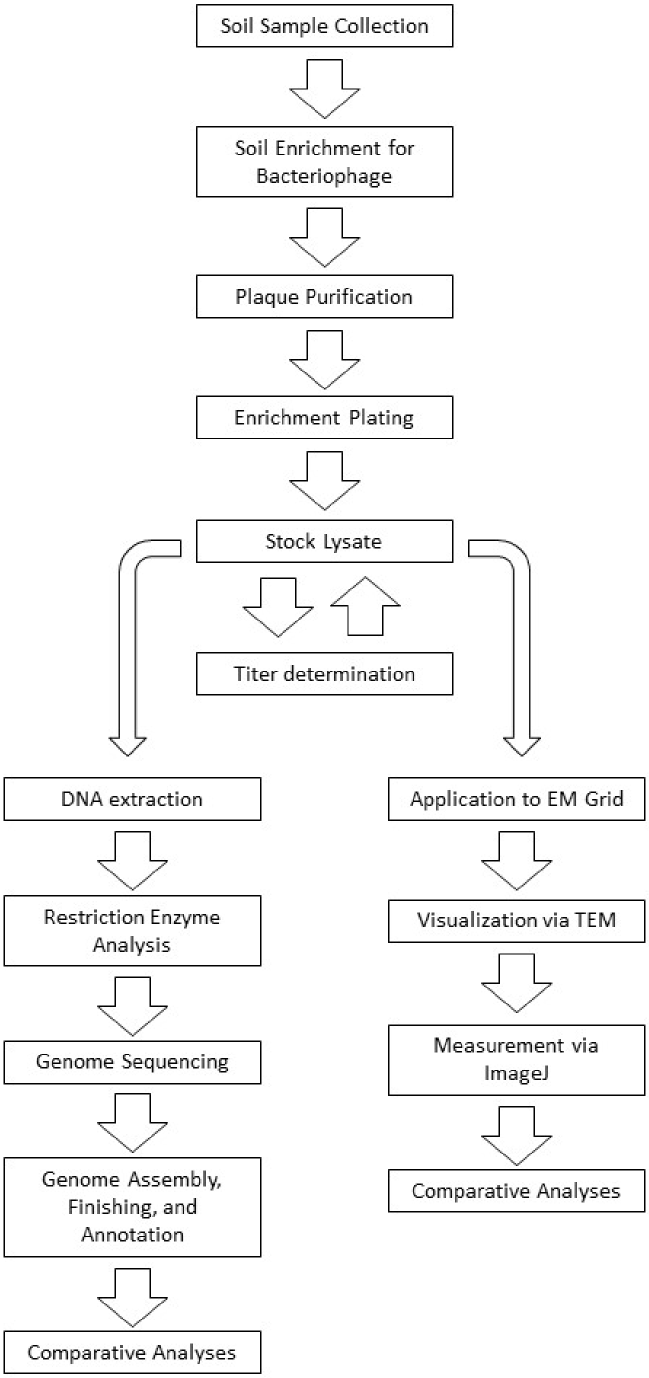


Figure 1.2: Workflow diagram for the genomic and physical characterization of 45 bacteriophages that infect the host *S. griseus* 10137. Once a stock lysate is produced, a titer is determined, and the lysate is then used to proceed to (1) genomic protocols (left side) and (2) particle visualization protocols (right side).

Despite the high percentage of reported isolates that have sequence and annotation information available to researchers, there is a surprising lack of published literature that compares genomic information across large groups of phages that infect hosts in the genus *Streptomyces*. More importantly, there is a virtual absence of published literature that addresses genomic comparisons across large groups of phages that infect a single *Streptomyces* host, e.g. *Streptomyces griseus* ATCC 10137.

1.6 Questions Addressed and General Workflow

This work is guided by a two, overarching questions: (1) what is the extent of genotypic and phenotypic diversity among 45 phages that infect a single host, i.e. *S. griseus* ATCC 10137, and (2) how does this diversity compare to other groups of actinobacteriophages from different hosts? A graphical representation of the workflow may be found in Figure 1.2.

CHAPTER 2

PHAGE ISOLATION AND SELECTION FOR STUDY

2.1 Methods and Materials

The methods and materials outlined below were modified from the *Mycobacterium* phage isolation protocols set forth by the Howard Hughes Medical Institute's Science Education Alliance (SEA) in 2013. While they were largely developed by the Hughes lab out of a desire to expand the efforts of the Pittsburgh Bacteriophage Institute to include *Streptomyces* phages in the growing body of phage genomics data, they were also developed with an eye on appropriateness for classroom instruction.

2.1.1 Bacterial Strains, Cultivation, and Maintenance

All phage isolations described herein were conducted using *S. griseus* subspecies *griseus*, American Type Culture Collection (ATCC) Number 10137, as host. The strain was maintained through storage of exospores. Exospore suspensions were prepared by streaking the organism on nutrient agar and incubating at 30° C for at least 96 hours. The exospores were then harvested by (1) flooding the plate with sterile water, and (2) using a sterile loop to scrape the exospores from the aerial hyphae of the organisms. The resulting suspension was aseptically collected, transferred to a sterile conical tube, centrifuged at 2200 x *g* for 10 minutes, and most of the supernatant removed. The resulting pellet and small amount of supernatant was re-suspended in 80% (v/v) glycerol, to a final concentration of about 25% (v/v) glycerol and stored at -20° C until use. Working cultures comprised liquid cultures, initially prepared by (1) inoculation of a 250 mL baffled flask containing 50 mL of nutrient broth and 50 g/L of PEG with 10 μL of spore suspension, and (2) allowing the culture to incubate on a rotary shaker at 225 RPM and 30° C for 96-100 hours. Subsequent cultures were prepared by single-colony inoculation of a 250 mL baffled flask with 50 mL of nutrient broth containing 50 g/L of PEG and allowing the culture to incubate on a rotary shaker at 225 RPM and at 30° C for 96-100 hours.

2.1.2 Phage Isolation and Purification

2.1.2.1 Sampling and Enrichment

Soil samples were collected from various locations by removing several tablespoons of soil from the top layer and depositing into either a sealable plastic bag or sterile 15 mL conical tube for transport to the lab. Soil samples were kept sealed and in low-light and low-heat conditions until needed for enrichment.

Soil samples were enriched for bacteriophages in a 250 mL baffled flask containing 50 mL nutrient broth that is brought to 30 mM glucose, 10 mM $MgCl_2$, and 8mM $Ca(NO_3)_2$ with the addition of a media supplement. Between one and five grams of soil and 1 mL of liquid *S. griseus* culture was added to the enrichment flask and the flask was incubated on a rotary shaker at 225 RPM and 30° C for 17-24 hours. Following incubation, the enrichment was allowed to settle and about 5 mL of supernatant removed via a 10 mL syringe and filtered through a 0.22 um filter. Fifty microliters of the filtrate were then spread plated on a nutrient agar plate (brought to 30mM glucose, 10mM $MgCl_2$, and 8mM $Ca(NO3)_2$ with the addition of a media supplement) with 100 µL of liquid *S. griseus* culture and allowed to incubate for 17-24 hours. The resulting plate was scanned for the presence of plaques, indicating a positive enrichment. Where necessary, the presence of putative plaques was confirmed via a 5 µL spot test on a lawn of *S. griseus*.

In the alternative, soil samples were screened for phages through direct plating. In this method, roughly 1-2 g of soil was combined with 10 µL of phage buffer (10 mM Tris, pH 7.5, 10 mM $MgCl_2$, and 68 mM NaCl) in a 15 mL conical tube. The tube was then vortexed for roughly 30 seconds and then allowed to settle until the soil in the tube settled at the bottom. A syringe was then used to remove 1 mL of buffer from the top of the tube and then forced through a 0.22 µm filter. Fifty microliters of the resulting filtrate were then combined with 100 µL of liquid *Streptomyces* culture and spread plated onto a nutrient agar plate containing media supplement. The plate was then incubated at 30° C for 17-24 hours and scanned for plaques in the manner described above.

2.1.2.2 Single-Phage Purification

Illustrated in Figure 2.1, bacteriophage purification was achieved through three successive rounds of purification plating, wherein a small agar plug was removed from a well-isolated plaque and transferred to 50 µL of phage buffer using a sterile micropipette tip. A 10-fold dilution series was prepared and then 50 µL of each dilution was combined with 100 µL of *S. griseus* liquid culture and then spread plated onto a new nutrient agar plate.

In some instances, purification in the first two rounds was performed by three-way isolation streak of the plug-Phage Buffer mixture onto a 100 µL lawn of *S. griseus* liquid culture on a nutrient agar plate. The different combinations of purification methods used are illustrated in Figure 2.1.

Following the third round of purification, the plate with the most web-like pattern of plaques was flooded with 8 mL of Phage Buffer and allowed to sit at room temperature for 4 hours. An example of a web-like pattern, i.e. a large number of plaques which cover the plate without causing complete lysis, is illustrated in Figure 2.2. The resulting lysate was then removed from the plate and filtered using a syringe and a 0.22 um syringe top filter. The lysate was stored at 4° C until further use.

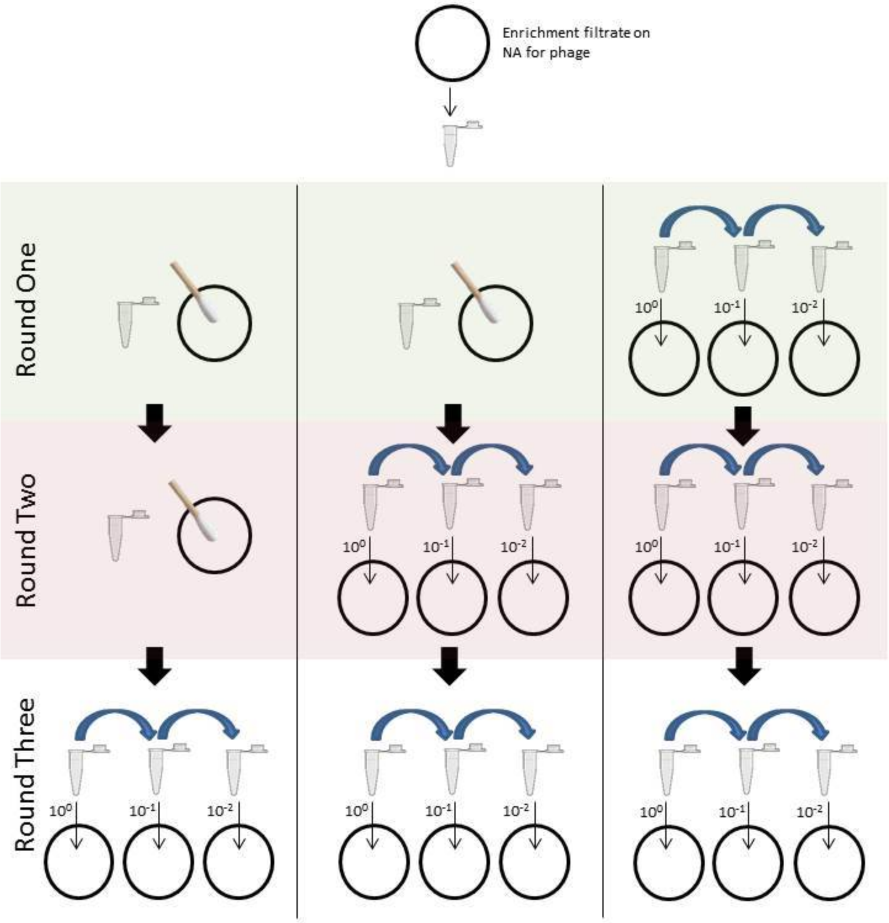


Figure 2.1: Diagram of three potential purification schemes for bacteriophage isolation.

Illustrated in Figure 2.3, phage particles were again increased via plating, wherein the lysate was combined with *S. griseus* liquid culture and spread plated onto ten nutrient agar plates. The appropriate volume of lysate needed for optimal enrichment was determine through an empirical assay using a 10-fold series of diluted lysate which was (1) combined with liquid bacterial culture, and then (2) spread plated to determine which combination yields the best web-like pattern of plaques. Once the optimal combination was determined, this combination was reproduced for each of the ten plates used in the enrichment plating. After incubation at 30° C for 17-24 hours, each of the appropriately-webbed plates was flooded with 8 mL of buffer and allowed to sit at room temperature for 4 hours. The lysate was then removed from the plates and filtered through a 0.22 um filter. Stock lysates were stored at 4° C until further use.

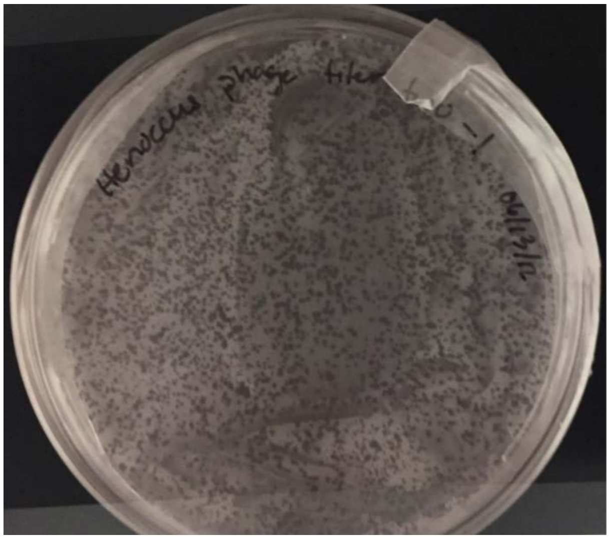


Figure 2.2: Sample web pattern on a bacterial lawn, created by the lysis of *S. griseus* ATCC 10137 cells following bacteriophage infection.

Titers of bacteriophage stock lysates were determined by spot test, wherein the lysate was used to create a 10-fold series of dilutions and 5 µL of each dilution was transferred to a single spot on top of a bacterial lawn consisting of 100 µL *S. griseus* liquid culture which had been spread plated onto a nutrient agar plate. Spot test plates were then incubated for 17-24 hours at 30° C, at which point the number of plaques was determined by manual counting and the titer of the stock lysate was calculated. An example spot test plate is illustrated in Figure 2.4.

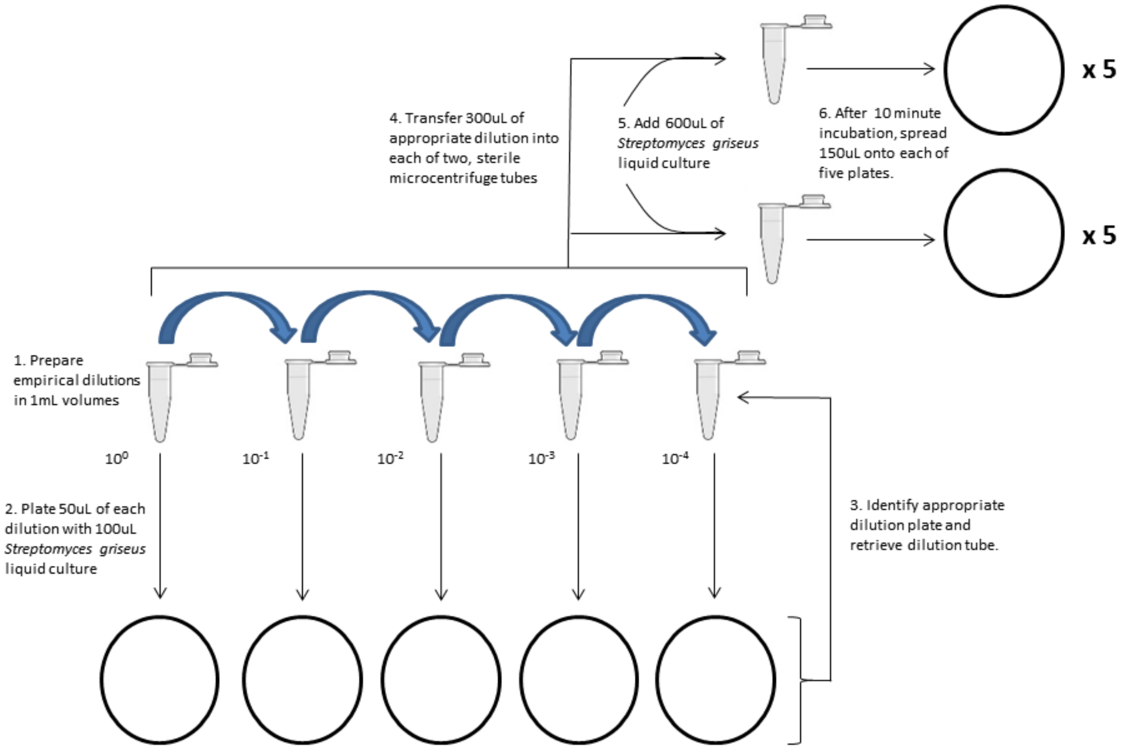

4. Transfer 300uL of appropriate dilution into each of two, sterile microcentrifuge tubes
5. Add 600uL of *Streptomyces griseus* liquid culture
6. After 10 minute incubation, spread 150uL onto each of five plates.
x 5
x 5
1. Prepare empirical dilutions in 1mL volumes
10^0
10^-1
10^-2
10^-3
10^-4
2. Plate 50uL of each dilution with 100uL *Streptomyces griseus* liquid culture
3. Identify appropriate dilution plate and retrieve dilution tube.

Figure 2.3: Plating procedure used in creating a large volume (~50 mL), high-titer lysate of bacteriophage particles.

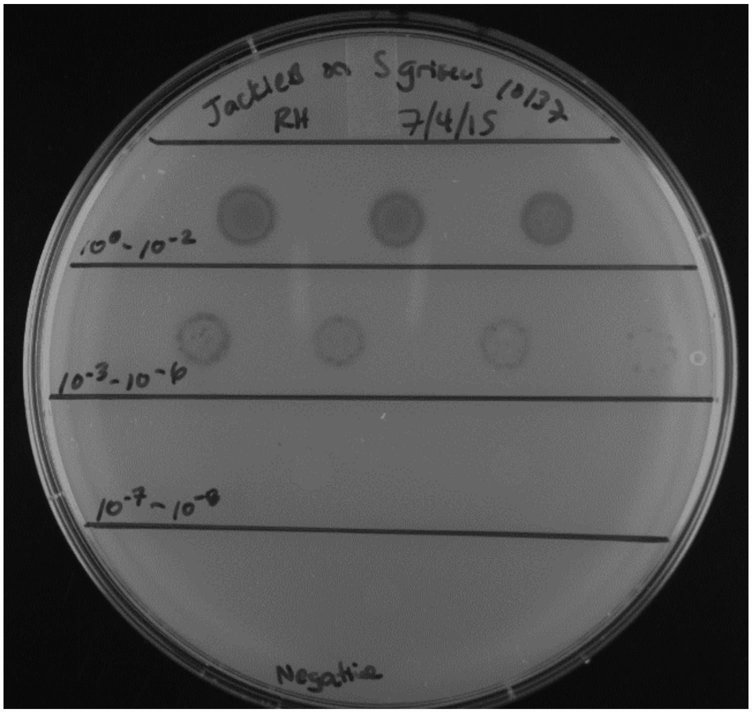


Figure 2.4: Sample spot test plate, used to determine titers of phage stock lysates.

2.1.2.3 Nomenclature of Isolates

Bacteriophage isolates were each named according to the convention set forth by the Pittsburgh Bacteriophage Institute. The rules governing nomenclature can be found on the Actinobacteriophage Database website (http://phagesdb.org/namerules/).

2.2 Results and Discussion

2.2.1 UNT and the *Streptomyces* Phages

Since 2009, students and researchers at the University of North Texas have isolated a total of 411 phages, as reported in the Actinobacteriophage Database. Of the 411 phages, 293 phages have been isolated on hosts in the *Streptomyces* genus. These hosts include (with the number of phages isolated on each host in parentheses): *Streptomyces azureus* NRRL B-2655 (7), *S. griseus* ATCC 10137 (193), *Streptomyces indigocolor* NRRL B-12366 (3), *Streptomyces toxytricini* NRRL B-5426 (12), *Streptomyces tricolor* NRRL B-16925 (1), *Streptomyces venezuelae* ATCC 10712 (42), *Streptomyces virginiae* ISP-5094 (1), and *Streptomyces xanthochromogenes* NRRL B-5410 (34). These 293 phages constitute about 54% of the 547 total phages reported in the database that have been isolated on hosts in the genus *Streptomyces*.

As noted above, 193 (~ 66%) of the 293 phages isolated at UNT on species in the genus *Streptomyces* have been on *S. griseus* ATCC 10137. As illustrated in Figure 2.5, the number of phages isolated using *S. griseus* ATCC 10137 as host has enjoyed an overall increase since 2012, both at UNT and in member schools depositing information in the database. During 2012 and 2013, UNT isolated the only phages on this host that were reported in the database. Beginning in 2015, other member schools began to isolate phages using the host, and the total number of phages using *S. griseus ATCC 10137* as host that were isolated over the year increased over two-fold both at UNT and among other schools using the database by 2017.

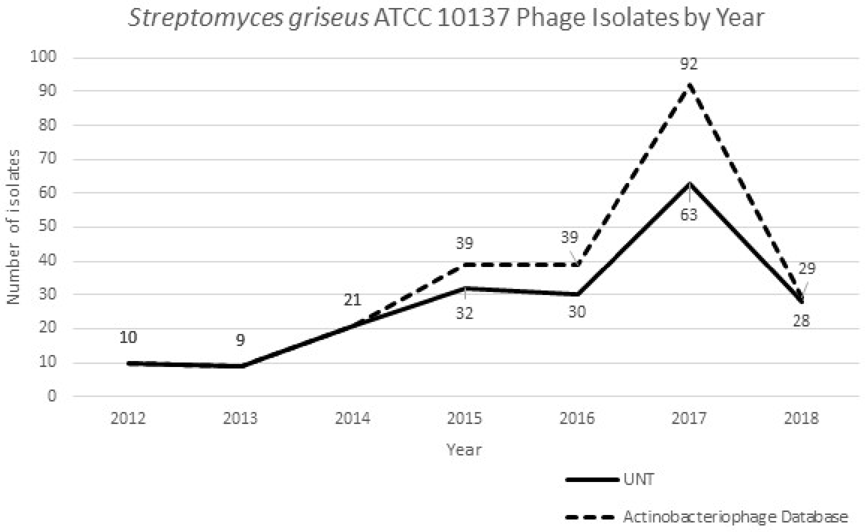


Figure 2.5: The total number of phages (by year) isolated at SEA-PHAGES member institutions and using *S. griseus* ATCC 10137 as host, as reported in the Actinobacteriophage Database (phagesbd.org).

2.2.2 Selection of Phages for Study

In order to be included in this study, phages must have met each of the following criteria enumerated below.

- Isolated at the University of North Texas
- Isolated using *S. griseus* ATCC 10137 as host
- Sequenced, with the genome finished and annotated as of June 1, 2018
- Submitted to GenBank prior to July 15, 2018

2.2.3 Phage Isolate Data

As shown in Table 2.1, there are currently 45 phages that meet the criteria delineated above. All but two of these were isolated from soil samples collected in the state of Texas. Phages OlympicHelado and Raleigh were isolated from samples collected in Brooklyn, NY and Jersey City, NJ, respectively. Of those phages isolated from soil samples collected in Texas, 25 (~ 58%) were isolated from samples collected in Denton, 39 of them (~ 91%) were isolated from samples collected in the North Texas region, and four were collected from Texas regions outside of North Texas. Of the 45 phages included in this study, 39 (~ 87%) were isolated from soil samples collected within a 55-mile radius of the University of North Texas.

A complete list of phages, year of isolation, sample collection location, and sample location coordinates can be found in Table 2.1.

Table 2.1: Summary of phages used for this study, including (1) name, (2) isolation year, (3) sample location, and (4) sample geolocation data.

Phage Name	Isolation Year	Sample Location	Sample Coordinates
Aaronocolus	2014	Denton, TX	33.213 N, 97.148 W
Annadreamy	2016	Denton, TX	33.2075 N, 97.1526 W
BabyGotBac	2016	Denton, TX	33.207 N, 97.152 W
BayC	2017	Keller, TX	32.9262 N, 97.2908 W
Blueeyedbeauty	2016	Denton, TX	33.211081 N, 97.146375 W
BryanRecycles	2016	Denton, TX	33.227837 N, 97.130108 W
Caliburn	2014	Grapevine, TX	32.887056 N, 97.094472 W
Comrade	2017	Denton, TX	33.209374 N, 97.136695 W
Crosby	2014	Denton, TX	33.21399 N, 97.146347 W
DrGrey	2016	Denton, TX	33.212584 N, 97.14542 W
Eddasa	2017	Denton, TX	33.2344 N, 97.104 W
HaugeAnator	2017	Denton, TX	33.197779 N, 97.132796 W
Henoccus	2012	Denton, TX	33.211624 N, 97.15459 W
Hydra	2014	Gainesville, TX	33.626671 N, 97.118444 W
Immanuel3	2012	Allen, TX	33.084722 N, 96.643889 W
Izzy	2014	San Antonio, TX	29.590722 N, 98.993167 W
JackieB	2012	Keller, TX	32.862146 N, 97.295193 W
Jash	2016	Denton, TX	33.213811 N, 97.151103 W
Karimac	2012	Fort Worth, TX	32.765556 N, 97.478611 W
LazerLemon	2017	Denton, TX	33.202295 N, 97.158154 W
Lorelei	2015	Beaumont, TX	30.099116 N, 94.214458 W

Phage Name	Isolation Year	Sample Location	Sample Coordinates
LukeCage	2017	Dallas, TX	32.749208 N, 96.95602 W
Maih	2014	Denton, TX	33.242857 N, 97.15677 W
Nabi	2015	Spurger, TX	30.79998 N, 94.213428 W
NootNoot	2016	Keller, TX	32.951878 N, 97.287514 W
OlympicHelado	2014	Brooklyn, NY	40.674761 N, 73.952306 W
Paradiddles	2016	Denton, TX	33.213968 N, 97.148256 W
Percastrophe	2017	Denton, TX	33.19747 N, 97.132884 W
Raleigh	2014	Jersey City, NJ	40.725426 N, 74.045229 W
Rana	2015	Spurger, TX	30.79998 N, 94.213428 W
Romero	2017	Venus, TX	32.416202 N, 97.100147 W
Salete	2017	Denton, TX	33.2121 N, 97.1474 W
Samisti12	2012	Denton, TX	33.235744 N, 97.157979 W
SparkleGoddess	2017	Denton, TX	33.21179 N, 97.152345 W
Spectropatronm	2016	Justin, TX	33.085432 N, 97.302528 W
Starbow	2017	Denton, TX	33.208357 N, 97.150035 W
StarPlatinum	2017	Denton, TX	33.202415 N, 97.16187 W
ToriToki	2017	Denton, TX	33.202555 N, 97.13444 W
TP1604	2012	Keller, TX	32.913909 N, 97.298122 W
UNTPL	2012	Denton, TX	33.210244 N, 97.152818 W
Wentworth	2017	Fort Worth, TX	32.98522 N, 97.285151 W
Wofford	2017	Denton, TX	33.2078 N, 97.1522 W
Xkcd426	2012	Valley View, TX	33.43543 N, 97.21993 W
YDN12	2012	Richardson, TX	32.923056 N, 96.7225 W
ZooBear	2017	Denton, TX	33.19747 N, 97.132884 W

CHAPTER 3

GENOME SEQUENCING AND CLUSTERING BY DOTPLOT AND ANI

3.1 Methods and Materials

The following Methods and Materials contain descriptions of methods used for genome sequencing on the Ion Torrent PGM. When this project started, all phage sequences were obtained using the Ion Torrent and, as time progressed, moved to the Illumina MiSeq. While a majority (28) of the phages described here were sequenced at the University of North Texas, 17 of the phages were sequenced by the Pittsburg Bacteriophage Institute. Because most of the phages (1) used in this study, and (2) sequenced using the Illumina MiSeq sequencing platform, were not sequenced at UNT, the methods and materials for an Illumina sequencing run are not detailed below.

3.1.1 Nucleic Acid Extraction and Storage

Stock lysates were treated with a nuclease mix prior to nucleic acid extraction to remove any exogenous nucleic acids present in the lysate. Sixty microliters of a nuclease mix (0.25 mg/mL DNase I, 0.25 mg/mL RNase A, 150 mM NaCl, and 50% (v/v) glycerol) was added to 15 mL of stored lysate that had been transferred to an Oak Ridge tube. The lysate-nuclease mix was incubated at 37° C for 30 minutes and then allowed to sit at room temperature for one hour.

Following nuclease treatment, 8 mL of phage precipitation solution (30% (w/v) PEG and 3.3 M NaCl) was added to the lysate for final concentrations of 10% (w/v) PEG and 1.1 M NaCl. The tube was then (1) placed at 4° C overnight, or (2) tightly packed on ice for at least 30 minutes. Samples were then centrifuged for 20 minutes at 4° C and

8500 RPM (10,000 x *g*) in a Sorvall RC-5C Superspeed centrifuge (Beckmann-Coulter) using an SA-600 fixed angle rotor (Thermo Scientific). Once the tubes were removed from the rotor, the supernatant was carefully decanted, and the tubes were inverted on a paper towel for 3-5 minutes to allow any excess supernatant to drain away from the pellet and out of the tube.

To purify and recover phage nucleic acids, the Wizard DNA Clean-up System (Promega) was used. Pellets containing phage particles were treated with 2 mL of DNA clean-up resin that had been pre-warmed to 37° C. Phage particles were then uncoated, (i.e. the capsid proteins removed) by pipetting up and down about 30 times. The resin mixture was then transferred to a DNA clean-up column and excess mobile phase passed through using a syringe. The DNA was then cleaned by passing 2 mL of ice-cold 80% (v/v) isopropanol over the column. The column was then placed atop a clean 1.5 mL microcentrifuge tube and transferred to a table-top centrifuge and centrifuged at 6700 x *g* for a total of six minutes. Fifty microliters of molecular grade water, pre-heated to 80° C, was then added to the column and allowed to sit for 30 seconds. The column was then centrifuged again at 6700 x *g* for one minute and the DNA-containing eluent was collected and transferred to a clean 1.5 mL microcentrifuge tube.

The quality of nucleic acid extracts was assessed via gel electrophoresis by combining 10 µL of recovered DNA (in molecular-grade water) with an appropriate volume of loading dye (bromophenol blue and xylene cyanol FF) and transferring to a 0.8% (w/v) agarose gel containing ethidium bromide and run at 80 V for approximately 1.5 hours. Agarose gels were approximately 7.5 cm wide and 10 cm long. Gels were then visualized in a gel documentation system at 302 nm. Quality was assessed by looking for RNA contamination that would collect at the bottom of the gel. DNA was quantified using 2 μL of DNA solution on a NanoDrop 2000c spectrophotometer. Additionally, quality was assessed for protein contamination using 260/280 ratios as determined by the spectrophotometer. Acceptable values for 260/280 ratios fell between 1.7 and 2.0. All nucleic acid samples were stored in molecular grade water and in DNA LoBind tubes at 4° C until further use.

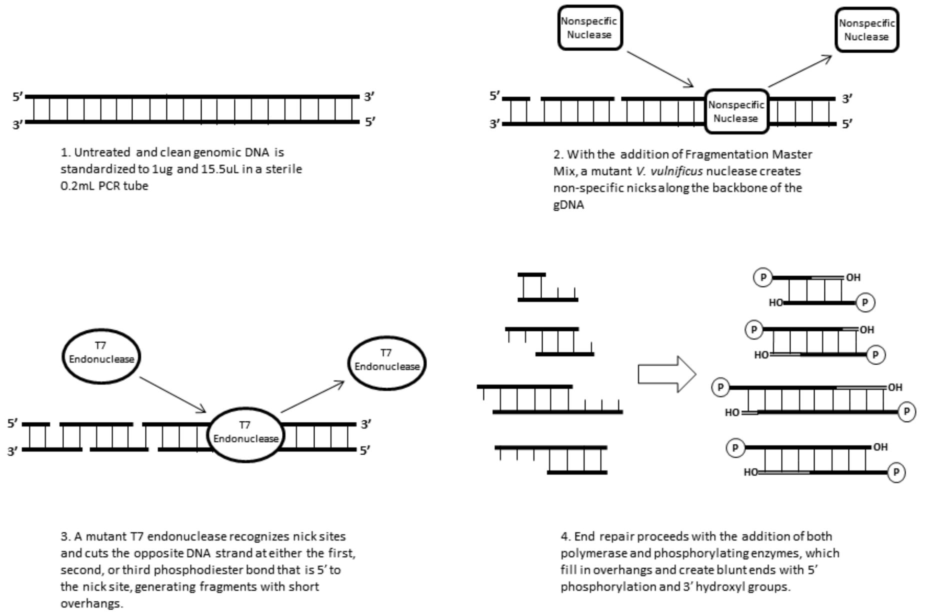
3.1.2 Genome Sequencing via Ion Torrent PGM

Immediately prior to library preparation, phage DNA samples were standardized via dilution to 1 μg of total DNA in 15.5 μL of molecular grade water. Where the DNA concentration of original sample was less than 64.5 ng/μL but greater than 32 ng/μL, 15.5 μL of original sample was used in the library preparation and the library preparation protocol was adjusted to reflect this change.

3.1.2.1 Preparing the Sequencing Library

All phage gDNA libraries for sequencing on Ion Torrent PGM were prepared using the NEBNext® Fast DNA Fragmentation & Library Prep Set for Ion Torrent™ (NEB). Illustrated in Figure 3.1, the library preparation began with treatment of the standardized gDNA with a Fragmentation Master Mix, which fragments the gDNA into 100-300bp fragments and then repairs the ends of the fragments, leaving 5'-phosphorylated blunt ends. Following this procedure, samples were held at room temperature prior to adaptor ligation.

Barcode adapters allow for the multiplex sequencing of several phage DNA samples in the same sequencing run. Therefore, each sample was treated with a unique barcode, each comprised of a different nucleotide sequence. Additionally, P1 adapters that allow for the attachment of the resulting fragments to particles for sequencing must also be added. As illustrated in Figure 3.2, both adapters were added to fragments simultaneously and through the use of T4 ligase.



1. Untreated and clean genomic DNA is standardized to 1ug and 15.5uL in a sterile 0.2mL PCR tube

2. With the addition of Fragmentation Master Mix, a mutant *V. vulnificus* nuclease creates non-specific nicks along the backbone of the gDNA

3. A mutant T7 endonuclease recognizes nick sites and cuts the opposite DNA strand at either the first, second, or third phosphodiester bond that is 5′ to the nick site, generating fragments with short overhangs.

4. End repair proceeds with the addition of both polymerase and phosphorylating enzymes, which fill in overhangs and create blunt ends with 5′ phosphorylation and 3′ hydroxyl groups.

Figure 3.1: Mechanisms of (1) fragmentation by a non-specific nuclease, and (2) end repair, used in the preparation of genomic DNA for sequencing via Ion Torrent PGM.

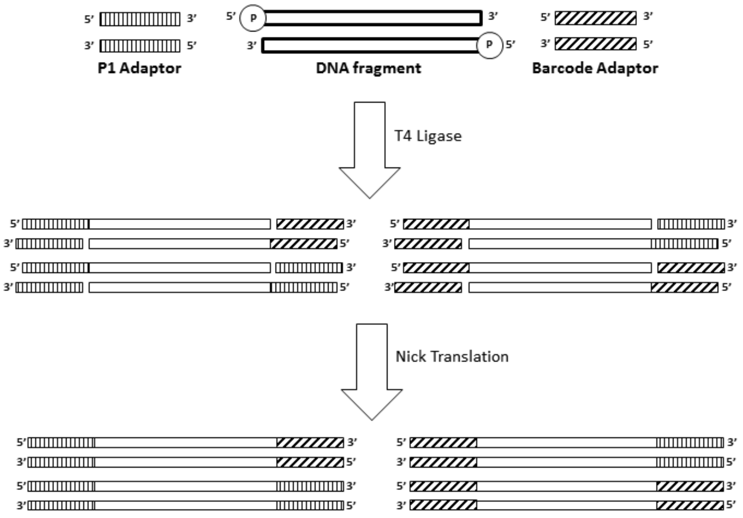


Figure 3.2: Addition of P1 and barcode adaptors to DNA fragments via T4 Ligase and nick translation. Barcode adaptors are specific to the sample being analyzed and allow for multiplex sequencing of many samples in the same sequencing run.

Following adaptor ligation, fragments were simultaneously cleaned and selected for a specific size range by using Agencourt AMPure XP Beads (Beckmann-Coulter), a magnetic rack, and 80% (v/v) ethanol, as illustrated in Figure 3.3. Resulting fragments contain an original DNA insert size of around 200 bp.

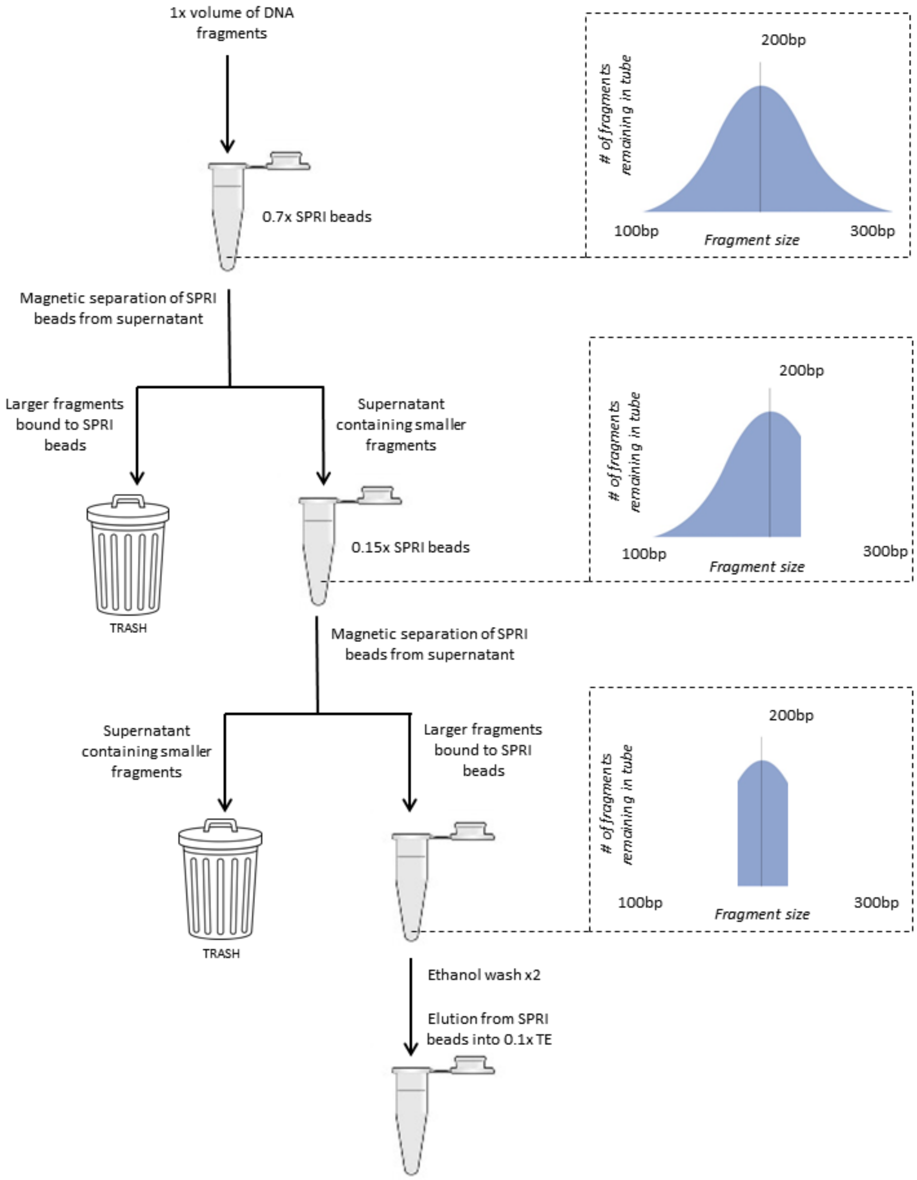

1x volume of DNA fragments
0.7x SPRI beads
200bp
# of fragments remaining in tube
100bp
Fragment size
300bp
Magnetic separation of SPRI beads from supernatant
Larger fragments bound to SPRI beads
Supernatant containing smaller fragments
TRASH
0.15x SPRI beads
200bp
# of fragments remaining in tube
100bp
Fragment size
300bp
Magnetic separation of SPRI beads from supernatant
Supernatant containing smaller fragments
Larger fragments bound to SPRI beads
TRASH
200bp
# of fragments remaining in tube
100bp
Fragment size
300bp
Ethanol wash x2
Elution from SPRI beads into 0.1x TE

Figure 3.3: Size selection and fragment cleaning using Agencourt solid phase reversible immobilization (SPRI) beads. Cleaning was achieved through the use of 80% (v/v) ethanol, while size selection was achieved by selectively removing target size ranges through the ionic association of fragments with beads. The retained fragments were around 200 bp long.

Fragments remaining after size selection were amplified via PCR under the following conditions: Initial denaturation at 98° C for 30 seconds; six cycles of (1) 10 seconds at 98° C for denaturation, (2) 30 seconds at 58° C for annealing, and (3) 30 seconds at 72° C for elongation; and a final elongation of five minutes at 72° C. Primer design preferentially amplified fragments that contain both the P1 and barcode adapters discussed earlier. Following amplification, the amplicons were cleaned using a single treatment of Agencourt AMPure XP Beads (Beckmann-Coulter), a magnetic rack, and 80% (v/v) ethanol.

Amplified fragments were quantified using one of two methods. The first was a TaqMan® assay performed on a Bio-Rad IQ5 RT-PCR machine (Bio-Rad) and using the Ion Library Quantitation Kit (Life Technologies). The second, preferred method used an Agilent 2100 Bioanalyzer, the accompanying 2100 Expert Software, and an Agilent High Sensitivity DNA kit (Agilent). The bioanalyzer method was preferred because it yields quantitative data concerning fragment size as well as total DNA concentration. Once DNA concentration was determined, each sample was diluted appropriately to 23 pM using molecular grade water. After dilution, 20 µL of each sample to be included in the final library was combined into a single, sterile DNA LoBind tube, constituting a single sequencing library. Where appropriate, the sequencing library was stored for up to four days at 4° C.

3.1.2.2 Particle Enrichment and Library Sequencing

Emulsion PCR was performed using an Ion Torrent One Touch 2 system. The sequencing library was combined in an aqueous medium with Ion Sphere Particles (ISPs) in a roughly 3:1 (ISP:DNA fragment) ratio, along with necessary enzymes and primers for a typical PCR reaction. The aqueous mixture was then transferred to an adaptor vessel and layered with an oil-based medium. The One Touch 2 system generates an emulsion of aqueous droplets in the oil and these structures are referred to as microreactors. As illustrated in Figure 3.4, each active microreactor should contain a single ISP with a single template (library fragment) and polymerase/primer/dNTPs. Clonal amplification occurred as the machine then acts as a thermal cycler, producing multiple copies of the same template on each ISP. Following emulsion PCR, the Ion Torrent One Touch-ES enriched the resulting solution for template-positive ISPs by removal of ISPs that lack template.

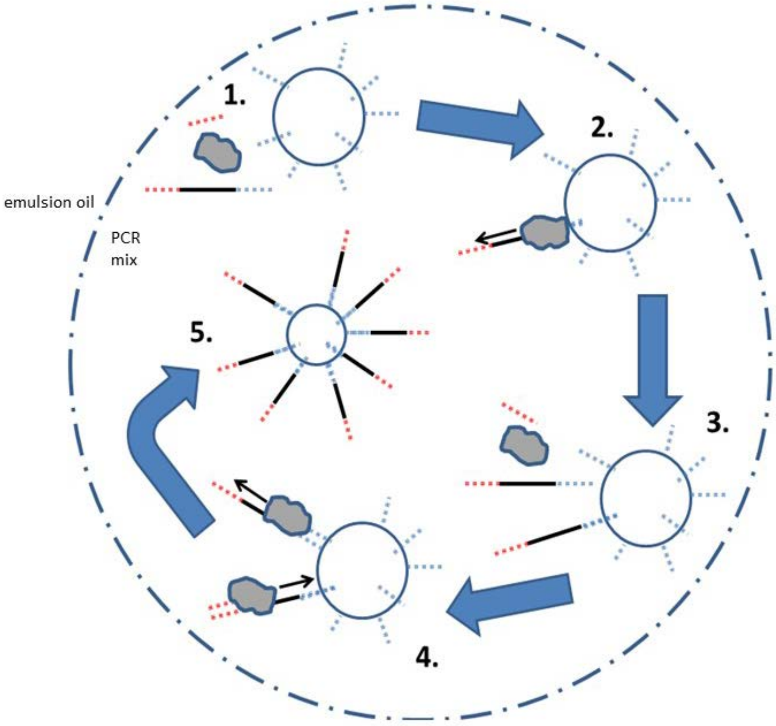


Figure 3.4: Clonal amplification via PCR inside of an aqueous microreactor. The reaction results in an ISP that is covered in identical copies of the original template.

Template-positive ISPs were loaded onto an Ion Semiconductor Sequencing Chip with polymerase and the chip was then placed on the Ion Personal Genome Machine™

(PGM). Each microwell in the chip contained a single, template-positive ISP. Sequencing of the template was by “sequencing by synthesis,” wherein the microwells were flooded by a single species of dNTP and, if the nucleotide was incorporated into the growing complementary DNA strand, the chip recorded the change in pH resulting from the hydrogen ion released during incorporation. In this reaction, hydrogen ions are released in a 1:1 (hydrogen ion: incorporated nucleotide) ratio. The machine then flooded the microwell with a different nucleotide, moving through the four nucleotides (A, C, G, and U) in a cyclical fashion throughout the run of the instrument.

Chemical information was converted to a digital signal (0 or 1) and then processed by an in-house server which assembled the corresponding sequences for each of the templates, compiled them, and converted them to a usable format. Sequencing information was retrieved from the server by accessing the server from an approved location and downloading the appropriate file(s).

3.1.3 Genome Assembly and Finishing

Initial de novo assembly of sequencing data was performed using 50,000 reads, randomly selected from the sequencing file using the program sff tools by Roche. Assembly was performed by Roche’s GS Assembler program, i.e. “Newbler,” version 2.6 (Roche). Where applicable, a different number of initial reads was randomly selected in order to provide the best initial assembly, e.g. the fewest and/or largest contiguous sequences (“contigs”).

Following the initial assembly, Consed [46] was used to generate a new fasta file from the initial assembly’s consensus sequence. This new fasta was then used to generate a new ace file using the sff tools program. All reads provided during sequencing were then added to the assembly in the programs Phred and Phrap [47, 48] using the consensus sequence as a scaffold.

Assemblies were then viewed in the program Consed. Where applicable, assemblies that resulted in multiple contigs were joined using the “crossmatch” function in the accompanying Phred/Phrap software package. Assemblies were also checked for an average coverage depth of at least 30x.

The program aceUtil [49] was used to identify any position in the consensus for which > 25% of the reads at that position were discrepant or if the reads at that position showed < 12x coverage. The positions fitting either of these criteria were flagged for further review. All positions tagged in this manner were then manually inspected in Consed. Most discrepancies were attributed to a phenomenon known as strand bias [50] or an overconfident call on the part of the Newbler assembly program. Where a positive nucleotide identity confirmation could not be made, primers were designed in Consed, generally about 100 bp upstream and downstream from the ambiguous base, and the region was amplified via PCR. The resulting amplicons were sent for direct sequencing at MWG Operon and the resulting reads incorporated into the assembly in order to make a positive identification. After the consensus sequence was complete, the file was either (1) sent to the Pittsburgh Bacteriophage Institute, or (2) retained in-house for quality control and the sequence was then deposited in the database (http://streptomyces.phagesdb.org/phages/).

Finished genome sequences were annotated using DNA Master (v5.0.2, http://cobamide2.bio.pitt.edu/computer.htm). Initial annotation was accomplished using the auto-annotate function in the program. Following an auto-annotation, each predicted gene was carefully examined, and a subjective analysis performed. Final gene calls were based on careful weighing of the pieces of information listed below (in no particular order).

- Agreement between Glimmer [51] and GeneMark [52]
- BLAST results
- Shine-Dalgarno scores
- Six-frame translations
- Coding potential as predicted by GeneMark HMM
- Length of ORF
- The Guiding Principles of Bacteriophage Genome Annotation (phagesdb.org)

Putative gene functions were assigned using the Phage Evidence Collection and Annotation Network (PECAAN, https://discover.kbrinsgd.org), an online platform which provides sequence alignment information from NCBI BLAST, an internal BLAST using available phages in the Actinobacteriophage Database, the Conserved Domain Database, and HHPRED, as well as transmembrane prediction using TmHmm.

Where possible, multiple people annotated each genome independently of one another and the resulting annotation files were merged. The final annotation for each phage was discussed and agreed upon by all those who participated in the annotation process. After a final annotation was available, the annotation was sent for quality control at either the Pittsburgh Bacteriophage Institute or by one of the institute's designated Quality Control Specialists, e.g. Dr. Lee Hughes at the University of North Texas. Once an annotation was finalized, it was submitted to GenBank [53] using proper submission procedures as outlined by the National Center for Biotechnology Information (NCBI).

3.1.4 Clustering via Dot Plot Analysis and ANI

Because of the mosaic nature of bacteriophage genomes, it is useful when analyzing them to group them into related clusters. Clusters, as they are presented here, do not represent phylogenetic or taxonomic groupings. They do, however, provide a framework for (1) analyzing overall genome relationships, and (2) identifying genes or groups of genes that have been recently exchanged. Overall, there are four methods of clustering that are currently used in bacteriophage genome clustering. The first two are described briefly below.

The first method of clustering is via dot plot analysis, using Gepard [54]. Once a genomic sequence is finished, it is the primary method of clustering. The accepted criterion for clustering via dot plot is contiguous nucleotide sequence similarity that is (1) evident on the dot plot, and (2) covers at least 50% of the smaller of the two genomes.

As illustrated in Figure 3.5, there are three distinct classes of relationships that are observable on a dot plot. The first two cases, i.e. the extremes of the relationships, include (1) those genomes which are obviously related and therefore easily placed into the same cluster (Genomes A and B, Figure 3.5.1), and (2) those which show no contiguous nucleotide sequence similarity and are clearly not able to be placed in the same cluster (Genomes A and D, Figure 3.5.3).

The third class of relationships make the process of clustering less clear-cut. This class comprises three subclasses of relationships, i.e. those where (1) there is weak, but evident, nucleotide identity that spans large segments of the genome (Genomes A and C, Figure 3.5.2), (2) there are segments of strong nucleotide sequence similarity that cover only very short segments, and (3) there is little or no evidence of sequence similarity at the nucleotide level, but the genomes share a large number of proteins which are related at the amino acid level. The second subclass may be addressed using average nucleotide identity (discussed immediately below), and the third subclass is addressed using the clustering method discussed in Chapter 4.

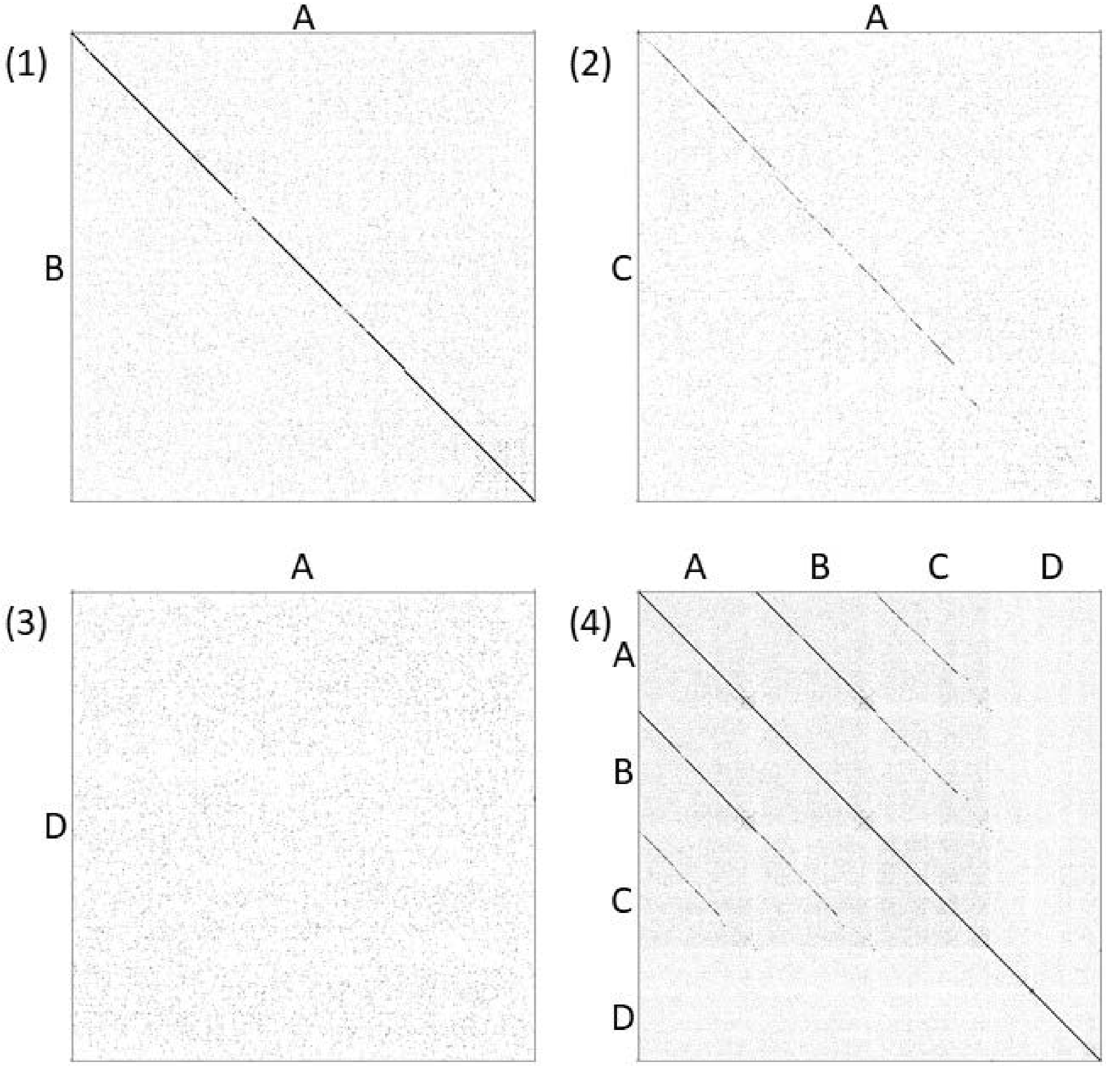
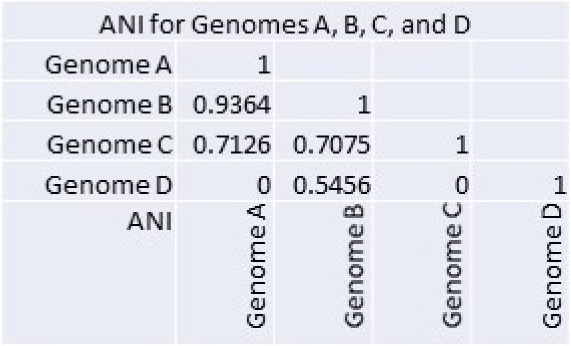


Figure 3.5: Dot plot representations of the three major classes of relationships observable between genomes on a dot plot. (1) An obvious relationship, where contiguous nucleotide identity is easily observable across greater than 50% of the smaller of the two genomes. (2) A less obvious relationship, where contiguous nucleotide identity is weak, but spans large segments of the genome. (3) No obvious relationship, where the two genomes share no contiguous nucleotide identity as observable on a dot plot. (4) A representation of the four genomes as they would appear when their sequences are concatenated into a single file and placed on each of the two axes.

The second approach to clustering uses average nucleotide identities (ANI) as computed using DNAMaster. As illustrated in Table 3.1, where two genomes fall into the extreme classes discussed above, i.e. those that show highly evident sequence relationship (Genomes A and B) and those that show little to know sequence similarity (Genomes A and D, Genomes C and D), the ANI values will correspond and corroborate the dot plot findings.

Where two genomes fall into the first subclass of relationships, i.e. those with weak similarity that spans large portions of the genome, ANI becomes extremely useful in parsing out the relationship between the two genomes. As illustrated in both Figure 3.5 and Table 3.1, the relationship between Genomes A and C is moderate, with a weak relationship that spans a great portion of the genome.

Table 3.1: ANI values, as calculated by DNA Master, for Genomes A, B, C, and D. Genomes A and D show a high degree (~94%) of ANI. Genomes A and C and Genomes B and C show a moderate level (~71%) of ANI. Genomes A and D show no observable ANI. Note the relationship between Genomes B and D, discussed further in the text below.



ANI for Genomes A, B, C, and D				
Genome A	1			
Genome B	0.9364	1		
Genome C	0.7126	0.7075	1	
Genome D	0	0.5456	0	1
ANI	Genome A	Genome B	Genome C	Genome D

The ANI value of 0.71 (~71%) aids in clearing up the relationship. In the immediate example, it is likely that Genomes A, B, and C are all able to be put into the same cluster, but that Genome C belongs in a different subcluster that Genomes A and B. For example, if Genomes A, B, and C were assigned to cluster “X,” then Genomes A and B would belong to subcluster X1 and Genome C would belong to subcluster X2.

Before moving on, it is important to speak to two limitations to using dot plot analysis and ANI to cluster phages. The first, as illustrated above, is that the use of ANI alone can be tricky with respect to clustering and should be used alongside a dot plot when assigning clusters. In the example above, Genomes B and D share an ANI of 0.55 (~55%). As shown below in Figure 3.6, the two genomes, while sharing a moderately high ANI, share little to no contiguous nucleotide identity and would therefore not be placed in the same cluster or subcluster.

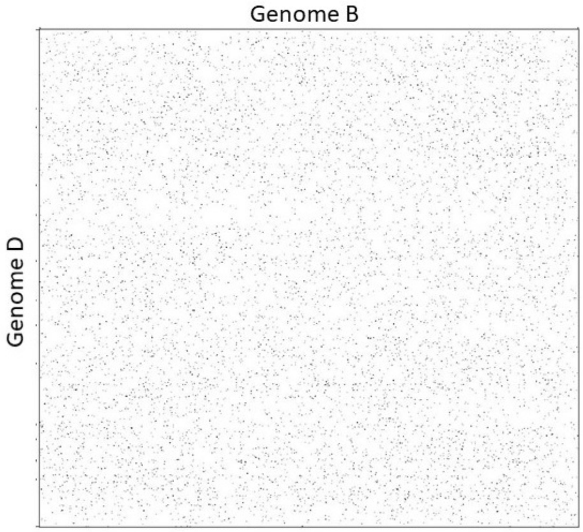


Figure 3.6: Dot plot of Genomes B and D. Note that although they share about 55% ANI (see Table 3.1), they show no contiguous nucleotide identity on a dot plot. Despite a moderate ANI value, Genomes B and D would not be clustered together.

The second limitation is that neither dot plot analysis nor ANI considers those relationships described above where two genomes share a large number of genes in common that are related at the amino acid level but show little to no nucleotide sequence similarity. It is important, therefore, to always use a holistic approach when clustering phages.

3.1.5 Restriction Fragment Pattern Analysis

Fifteen restriction enzymes were considered for use in creating restriction digests for bacteriophage isolates. From those fifteen, six were selected for the panel used for this study. These enzymes consist of HaeIII, KpnI, PmlI, SacI, SfiI, and StyI. Selection was based on the enzyme’s ability to produce fragments of an appropriate size for analysis.

A master mix was prepared by combining 500 ng of DNA and needed components for digestion according to protocols supplied with enzymes of interest (New England Biolabs). Samples were incubated at appropriate conditions in a T100 Thermal Cycler (BioRad) and then placed on ice until ready to visualize.

Digestion fragments were visualized by combining the entire contents of each digestion reaction tube with an appropriate volume of loading dye (bromophenol blue and xylene cyanol FF) and loading into a separate well of a 0.8% (w/v) agarose gel containing ethidium bromide. Agarose gels were approximately 7.5 cm wide and 10 cm long. The gel was then subjected to 80 V for approximately 1.5 hours and then visualized at 302 nm in a gel documentation system. Images were captured and used for further analysis.

Based on patterns in the resulting fragments from restriction enzyme digestion, new, unsequenced bacteriophage isolates were preliminarily grouped into provisional “clusters.” Genome size estimations are also performed using digests that contain fragments that range exclusively from 500-10,000 base pairs in length. This is also helpful in assessing potential cluster designations.

3.1.6 Phage Visualization via TEM

Phage isolates were structurally characterized by analysis of images obtained through transmission electron microscopy (TEM). Phage lysates were applied to a 400 Mesh, support film, carbon coated EM grid (Ted Pella) by “floating” the grid on a 10 μL drop of lysate. The sample was then stained and washed using a filtered 1% (w/v) uranyl acetate solution and molecular grade water according to proper procedure. Samples were then observed on an FEI Tecnai F20 scanning/transmission electron microscope, a field emission 200 kV microscope with a high brightness field emission electron gun. Resulting images were analyzed via ImageJ (https://imagej.nih.gov/ij/) and characterized by observing the gross morphology of the phage isolate and measurement of the (1) tail length and (2) capsid dimensions.

3.2 Results and Discussion

Much of the discussion below is framed around how the genometric diversity measured among the 45 *Streptomyces* phages here compares to phages that infect other hosts outside of the genus *Streptomyces*. Because of (1) the relatedness of the hosts, and (2) the expanded host range of many actinobacteriophages, it is likely that many of the phages here have access to a common gene pool with phages that infect hosts outside of the *Streptomyces* genus. As such, a comparison with other actinophage groups which have been studied is likely an appropriate frame for the discussion of genometric diversity found within those phages infecting the host *S. griseus* ATCC 10137. Other groups of phages included here include those that infect: (1) *M. smegmatis* mc^{2}155, 60 phages [55]; (2) *Arthrobacter sp.* ATCC 21022, 46 phages [56]; (3) *Gordonia spp.*, 79 phages [57] and *Microbacterium foliorum* NRRL B-24224 SEA, 67 phages (genome sequences retrieved from phagesdb.org).

3.2.1 Genometrics of the 45 Phages

Genometrics evaluated below include the (1) genome length, (2) G+C content, (3) characterization of genome termini, (4) number of predicted genes, and (5) number of tRNAs and mtRNAs. A complete summary of the information reported and discussed in the following sections may be found at the end of the chapter.

3.2.1.1 Genome Length and G+C Content

As illustrated in Table 3.2, genome lengths range considerably, from 40,785 bp (Raleigh) to 133,886 bp (StarPlatinum) and have an average genome length of 73,654 bp. G+C content also varies, spanning a range from 47.1% (Comrade) to 71.8% (Raleigh). The average G+C content of the reported phages is 60.9%.

Table 3.2: Genome lengths and G+C content as reported for 45 phages that infect *S. griseus* ATCC 10137. Phages are ordered from smallest genome (40785 bp, Raleigh) to largest (133886 bp, StarPlatinum).

Phage Name	Genome Length (bp)	% GC	Phage Name	Genome Length (bp)	% GC
Raleigh	40785	71.8	DrGrey	56076	59.5
Percastrophe	45999	59.7	OlympicHelado	56189	59.5
ToriToki	46077	59.7	YDN12	56528	69.2
Romero	46079	59.7	BabyGotBac	57165	69.2
Immanuel3	46094	59.6	TP1604	57168	69.2
HaugeAnator	46135	59.6	Salete	57243	69.2
ZooBear	46135	59.7	BayC	57243	69.2
Aaronocolus	49562	66.2	Maih	57256	69.3
Caliburn	49949	66.2	Xkcd426	64477	68.8

Phage Name	Genome Length (bp)	% GC	Phage Name	Genome Length (bp)	% GC
BryanRecycles	50066	65.9	Wentworth	68260	64.1
Jash	50066	65.9	Annadreamy	125726	47.6
Izzy	50113	65.9	Comrade	129015	47.1
Lorelei	50558	65.8	SparkleGoddess	129742	47.1
Eddasa	50605	65.9	Blueeyedbeauty	130473	47.9
Hydra	50727	66.2	NootNoot	131086	50.2
Rana	50980	65.8	Starbow	131427	49.5
Nabi	51127	65.8	Karimac	131909	49.4
Crosby	54036	68.3	Wofford	133007	47.7
UNTPL	54495	68.3	LukeCage	133195	49.0
LazerLemon	54798	68.1	Paradiddles	133486	49.5
JackieB	54912	68.2	Samisti12	133710	49.9
Henoccus	55137	68.2	StarPlatinum	133886	49.5
Spectropatronm	55707	59.5			

The average genome size of the *Streptomyces* phages reported here (73654 ± 35554 bp) is closer to the reported average genome size for 60 phages that infect *Mycobacterium* (72583 ± 32488 bp) than to that of 79 *Gordonia* phages, 67 *Microbacterium* phages, and 46 *Arthrobacter* phages, reported at 59939 ± 18961 bp, 43015 ± 12453 bp, and 43485 ± 18088 bp, respectively. Further, and as Figure 3.7 illustrates, a majority (33 genomes, ~73%) of the 45 sequenced phages have genome lengths that range from 40,785 bp (Raleigh) to 68,260 bp (Wentworth). The remaining 12 phages have genome lengths ranging from 125,726 bp (Annadreamy) to 133,886 bp (StarPlatinum), with no reported genome lengths in the 57,466 bp gap between 68,260 bp and 125,726 bp.

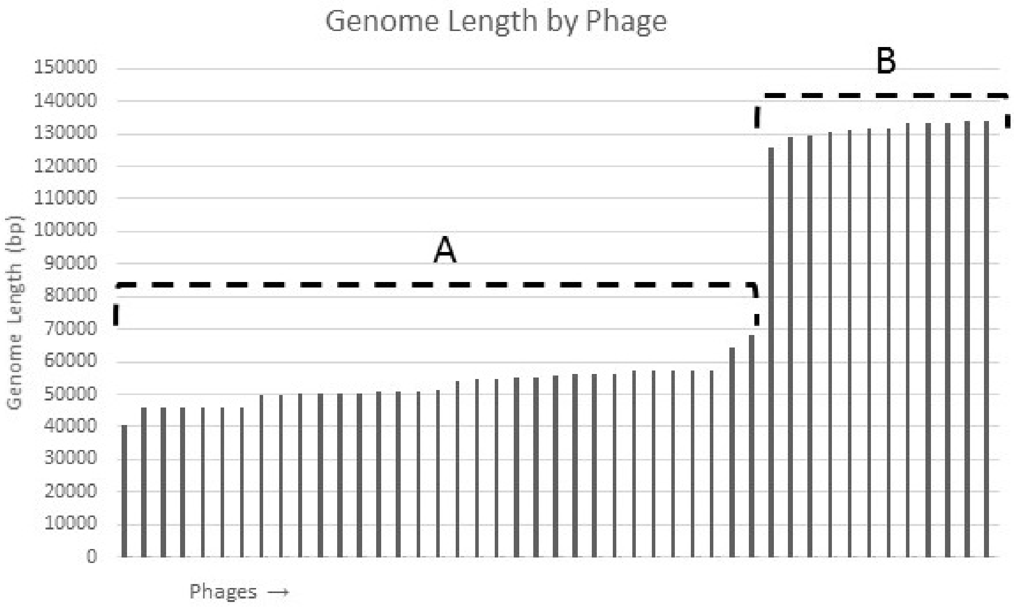


Figure 3.7: Distribution of genome lengths across 45 phages that infect *S. griseus* ATCC 10137. The x-axis represents phages, ordered from smallest to largest (see Table 3.2). Two distinct size groups emerge: group A, with genomes ranging from 40785 bp to 68260 bp, and group B, with genomes ranging from 125726 bp to 133886 bp. A clear gap between the two groups is apparent.

With respect to other actinobacteriophages that have been studied in groups this size, the distribution of genome sizes among these 45 *Streptomyces* phages shows distinct differences from its actinobacteriophage counterparts. As shown in Figure 3.8, the distribution of genome sizes among the *Streptomyces phages* here most closely resembles the distribution of *Mycobacterium* phages, as well, the latter showing two groups of genome sizes, i.e. one ranging from around 40,000 bp to around 80,000 bp, and a second, larger group ranging from around 155,000 bp to around 165,000 bp.

It should be noted here that the second group of *Mycobacterium* phages above consists of phages with myoviral morphology, a morphology that is absent from the *Streptomyces* phages in this study (discussed below). Also note the presence of a single *Mycobacterium* phage (Omega, genome length = 110,865 bp) in the gap between the two groups of genome sizes in the *Mycobacterium* phage, a feature that is also missing in the distribution of genome sizes among the *Streptomyces* phages examined here. The other three groups of actinobacteriophages shown in Figure 3.8 appear to also have two distinct groups of genome sizes, however, the smaller groups are less than 20,000 bp in size and the larger groups begin at around 40,000 bp and have a maximum size around 100,000 bp (*Gordonia* and *Microbacterium*) and 70,000 bp (*Arthrobacter*). Additionally, both standard deviations and the distributions shown in Figure 3.8 suggest that the genome lengths of the latter three groups are more tightly distributed around their respective average lengths than those in either the *Streptomyces* phages or the *Mycobacterium* phages. Although patterns are beginning to emerge, it is unclear what the determinates of genome length are.

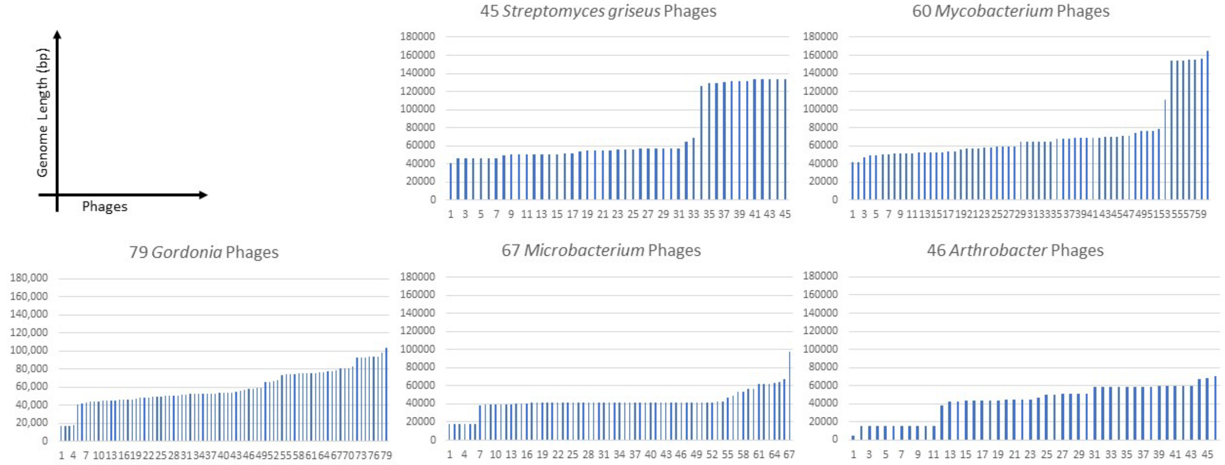


Figure 3.8: Distribution of genome lengths among phage isolates from other Actinomycete hosts. Included are phage groups from the genera *Streptomyces*, *Mycobacterium*, *Gordonia*, *Microbacterium*, and *Arthrobacter*. The axis labels for all distributions is located in the top left corner.

As reported above, the average G+C content for the reported phages is 60.9%, about 11% lower than the isolation host, *S. griseus* 10137 (reported around 72%). This is a low average G+C when compared to the findings among the *Mycobacterium* phages, i.e. an average G+C of 63.7% compared to the host's 63%; the *Arthrobacter* phages, i.e. an average G+C of 59.1% compared to the host's 63.4%; the *Gordonia* phages, i.e. an average G+C of 62.7% compared to the host's 67.8%; and the Microbacterium phages, i.e. an average G+C of 64.5% compared to the host's 67%. However, variation among the phages isolated on each host does exist, and within the range exhibited by those phage infecting *Streptomyces*. Of the *Mycobacterium* phages studied, G+C ranged from 59-69%. Of the *Arthrobacter* phages, the G+C ranged from 45.1-68.4%. Of the *Gordonia* and *Microbacterium* phages, G+C ranged from 47-68.8% and from 58.3 to 69.7%, respectively. Further, it should be noted that the group of *Streptomyces* phages presented here comprised 11 phages with G+C lower than 50%, a percentage of total phages observed that was not seen in any other group of phages.

Although the exact reason for the variation in G+C both (1) between the *Streptomyces* phages observed, and (2) among the *Streptomyces* phages and the other groups of actinobacteriophages reported here remains unknown, one potential explanation is that some of the phages (both within and among each group) have a substantially different host range than the others. A G+C content that varies significantly from the isolation host could indicate that the phage has another "preferred host," e.g. a host that they most frequently infect in a natural environment, and that the G+C content of the phage more closely reflects the G+C content of that preferred host. However, a substantially different G+C content does not preclude infection between phage and host.

3.2.1.2 Genome Termini

As shown in Table 3.3, the nature of the genome termini varies across the 45 sequenced phages. Genome termini of all of the 45 phages here were readily ascertainable from the genome assemblies. Of the 45 phages here, 18 appear to have Direct Terminal Repeats (DTR), 13 appear to have 3’ single-stranded DNA extensions, and 14 of them show evidence of being circularly permuted.

Table 3.3: Nature of genome termini for 45 phages that infect *S. griseus* ATCC 10137. All termini were predicted via analysis of genome assemblies in Consed. All phages are sorted according to nature of termini, e.g. cohesive (COS) ends (9 bp, 11 bp), circularly permuted and terminally redundant (circ perm), and direct terminal repeats (DTR). Length of DTRs are reported in parentheses.

Phage Name	Genome Termini	overhang	Phage Name	Genome End Structure
OlympicHelado	9 bp cos	CGCCCGCCT	JackieB	Circ Perm
DrGrey	9 bp cos	CGCCCGCCT	Percastrophe	DTR (264)
Spectropatronm	9 bp cos	CGCCCGCCT	ToriToki	DTR (264)
Lorelei	11 bp cos	CGGCCAGTCAT	Wentworth	Circ Perm
Aaronocolus	11 bp cos	CGGGCAGTGAT	LazerLemon	Circ Perm
BryanRecycles	11 bp cos	CGGGCAGTGAT	Romero	DTR (264)
Eddasa	11 bp cos	CGGGCAGTGAT	HaugeAnator	DTR (274)
Jash	11 bp cos	CGGGCAGTGAT	ZooBear	DTR (274)
Hydra	11 bp cos	CGGGCAGTGAT	Immanuel3	DTR (275)
Caliburn	11 bp cos	CGGGCAGTGAT	Comrade	DTR (734)
Izzy	11 bp cos	CGGGCAGTGAT	SparkleGoddess	DTR (734)
Nabi	11 bp cos	CGGCCAGTCAT	Blueeyedbeauty	DTR (788)
Rana	11 bp cos	CGGCCAGTCAT	Annadreamy	DTR (789)
Raleigh	Circ Perm		Samisti12	DTR (10666)
BabyGotBac	Circ Perm		Paradiddles	DTR (10778)
Maih	Circ Perm		NootNoot	DTR (10787)
Salete	Circ Perm		Wofford	DTR (11214)
BayC	Circ Perm		StarPlatinum	DTR (12199)
TP1604	Circ Perm		LukeCage	DTR (12291)
YDN12	Circ Perm		Starbow	DTR (12579)
Xkcd426	Circ Perm		Karimac	DTR (12590)
Crosby	Circ Perm			
Henoccus	Circ Perm			

Phage Name	Genome Termini	overhang	Phage Name	Genome End Structure
UNTPL	Circ Perm			

DTRs range from 264 bp (Percastrophe, ToriToki, and Romero) to 12,590 bp (Karimac). As illustrated in Figure 3.9, three groups of DTRs emerge based on length. The first group comprises phages Percastrophe, ToriToki, Romero, HaugeAnator, ZooBear, and Immanuel3 and ranges from 264 – 275 bp in length. The second group comprises phages Comrade, SparkleGoddess, Blueeyedbeauty, and Annadreamy and ranges from 734 – 789 bp in length. The third and final group comprises phages Samisti12, Paradiddles, NootNoot, Wofford, StarPlatinum, LukeCage, Starbow, and Karimac and ranges from 10,666 – 12,590 bp in length.

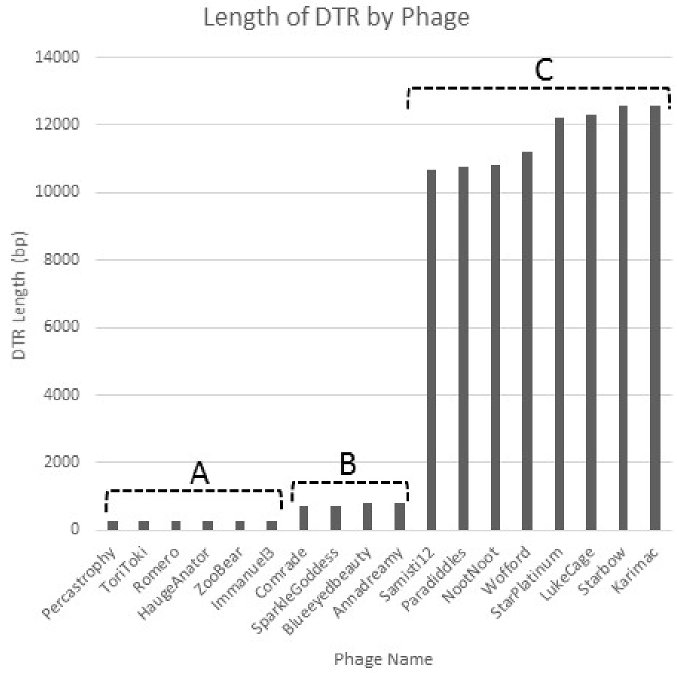


Figure 3.9: Distribution of DTR lengths among 18 phages that infect *S. griseus* ATCC 10137. Each of the two known groups of DTRs are represented, i.e. short DTRs (groups A and B) and long DTRs (group C). The short DTRs may be further divided into two groups, i.e. group A (264-275 bp) and group B (734-789 bp).

The phage assemblies that indicate 3' single-stranded DNA extensions can be divided into two groups based on the length of those extensions. The first group comprises phages Spectropatronm, DrGrey, and OlympicHelado and have extensions that are 9 bp long. The second group comprises phages Aaronocolus, Caliburn, BryanRecycles, Jash, Izzy, Lorelei, Eddasa, Hydra, Rana, and Nabi and have extension that are 11 bp long. As illustrated in Table 3.3, all phages with a 9 bp extension have the same nucleotide sequence comprising that extension, i.e. CGCCCGCCT.

As illustrated in both Table 3.3 and Figure 3.10, the phages having the 11 bp extension have largely the same sequence with variations in the sequence occurring at positions four and nine. Of the 10 phages in this group, a majority, i.e. seven, have a G at both the four and nine positions. Phages Lorelei, Nabi, and Rana all have a C at both the four and nine positions.

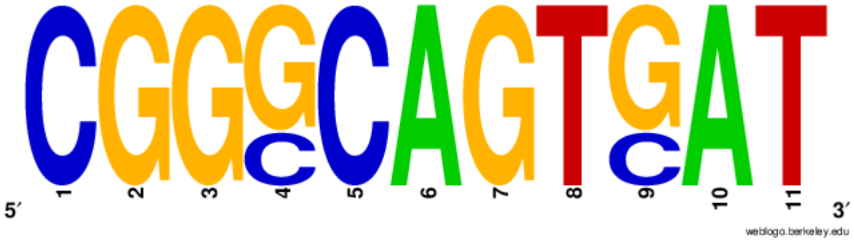


Figure 3.10: WebLogo representation of overhang sequence among 10 phages infecting *S. griseus* ATCC 10137. All nucleotide identities are the same except in two positions, i.e. positions four and nine. Graphic generated at https://weblogo.berkeley.edu/logo.cgi.

With respect to the distribution of terminus types, the *Streptomyces* phages here have a relatively even distribution of termini when compared to other groups of actinophages. About 24% have short direct terminal repeats (Short DTR), i.e. those that are several hundred bp long; about 19% have long direct terminal repeats, i.e. those that are 1000 bp or longer; about 31% have either 3' or 5' sticky (cos) ends, and about 26% are circularly permuted and terminally redundant. This is in contrast to all other groups of phages, shown in Figure 3.11, who have a clearly predominate terminus among their phage isolates, that terminus being either COS ends (*Mycobacterium* phages, *Arthrobacter* phages, and *Gordonia* phages) or a genome which is circularly permuted and terminally redundant (*Microbacterium* phages).

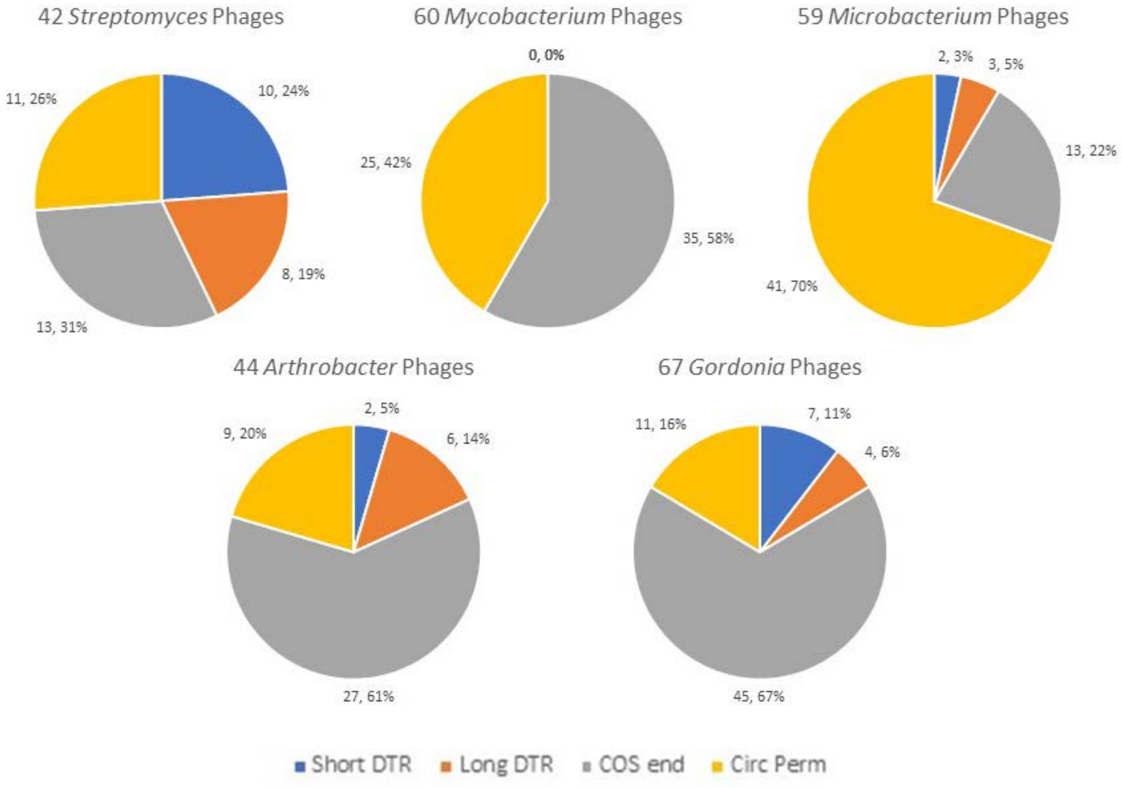


Figure 3.11: Distributions of genome terminus types across phages that infect Actinomycete hosts.

3.2.1.3 ORFs and Other Features

Annotation of the 45 phages reveals a total of 5243 open reading frames (ORFs), 621 tRNAs, and 12 tmRNAs. As shown in Table 3.4, the number of ORFs varies across the phages, ranging from 52 (Raleigh) to 251 (StarPlatinum). Without surprise, and similar to the distribution in genome lengths, there is a large gap between those phages having

88 ORFs or less and those having 216 ORFs or more, with a lone phage (Wentworth) with a total of 103 ORFs in the middle.

Table 3.4: Number of predicted ORFs, tRNAs, and tmRNAs for 45 phages that infect *S. griseus* ATCC 10137. The phages are arranged in order from the lowest number of predicted ORFs (Raleigh) to the highest (StarPlatinum).

Phage	# ORFs	# tRNA	# tmRNA	Phage	# ORFs	# tRNA	# tmRNA
Raleigh	52			Xkcd426	78		
Immanuel3	60	17		DrGrey	80		
HaugeAnator	63	22		LazerLemon	81		
Percastrophe	64	22		UNTPL	81		
Romero	64	22		Crosby	82		
ToriToki	64	22		Henoccus	82		
ZooBear	64	22		JackieB	82		
Lorelei	65			Spectropatronm	84		
Maih	70			OlympicHelado	88		
Salete	70			Wentworth	103		
BayC	71			Paradiddles	216	46	1
TP1604	71			NootNoot	221	45	1
YDN12	71			Samisti12	227	44	1
BabyGotBac	72			Comrade	229	34	1
Caliburn	72			Annadreamy	230	35	1
Aaronocolus	73			SparkleGoddess	232	35	1
Izzy	74			Wofford	235	45	1
BryanRecycles	75			Starbow	238	44	1
Jash	75			Blueeyedbeauty	240	37	1
Eddasa	76			Karimac	241	44	1
Hydra	76			LukeCage	248	41	1
Nabi	76			StarPlatinum	251	44	1
Rana	76			TOTALS	5243	621	12

Eighteen phages have predicted tRNAs and fall into two groups, i.e. those with 17-22 tRNAs (six phages) and those with 34-46 tRNAs (12 phages). The phages with the higher number of tRNAs also have a single predicted tmRNA as well. It should be noted that those phages have genomes greater than 100 kbp in length. This is a trend mirrored in the *Mycobacterium* phages. Of the 60 *Mycobacterium* phages discussed throughout this chapter, 19 have predicted tRNAs and, of those 19, seven have predicted tmRNAs. As observed in the *Streptomyces* phages here, there is a disparity between two groups of *Mycobacterium* phages that carry tRNAs, i.e. a group of smaller (< 100 kbp) phages that carry 1-2 tRNAs and a group of larger (>100 kbp) phages that carry between 35 and 41 tRNAs. Almost all of those that have the higher number of tRNAs also carry a single tmRNA. All tRNA and tmRNA predictions were made using Aragorn [58].

3.2.2 Clustering via Dot Plot Analysis and ANI

The first two methods of clustering phages include dot plot analysis and ANI comparisons. Figure 3.12 illustrates a dot plot generated using Gepard, wherein the individual genome sequences, in FASTA format, for each of the 45 phages were concatenated into a single file and placed onto each of the axes of the plot. As discussed above, clustering occurs where there is evident sequence similarity that spans greater than 50% of the smaller of the two genomes. Dot plot examination of the 45 phages included here readily reveals seven distinguishable clusters based on contiguous nucleotide identity that is observable as lines where the sequences of two phages converge on the dot plot. Two phages, Raleigh and Wentworth, appear to have no observable contiguous nucleotide identity that spans greater than 50% of their respective genomes.

These readily distinguishable clusters comprise (1) Cluster BD, further comprising phages Aaronocolus, BryanRecycles, Caliburn, Eddasa, Hydra, Izzy, Jash, Lorelei, Nabi, and Rana; (2) Cluster BE, further comprising phages Karimac, LukeCage, NootNoot, Paradiddles, Samisti12, Starbow, StarPlatinum, and Wofford; (3) Cluster BF, further comprising phages HaugeAnator, Immanuel3, Percastrophe, Romero, ToriToki, and ZooBear; (4) Cluster BG, further comprising phages BabyGotBac, BayC, Maih, Salete, TP1604, Xkcd426, and YDN12; (5) Cluster BH, further comprising phages Crosby, Henoccus, JackieB, LazerLemon, and UNTPL; and (6) Cluster BK, further comprising phages DrGrey, OlympicHelado, and Spectropatronm. As noted above, two phages (Raleigh and Wentworth) are treated as individual phages here, although it should be noted again that they belong to clusters comprising phages isolated on hosts other than *S. griseus*.

The second method of clustering used was the comparison of average nucleotide identities (ANI) and using them to confirm, reject, or clarify relationships that were readily observable via dot plot. As illustrated in Figures 3.13 through 3.19, all ANI values agree well with the cluster assignments made through dot plot analysis. As a whole, the most closely related cluster of phages appears to be Cluster BF (Figure 3.15), with ANI values ranging from 0.9835 to 0.9984, indicating a high degree of shared nucleotide identity between all phages in that cluster.

The least-related cluster of phages appears to be Cluster BE (Figure 3.14), with ANI values ranging from 0.7059 to 0.9715. Indeed, a closer look reveals that two groups of phages emerge based on both ANI values and analysis of contiguous nucleotide sequence similarity on a dot plot. The first group of Cluster BE phages, i.e. NootNoot, Paradiddles, and Samisti12, appear to be more closely related to each other than to the other phages in the cluster. ANI values for these three phages range from 0.8543 (Samisti12 and NootNoot) to 0.9715 (Paradiddles and NootNoot).

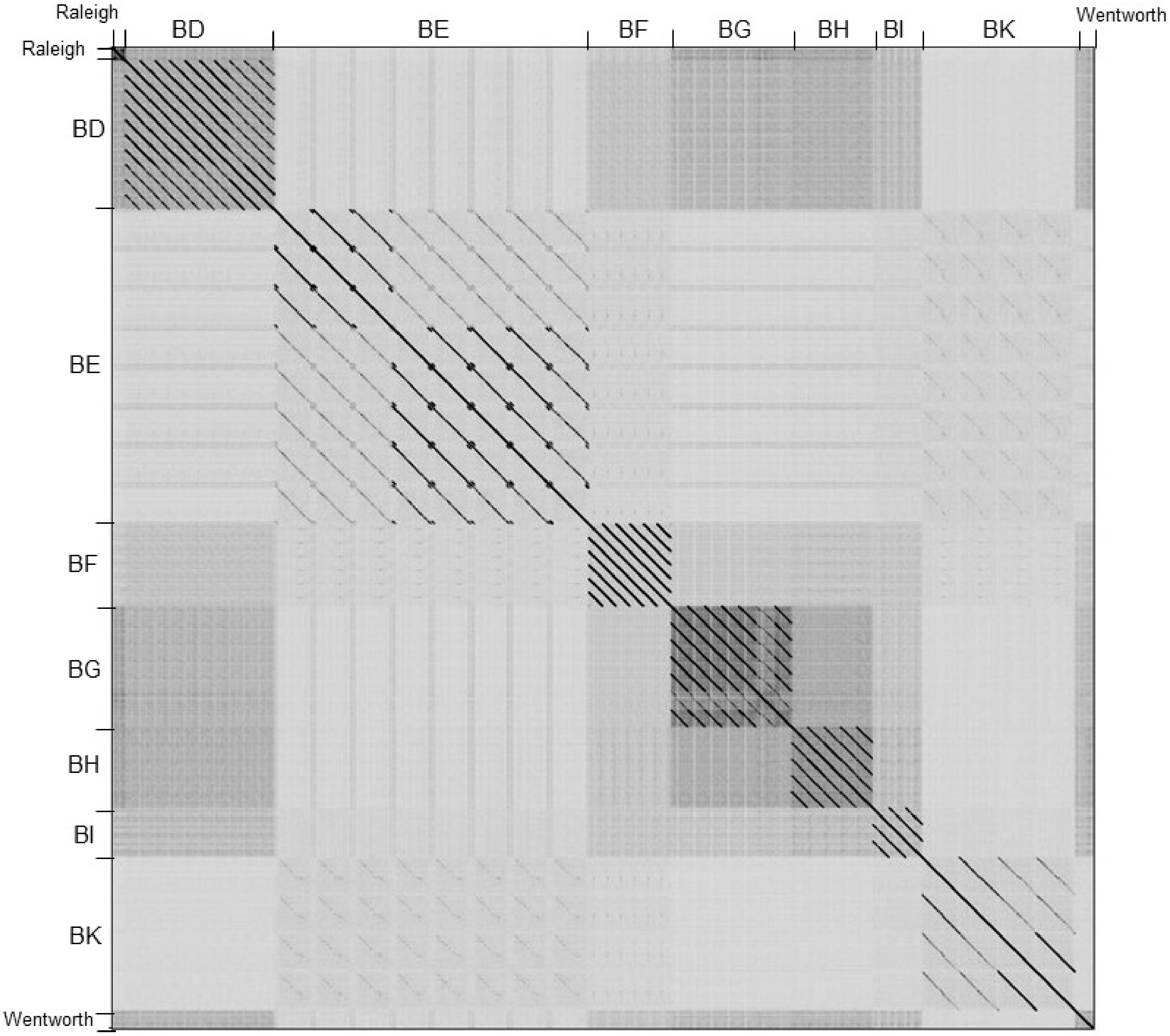
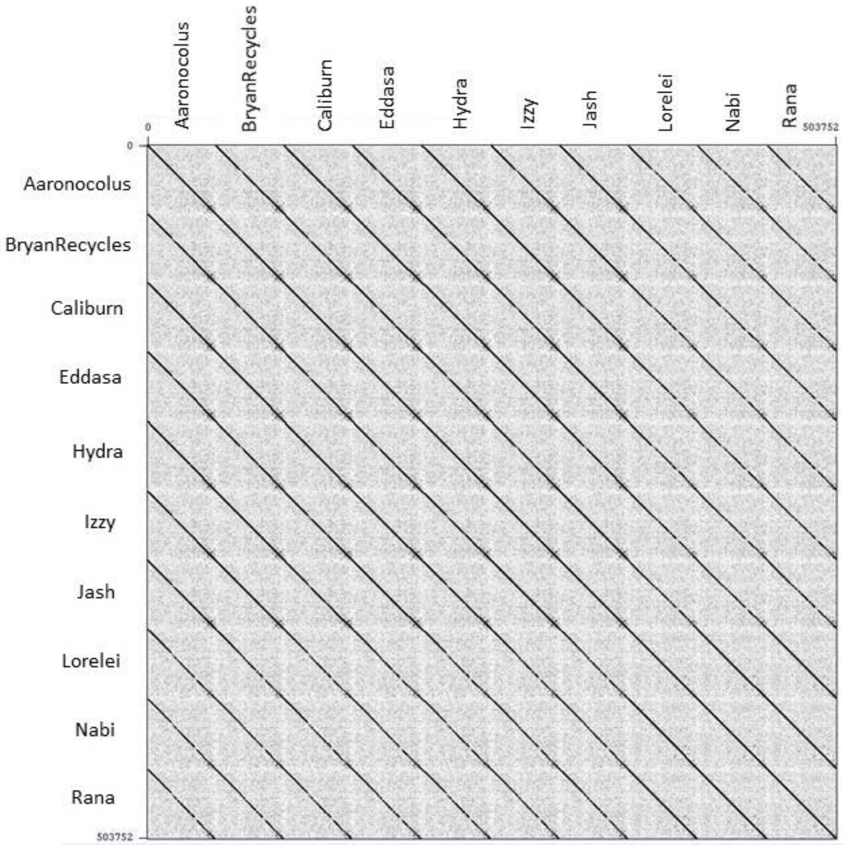
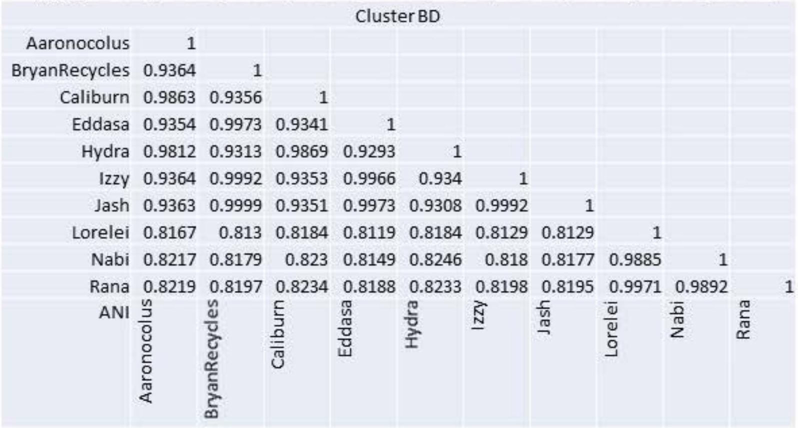


Figure 3.12: Dot plot of the genome sequences of 45 *S. griseus* phages generated using Gepard. The plot reveals seven identifiable clusters (BD, BE, BF, BG, BH, BI, and BK) and two singleton phages (Raleigh and Wentworth).

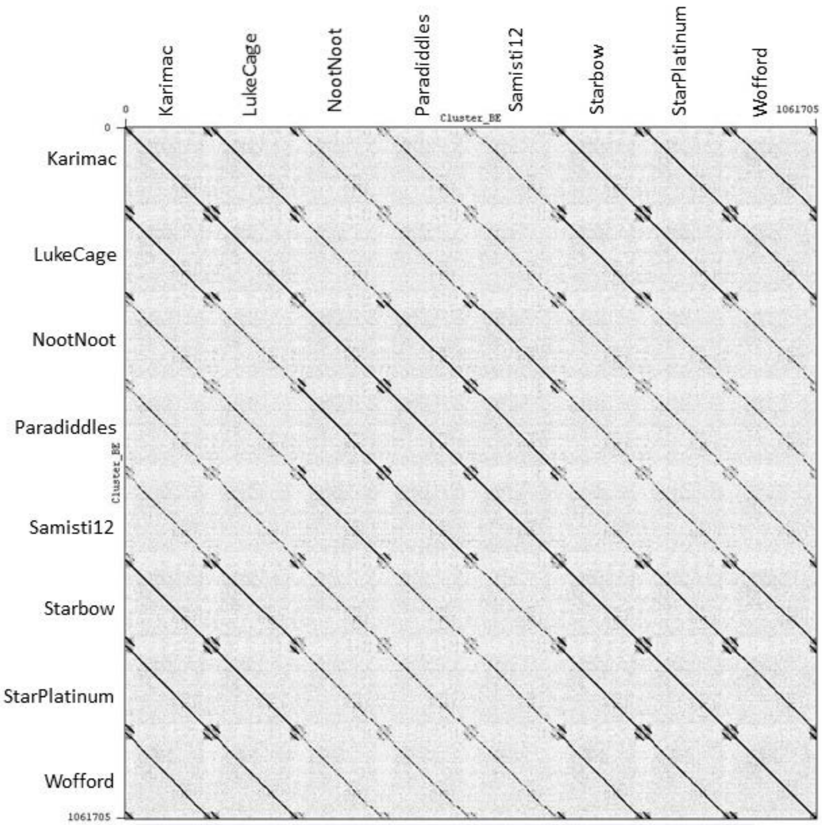


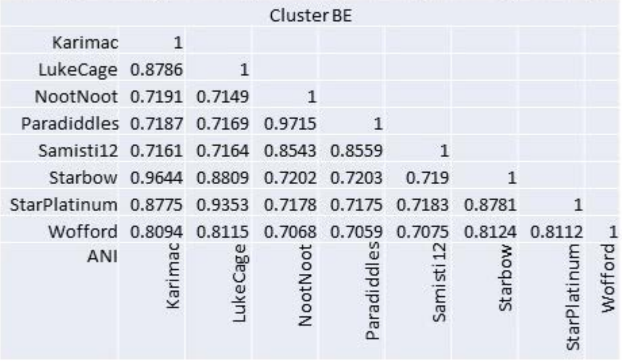


Cluster BD

Aaronocolus	1									
BryanRecycles	0.9364	1								
Caliburn	0.9863	0.9356	1							
Eddasa	0.9354	0.9973	0.9341	1						
Hydra	0.9812	0.9313	0.9869	0.9293	1					
Izzy	0.9364	0.9992	0.9353	0.9966	0.934	1				
Jash	0.9363	0.9999	0.9351	0.9973	0.9308	0.9992	1			
Lorelei	0.8167	0.813	0.8184	0.8119	0.8184	0.8129	0.8129	1		
Nabi	0.8217	0.8179	0.823	0.8149	0.8246	0.818	0.8177	0.9885	1	
Rana	0.8219	0.8197	0.8234	0.8188	0.8233	0.8198	0.8195	0.9971	0.9892	1
ANI	Aaronocolus	BryanRecycles	Caliburn	Eddasa	Hydra	Izzy	Jash	Lorelei	Nabi	Rana

Figure 3.13: Dot plot and ANI table for Cluster BD phages. ANI values range from 0.8119 (Lorelei and Eddasa) to 0.999 (Jash and BryanRecycles), where those phages with the higher ANI value share a higher percentage of ANI across their genomes.

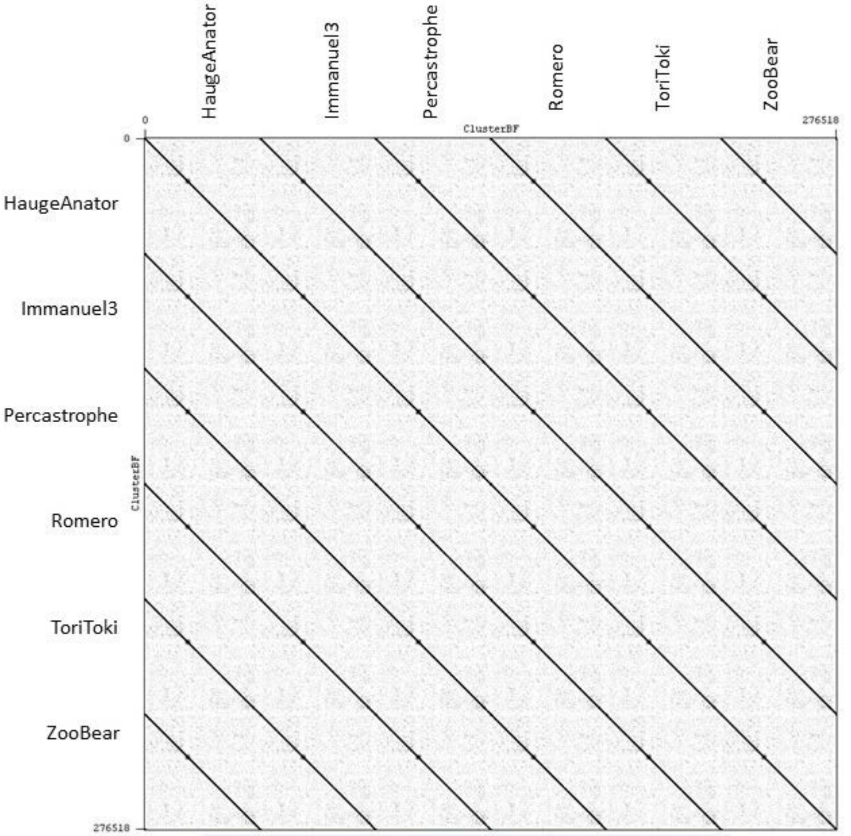


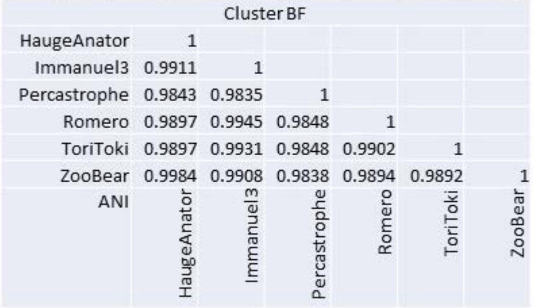


Cluster BE

ANI	Karimac	LukeCage	NootNoot	Paradiddles	Samisti12	Starbow	StarPlatinum	Wofford
Karimac	1							
LukeCage	0.8786	1						
NootNoot	0.7191	0.7149	1					
Paradiddles	0.7187	0.7169	0.9715	1				
Samisti12	0.7161	0.7164	0.8543	0.8559	1			
Starbow	0.9644	0.8809	0.7202	0.7203	0.719	1		
StarPlatinum	0.8775	0.9353	0.7178	0.7175	0.7183	0.8781	1	
Wofford	0.8094	0.8115	0.7068	0.7059	0.7075	0.8124	0.8112	1

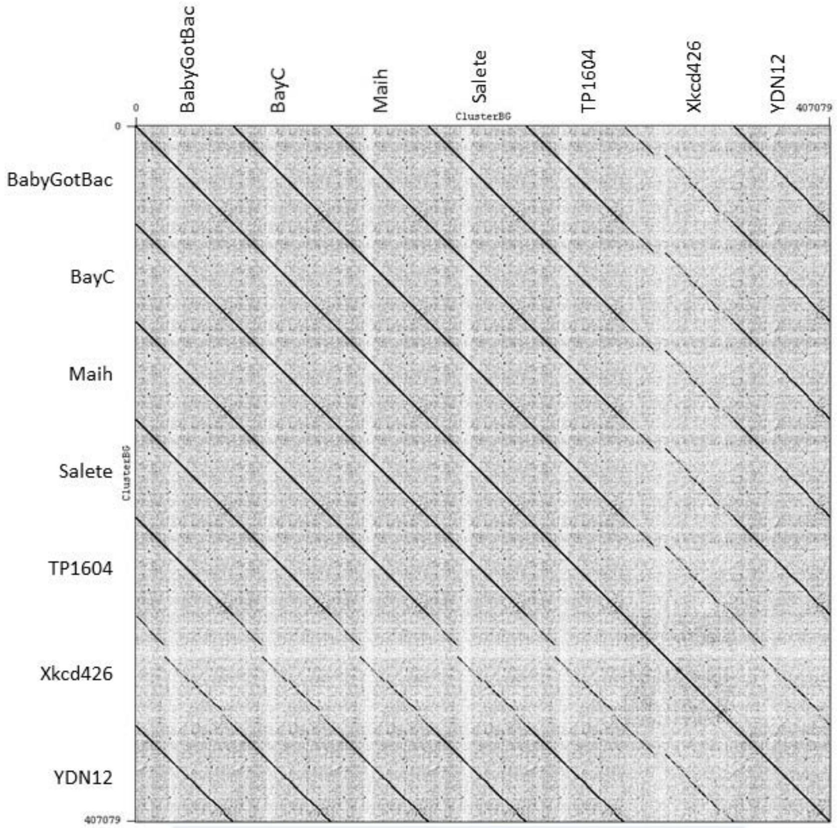
Figure 3.14: Dot plot and ANI table for Cluster BE phages. ANI values range from 0.7059 (Wofford and Paradiddles) to 0.9715 (Paradiddles and NootNoot). Note the two groups of phages that emerge based on ANI value, i.e. each phage shares low ANIs (0.70-0.72) with some cluster members, while sharing moderate-to-high ANIs (>0.8) with others. Cluster BE is likely a candidate for subclustering, discussed more in the text.

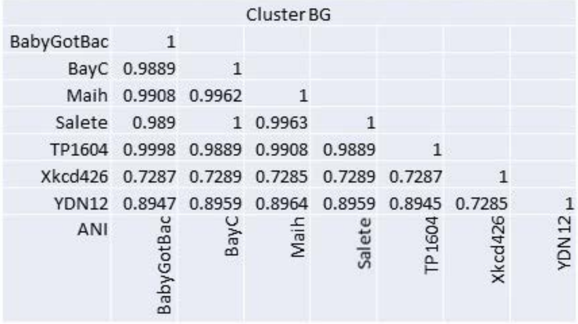




Cluster BF						
HaugeAnator	1					
Immanuel3	0.9911	1				
Percastrophe	0.9843	0.9835	1			
Romero	0.9897	0.9945	0.9848	1		
ToriToki	0.9897	0.9931	0.9848	0.9902	1	
ZooBear	0.9984	0.9908	0.9838	0.9894	0.9892	1
ANI	HaugeAnator	Immanuel3	Percastrophe	Romero	ToriToki	ZooBear

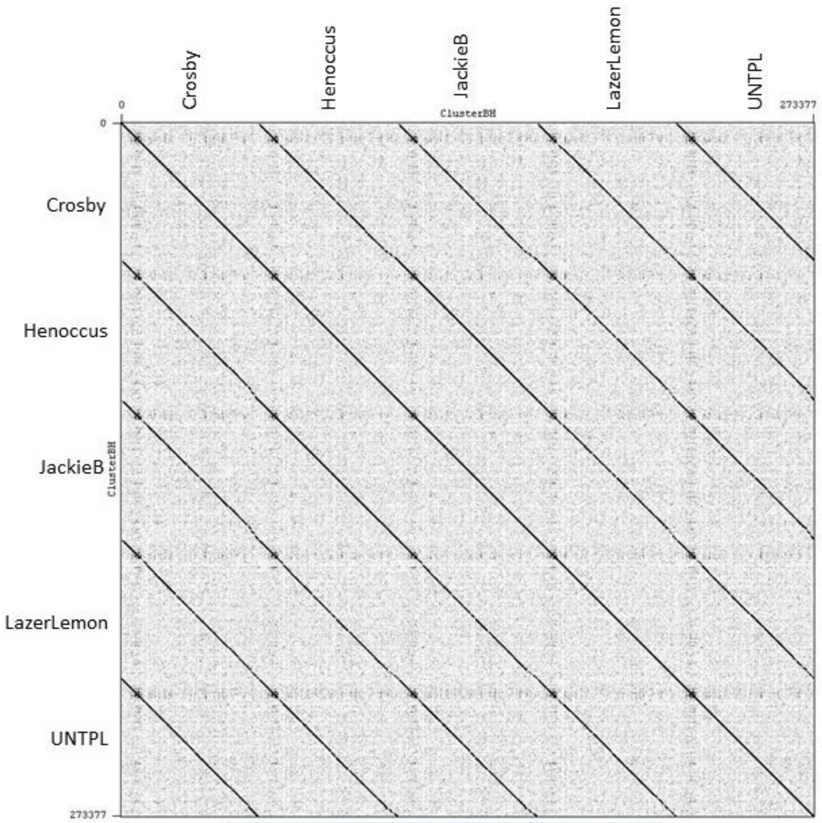
Figure 3.15: Dot plot and ANI table for Cluster BF phages. ANI values range from 0.9835 (Percastrophe and Immanuel3) to 0.9984 (ZooBear and HaugeAnator). Note that both the low- and high-end ANI values indicate that the phages in the Cluster are all highly related at the nucleotide level.

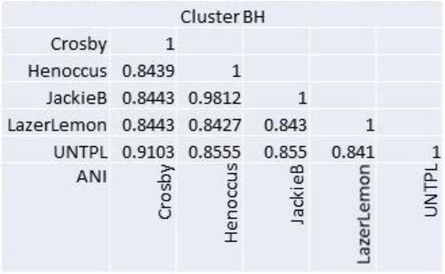




Cluster BG							
BabyGotBac	1						
BayC	0.9889	1					
Maih	0.9908	0.9962	1				
Salete	0.989	1	0.9963	1			
TP1604	0.9998	0.9889	0.9908	0.9889	1		
Xkcd426	0.7287	0.7289	0.7285	0.7289	0.7287	1	
YDN12	0.8947	0.8959	0.8964	0.8959	0.8945	0.7285	1
ANI	BabyGotBac	BayC	Maih	Salete	TP1604	Xkcd426	YDN12

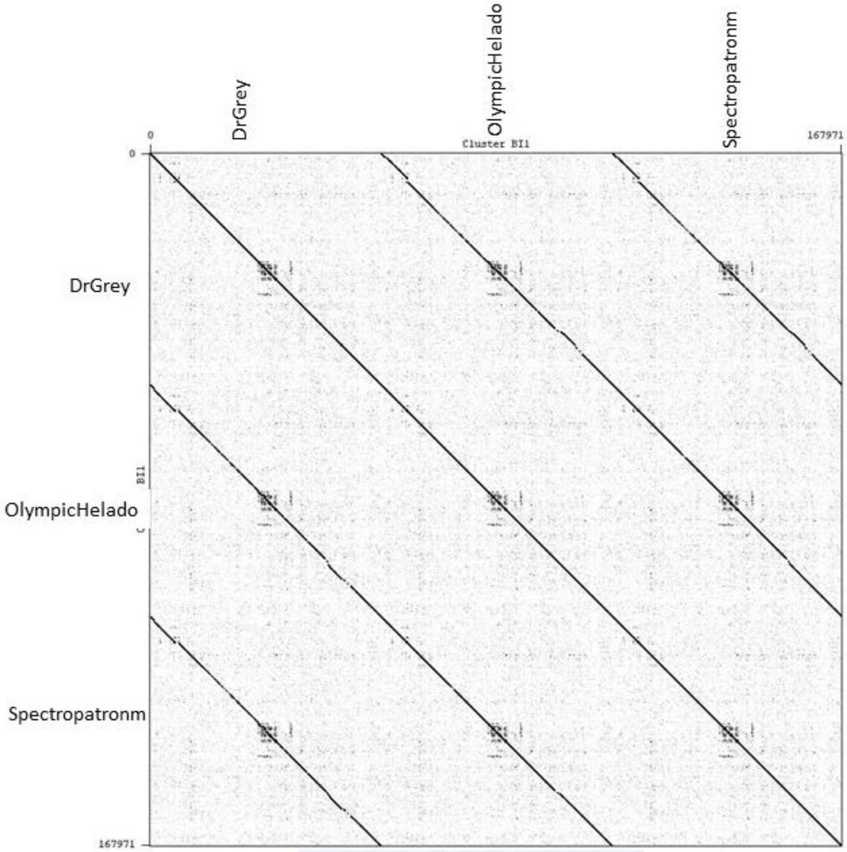
Figure 3.16: Dot plot and ANI table for Cluster BG phages. ANI values range from 0.7285 (Xkcd426 and YDN12, Xkcd426 and Maih) to 1.0 (Salete and BayC). Note that the ANI value of 1.0 is because the two phages (Salete and BayC) differ by only a single nucleotide. Also note that Xkcd426 has ANI values in the low range (0.72-0.73) with all other phages, which is discussed further in the text.

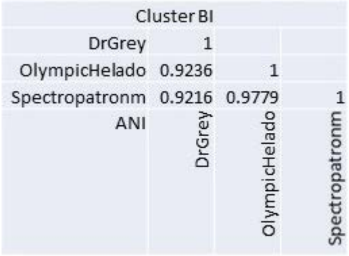




Cluster BH					
Crosby	1				
Henoccus	0.8439	1			
JackieB	0.8443	0.9812	1		
LazerLemon	0.8443	0.8427	0.843	1	
UNTPL	0.9103	0.8555	0.855	0.841	1
ANI	Crosby	Henoccus	JackieB	LazerLemon	UNTPL

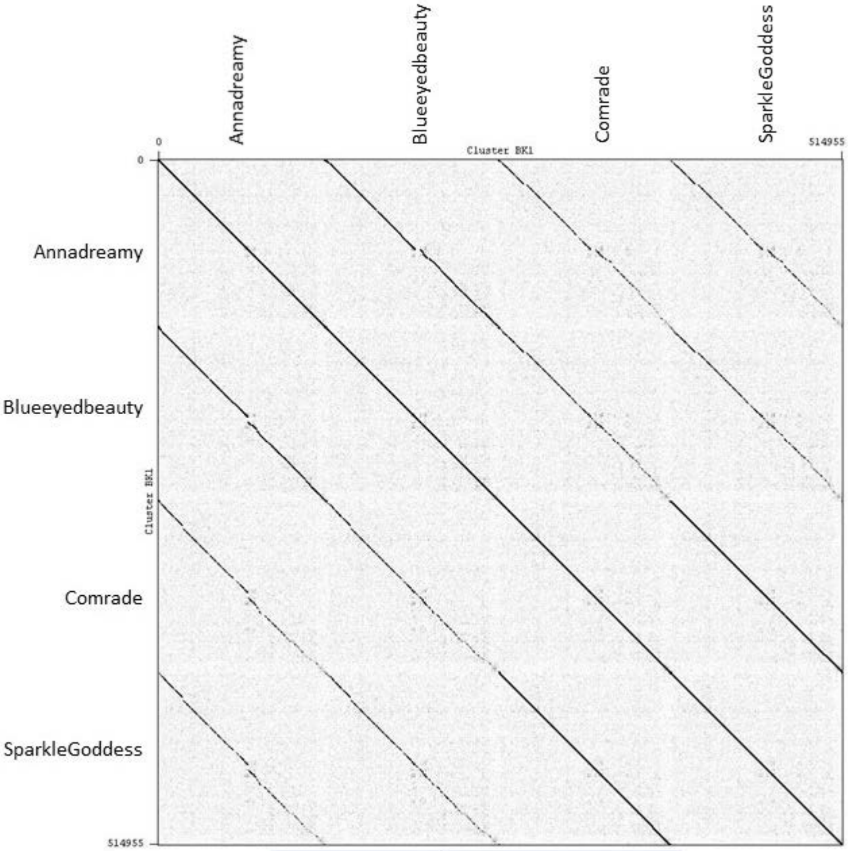
Figure 3.17: Dot plot and ANI table for Cluster BH phages. ANI values range from 0.841 (UNTPL and LazerLemon) to 0.9812 (JackieB and Henoccus).

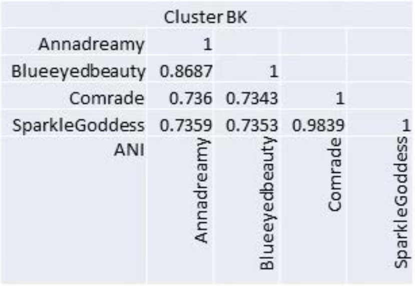




	Cluster BI		
DrGrey	1		
OlympicHelado	0.9236	1	
Spectropatronm	0.9216	0.9779	1
ANI	DrGrey	OlympicHelado	Spectropatronm

Figure 3.18: Dot plot and ANI table for Cluster BI phages. ANI values range from 0.9216 (Spectropatronm and DrGrey) to 0.9779 (Spectropatronm and OlympicHelado).





Cluster BK				
Annadreamy	1			
Blueeyedbeauty	0.8687	1		
Comrade	0.736	0.7343	1	
SparkleGoddess	0.7359	0.7353	0.9839	1
ANI	Annadreamy	Blueeyedbeauty	Comrade	SparkleGoddess

Figure 3.19: Dot plot and ANI table for Cluster BK phages. ANI values range from 0.7343 (Comrade and Blueeyedbeauty) to 0.9839 (Comrade and SparkleGoddess).

The second group of Cluster BE phages, i.e. Karimac, LukeCage, Starbow, StarPlatinum, and Wofford, also appear to be more closely related to each other than to the phages in the first group. ANI values among these five phages range from 0.8094 (Karimac and Wofford) to 0.9644 (Karimac and Starbow). When ANI values are compared across the two groups, i.e. group one against group 2, the ANI values range from 0.7059 (Paradiddles and Wofford) to only 0.7203 (Paradiddles and Starbow). The relationship among phages in this cluster is representative of a scenario where a dot plot analysis, shown here in Figure 3.14, indicates a weak relationship that spans large segments of the genomes and ANI values reflect a divide in the phages comprising the cluster, indicating that phages may be segregated into subclusters. Through the combined analysis of dot plot and ANI values, Cluster BE may be further divided into two subclusters, i.e. BE1, comprising phages NootNoot, Paradiddles, and Samisti12; and BE2, comprising phages Karimac, LukeCage, Starbow, StarPlatinum, and Wofford.

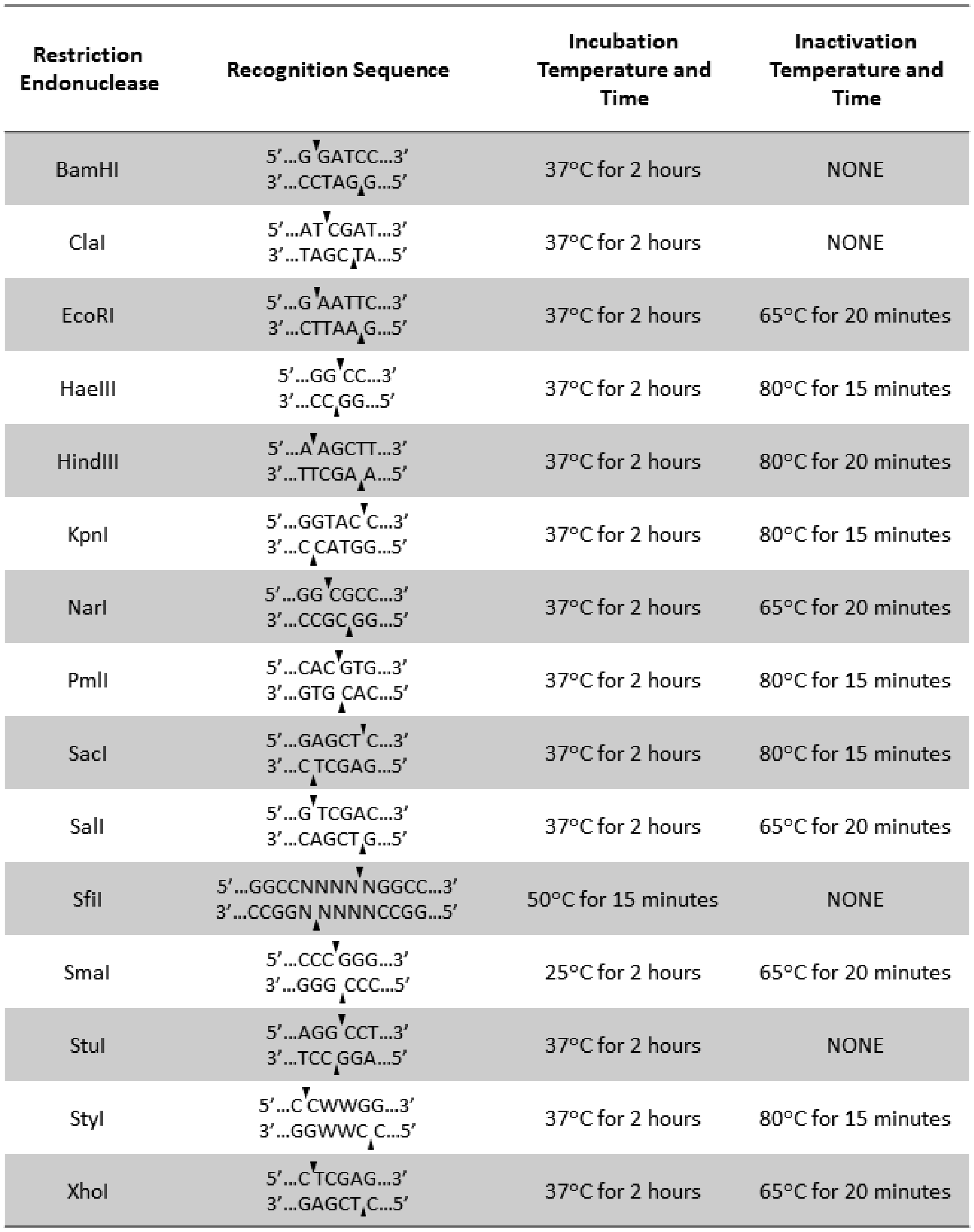
Note that Figure 3.16 illustrates a case where a single Cluster BG phage, Xkcd426, shows a weak relationship with the other phages in the cluster on both dot plot analysis and ANI values (ranging from 0.72-0.73). This case is distinguishable from the scenario discussed above for Cluster BE in that Xkcd426 is a lone phage indicating this relationship and phages may not be clustered (or subclustered) by themselves. It is reasonable to anticipate that when and if a Cluster BG phage is isolated and described that (1) shares a higher degree of ANI with Xkcd426, e.g. around 0.80 or higher, (2) shares a similar degree of ANI with the other phages in the cluster, e.g. around 0.72 or lower, and (3) shows a similar dot plot relationship with the phages in the cluster, i.e. one that is weak, but spans large segments of the genome; then the newly isolated and described phage will be subclustered with Xkcd426. For now, Xkcd426 remains clustered with the BG phages without a subcluster designation.

3.2.3 Pre-Sequencing Cluster Prediction

Genome sequencing has drastically decreased in price over the past decade. Advances in technology, coupled with the ability to “barcode” individual samples with unique nucleotide-sequence adapters, have made sequencing more accessible to researchers in every field. The relatively small size of phage genomes additionally increases the number of samples that may be sequenced in a single run. Nonetheless, sequencing for a single phage genome still currently costs around $200, and sequencing every phage isolated becomes cost-prohibitive.

In 2012, isolation of phages on *Streptomyces* hosts expanded outside of the Hughes lab and undergraduate students began to add to the numbers of isolates in increasing numbers. Because of (1) the per-sample cost associated with sequencing, and (2) a bottleneck in the assembly/finishing/annotation pipeline; a method for screening isolates for potential cluster assignments needed to be developed. Borrowing from the work done with *Mycobacterium* phages, an analysis of the fragment patterns of restriction digests was the most obvious route to achieve this. Unfortunately, when the established enzymes used for the *Mycobacterium* phages, i.e. BamHI, ClaI, EcoRI, HaeIII, and HindIII, were tried, the results were unusable. generally speaking, the enzymes would either (1) produce no fragments, or (2) produce exclusively fragments smaller than 500 bp and, therefore, indistinguishable on a gel under the conditions used for this study.

Table 3.5: Restriction endonucleases (REs) tested for a screening panel, used to predict cluster affiliation among newly isolated phages that infect *S. griseus* ATCC 10137. Shown above are the (1) enzyme, (2) recognition sequence, and (3) digestion conditions for each tested RE.

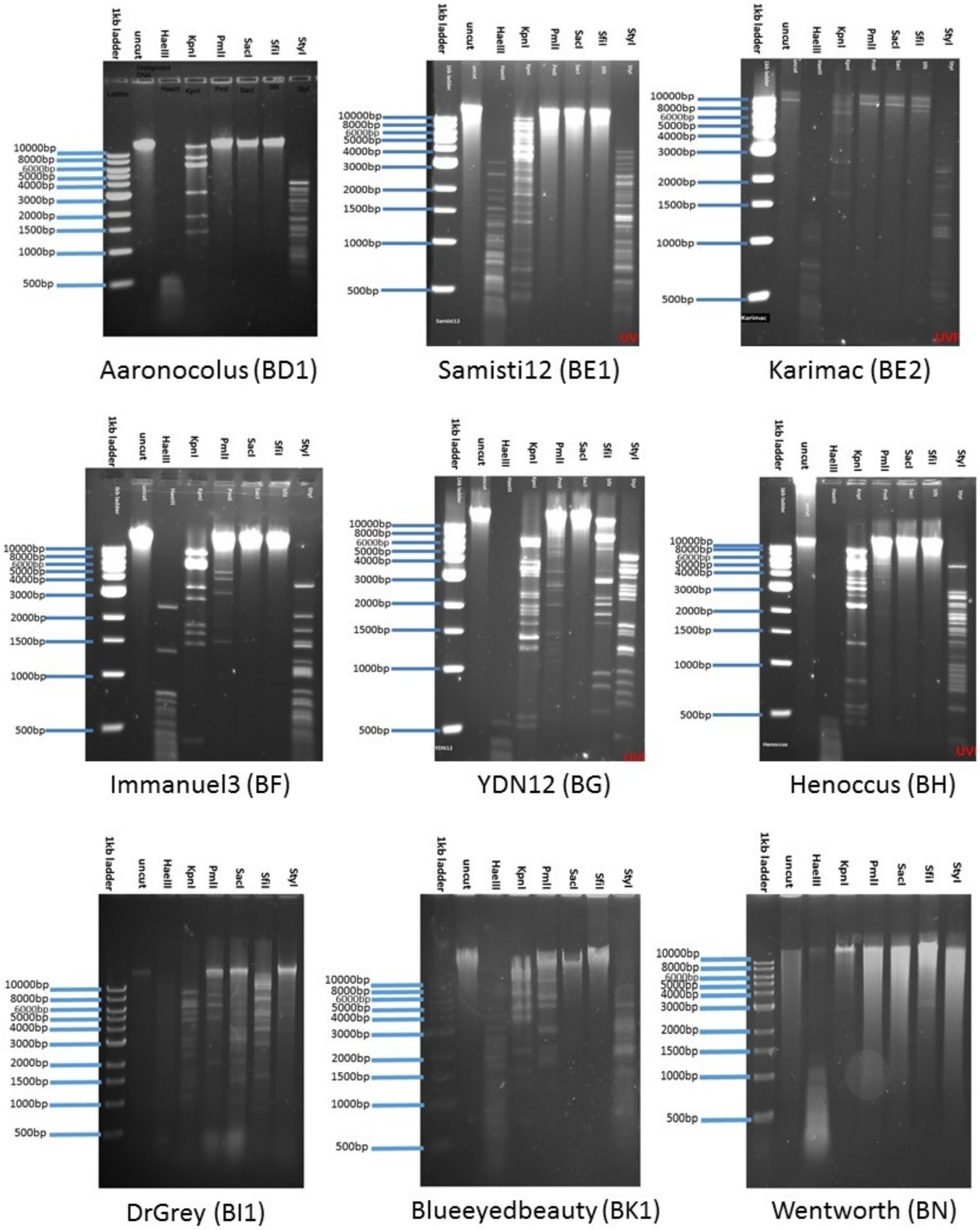


Restriction Endonuclease	Recognition Sequence	Incubation Temperature and Time	Inactivation Temperature and Time
BamHI	5'…G▼GATCC…3' 3'…CCTAG▲G…5'	37°C for 2 hours	NONE
ClaI	5'…AT▼CGAT…3' 3'…TAGC▲TA…5'	37°C for 2 hours	NONE
EcoRI	5'…G▼AATTC…3' 3'…CTTAA▲G…5'	37°C for 2 hours	65°C for 20 minutes
HaeIII	5'…GG▼CC…3' 3'…CC▲GG…5'	37°C for 2 hours	80°C for 15 minutes
HindIII	5'…A▼AGCTT…3' 3'…TTCGA▲A…5'	37°C for 2 hours	80°C for 20 minutes
KpnI	5'…GGTAC▼C…3' 3'…C▲CATGG…5'	37°C for 2 hours	80°C for 15 minutes
NarI	5'…GG▼CGCC…3' 3'…CCGC▲GG…5'	37°C for 2 hours	65°C for 20 minutes
PmlI	5'…CAC▼GTG…3' 3'…GTG▲CAC…5'	37°C for 2 hours	80°C for 15 minutes
SacI	5'…GAGCT▼C…3' 3'…C▲TCGAG…5'	37°C for 2 hours	80°C for 15 minutes
SalI	5'…G▼TCGAC…3' 3'…CAGCT▲G…5'	37°C for 2 hours	65°C for 20 minutes
SfiI	5'…GGCCNNNN▼NGGCC…3' 3'…CCGGN▲NNNNCCGG…5'	50°C for 15 minutes	NONE
SmaI	5'…CCC▼GGG…3' 3'…GGG▲CCC…5'	25°C for 2 hours	65°C for 20 minutes
StuI	5'…AGG▼CCT…3' 3'…TCC▲GGA…5'	37°C for 2 hours	NONE
StyI	5'…C▼CWWGG…3' 3'…GGWWC▲C…5'	37°C for 2 hours	80°C for 15 minutes
XhoI	5'…C▼TCGAG…3' 3'…GAGCT▲C…5'	37°C for 2 hours	65°C for 20 minutes

Because restriction fragment pattern analysis was the most cost-effective method of screening new isolates for potential cluster affiliations, several different restriction enzymes were tested for patterns that were distinguishable between clusters and, therefore, useful. A complete list of enzymes tested may be found in Table 3.5.

At the time, isolates were available for what would become Clusters BE, BF, BG, and BH. A panel of six enzymes was selected based on restriction fragment patterns produced, which included HaeIII, KpnI, PmlI, SacI, SfiI, and StyI. In the *Mycobacterium* phage panel (discussed above), HaeIII was included primarily as a positive control because of its tendency to produce many fragments smaller than 500 bp and rarely, if ever, leave fragments larger than that. In that sense, it became a positive control for the digest panel. Because of this, HaeIII was left on the *Streptomyces* panel and produced surprising results. In two of the four known clusters (BE and BF), HaeIII produced distinct restriction fragment patterns that included fragments larger than 500 bp. Importantly, these patterns were distinguishable between the two clusters. As shown in Figure 3.20, phages in Cluster BE produce many fragments between 1000 and 3000 bp, while those in Cluster BF produce only a few.

As shown in Figure 3.20, each of the clusters which include a phage in this study produces a restriction fragment pattern with at least one “signature” feature and readily distinguishable from the other clusters. As the number of phage isolates increases, the number of clusters also increase along with the number of phages with nucleotide identity across multiple clusters. Because of this, using the restriction fragment pattern analysis for *Streptomyces* phages has become limited in its utility to segregate between phages that are likely clustered.



Aaronocolus (BD1)

Samisti12 (BE1)

Karimac (BE2)

Immanuel3 (BF)

YDN12 (BG)

Henoccus (BH)

DrGrey (BI1)

Blueeyedbeauty (BK1)

Wentworth (BN)

Figure 3.20: Representative RE digests from each of the clusters represented among 45 *Streptomyces* phages. Note that each cluster produces a distinct pattern when subjected to RE digestion.

3.2.4 Distribution of Morphotypes

Suitable electron micrographs for 32 of the 45 phages were produced, providing a total of 133 phage particles for analysis. As shown in Figure 3.21, all of the particles observed were classifiable in the phage order Caudovirales, with a vast majority having long, flexible, non-contractile tails indicative of the family Siphoviridae. Phages in single cluster (Cluster BF) appear to have a short tail, and likely belong in the family *Podoviridae*. There is no evidence that any of the isolates belong in the family Myoviridae. One group (Cluster BG) has a siphoviral morphology and possesses a prolate icosahedral capsid structure as opposed to the more equilateral icosahedral capsid present on all phages observed in the other clusters. Working on the assumption that phages that are closely related at the genomic level will produce particles of the same morphotype, i.e. phages that are clustered will have the same morphotype, then among the 45 phages included here, roughly 87% (39 phages) of the isolates have siphoviral morphology and roughly 13% (6 phages) are podoviruses. As discussed in Chapter 1 and illustrated in Figure 1.1, about 94% of the known phages classified to date belong in the order Caudovirales, with about 55% of those having siphoviral morphology. The phages presented here fit comfortably within those percentages.

Variation in morphotype is common among the actinobacteriophages. Among 46 recently described phages that infect hosts in the genus *Arthrobacter*, roughly 83% (38 phages) were siphoviral, while 15% (7 phages) were myoviral and 2% (1 phage) was podoviral [56]. Similarly, when a study of 60 *Mycobacterium* phages reported morphotypes, about 88% (53 phages) were reported as siphoviral, while 12% (7 phages) were myoviral, and none were podoviral [55].

A survey of 36 phages isolated using a single host in the genus *Gordonia* (*G. terrae* 3612), as reported in the Actinobacteriophage Database, revealed that 100% of the isolates were siphoviral. The finding among *Gordonia* phages has been supported in published literature, although those studies involved fewer total phages (13 phages) [59, 60]. A similar survey of 61 phages that infect a single *Microbacterium* species, all pulled from the Actinobacteriophage Database, revealed that, like the *Gordonia* phages, 100% of the isolates surveyed were siphoviral.

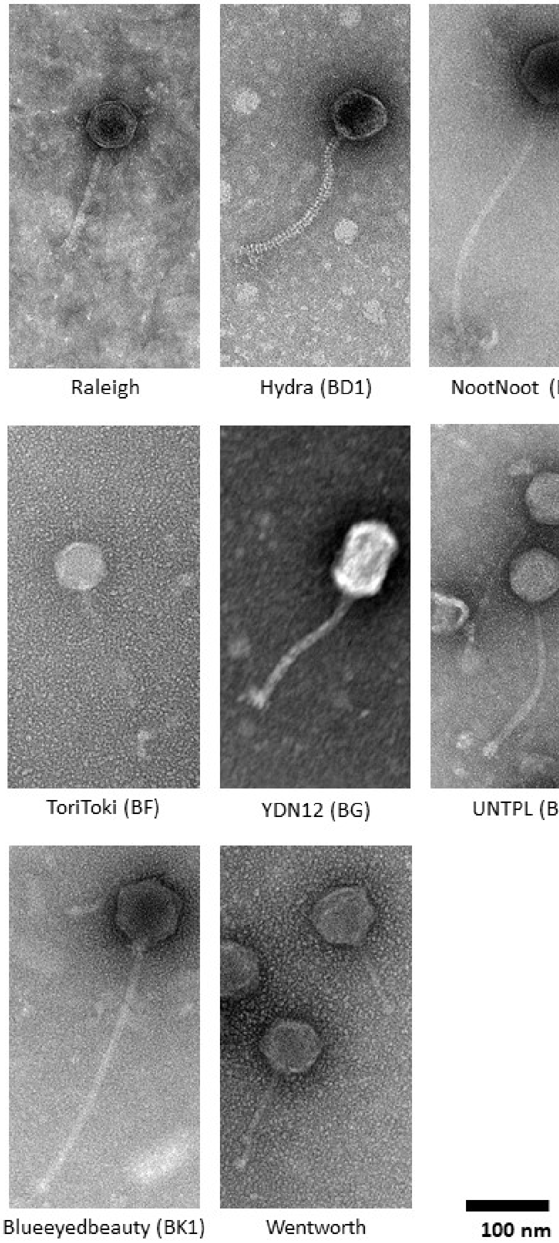


Figure 3.21: Morphotypes, by cluster, represented by 45 phages isolated on *S. griseus* ATCC 10137. Two of the three morphotypes are represented, with no myoviral particles observed. Note the prolate capsid of phage YDN12 (Cluster BG).

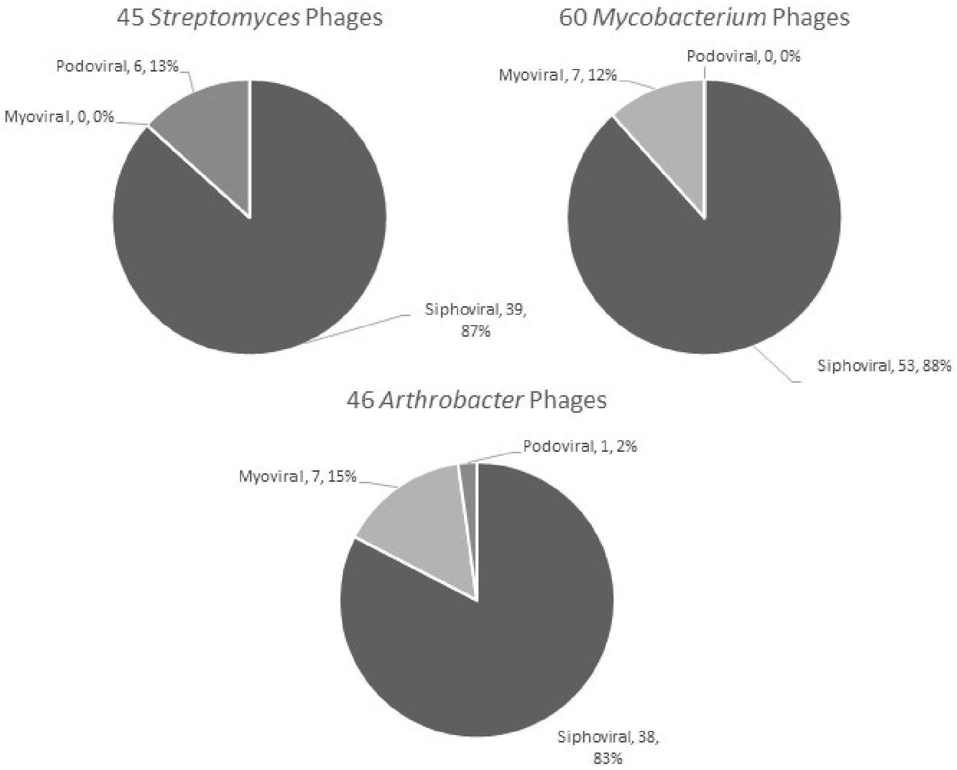


Figure 3.22: Distributions of morphotype among actinobacteriophages by isolation host. Note that the distribution favors those phages with siphoviral morphology. Only a single host group (*Arthrobacter* spp.) contains phages displaying all three morphotypes.

The relative distribution of morphotypes among these actinobacteriophages reveals a couple of prevailing trends. First, a vast majority (166 phages or about 92%) of the 181 total phages examined are siphoviral. Indeed, two groups of phages were lacking diversity in morphotype altogether, i.e. those that infect *Gordonia* and those that infect *Microbacterium*. Secondly, and as illustrated in Figure 3.22, among those actinobacteriophage groups that do have diverse morphotypes, the distribution of morphotypes tends to be concentrated around two of the morphotypes and the third is either a very low percentage, e.g. podoviruses in *Arthrobacter* phages, or missing altogether (myoviruses in *Streptomyces* phages or podoviruses in *Mycobacterium* phages).

Roughly 18% of these 45 phages have prolate capsids, a high percentage when compared to other groups of actinobacteriophages. Of the *Arthrobacter* phages discussed earlier, about 5% (2 phages) had prolate capsids. Of the *Mycobacterium* phages discussed, only about 6% (3 phages) were prolate. As mentioned earlier, all of the seven phages here that have prolate capsids belong to a single cluster, i.e. Cluster BG. As such, it is unclear at this time if the percentage of prolate capsids observed here is actually representative of the *S. griseus* phages as a whole, or if the high percentage is due to some other factor, e.g. favorable conditions for isolating phages belonging to Cluster BG.

3.2.5 Brief Conclusions

The 45 phages presented here display a tremendous amount of diversity in (1) genome size, (2) G+C content, (3) nature of genome termini, (4) number of ORFs, and (5) presence and number of tRNAs. While all are shown to be tailed phages, they show diversity in morphology, including (1) the nature of their tail, and (2) their capsid dimensions. As discussed above in both this chapter and Chapter 1, the most useful framework to discuss and measure this diversity is through clusters. A summary of the data presented above may be found in Table 3.6.

Sequencing of these 45 phages produced over 3.3 million nucleotides of sequence. Dot plot analysis, combined with analysis of the average nucleotide identity (ANI) values for each of the phages supported the sorting of the 45 phages into a total of seven clusters. ANI values, combined with the observation of the moderate contiguous nucleotide identity spanning the entire length of the genomes, support the further division of one cluster (BE) into two subclusters (BE1 and BE2).

As shown in Table 3.6, phages Raleigh and Wentworth are treated as stand-alone phages for the purposes of this study, as they have no closely related phages in this dataset. However, it is noted again that there are genomes present in the Actinobacteriophage Database which are closely enough related at the nucleotide level to support being clustered with each of these phages. Raleigh, for example, is currently classified as a Cluster BC phage, along with ten other phages in the database. Similarly, Wentworth is classified as a Cluster BN phage, along with two other phages in the database. Because none of the phages in either Clusters BC or BN fit the criteria for inclusion in this study, phages Raleigh and Wentworth are treated, here, as stand-alone phages.

Shown in Table 3.6, Cluster BD is comprised here of phages Lorelei, Aaronocolus, BryanRecycles, Eddasa, Jash, Hydra, Caliburn, Izzy, Nabi, and Rana (GenBank Acc. Nos. KX507343, KT124227, MF541404, MH171096, MF541408, KT124229, KT152029, KT184390, MH171094, and MH171093, respectively).

Table 3.6: Summary of genometrics of the 45 bacteriophages included in this study. Phages are organized by cluster, where appropriate.

Cluster	Phage	GenBank Acc. No.	Genome Length (bp)	G+C	# genes/ tRNA/ tmRNA	virion	Termini
	Raleigh	pending	40785	71.8	53	sipho	Circ Perm
BD	Lorelei	KX507343	50558	65.8	75	sipho	11 bp over
	Aaronocolus	KT124227	49562	66.2	72	sipho	11 bp over
	BryanRecycles	MF541404	50066	65.9	73	sipho	11 bp over
	Eddasa	MH171096	50605	65.9	76	sipho	11 bp over
	Jash	MF541408	50066	65.9	74	sipho	11 bp over
	Hydra	KT124229	50727	66.2	76	sipho	11 bp over
	Caliburn	KT152029	49949	66.2	72	sipho	11 bp over
	Izzy	KT184390	50113	65.9	74	sipho	11 bp over
	Nabi	MH171094	51127	65.8	76	sipho	11 bp over
	Rana	MH171093	50980	65.8	76	sipho	11 bp over

Cluster	Phage	GenBank Acc. No.	Genome Length (bp)	G+C	# genes/ tRNA/ tmRNA	virion	Termini
BE1	Samisti12	MF347639	133710	49.9	227/44/1	sipho	10666 DTR
	Paradiddles	MF347637	133486	49.5	216/46/1	sipho	10778 DTR
	NootNoot	MF347636	131086	50.2	221/45/1	sipho	10787 DTR
BE2	Wofford	MH576968	133007	47.7	235/45/1	sipho	11214 DTR
	StarPlatinum	MH576965	133886	49.5	251/44/1	sipho	12199 DTR
	LukeCage	MH590597	133195	49.0	248/41/1	sipho	12291 DTR
	Starbow	MH576964	131427	49.5	238/44/1	sipho	12579 DTR
	Karimac	MH590599	131909	49.4	241/44/1	sipho	12590 DTR
BF	Percastrophe	MG663583	45999	59.7	64/22	podo	264 DTR
	ToriToki	MG663585	46077	59.7	64/22	podo	264 DTR
	Romero	MG663584	46079	59.7	64/22	podo	264 DTR
	HaugeAnator	MG663582	46135	59.6	63/22	podo	274 DTR
	ZooBear	MG663586	46135	59.7	64/22	podo	274 DTR
	Immanuel3	MG518520	46094	59.6	60/17	podo	275 DTR
BG	BabyGotBac	KY365739	57165	69.2	72	sipho	Circ Perm
	Maih	KU189325	57256	69.3	70	sipho	Circ Perm
	Salete	MH178382	57243	69.2	70	sipho	Circ Perm
	TP1604	KP876466	57168	69.2	71	sipho	Circ Perm
	YDN12	KP876465	56528	69.2	71	sipho	Circ Perm
	Xkcd426	KU530220	64477	68.8	78	sipho	Circ Perm
	BayC	MH178381	57243	69.2	71	sipho	Circ Perm
BH	Crosby	MH536815	54036	68.3	82	sipho	Circ Perm
	Henoccus	MH229862	55137	68.2	82	sipho	Circ Perm
	UNTPL	MH229864	54495	68.3	81	sipho	Circ Perm
	JackieB	MH229863	54912	68.2	82	sipho	Circ Perm
	LazerLemon	MH229865	54798	68.1	81	sipho	Circ Perm
BI	OlympicHelado	KX670789	56189	59.5	88	sipho	9 bp over
	DrGrey	MF467948	56076	59.5	80	sipho	9 bp over
	Spectropatronm	MF467949	55707	59.5	84	sipho	9 bp over
BK1	Comrade	pending	129015	47.1	229/34/1	sipho	734 DTR
	SparkleGoddess	pending	129742	47.1	232/35/1	sipho	734 DTR
	Blueeyedbeauty	MH536814	130473	47.9	240/37/1	sipho	788 DTR
	Annadreamy	MH536811	125726	47.6	230/35/1	sipho	789 DTR
	Wentworth	MH019216	68260	64.1	103	sipho	Circ Perm

Genome sizes in Cluster BD range from 49,562 bp to 51,127 bp with an average genome length of 50,375 (± 500) bp. The percentages G+C range from 65.8 to 66.2, with a cluster average of 66.0 (± 0.2) percent. Overall, genome annotations reveal that the number of ORFs range from 72 to 76, with a cluster average of 74 (± 2) ORFs per phage genome. There is evidence that all ten have 11 bp overhangs at their termini. All Cluster BD phages observed via TEM have siphoviral morphology with icosahedral heads.

Cluster BE is comprised here of eight phages which have been classified into two subclusters based on both dot plot and ANI value analysis. Cluster BE1 is made up of phages Samisti12, Paradiddles, and NootNoot (GenBank Acc. Nos. MF347639, MF347637, and MF347636, respectively). Genome sizes range from 131086 bp to 133710 bp, with an average genome length of 132761 (± 1455) bp. The percentages G+C range from 49.5 to 50.2, with an average G+C of 49.9 (± 0.4) percent. Genome annotations predict numbers of ORFs ranging from 216 to 227, with a cluster average of 221 (± 6) ORFs per phage genome. There is evidence that all Cluster BE1 phages have large direct terminal repeats (DTRs), ranging in size from 10666 bp to 10787 bp, with an average length of 10745 bp. All of the Cluster BE1 phages code for tRNAs, with numbers ranging from 44 to 46 tRNAs per genome, and an average of 45 tRNAs per genome. Additionally, each of these phages code for a single tmRNA.

Cluster BE2 is comprised here of five phages: Wofford, StarPlatinum, LukeCage, Starbow, and Karimac (GenBank Acc. Nos. MH576968, MH576965, MH590597, MH576964, and MH590599, respectively). Genome sizes range from 131427 bp to 133886 bp, with a cluster average of 132658 (± 999) bp. The percentages G+C range from 47.7 to 49.5, with a cluster average of 49.0 (± 0.8) percent. Genome annotations predict numbers of ORFs ranging from 235 to 251 ORFs, with a cluster average of 243 (± 7) ORFs per phage genome. There is evidence that all Cluster BE2 phages, like their BE1 counterparts, have large DTRs at their termini, ranging in size from 11214 bp to 12590 bp, with a cluster average of 12175 (± 564) bp. All of the Cluster BE2 phages code for tRNAs, with numbers ranging from 41 to 45 tRNAs per genome and an average tRNA composition of 43 (± 2). Again, like their BE1 counterparts, all BE2 phages presented here code for a single tmRNA. A summary of all BE phages may be found at the end of the chapter in Table 3.7.

Cluster BF is comprised here of six phages: Percastrophe, ToriToki, Romero, HaugeAnator, ZooBear, and Immanuel3 (GenBank Acc. Nos. MG663583, MG663585, MG663584, MG663582, MG663586, and MG518520, respectively). Genome sizes range from 45999 bp to 46135 bp, with a cluster average of 46087 (± 50) bp. The percentages G+C range from 59.6 to 59.7 with an average of 59.7 (± 0.05) percent. Genome annotations predict numbers of ORFs ranging from 60 to 64, with a cluster average of 63 (± 2) ORFs per phage genome. There is evidence that all Cluster BF phages presented here have small DTRs at their termini, with lengths ranging from 264 to 275, with a cluster average of 269 (± 6) bp. Cluster BF phages also code for tRNAs, with numbers ranging from 17 to 22 and a cluster average of 21 (± 2) tRNAs per genome. It should be noted that all Cluster BF phages here code for 22 tRNAs with the exception of Immanuel3, which codes for only 17. Additionally, Cluster BF phages all have podoviral morphology and all phages in this study with podoviral morphology belong to Cluster BF.

Cluster BG is comprised here of seven phages: BabyGotBac, Maih, Salete, TP1604, YDN12, Xkcd426, and BayC (GenBank Acc. Nos. KY365739, KU189325, MH178382, KP876466, KP876465, KU530220, and MH178381, respectively). Genome sizes range from 56528 bp to 64477 bp, with a cluster average of 58154 (± 2800) bp. The percentages G+C range from 68.8 to 69.2, with an average G+C of 69.2 (± 0.2) percent. Genome annotations predict numbers of ORFs ranging from 70 to 78, with an average of 72 (± 3) ORFs per phage genome. There is evidence that all phages in this cluster are circularly permuted. Additionally, Cluster BG phages comprise phages with prolate icosahedral capsids, i.e. elongated capsids as opposed to the more typical, near-equilateral icosahedral capsids of most known siphoviral phages. The ratio of capsid length (from tail connector to apex of capsid) to width (from the left vertex of the capsid to the right vertex of the capsid) has been previously reported for these phages at 1.6:1 [64].

Cluster BH is comprised here of five phages: Crosby, Henoccus, UNTPL, JackieB, and LazerLemon (GenBank Acc. Nos. MH536815, MH229862, MH229864, MH229863, and MH229865, respectively). Genome sizes range from 54036 bp to 55137 bp, with a cluster average of 54676 (± 426) bp. The percentages G+C range from 68.1 to 68.3 with an average G+C of 68.2 (± 0.1) percent. Genome annotations predict numbers of ORFs ranging from 81 to 82, with an average of 81.6 (± 0.5) ORFs per phage genome. There is evidence that all Cluster BH phages presented here, like the Cluster BG phages, are circularly permuted.

Cluster BI is comprised here of three phages: OlympicHelado, DrGrey, and Spectropatronm (GenBank Acc. Nos. KX670789, MF467948, and MF467949, respectively). Genome sizes range from 55707 bp to 56189 bp, with an average genome size of 55991 (± 252) bp. The percentages G+C for all three Cluster BI phages is equal to 59.9%. Genome annotations predict number of ORFs ranging from 80 to 88, with a cluster average of 84 (± 4) ORFs per phage genome. There is evidence that all phages in Cluster BI have 9 bp cos ends, with all three phages having the same sequence (reported above) at this location.

Cluster BK is comprised here of four phages: Comrade, SparkleGoddess, Blueeyedbeauty, and Annadreamy (GenBank Acc. Nos. MH536814 and MH536811 for Blueeyedbeauty and Annadreamy, respectively; GenBank Acc. Nos. pending for phages Comrade and SparkleGoddess). Genome lengths range from 125726 bp to 130473 bp, with an average genome length of 128739 (± 2095) bp. Percentages G+C range from 47.1 to 47.9 with a cluster average of 47.4 (± 0.4) percent. Genome annotations predict numbers of ORFs ranging from 229 to 240 with a cluster average of 233 (± 5) ORFs per phage genome. There is evidence that all phages in Cluster BK presented here have short DTRs at their termini with lengths ranging from 734 bp to 789 bp, and an average length of 761 (± 32) bp. Each of the Cluster BK phages here codes for tRNAs, with numbers ranging from 34 to 37 and a cluster average of 35.3 (± 1.3) tRNAs per phage genome. Like the phages in Cluster BE, each of the phages in Cluster BK presented here code for a single tmRNA.

Table 3.7 summarizes the by-cluster findings discussed above.

Table 3.7: Summary of cluster genometrics of 45 phages that infect *S. griseus*. Note that two subclusters, BE1 and BE2, appear in parentheses.

Cluster (subcluster)	Avg Genome Length (bp)	σ	Avg G+C (%)	σ	ORFs	σ	tRNA (tmRNA)	σ	Termini	Terminus Length	σ
BD	50375	500	66.0	0.2	75	2	-	-	11 bp cos	-	-
BE	132713	1085	49.3	0.8	235	13	44.1 (1)	1.5	DTR	11638	855
(BE1)	132760	1455	50.0	0.4	221	6	45 (1)	1	DTR	10744	67

Cluster (subcluster)	Avg Genome Length (bp)	σ	Avg G+C (%)	σ	ORFs	σ	tRNA (tmRNA)	σ	Termini	Terminus Length	σ
(BE2)	132685	49	49.0	0.8	243	7	44 (1)	2	DTR	12175	564
BF	46087	50	59.7	0.05	63	2	21	2	DTR	269	6
BG	58154	2800	69.2	0.2	72	3	-	-	circ perm	-	-
BH	54676	426	68.2	0.1	81.6	0.6	-	-	circ perm	-	-
BI	55991	252	59.5	0	84	4	-	-	9 bp cos	-	-
BK	128739	2095	47.4	0.4	233	5	35 (1)	1.3	DTR	761	32
Raleigh	40785	-	71.8	-	53	-	-	-	circ perm	-	-
Wentworth	68260	-	64.1	-	103	-	-	-	circ perm	-	-

CHAPTER 4

CLUSTERING VIA GCS AND THE DIVERSITY OF PHAGE CLUSTERS

As discussed in Chapter 1, the mosaic nature of bacteriophage genomes makes drawing concrete phylogenetic relationships between phages a difficult, cumbersome, and often inexact endeavor. These complex relationships, reflected at the nucleotide level where entire sections of genomes appear to be completely unrelated in phages that show an otherwise high degree of contiguous nucleotide identity, can be difficult to trace and even more difficult to accurately describe in a meaningful way. The idea that phages are largely comprised of different segments, acquired through horizontal gene transfer, and each with a distinct and noncongruent phylogenetic history has been well documented in the age of genomics.

4.1 Methods and Materials

4.1.1 Phamerator and Assigning Genes to Phamilies

To address the need for a computational way to arrange and analyze genes that are apparently related at the amino acid level, researchers at the University of Pittsburgh developed a program called Phamerator [45]. Using a combination of ClustalW and BLASTp, Phamerator groups genes that are related by either (1) a BLASTp score of 0.001 or greater, or (2) greater than 27.5% amino acid sequence identity; into groups of gene “phamilies,” or “phams.” Once a gene is assigned to a pham, Phamerator provides an efficient way to compare across all sequenced bacteriophages that infect an actinobacterial host, as long as the sequenced phage is present in the Actinobacteriophage Database.

Illustrated in Figure 4.1, Phamerator produces “maps” of genomes which may be compared across different phages, with such identifying information as (1) the gene product number of the specific genome; (2) the pham identifier, indicating which pham the gene has been classified into; (3) the number of genes that have been classified into the pham, such genes being present across the entire spectrum of sequenced phages that (i) infect an actinobacterial host, and (ii) have sequence information available in the Actinobacteriophage Database; (4) the position of the gene along the length of the phage genome; (5) the predicted function, where that information is available; and (6) a graphical representation of the degree of relatedness, at the nucleotide level, across contiguous genomes, e.g. genomes that are positioned vertically adjacent to one other. Each of these elements is explained further in Figure 4.1.

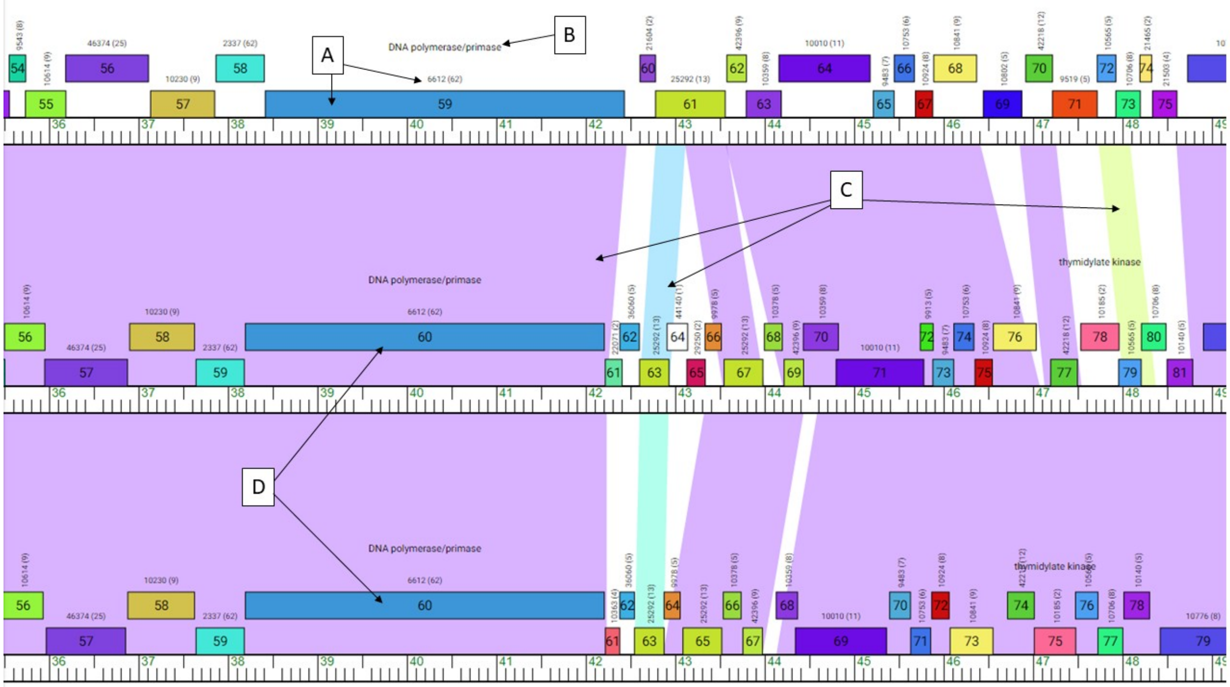


Figure 4.1: Reading a Phamerator Map. (A) Pham information for a particular gene may be found in both text, located above the graphical representation of the gene, as well as in the color of the gene box. Genes that belong in the same pham will be the same color. (B) Predicted function, where available, is displayed above the gene and pham identifier. (C)

When comparing two genomes, areas of nucleotide identity, as determined by BLASTn, are highlighted between the two genomes. Violet regions have E values of 0.0, and the color of the shading progresses through the color spectrum up to red, indicating an E value of around 0.0001. Regions with E values greater than 0.001 have no shading. (D) Two genes, gp 60 in two different phage genomes, that belong to the same pham. Note that that the two genes are the same color, have the same pham identifier located above them, the same function, and show a very high degree of nucleotide identity, indicated by the purple shading.

In addition to providing information about the relatedness of genes through the placement into phams, Phamerator also provides information when a particular gene has no related genes in the database. In this case, the gene would belong to an "orpham."

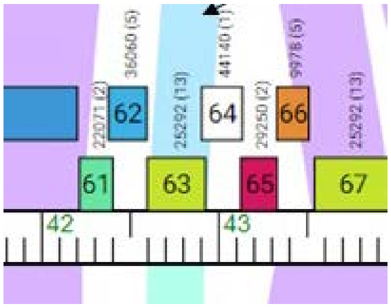


Figure 4.2: Graphical representation of an orpham, e.g. a gene product that has no close relatives in the Actinobacteriophage Database. Here, the Phamerator map indicates that gp 64 is an orpham, recognizable by the white color of the gene. Note that orphams, while they are the only member of their respective phams, are still assigned a pham identifier.

4.1.2 Gene Content Similarity and Splitstree Analysis

Because of the sometimes ambiguous nature of grouping phages into clusters, it is prudent to employ several methods when determining relationships between phages, particularly where those phages are of the type that show areas of weak correlation, spanning the length of a genome in a dot plot, or where there are several areas of strong contiguous nucleotide similarities that are localized to only certain regions of two phage genomes. While dot plot analysis and ANI (discussed in Chapter 3) can be powerful indicators of phage relatedness, it is not entirely uncommon for two phages to share genes with protein products that are related at the amino acid level but show only weak correlations in their nucleotide sequence.

Gene content similarity (GCS) addresses this particular scenario and provides further insight into how groups of phages may be related. GCS is calculated on a phage-to-phage basis and is a measure of the relatedness of any two phages at the gene content level. In other words, it measures the extent to which any two phages share related gene products. To calculate GCS, the number of phams shared between two phage genomes must be determined, as well as the total number of phams present in each of the individual phages. Then, calculate the GCS using the equation below.

$$GCS = \frac{\frac{\#\ of\ shared\ phams}{total\ phams\ phage\ one} + \frac{\#\ of\ shared\ phams}{total\ phams\ phage\ two}}{2}$$

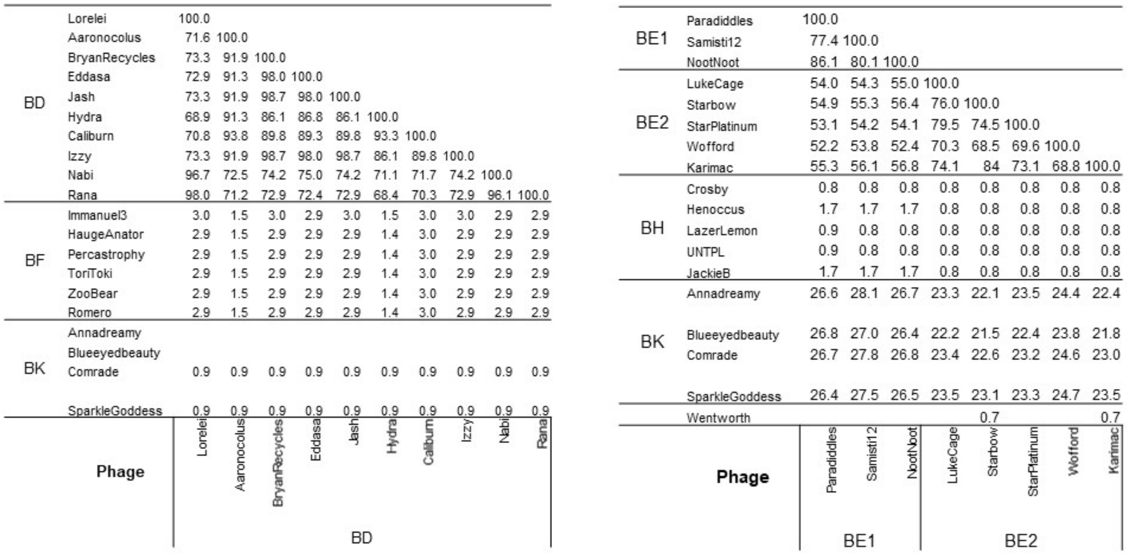
Note that, under the convention set forth by the Pittsburgh Bacteriophage Institute, two phages that share a GCS value of 35% or higher are likely to be clustered together.

Using a truncated Phamerator Database, new pham identifiers were assigned to each pham in each phage and a relative distribution of phams was calculated for each of the 45 phages. A value was assigned to each phage based on the presence or absence of each pham in the truncated database and used to generate a Splitstree [61] representation.

4.2 Results and Discussion

4.2.1 Clustering via Gene Content Similarity and Splitstree

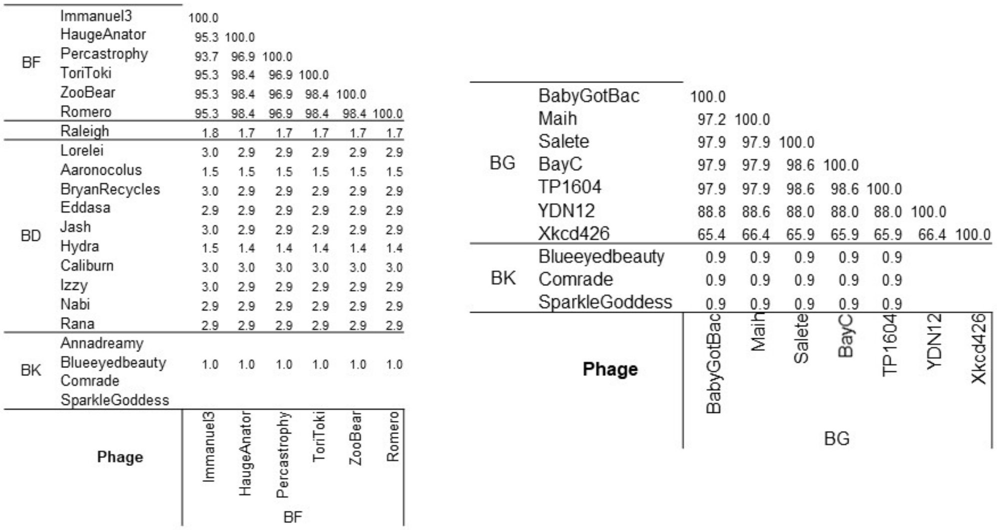
As illustrated in Figures 4.3 through 4.8, comparisons of both gene content similarity (GCS) values and Splitstree analysis agree with the clustering assignments made using dot plot analysis and comparison of average nucleotide identity (ANI) values, including the subdivision of Cluster BE phages into two subclusters, BE1 and BE2.



		BD									
	Phage	Lorelei	Aaronocolus	BryanRecycles	Eddasa	Jash	Hydra	Caliburn	Izzy	Nabi	Rana
BD	Lorelei	100.0									
	Aaronocolus	71.6	100.0								
	BryanRecycles	73.3	91.9	100.0							
	Eddasa	72.9	91.3	98.0	100.0						
	Jash	73.3	91.9	98.7	98.0	100.0					
	Hydra	68.9	91.3	86.1	86.8	86.1	100.0				
	Caliburn	70.8	93.8	89.8	89.3	89.8	93.3	100.0			
	Izzy	73.3	91.9	98.7	98.0	98.7	86.1	89.8	100.0		
	Nabi	96.7	72.5	74.2	75.0	74.2	71.1	71.7	74.2	100.0	
	Rana	98.0	71.2	72.9	72.4	72.9	68.4	70.3	72.9	96.1	100.0
BF	Immanuel3	3.0	1.5	3.0	2.9	3.0	1.5	3.0	3.0	2.9	2.9
	HaugeAnator	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9
	Percastrophy	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9
	ToriToki	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9
	ZooBear	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9
	Romero	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9
BK	Annadreamy										
	Blueeyedbeauty										
	Comrade	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
	SparkleGoddess	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9

		BE1			BE2				
	Phage	Paradiddles	Samisti12	NootNoot	LukeCage	Starbow	StarPlatinum	Wofford	Karimac
BE1	Paradiddles	100.0							
	Samisti12	77.4	100.0						
	NootNoot	86.1	80.1	100.0					
BE2	LukeCage	54.0	54.3	55.0	100.0				
	Starbow	54.9	55.3	56.4	76.0	100.0			
	StarPlatinum	53.1	54.2	54.1	79.5	74.5	100.0		
	Wofford	52.2	53.8	52.4	70.3	68.5	69.6	100.0	
	Karimac	55.3	56.1	56.8	74.1	84	73.1	68.8	100.0
BH	Crosby	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Henoccus	1.7	1.7	1.7	0.8	0.8	0.8	0.8	0.8
	LazerLemon	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	UNTPL	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	JackieB	1.7	1.7	1.7	0.8	0.8	0.8	0.8	0.8
BK	Annadreamy	26.6	28.1	26.7	23.3	22.1	23.5	24.4	22.4
	Blueeyedbeauty	26.8	27.0	26.4	22.2	21.5	22.4	23.8	21.8
	Comrade	26.7	27.8	26.8	23.4	22.6	23.2	24.6	23.0
	SparkleGoddess	26.4	27.5	26.5	23.5	23.1	23.3	24.7	23.5
	Wentworth					0.7			0.7

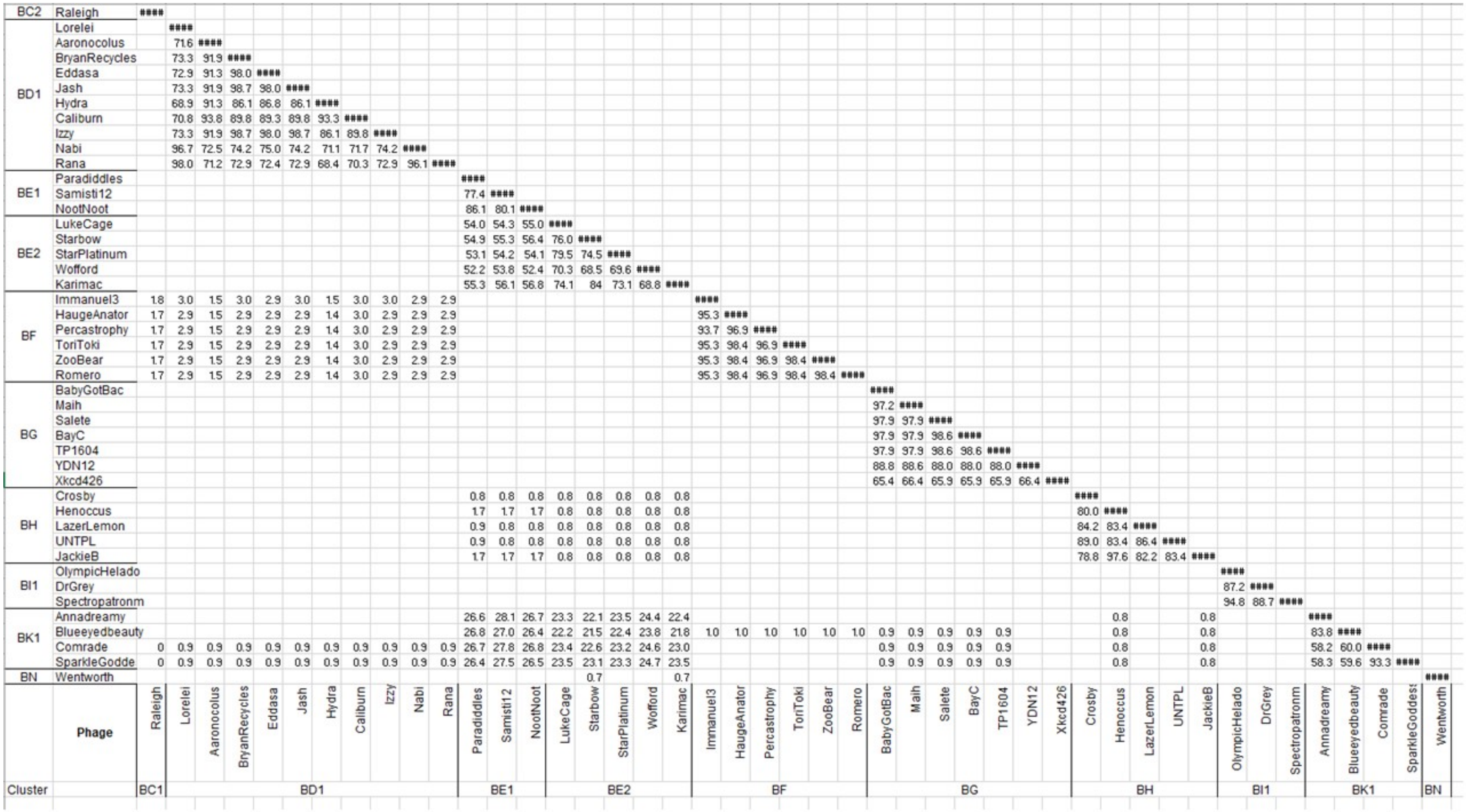
Figure 4.3: Calculated GCS values, where greater than zero, for Cluster BD (left) and Cluster BE (right) versus other phage clusters.



		BF					
	Phage	Immanuel3	HaugeAnator	Percastrophy	ToriToki	ZooBear	Romero
BF	Immanuel3	100.0					
	HaugeAnator	95.3	100.0				
	Percastrophy	93.7	96.9	100.0			
	ToriToki	95.3	98.4	96.9	100.0		
	ZooBear	95.3	98.4	96.9	98.4	100.0	
	Romero	95.3	98.4	96.9	98.4	98.4	100.0
	Raleigh	1.8	1.7	1.7	1.7	1.7	1.7
BD	Lorelei	3.0	2.9	2.9	2.9	2.9	2.9
	Aaronocolus	1.5	1.5	1.5	1.5	1.5	1.5
	BryanRecycles	3.0	2.9	2.9	2.9	2.9	2.9
	Eddasa	2.9	2.9	2.9	2.9	2.9	2.9
	Jash	3.0	2.9	2.9	2.9	2.9	2.9
	Hydra	1.5	1.4	1.4	1.4	1.4	1.4
	Caliburn	3.0	3.0	3.0	3.0	3.0	3.0
	Izzy	3.0	2.9	2.9	2.9	2.9	2.9
	Nabi	2.9	2.9	2.9	2.9	2.9	2.9
	Rana	2.9	2.9	2.9	2.9	2.9	2.9
BK	Annadreamy						
	Blueeyedbeauty	1.0	1.0	1.0	1.0	1.0	1.0
	Comrade						
	SparkleGoddess						

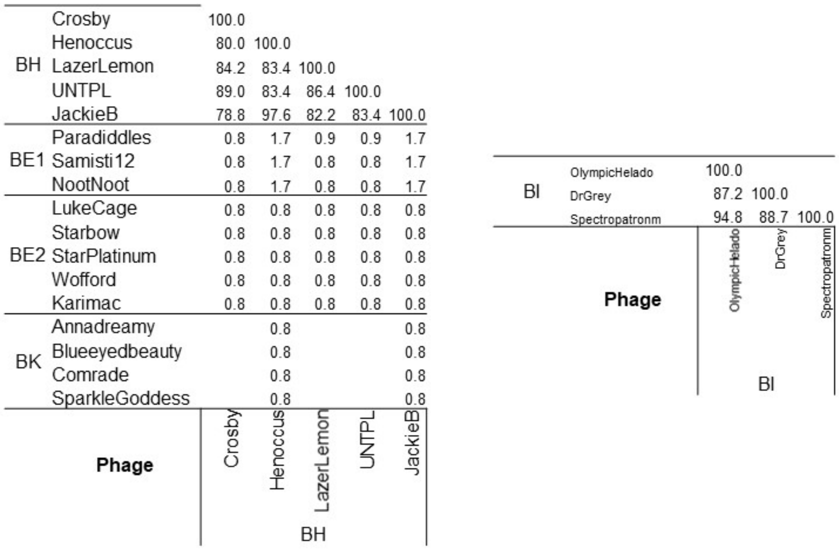
		BG						
	Phage	BabyGotBac	Maih	Salete	BayC	TP1604	YDN12	Xkcd426
BG	BabyGotBac	100.0						
	Maih	97.2	100.0					
	Salete	97.9	97.9	100.0				
	BayC	97.9	97.9	98.6	100.0			
	TP1604	97.9	97.9	98.6	98.6	100.0		
	YDN12	88.8	88.6	88.0	88.0	88.0	100.0	
	Xkcd426	65.4	66.4	65.9	65.9	65.9	66.4	100.0
BK	Blueeyedbeauty	0.9	0.9	0.9	0.9	0.9		
	Comrade	0.9	0.9	0.9	0.9	0.9		
	SparkleGoddess	0.9	0.9	0.9	0.9	0.9		

Figure 4.4: Calculated GCS values, where greater than zero, for Cluster BF (left) and Cluster BG (right) versus other phage clusters.



	Phage	Raleigh	Lorelei	Aaronocolus	BryanRecycles	Eddasa	Jash	Hydra	Caliburn	Izzy	Nabi	Rana	Paradiddles	Samisti12	NootNoot	LukeCage	Starbow	StarPlatinum	Wofford	Karimac	Immanuel3	HaugeAnator	Percastrophy	ToriToki	ZooBear	Romero	BabyGotBac	Maih	Salete	BayC	TP1604	YDN12	Xkcd426	Crosby	Henoccus	LazerLemon	UNTPL	JackieB	OlympicHelado	DrGrey	Spectropatronm	Annadreamy	Blueeyedbeauty	Comrade	SparkleGoddes	Wentworth
BC2	Raleigh	####																																											
BD1	Lorelei		####																																										
	Aaronocolus		71.6	####																																									
	BryanRecycles		73.3	91.9	####																																								
	Eddasa		72.9	91.3	98.0	####																																						
	Jash		73.3	91.9	98.7	98.0	####																																					
	Hydra		68.9	91.3	86.1	86.8	86.1	####																																				
	Caliburn		70.8	93.8	89.8	89.3	89.8	93.3	####																																			
	Izzy		73.3	91.9	98.7	98.0	98.7	86.1	89.8	####																																		
	Nabi		96.7	72.5	74.2	75.0	74.2	71.1	71.7	74.2	####																																	
	Rana		98.0	71.2	72.9	72.4	72.9	68.4	70.3	72.9	96.1	####																																
BE1	Paradiddles												####																																
	Samisti12												77.4	####																															
	NootNoot												86.1	80.1	####																														
BE2	LukeCage												54.0	54.3	55.0	####																													
	Starbow												54.9	55.3	56.4	76.0	####																												
	StarPlatinum												53.1	54.2	54.1	79.5	74.5	####																											
	Wofford												52.2	53.8	52.4	70.3	68.5	69.6	####																										
	Karimac												55.3	56.1	56.8	74.1	84	73.1	68.8	####																									
BF	Immanuel3	1.8	3.0	1.5	3.0	2.9	3.0	1.5	3.0	3.0	2.9	2.9									####																								
	HaugeAnator	1.7	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9									95.3	####																							
	Percastrophy	1.7	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9									93.7	96.9	####																						
	ToriToki	1.7	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9									95.3	98.4	96.9	####																					
	ZooBear	1.7	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9									95.3	98.4	96.9	98.4	####																				
	Romero	1.7	2.9	1.5	2.9	2.9	2.9	1.4	3.0	2.9	2.9	2.9									95.3	98.4	96.9	98.4	98.4	####																			
BG	BabyGotBac																										####																		
	Maih																										97.2	####																	
	Salete																										97.9	97.9	####																
	BayC																										97.9	97.9	98.6	####															
	TP1604																										97.9	97.9	98.6	98.6	####														
	YDN12																										88.8	88.6	88.0	88.0	88.0	####													
	Xkcd426																										65.4	66.4	65.9	65.9	65.9	66.4	####												
BH	Crosby												0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8														####											
	Henoccus												1.7	1.7	1.7	0.8	0.8	0.8	0.8	0.8														80.0	####										
	LazerLemon												0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8														84.2	83.4	####									
	UNTPL												0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8														89.0	83.4	86.4	####								
	JackieB												1.7	1.7	1.7	0.8	0.8	0.8	0.8	0.8														78.8	97.6	82.2	83.4	####							
BI1	OlympicHelado																																						####							
	DrGrey																																						87.2	####						
	Spectropatronm																																						94.8	88.7	####					
BK1	Annadreamy												26.6	28.1	26.7	23.3	22.1	23.5	24.4	22.4															0.8			0.8				####				
	Blueeyedbeauty												26.8	27.0	26.4	22.2	21.5	22.4	23.8	21.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9				0.8			0.8				83.8	####			
	Comrade	0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	26.7	27.8	26.8	23.4	22.6	23.2	24.6	23.0							0.9	0.9	0.9	0.9	0.9				0.8			0.8				58.2	60.0	####		
	SparkleGodde	0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	26.4	27.5	26.5	23.5	23.1	23.3	24.7	23.5							0.9	0.9	0.9	0.9	0.9				0.8			0.8				58.3	59.6	93.3	####	
BN	Wentworth																0.7			0.7																										####
Cluster		BC1	BD1										BE1			BE2					BF						BG							BH					BI1			BK1				BN

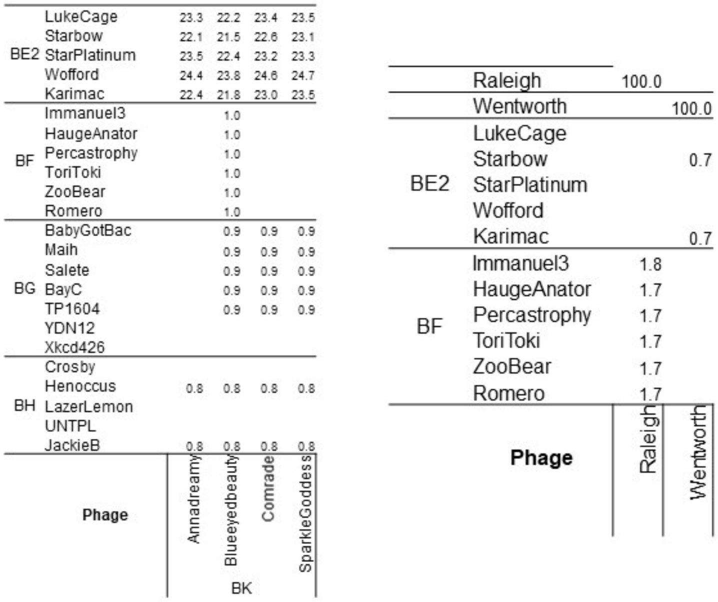
Figure 4.5: Calculated GCS values for the 45 phages included in this study. Phage names run along the x and y axis, according to cluster assignments made using dot plot analysis and average nucleotide identity (ANI) values. Note that the chart contains only non-zero values for GCS and intersections that are blank are to be considered as having a value of zero. A detailed view of each non-zero intersection is available, arranged by cluster, in Figures 4.4 through 4.7.



		BH				
	Phage	Crosby	Henoccus	LazerLemon	UNTPL	JackieB
BH	Crosby	100.0				
	Henoccus	80.0	100.0			
	LazerLemon	84.2	83.4	100.0		
	UNTPL	89.0	83.4	86.4	100.0	
	JackieB	78.8	97.6	82.2	83.4	100.0
BE1	Paradiddles	0.8	1.7	0.9	0.9	1.7
	Samisti12	0.8	1.7	0.8	0.8	1.7
	NootNoot	0.8	1.7	0.8	0.8	1.7
BE2	LukeCage	0.8	0.8	0.8	0.8	0.8
	Starbow	0.8	0.8	0.8	0.8	0.8
	StarPlatinum	0.8	0.8	0.8	0.8	0.8
	Wofford	0.8	0.8	0.8	0.8	0.8
	Karimac	0.8	0.8	0.8	0.8	0.8
BK	Annadreamy		0.8			0.8
	Blueeyedbeauty		0.8			0.8
	Comrade		0.8			0.8
	SparkleGoddess		0.8			0.8

		BI		
	Phage	OlympicHelado	DrGrey	Spectropatronm
BI	OlympicHelado	100.0		
	DrGrey	87.2	100.0	
	Spectropatronm	94.8	88.7	100.0

Figure 4.6: Calculated GCS values, where greater than zero, for Cluster BH (left) and Cluster BI (right) versus other phage clusters.



		BK			
	Phage	Annadreamy	Blueeyedbeauty	Comrade	SparkleGoddess
BE2	LukeCage	23.3	22.2	23.4	23.5
	Starbow	22.1	21.5	22.6	23.1
	StarPlatinum	23.5	22.4	23.2	23.3
	Wofford	24.4	23.8	24.6	24.7
	Karimac	22.4	21.8	23.0	23.5
BF	Immanuel3		1.0		
	HaugeAnator		1.0		
	Percastrophy		1.0		
	ToriToki		1.0		
	ZooBear		1.0		
	Romero		1.0		
BG	BabyGotBac		0.9	0.9	0.9
	Maih		0.9	0.9	0.9
	Salete		0.9	0.9	0.9
	BayC		0.9	0.9	0.9
	TP1604		0.9	0.9	0.9
	YDN12				
	Xkcd426				
BH	Crosby				
	Henoccus	0.8	0.8	0.8	0.8
	LazerLemon				
	UNTPL				
	JackieB	0.8	0.8	0.8	0.8

	Phage	Raleigh	Wentworth
	Raleigh	100.0	
	Wentworth		100.0
BE2	LukeCage		
	Starbow		0.7
	StarPlatinum		
	Wofford		
	Karimac		0.7
BF	Immanuel3	1.8	
	HaugeAnator	1.7	
	Percastrophy	1.7	
	ToriToki	1.7	
	ZooBear	1.7	
	Romero	1.7	

Figure 4.7: Calculated GCS values, where greater than zero, for Cluster BK (left) and phages Raleigh and Wentworth (right) versus other phage clusters.

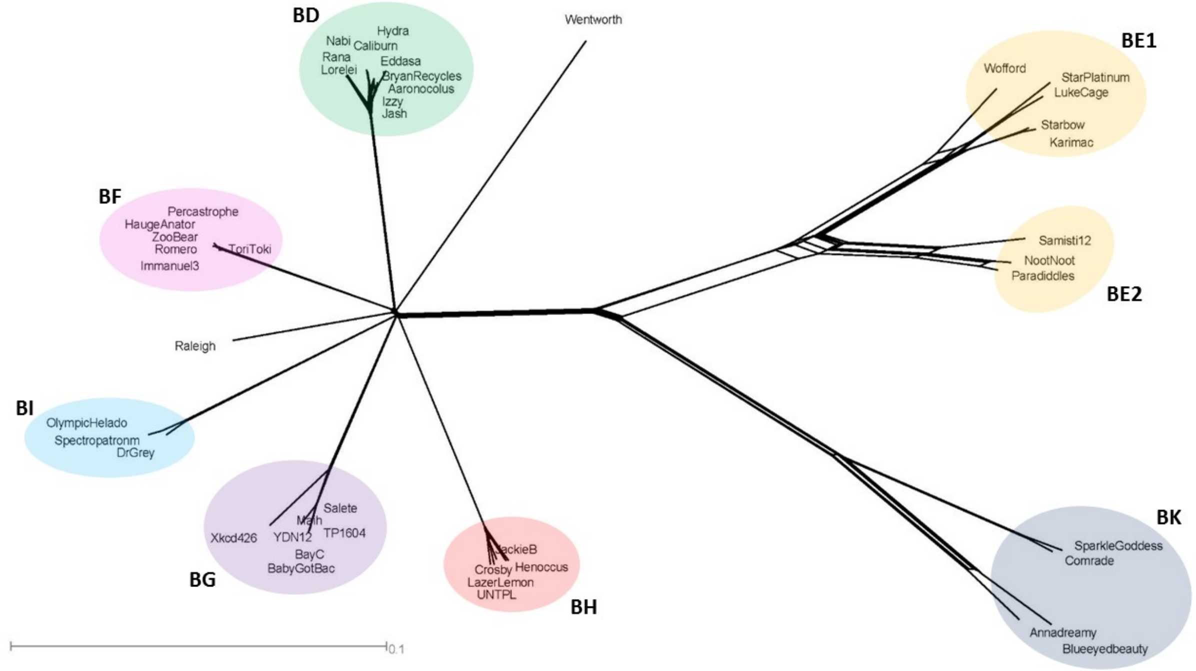


Figure 4.8: Splitstree representation of the 45 *S. griseus*-infecting phages included in this study. The 5250 predicted genes were assorted into 1300 phams according to shared amino acid sequences. Based on the presence or absence of each pham member, each genome was assigned a value reflecting its pham composition. Those values were displayed using Splitstree. The scale bar indicates 0.1 substitutions per site.

Largely, both the GCS values and Splitstree representation align with and reinforce the clustering assignments made through dot plot analysis and comparison of ANI values. Notably, some relationships were highlighted that may have been overlooked using just the nucleotide-based methods. Most notably, and shown in Figure 4.4, is the relationship between the Cluster BE phages and the Cluster BK phages. They share GCS values ranging from 21.5% (Blueeyedbeauty and Starbow) to 28.1% (Annadreamy and Samisti12), with an average GCS of 25.0%. The Splitstree representation in Figure 4.8 also reflects this relationship, as Clusters BE and BK share a prominent branch point on the tree. This relationship is discussed in greater detail below.

Another interesting observation is the seemingly isolated nature of the Cluster BK phages. All phage clusters share at least a low GCS with at least one other cluster included in this study, with the exception of the Cluster BK phages, which share a high degree of GCS with one another (87.2% - 94.8%), but no non-zero GCS value when compared to any other phage included here.

A third observation based is the divergence of GCS values within the Cluster BK phages. Like the Cluster BE phages, those belonging to Cluster BK appear to sort into two groups of phages that are more closely related to one another than they are to phages in the other group. Annadreamy and Blueeyedbeauty share a GCS value of 83.8% and SparkleGoddess and Comrade share a GCS value of 93.3%. However, when Annadreamy is compared to SparkleGoddess and Comrade, the GCS values drop to 58.3% and 58.2%, respectively. Similarly, when Blueeyedbeauty is compared to SparkleGoddess and Comrade, the GCS values drop to 59.6% and 60.0%. This mirrors a trend in the Cluster BE phages, where the GCS values between the phages in Cluster BE1 range from 77.4% to 86.1% and Cluster BE2 range 68.5% to 84%. However, when comparing phages across these subclusters, GCS values drop to between 52.2% and 56.8%. Arguably, the values between the two groups of BK phages are moderately higher than those between the two BE subclusters (58-60% vs. 52-56%), however, the Cluster BK phages may be candidates for reclassification into subclusters.

4.2.2 Cluster Diversity

4.2.2.1 Cluster BD

In the Actinobacteriophage database, there are a total of 59 phages comprising Cluster BD, two of which (Amela and Verse) have been previously described in the literature [62]. The cluster is further divided into five subclusters (BD1 – BD5) and phages within this cluster have been isolated on nine different species in the genus *Streptomyces*: *S. venezuelae*, *S. xanthochromogenes*, *S. griseus*, *S. toxytricini*, *S. lividans*, *S. azureus*, *S. platensis*, *S. sp.*, and *S. coelicolor*. As discussed in Chapter 3, there are 10 phages in this study that make up Cluster BD, i.e. Aaronocolus, BryanRecycles, Caliburn, Eddasa, Hydra, Izzy, Jash, Lorelei, Nabi, and Rana, none of which have been previously described. All belong in subcluster BD1. The 10 phages have an average genome length of 50375 (± 499) bp and an average G+C of 66.0 (± 0.2) percent. All have an 11 bp overhang at their genome termini.

The Cluster BD phages in this study have a total of 748 predicted genes, sorted into 86 different phams. Of those 86, 38 (~ 44%) can be classified as cluster-identifier phams, i.e. phams which are present in all members of a cluster (here, all cluster members included in this study) and not found in phages that belong to any other cluster. There are no orphams present in any of the ten phages presented here.

As Figure 4.9(a) illustrates, the 10 phages show a high degree of nucleotide identity across the length of their genomes, with the highest concentration of differences appearing between positions 17500 and 20000. A cluster representative, Hydra, is shown in Figure 4.9(b) with predicted functions.

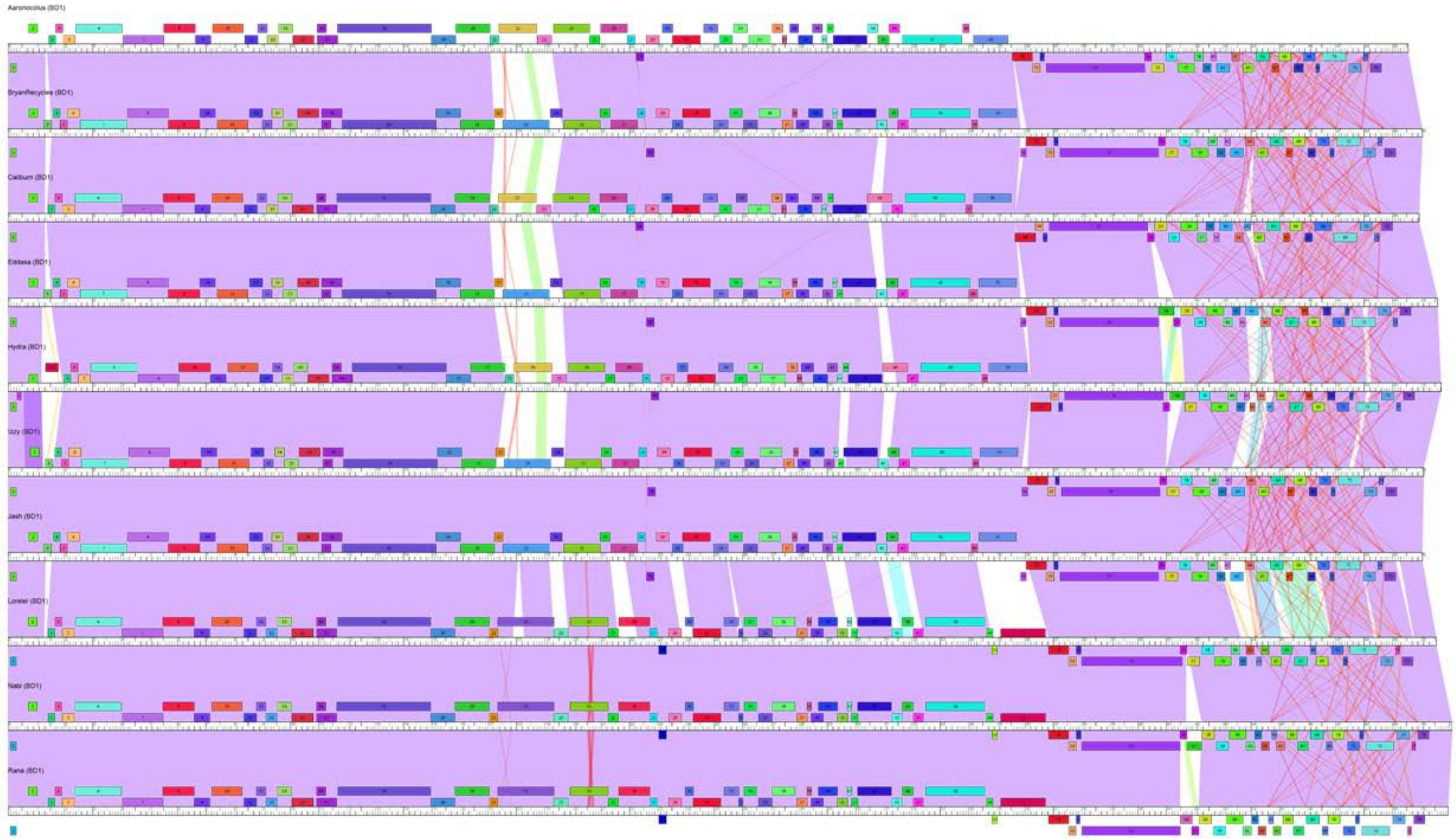
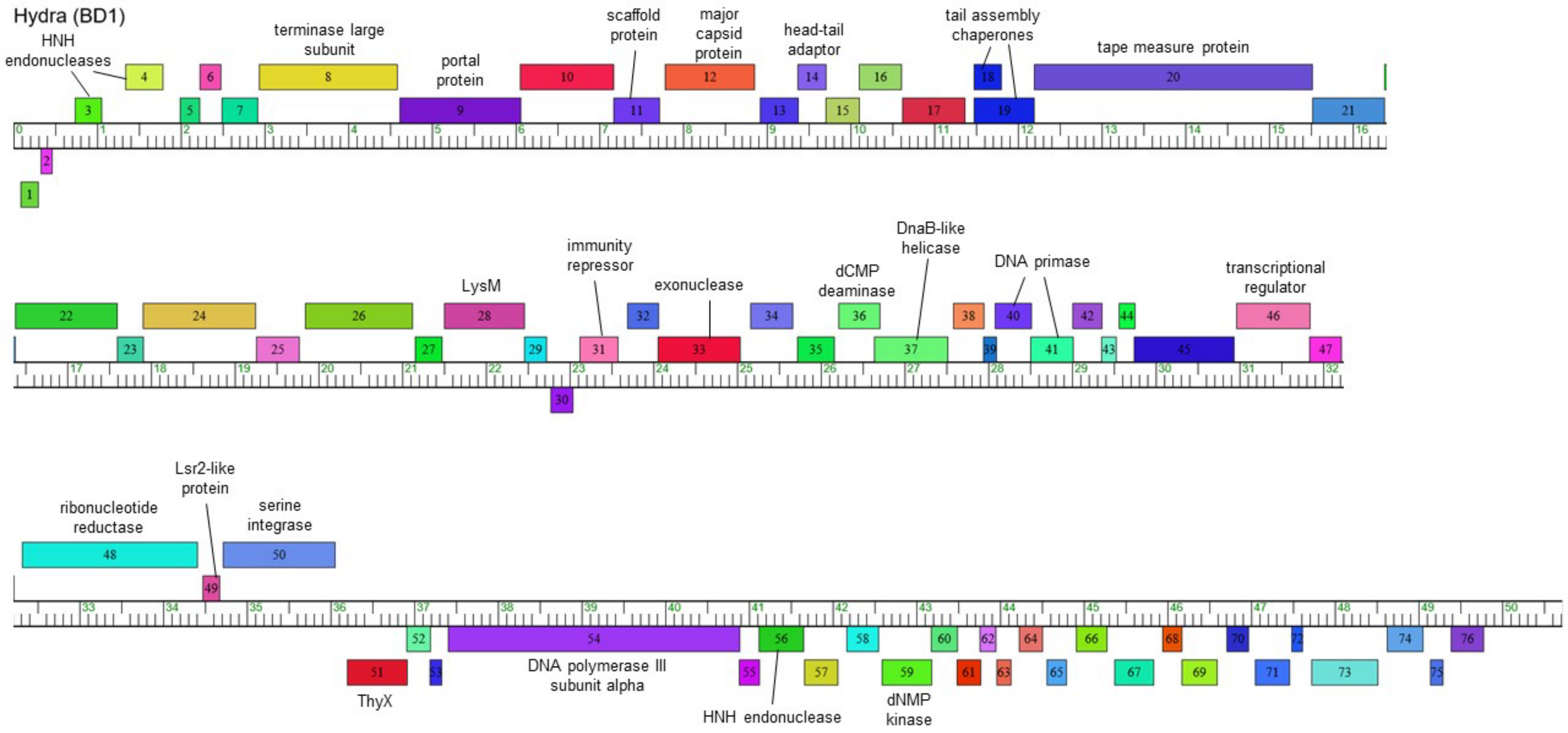


Figure 4.9: (a) Pairwise alignment of 10 Cluster BD phages included in this study. From top to bottom, the pairwise alignment illustrates phages Aaronocolus, BryanRecycles, Caliburn, Eddasa, Hydra, Izzy, Jash, Lorelei, Nabi, and Rana.



(b) Genome organization of *Streptomyces* phage Hydra (Cluster BD).

4.2.2.2 Cluster BE

In the Actinobacteriophage Database, there are a total of 21 phages that comprise Cluster BE. The cluster is further divided into two subclusters, BE1 and BE2, which contain 11 and 10 phages, respectively. Eight phages belonging to subcluster BE1, i.e. Jay2Jay, Mildred21, NootNoot, Paradiddles, Peebs, Samisti12, Sushi23, and Warpy, have been previously described in the literature [63]. A clear majority of the phages in this cluster (16 of the 21 phages, ~ 76%) were isolated using *S. griseus* as host. The other five phages were isolated using *S. lividans* (three phages), *S. viridochromogenes* (one phage), and *S. griseofuscus* (one phage) as host. As discussed in Chapter 3, there are eight phages that make up Cluster BE in this study. Further, three of these phages (Samisti12, Paradiddles, and NootNoot) are classified in subcluster BE1, and five (Karimac, LukeCage, Starbow, StarPlatinum, and Wofford) are classified in subcluster BE2. The eight phages of Cluster BE have an average genome size of 132713 (± 1085) bp and an average G+C of 49.34 (± 0.75) percent. All have large (10666 – 12590 bp) direct terminal repeats at their termini. Further, the eight members of this cluster code for between 41 (LukeCage) and 46 (Paradiddles) tRNAs, with a cluster average of 44.1.

The Cluster BE phages in this study have a total of 1877 predicted genes sorted into 347 different phams. Of those 347 phams, 43 (~ 12%) can be classified as cluster-identifier phams. Further, there are a total of 33 orphams present in the eight phages of this cluster.

Figures 4.10(a), 4.10(b), 4.11(a), and 4.11(b) illustrate Phamerator maps and representative phage annotations for subclusters BE1 and BE2, respectively.

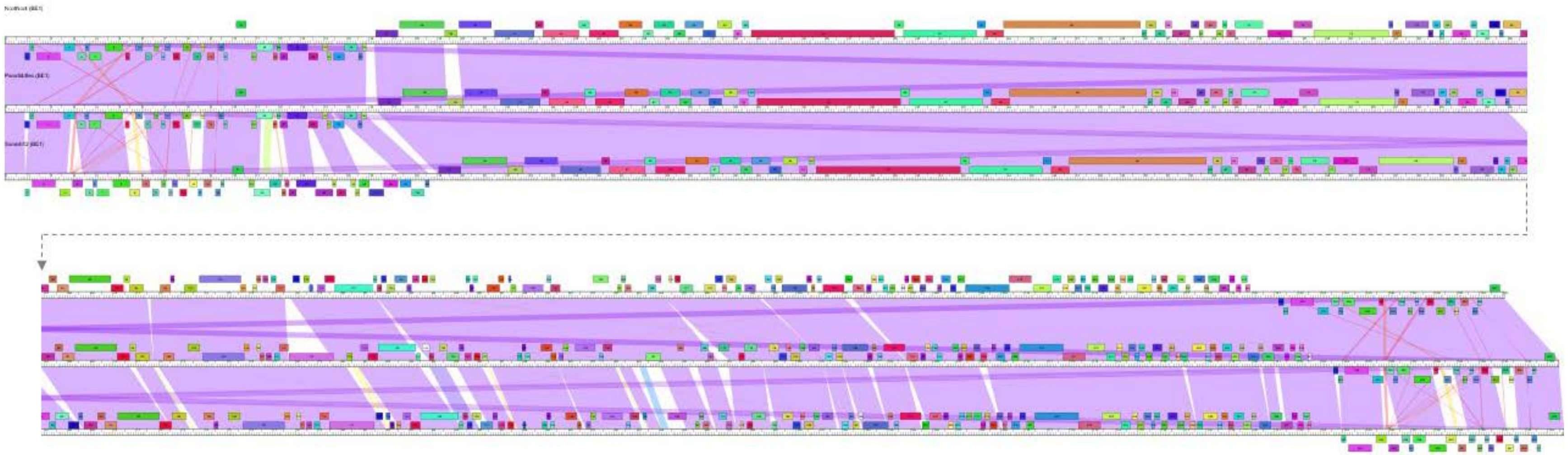
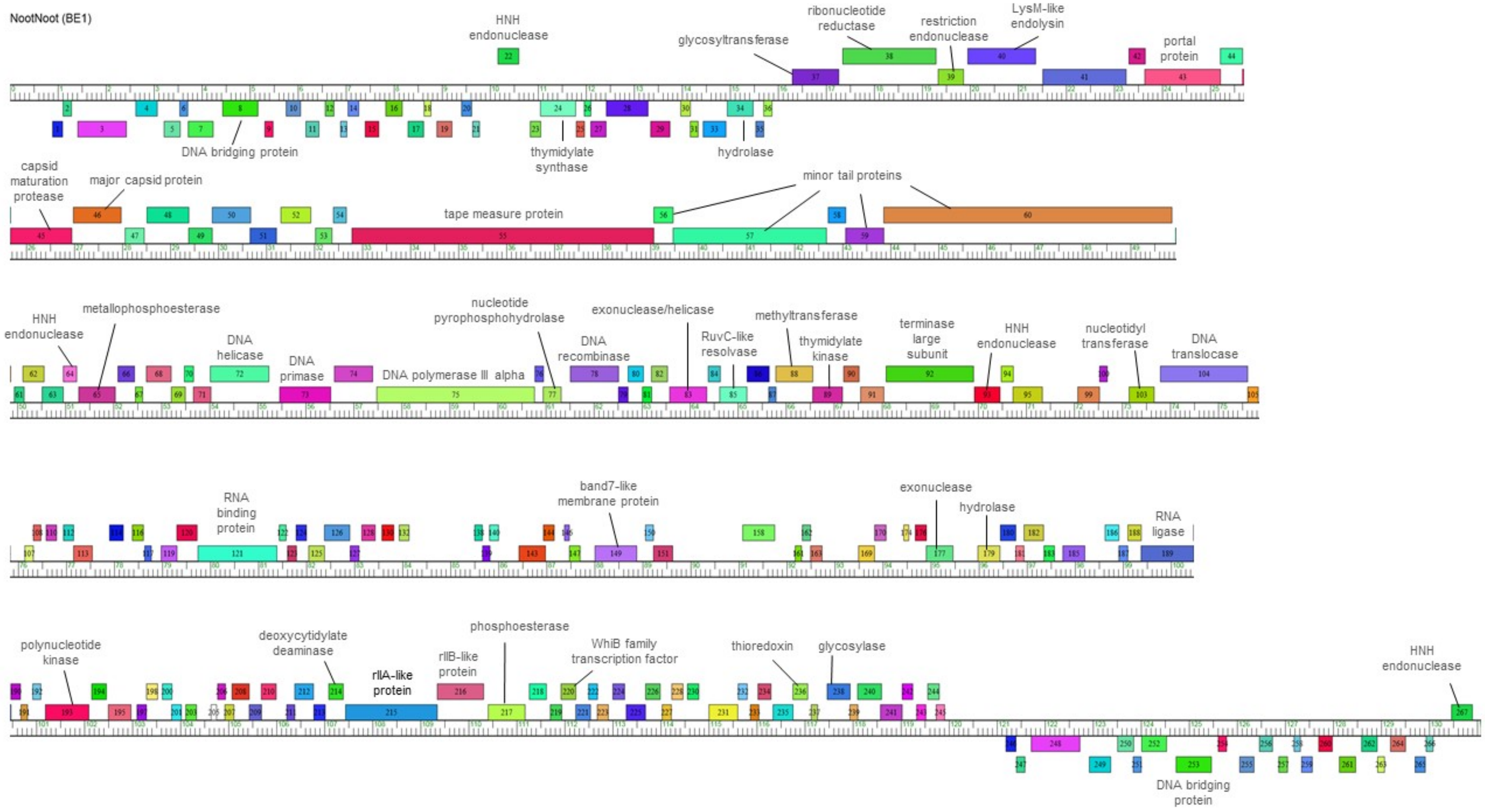


Figure 4.10: (a) Pairwise alignment of three Cluster BE1 phages included in this study. From top to bottom, the pairwise alignment illustrates phages NootNoot, Paradiddles, and Samisti12.



(b) Genome organization of *Streptomyces* phage NootNoot (Cluster BE1).

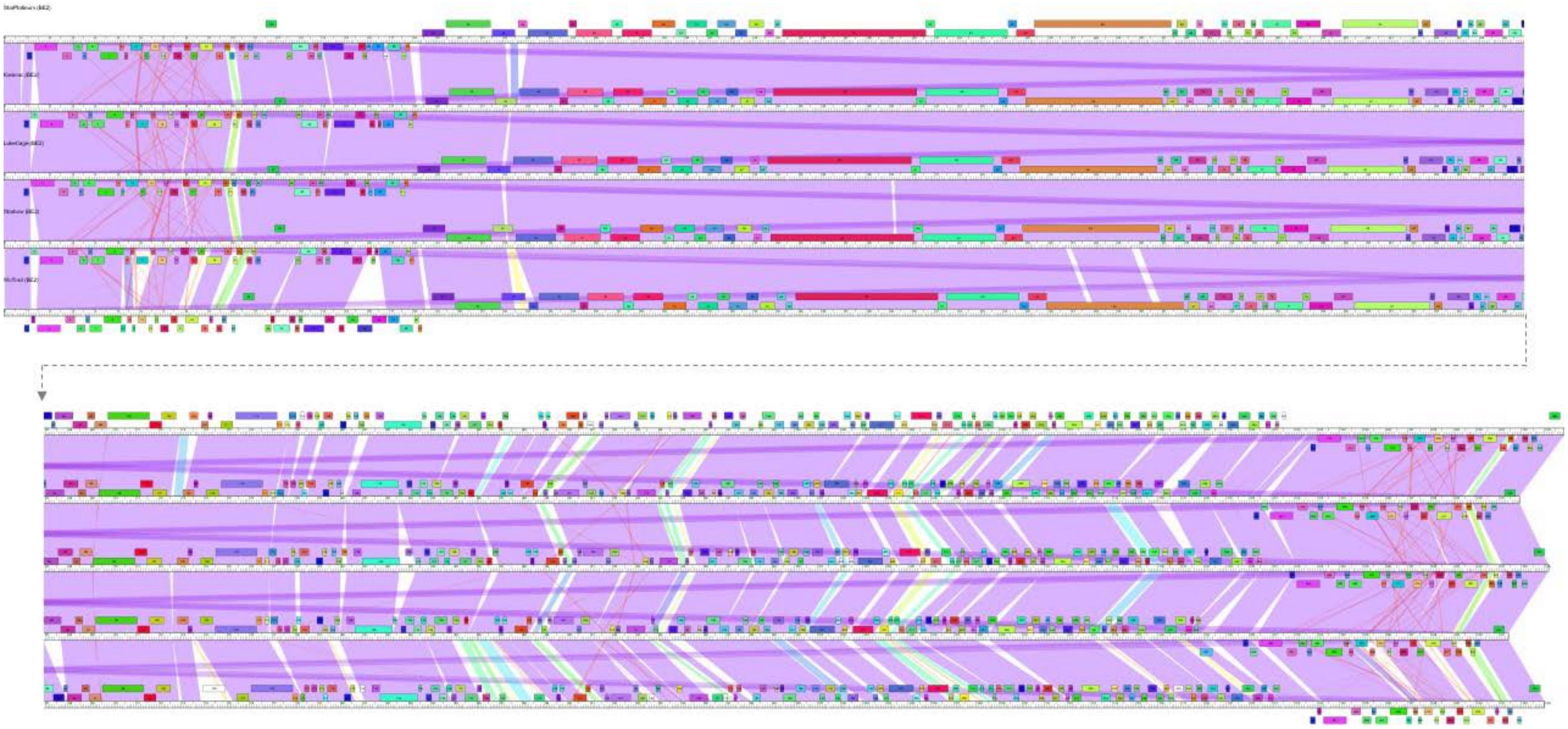
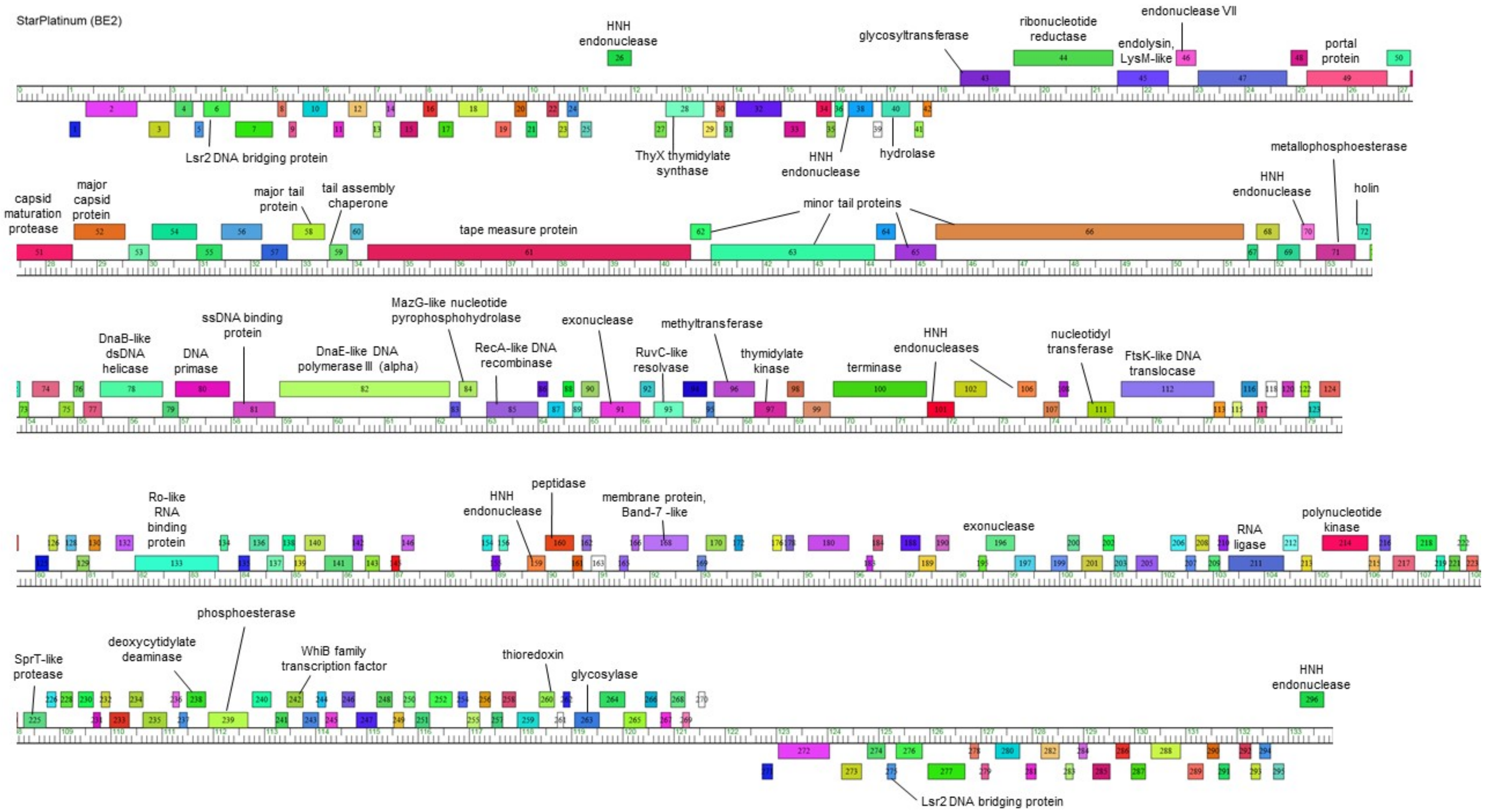


Figure 4.11: (a) Pairwise alignment of five Cluster BE2 phages included in this study. From top to bottom, the pairwise alignment illustrates phages StarPlatinum, Karimac, LukeCage, Starbow, and Wofford.



(b) Genome organization of *Streptomyces* phage StarPlatinum (Cluster BE2).

4.2.2.3 Cluster BF

In the Actinobacteriophage Database, there are a total of ten phages comprising Cluster BF. There are no Cluster BF subclusters. Phages in this cluster have been isolated from four different *Streptomyces* species, including *S. griseus* (~ 70% of isolates), S. *scabiei*, *S. lividans*, and *S. viridochromogenes* (each ~ 10% of isolates). As discussed in Chapter 3, there are six phages that make up Cluster BF for this study, i.e. HaugeAnator, Immanuel3, Percastrophe, Romero, ToriToki, and ZooBear. These six phages have an average genome length of 46087 (± 50) bp and an average G+C of 59.7 (± 0.1) percent. All have a short (between 264 and 275 bp) direct terminal repeat that comprises their termini and have podoviral morphology. To date, Cluster BF phages are the only phages that infect *Streptomyces* that have been described as having podoviral morphology. Additionally, each of these six phages code for a host of tRNAs, with all but one (Immanuel3) coding for 22 tRNAs; Immanuel3 codes for 17.

The Cluster BF phages in this study have a total of 375 predicted genes, sorted into 67 different phams. Of those 67, 56 (~ 84%) can be classified as cluster-identifier phams. The Cluster BF phages included here have a total of two orphams, gene 67 in phage Immanuel3 (pham 47730) and gene 86 in phage Percastrophe (pham 47782).

Illustrated in Figure 4.12(a), the six phages of Cluster BF display an extremely high degree of contiguous nucleotide identity across each of their genomes, a characteristic universally predicted by dot plot analysis, ANI (ranging from 0.9835 to 0.9984), and GCS values (ranging from .937 to .984). A cluster representative, ToriToki, is shown in Figure 4.12(b) with predicted functions. Of the 86 total phams present in ToriToki, putative functions were only assigned for 14 (~ 16%).

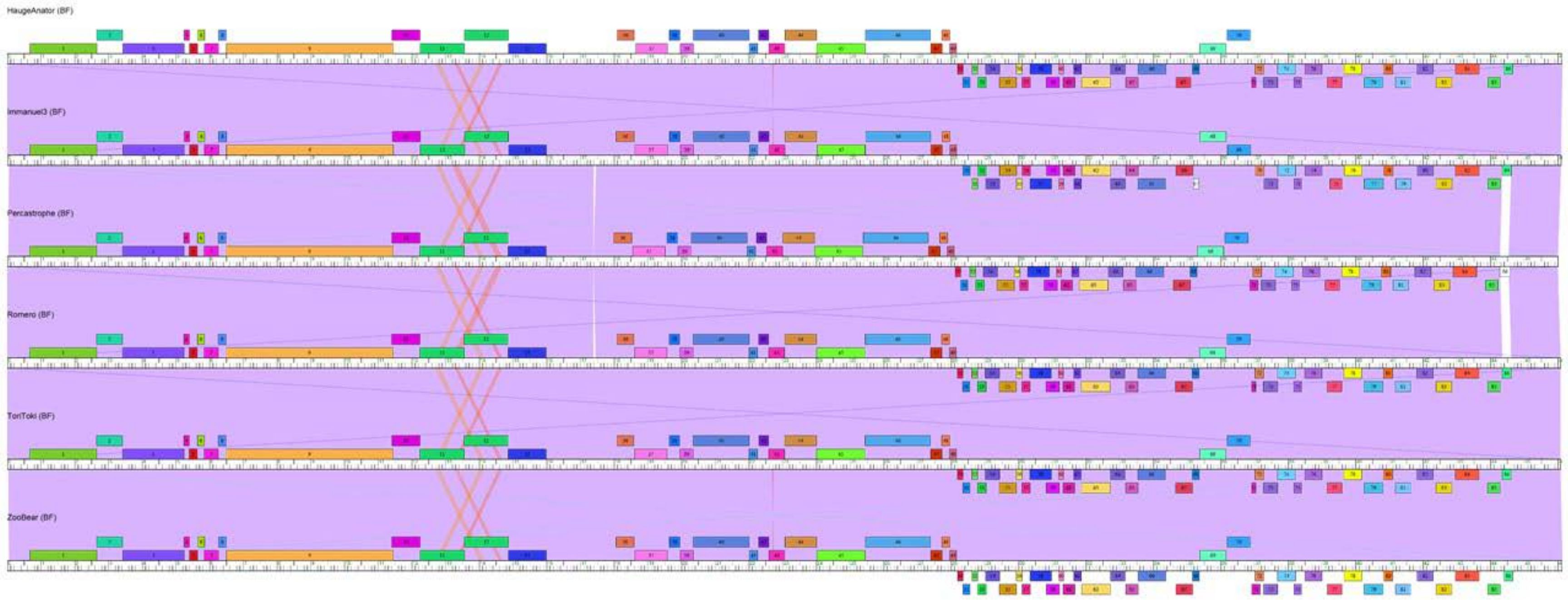
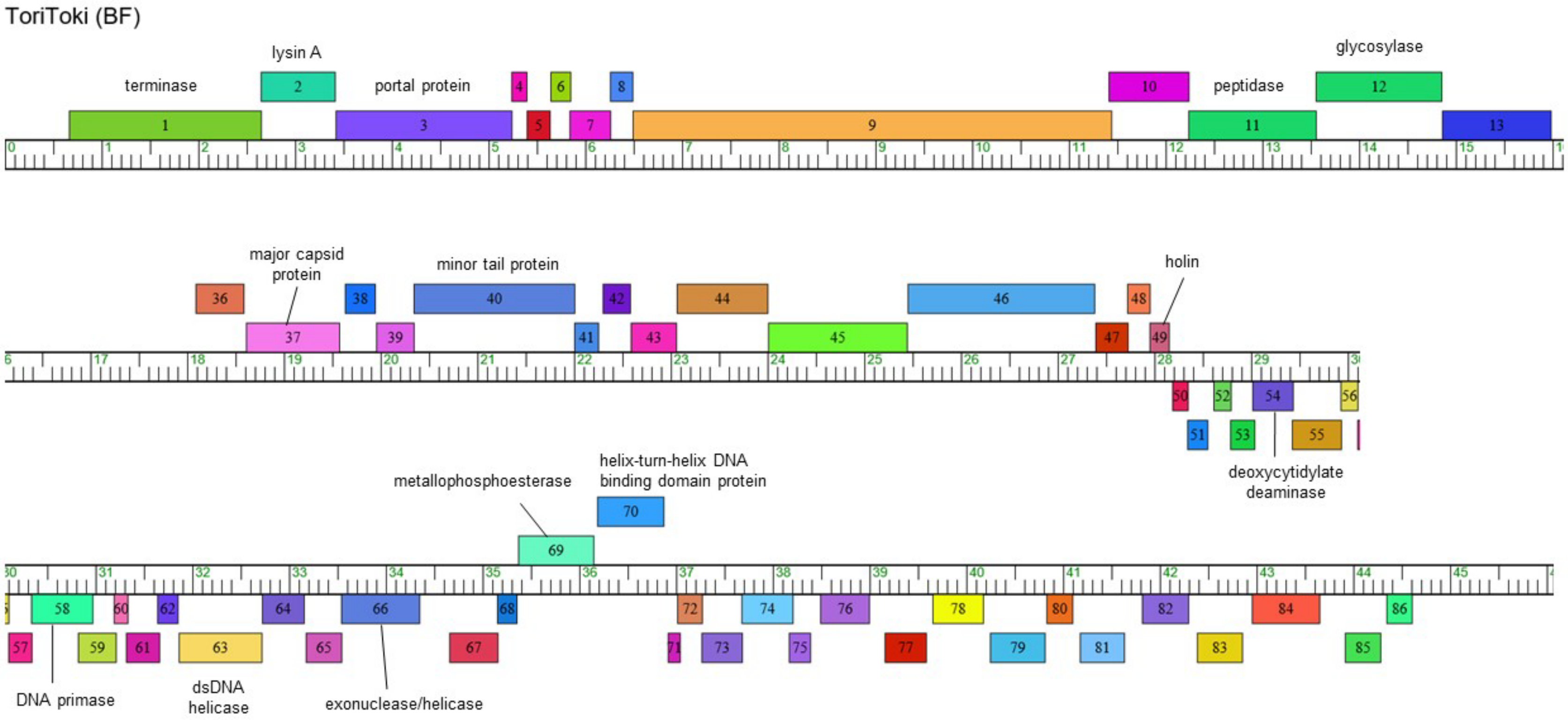


Figure 4.12: (a) Pairwise alignment of six Cluster BF phages included in this study. From top to bottom, the pairwise alignment illustrates phages HaugeAnator, Immanuel3, Percastrophe, Romero, ToriToki, and ZooBear.



(b) Genome organization of *Streptomyces* phage ToriToki (Cluster BF).

4.2.2.4 Cluster BG

In the Actinobacteriophage Database, there are a total of eight phages comprising Cluster BG, five of which (BabyGotBac, Maih, TP1604, Xkcd426, and YDN12) have previously described in the literature [64]. All eight phages were isolated using *S. griseus* as host and, to date, Clusters BG and BH (discussed below) are the only *Streptomyces* phage clusters in the database to be entirely comprised of phages isolated on a single host. As discussed in Chapter 3, seven phages make up Cluster BG for the purposes of this study, i.e. BabyGotBac, BayC, Maih, Salete, TP1604, Xkcd426, and YDN12. These seven phages have an average genome length of 58154 (± 2800) bp and an average G+C of 69.2 (± 0.2) percent. All are circularly permuted with siphoviral morphology and prolate capsids (described in Chapter 3). It is accepted convention in bacteriophage genome annotation to begin the annotation of circularly permuted phages by assigning an arbitrary point upstream from the terminase gene as the +1 position. All Cluster BG phage genomes annotated and included in this study follow this convention.

The seven Cluster BG phages included here have a total of 503 predicted genes sorted into 104 different phams. Of those 104 phams, 43 (~ 41%) can be identified as cluster-identifier phams. There are 27 orphams present throughout genomes of the phage cluster, comprising about 26% of the total phams present. A clear majority of these orphams are found in phage Xkcd426, which contains 24 of the 27 total orphams.

As shown in Figure 4.13(a), six of the seven phages in the cluster show a high degree of nucleotide identity across the length of their genomes. The seventh, Xkcd426, shows enough similarity at the nucleotide level (ANI = 0.7285 – 0.7289) to be clustered with the other six phages, yet has several large regions of dissimilarity spanning the genome. In fact, Xkcd426 is over 7 kbp larger than the other six members of the cluster, a difference that seems largely due to a roughly 6.5 kbp apparent insertion between about positions 17,800 and 24,300. This region contains five of the 24 orphams present in phage Xkcd426. Figure 4.13(b) shows a cluster representative, YDN12, with annotated functions. Of the 71 phams present in this phage, functions were predicted for 20 (about 28%).

4.2.2.5 Cluster BH

In the Actinobacteriophage Database, there are a total of six phages comprising Cluster BH, none of which have been previously described in the literature. As with Cluster BG, this cluster is not divided into subclusters and all of the phages belonging to this cluster have been isolated using *S. griseus* as host. As discussed in Chapter 3, there are five phages making up Cluster BH for this study, i.e. Crosby, Henoccus, JackieB, LazerLemon, and UNTPL. The five phages have an average genome length of 54676 (± 456) bp and an average G+C of 68.22 (± 0.08) percent. All are circularly permuted.

The Cluster BH phages in this study have a total of 408 predicted genes sorted into 106 different phams. Of those 106, 56 (~ 53%) can be classified as cluster-identifiers. There are nine orphams present in these five phages, with the orphams being relatively evenly distributed among the cluster members. Henoccus is the only cluster member lacking orphams.

As illustrated in Figure 4.14(a), the genomes largely align at both the nucleotide and predicted gene product levels. A cluster representative, UNTPL, is also shown in Figure 4.14(b).

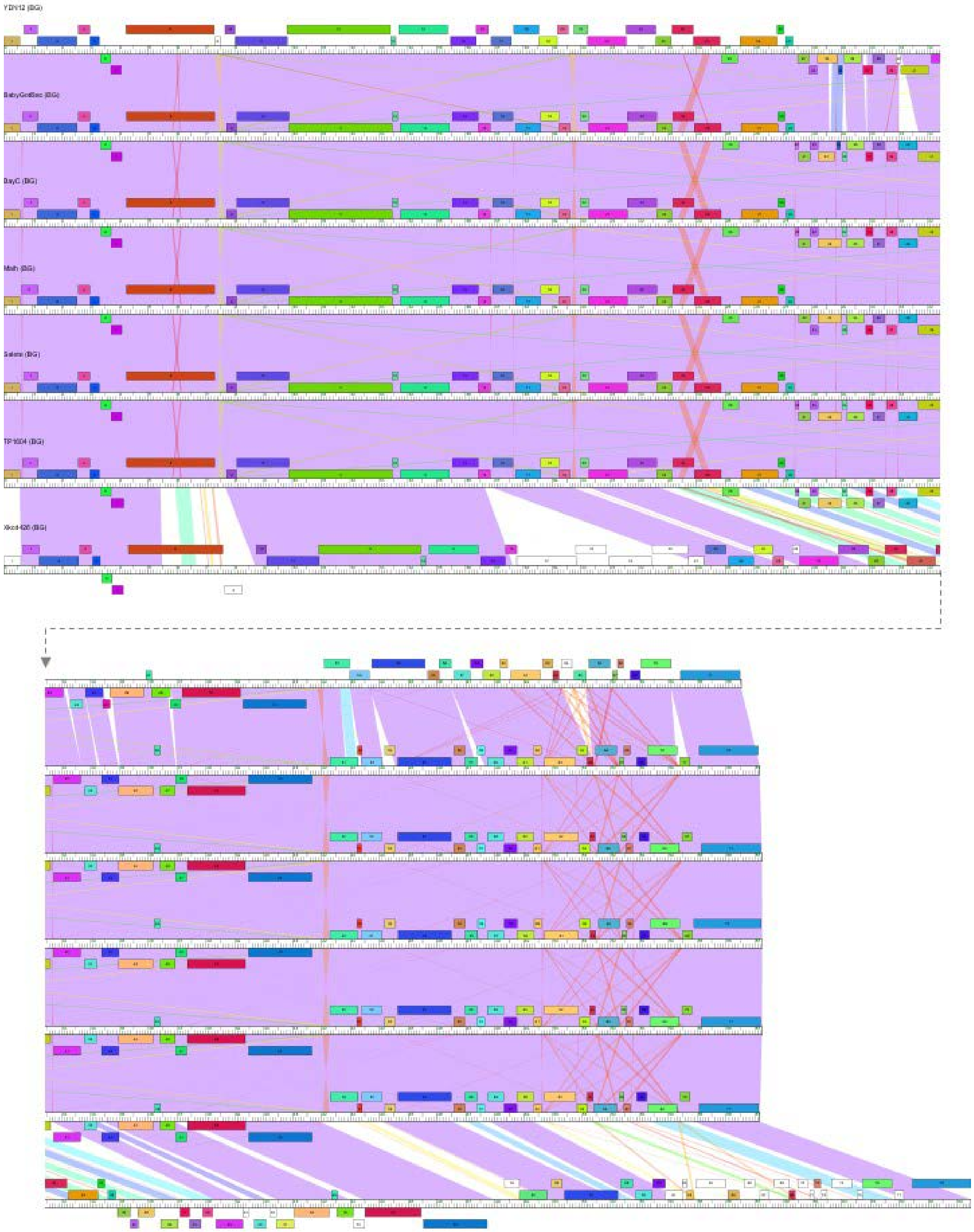
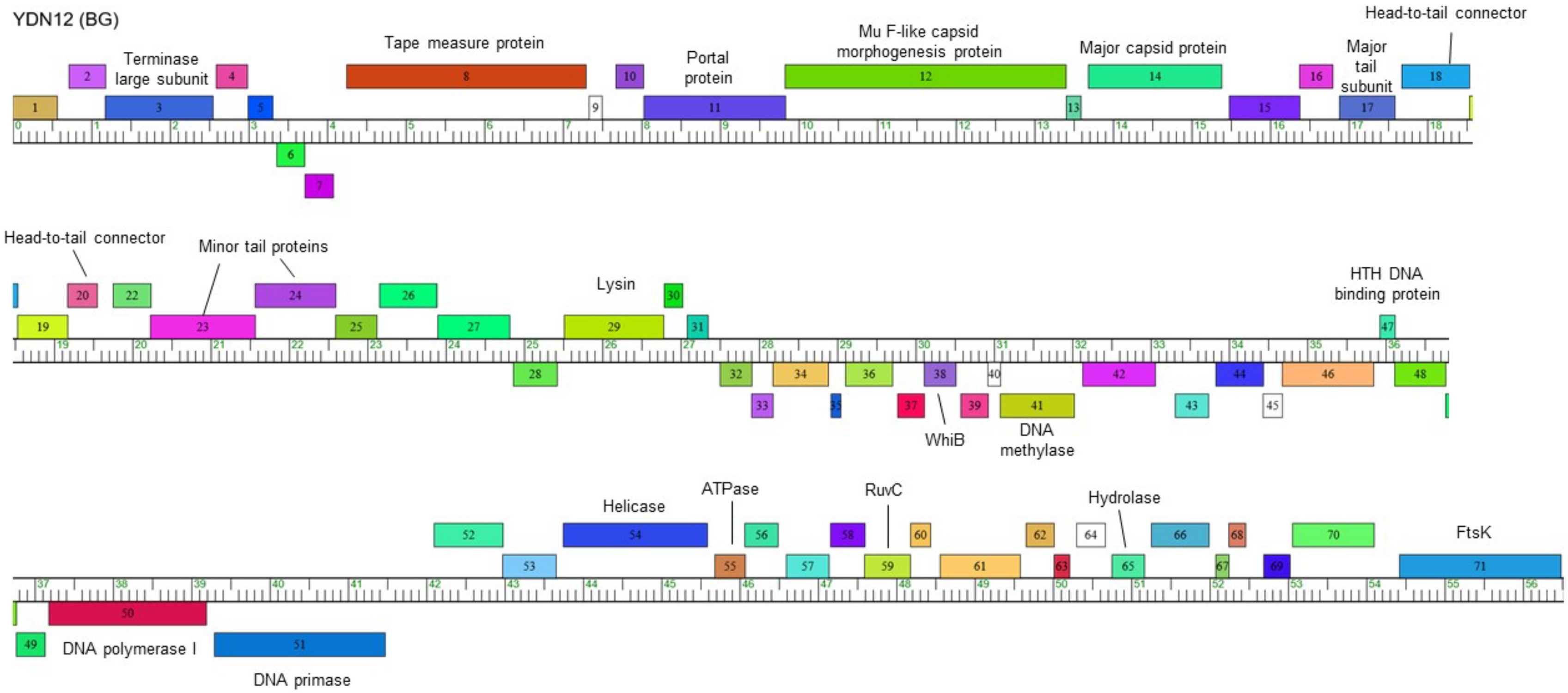


Figure 4.13: (a) Pairwise alignment of seven Cluster BG phages included in this study. From top to bottom, the pairwise alignment illustrates phages YDN12, BabyGotBac, BayC, Maih, Salete, TP1604, and Xkcd426.



(b) Genome organization of *Streptomyces* phage YDN12 (Cluster BG).

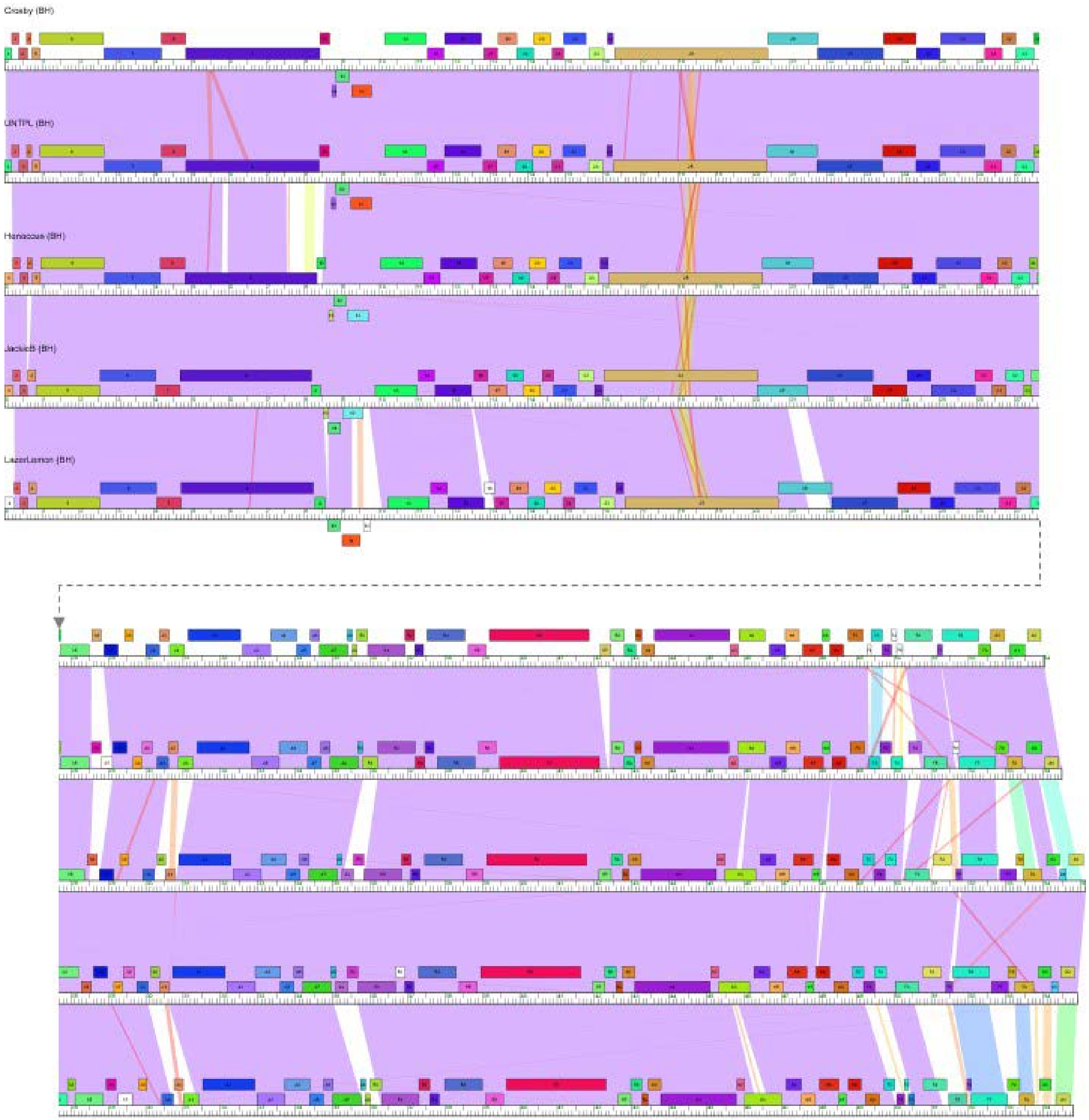
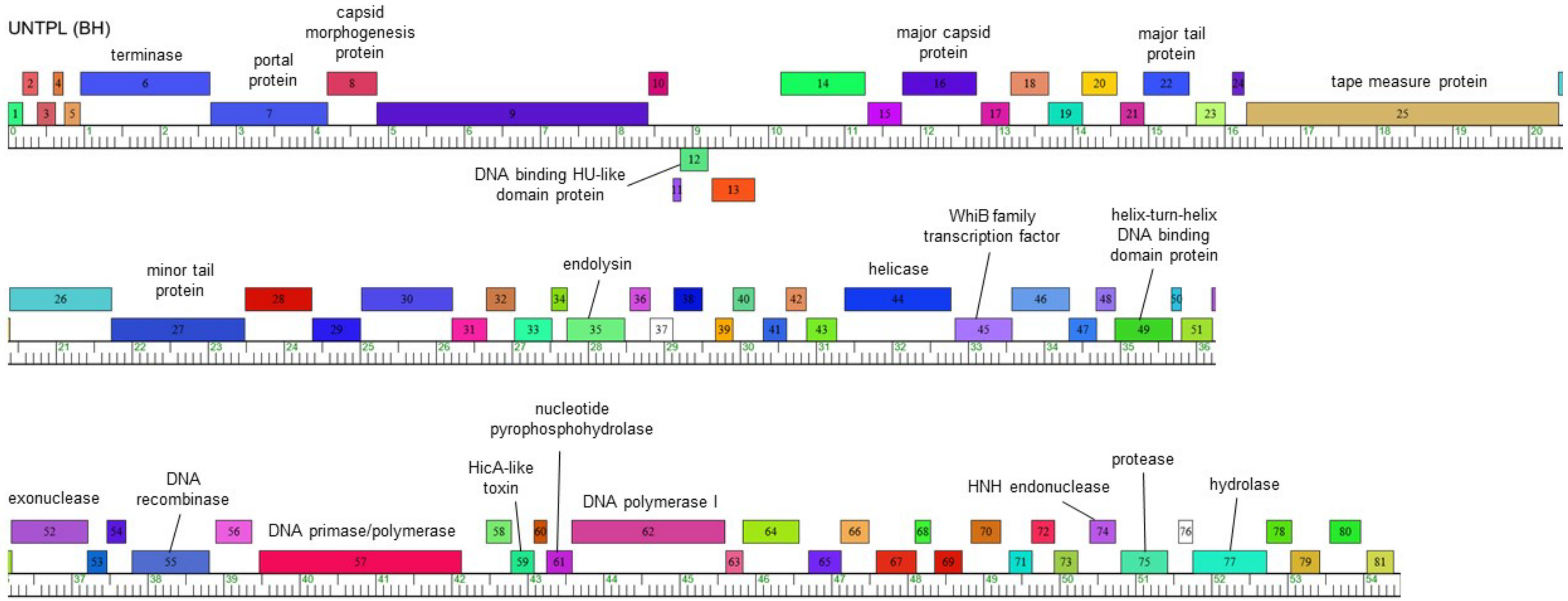


Figure 4.14: (a) Pairwise alignment of five Cluster BH phages included in this study. From top to bottom, the pairwise alignment illustrates phages Crosby, UNTPL, Henoccus, JackieB, and LazerLemon.



(b) Genome organization of *Streptomyces* phage UNTPL (Cluster BH).

4.2.2.6 Cluster BI

In the Actinobacteriophage Database, there are a total of 16 phages comprising Cluster BI. The cluster further divided into four subclusters (BI1 – BI4); the subclusters containing nine, four, one, and two members, respectively. The phages in this cluster have been isolated using six different species of *Streptomyces*, including *S. griseus*, *S. scabiei*, *S. platensis*, *S. virginiae*, *S. lividans*, and *S. azureus*. As discussed in Chapter 3, there are three phages in this study that belong to Cluster BI, i.e. DrGrey, OlympicHelado, and Spectropatronm. All three phages belong to the BI1 subcluster and none have been previously described in the literature. These three phages have an average genome length of 55991 (± 252) bp and an average G+C of 59.5 (± 0). All have a 9 bp overhang at their termini.

The Cluster BI phages in this study have a total of 252 predicted genes sorted into 95 different phams. Of those 95, 47 (~ 49%) can be classified as cluster-identifier phams. There is a single orpham present in the cluster, specifically in phage OlympicHelado (gp 64).

Illustrated in Figure 4.15(a), the three phages in this cluster share a high degree of similarity, spanning the entire length of their genomes, at both the nucleotide and amino acid levels. Most of the dissimilarity may be attributed to three regions. The first appears between positions 6000 and 7000, where phage DrGrey has an apparent 693 bp gene that the other two cluster members lack (gp 10). This gene product has been assigned to pham 21494, sharing similarity with genes found in two *Arthrobacter* phages, Ingrid and Loretta (Cluster AU3). No function has been predicted for this gene. The second appears between positions 42900 and 43400, where OlympicHelado appears to have two genes not shared by the other two phages. The first is the orpham discussed above (gp 64) and the second (gp 65) belongs to pham 29250 sharing similarity with only a single other gene found in phage Skog (Cluster DO, isolated using *Gordonia terrae* as host). The third, and final, region of dissimilarity is found between positions 46400 and 48600 of phage DrGrey, where this phage appears to have two genes not found in the other two cluster members, i.e. one belonging in the pham 10802 (also found in other phages in the cluster that do not appear in this study) and the other belonging to pham 9519, appearing in the same phages as pham 10802, discussed immediately above.

Also found in Figure 4.15(b) is a cluster representative, phage DrGrey. Of the 82 phams present in this phage, putative functions have been assigned for 14 (~ 17%) of these phams. Note that gps 17 and 19 (assigned as major tail proteins in Figure 4.15) belong to the same pham, i.e. pham 7626.

4.2.2.7 Cluster BK

In the Actinobacteriophage Database, there are a total of seven phages that comprise Cluster BK, none of which have been previously described in the literature. The cluster is further divided into two subclusters (BK1 and BK2); the subclusters being made up of five and two phage isolates, respectively. A majority of these phages (four of the seven) were isolated using *S. griseus* as host, and the other three were isolated using *S. griseofuscus* (two phages) and *S. viridochromogenes* (one phage) as host. As discussed in Chapter 3, there are four phages that make up Cluster BK for the purposes of this study, i.e. Annadreamy, Blueeyedbeauty, Comrade, and SparkleGoddess, all of which belong in subcluster BK1. These four phages have an average genome length of 128739

(± 2095) bp and an average G+C of 47.23 (± 0.4) percent. All have short (734 – 789 bp) direct terminal repeats at their termini.

The Cluster BK phages in this study have a total of 931 predicted genes sorted into 341 different phams. Of those 341 phams, 58 (~ 17%) can be classified as cluster-identifier phams. There are 34 orphams present throughout the four phages, with phage Blueeyedbeauty having the most (14 orphams, ~ 42%) and SparkleGoddess having the fewest (three orphams, ~ 9%).

As illustrated in Figure 4.16(a), the phages in this cluster share a high degree of similarity across their left arms, i.e. the left-hand side of the genomes, and diverge across their right arms, i.e. the right-hand side of the genome. The left arms of these phages are predicted to contain the structural genes for the phages, e.g. capsid and tail assembly genes, as well as the DNA/control genes, e.g. helicases and primases. The right arms of each of these phages contain genes with a variety of functions and do not appear to be well conserved across the length of the arm. Also illustrated in Figure 4.16(b) is a representative phage, Blueeyedbeauty. Of the 276 phams present in this phage, putative functions have been assigned for 50 (~ 18%) of these phams. Notably, gp 8 (an orpham) has been classified as an HTH nuclease protein although it does not appear to be significantly related to any phage protein found in the Actinobacteriophage Database. The predicted protein product here was assigned this function based on hits to other HTH endonucleases as determined in HHPRED. This is illustrative of the predictive power of using multiple programs (here, HHPRED, NCBI BLAST, and the Conserved Domain Database (CDD)) to predict functions in phage isolates.

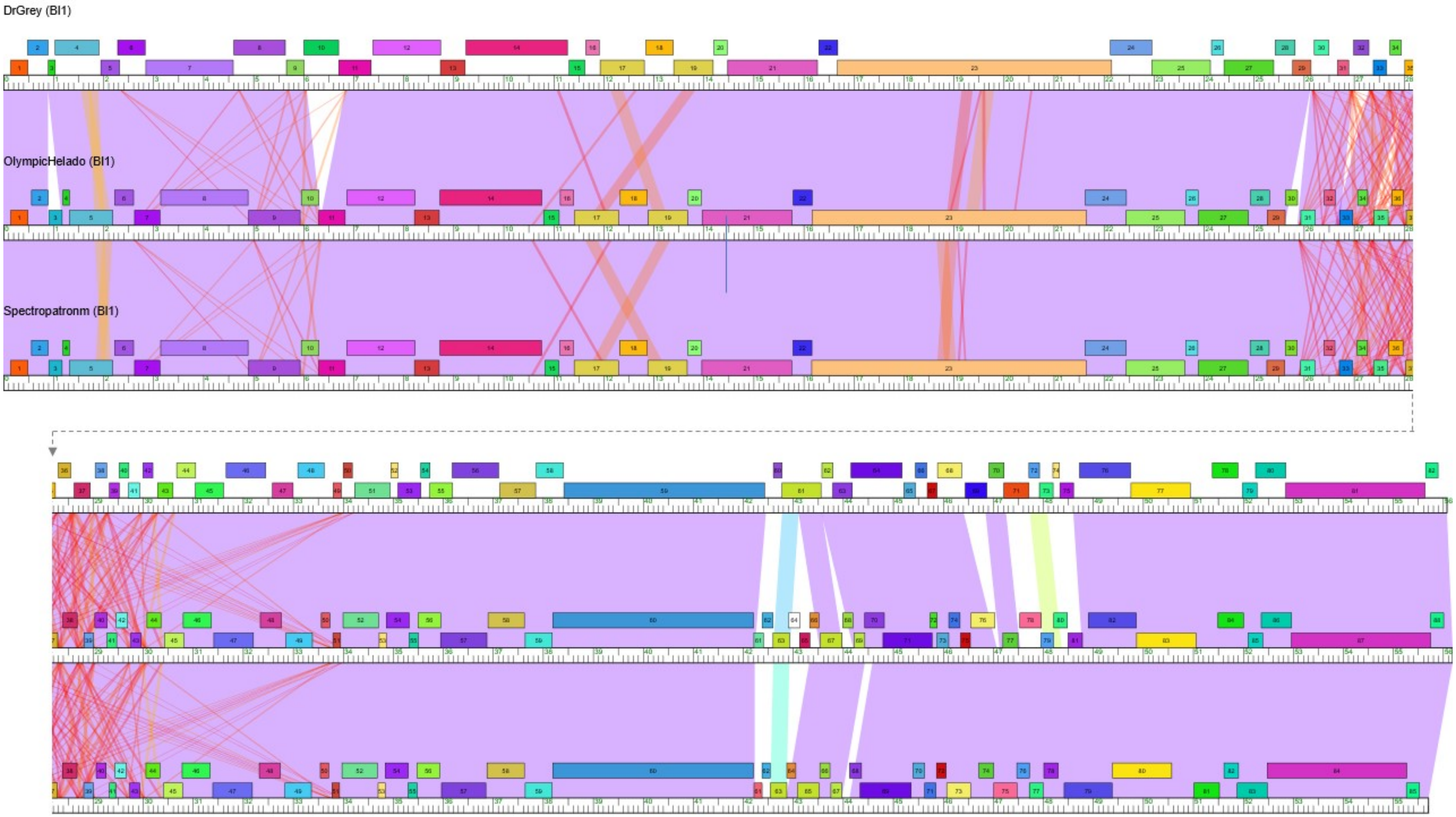
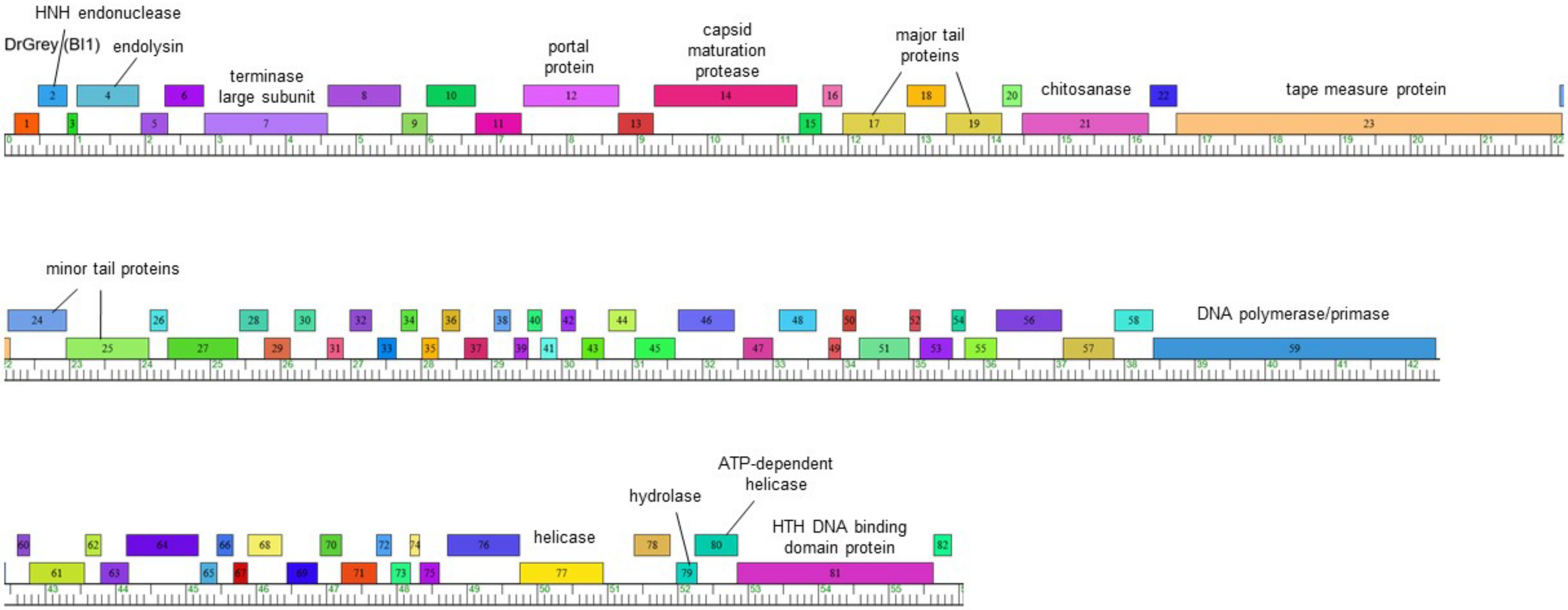


Figure 4.15: (a) Pairwise alignment of three Cluster BI phages included in this study. From top to bottom, the pairwise alignment illustrates phages DrGrey, OlympicHelado, and Spectropatronm.



(b) Genome organization of *Streptomyces* phage DrGrey (Cluster BI).

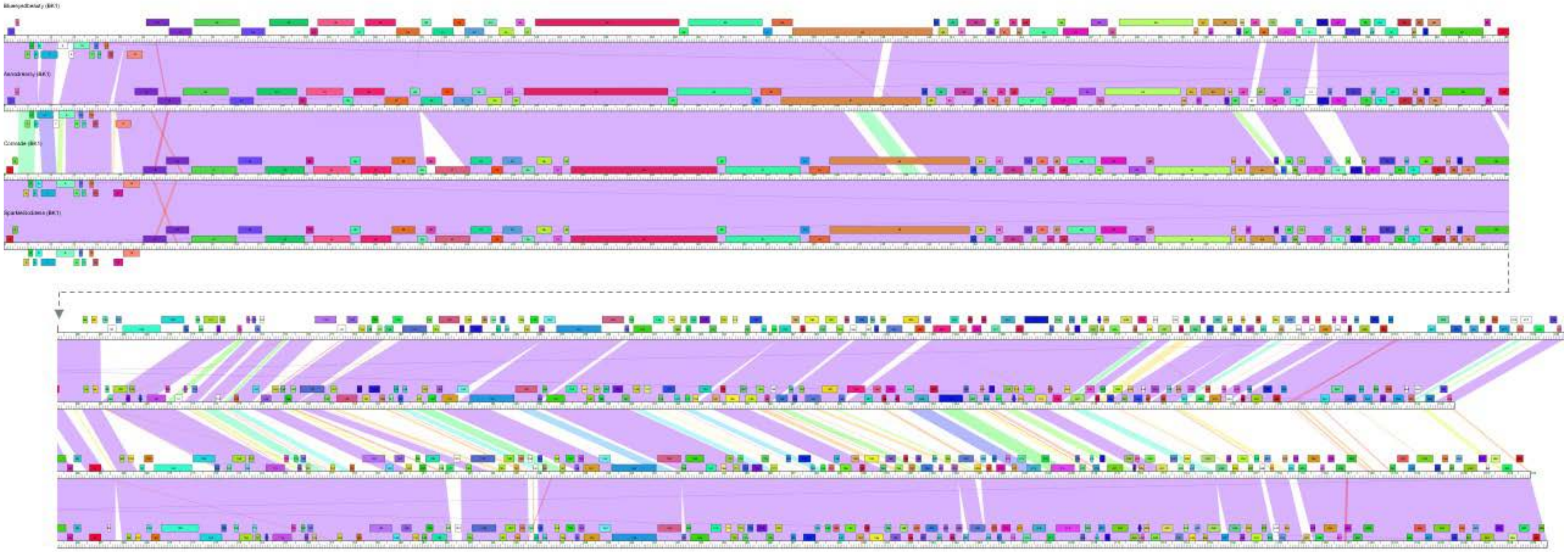
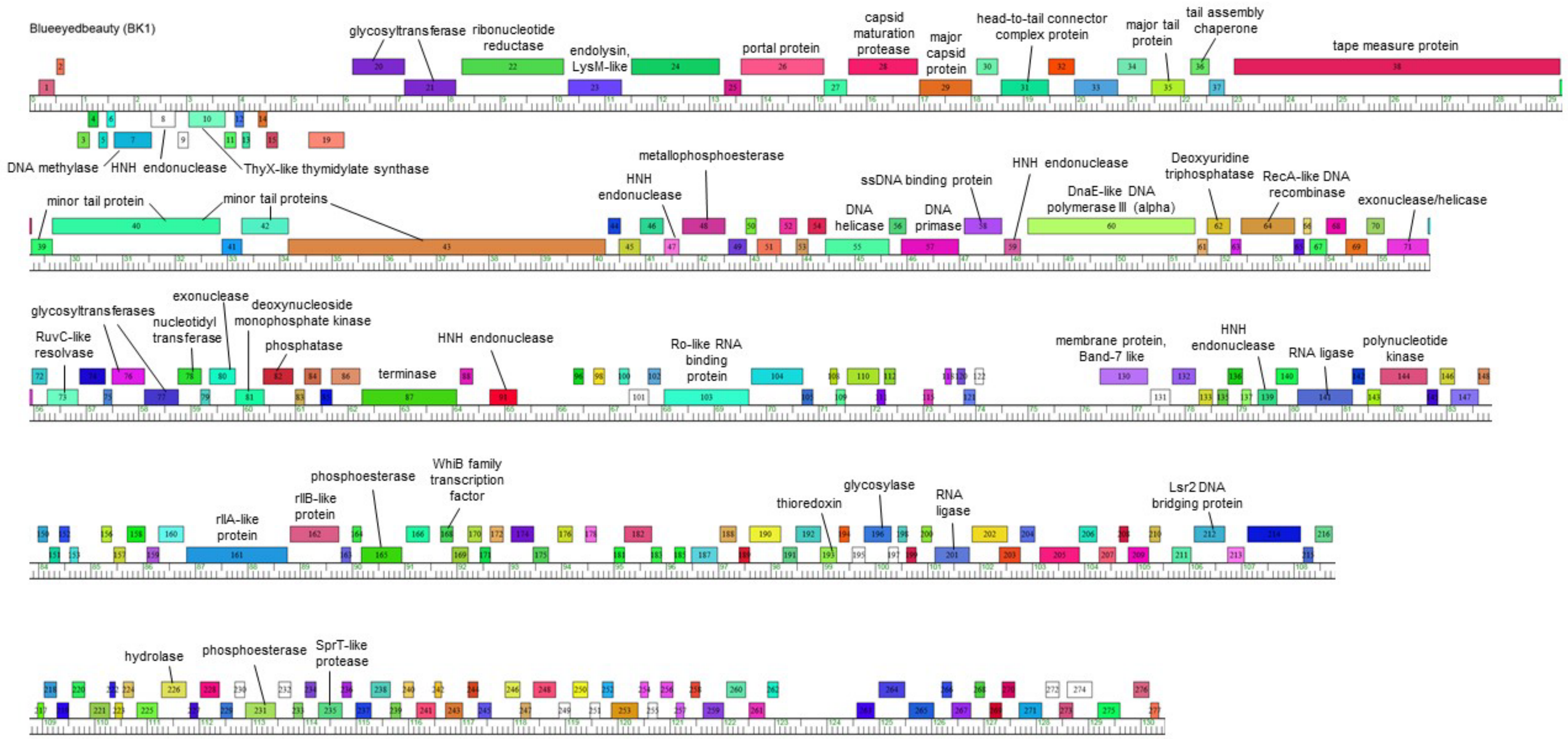


Figure 4.16: (a) Pairwise alignment of four Cluster BK phages included in this study. From top to bottom, the pairwise alignment illustrates phages Blueeyedbeauty, Annadreamy, Comrade, and SparkleGoddess.



(b) Genome organization of *Streptomyces* phage Blueeyedbeauty (Cluster BK).

4.2.2.8 Raleigh and Wentworth

Taking a momentary, higher-level view of phages Raleigh and Wentworth is necessary here, as context is important, and the phages included in this study often do not exist in the artificial “black box” created when selecting phages for a study based on fixed criteria (here, a single isolation host and founding institution, etc.).

In the Actinobacteriophage Database, phages Raleigh and Wentworth belong to Clusters BC and BN, respectively. Cluster BC is made up of 12 phages, further classified into three subclusters (BC1 – BC3). Two Cluster BC phages, belonging to subcluster BC1, have been described previously in the literature [56]. Raleigh is classified in subcluster BC2 and is the first of this subcluster to be described in detail. Phages in this cluster were isolated using six different species of *Streptomyces*, including *S. venezuelae* (seven phages, ~ 58%), *S. himastatinicus* (one phage, ~ 8%), *S. viridochromogenes* (one phage, ~ 8%), *S. platensis* (one phage, ~ 8%), *S. scabiei* (one phage, ~ 8%), and *S. griseus* (one phage, ~ 8%). Raleigh is the only cluster member isolated using *S. griseus* as host. The phages in this cluster have an average genome size of 38822 bp and an average G+C of 72.2. Of the clusters included in this study, Cluster BC has the closest G+C content to the host organism (*S. griseus*, G+C = 72%) of all of the clusters. All cluster members are circularly permuted.

Cluster BN (Wentworth) is made up of three phages and is not further classified into subclusters. Phages in this cluster were isolated using *S. griseus* (Wentworth and Gibson) and *Streptomyces toxytricini* (Yara) as host. All three phages were isolated at the University of North Texas, but Gibson is not included in this study because the annotation has not yet been submitted to GenBank. Overall, the phages in this cluster have an average genome size of 68790 bp and an average G+C of 64.1%. The character of the genome ends is circularly permuted. As such, the +1 position has been arbitrarily set as described above.

Focusing now on the phages in this study, Figure 4.17 illustrates the genome of Raleigh. Of the 53 predicted genes present, 52 (~ 98%) have been assigned to phams. Note that gps 14 and 15 (a) result from a programmed translational frameshift and are alternatively translated gene products, (b) have been assigned to the same pham (pham 24631), and (c) are predicted to be tail assembly chaperones. There is a single orpham (gp 28) present and there is no evidence of function for this gene. Because Raleigh is being treated in isolation from the rest of its cluster members for this study, cluster-identifier phams are not reported. Of the 52 phams present, putative functions have been assigned for 20 (~ 38%) of them.

Moving now to Wentworth, illustrated in Figure 4.18, the 103 total predicted genes have been assigned to 102 different phams. Like Raleigh, discussed above, gps 25 and 26 of Wentworth belong to the same pham, however no function has been assigned for this pham. There are 10 total orphams present in this phage and are distributed throughout its genome.

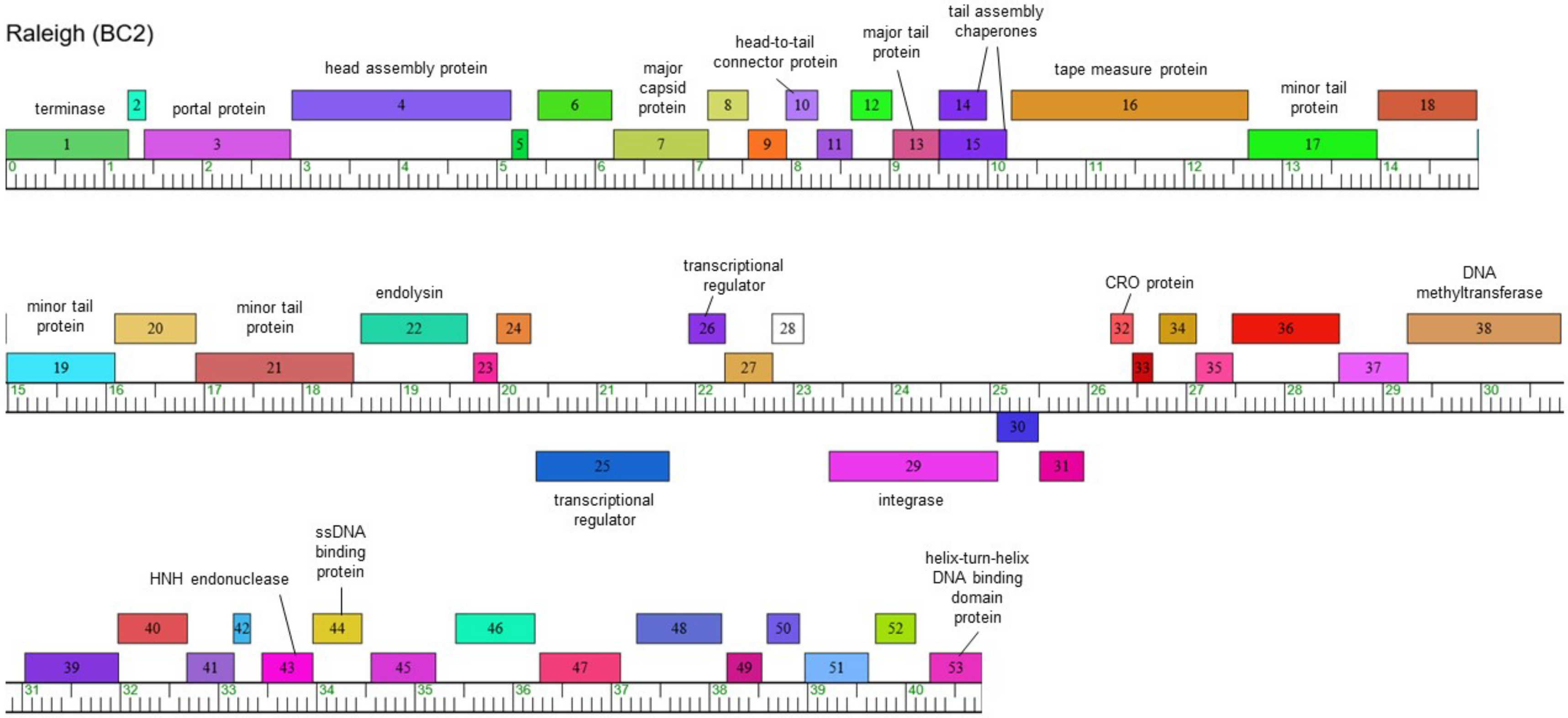


Figure 4.17: Genome organization of *Streptomyces* phage Raleigh.

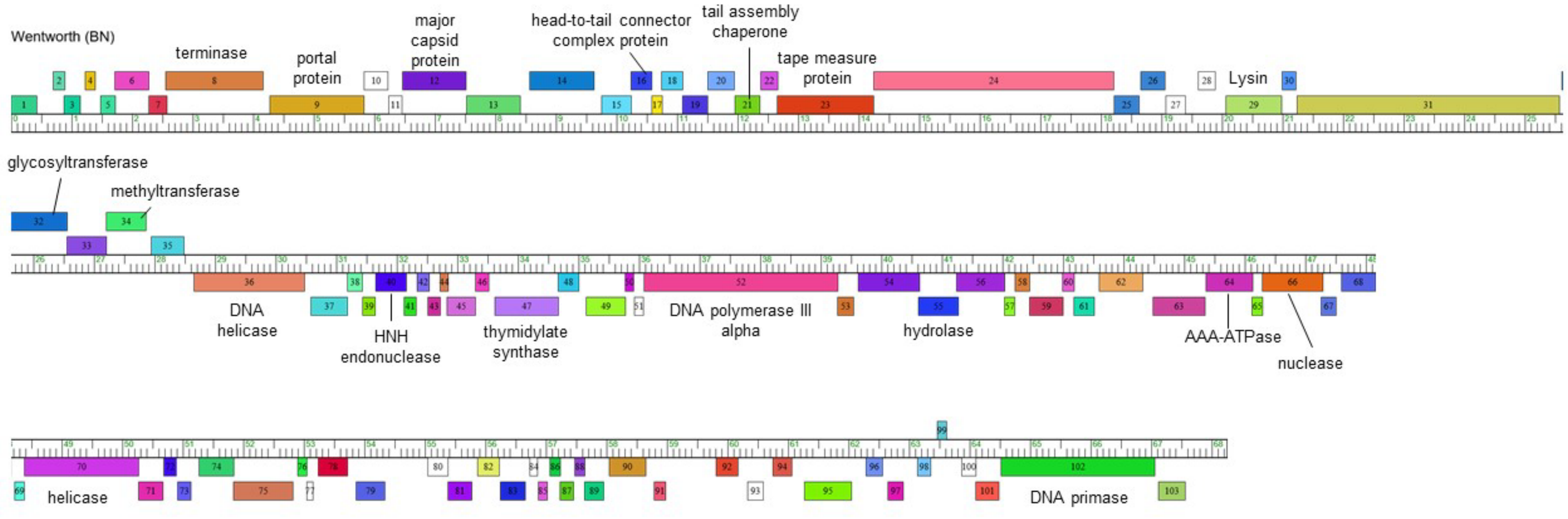


Figure 4.18: Genome organization of *Streptomyces* phage Wentworth.

4.2.3 Measures of Intra- and Inter-Cluster Diversity

The collection of 45 sequenced phages included in this study provide insight into the spectrum of diversity among themselves, both within and among clusters, as well as their relative diversity compared to phages of other hosts. Several metrics were used to describe the diversity within this group of phages, primarily centered around the distribution of phams within and among clusters.

The phages here have a total of 5250 genes which have been grouped into a total of 1300 phams, of which a total of 98 (~ 8%) are orphams. As illustrated in Figure 4.19, the highest concentration of orphams (orange column) is in Cluster BG with about 26% of the 72 total phams being orphams with no close relatives in the database. This is likely due to the presence of phage Xkcd426 in the cluster, as discussed above. With respect to Clusters BD, BE, BF, BH, BI, BK, Raleigh, and Wentworth, the relative number of orphams are all below 10%, with Wentworth and Cluster BK being at 10% and Clusters BE, BF, BH, BI, and Raleigh being at 4%, 3%, 8%, 1%, and 2%, respectively. Notably, there were no evident orphams present in the Cluster BD phages.

Illustrated in Figure 4.19, the diversity within clusters varies greatly. As a measure of diversity within (and ultimately among) clusters, the proportion of cluster-identifier phams was measured. Cluster-identifier phams are those phams which are present in every member of a cluster (here, excluding cluster members which are not included in this study) and not present in any phage belonging to another cluster. The higher the percentage of cluster-identifier phams, the more homogenous (and less diverse) the cluster members are with respect to one another.

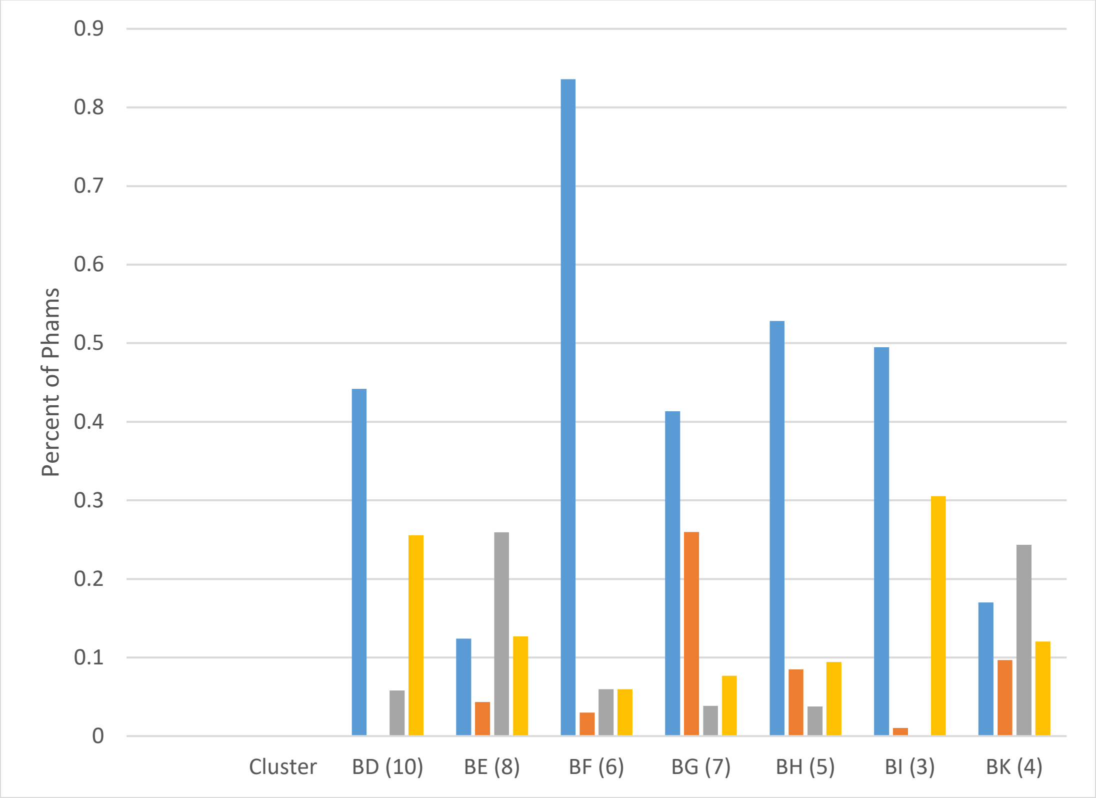


Figure 4.19: Cluster diversity and inter-cluster relationships among 45 phages that infect *S. griseus*. Intra-cluster diversity was measured by (1) the percent of "cluster-identifier" phams (blue bars) present in each of the phage clusters, and (2) the percentage of orphams within the cluster (orange bars). Cluster-identifier phams were not determined for phages Raleigh and Wentworth. Inter-cluster relationships are displayed as (1) the relative frequency of phams found in at least one other *Streptomyces*-infecting cluster (gray bars), and (2) the relative frequency of phams found in at least one phage of a cluster that infects a host outside of the genus *Streptomyces*. The total number of phages, included in this study, found in each cluster is indicated in brackets along with the cluster name.

The proportion of cluster-identifier phams varies from around 12% (Cluster BE) and 17% (Cluster BK) to 84% (Cluster BF) and appears to be independent of the number of phages representing each cluster. The low proportion of cluster-identifier phams present in Clusters BE and BK is not surprising, as these clusters share a high percentage of phams with one another. Discussed above, the gene-content similarities (GCS) between members of these two clusters indicate that, on average, they share about 25% of their genes. The high proportion of cluster-identifier genes present in Cluster BF is also not surprising, as it is the only cluster whose members share GCS values exclusively above 90%. The remainder of the clusters, i.e. Clusters BD, BE, BH, and BI, have proportions of cluster-identifier phams of 44%, 41%, 53%, and 49%, respectively. Because they are the only representatives of their respective clusters included in this study, the number of cluster-identifier phams was not determined for phages Raleigh and Wentworth.

Another metric of diversity measured here is the extent to which these 45 phages appear to exchange genes between clusters, an indication of the isolation of certain clusters from others. The number of phams in each cluster that is also present in at least one phage of another cluster reflects this relationship. Shown in Figure 4.19, five of the seven clusters, as well as Wentworth and Raleigh, share 10% or less of their phams with other related clusters. Clusters BG and BH each share about 4% of their phams with members of other *Streptomyces*-infecting clusters, while Clusters BD and BF share about 6%. Wentworth and Raleigh share 7% and 10%, respectively. Remarkably, the phages comprising Cluster BI (DrGrey, OlympicHelado, and Spectropatronm) do not share a single gene pham with another phage identified as belonging to a *Streptomyces*-infecting cluster. These clusters and the 32 phages that comprise them (about 71% of the phages included in this study) reflect a high degree of cluster isolation. This phenomenon mirrors the *Arthrobacter* phages, where six of the 10 clusters identified when characterizing 48 novel phages shared less than 10% of their gene phams with another phage in an adjacent *Arthrobacter*-infecting phage cluster [56]. Phages infecting the *Mycobacteria* ranged greatly when the same metric was applied [55]. Values for the proportion of phams shared between different clusters of phages infecting that host ranged from 16% to around 77%, with an average percentage of 60.8 [57].

Two clusters, Clusters BE and BK, share a relatively moderate percentage of their gene phams with phages in another *Streptomyces*-infecting cluster. Cluster BE phages share about 26% overall, while Cluster BK phages share about 24%. That this would be the case is not surprising because, as discussed earlier in the text, these two clusters share a moderate degree of both nucleotide and gene similarity with one another; roughly 25%. Indeed, of the 56 total phams in Cluster BK that are shared with a phage belonging to another *Streptomyces*-infecting cluster, only a single pham (Blueeyedbeauty gp 226, pham 32954), or about 0.3% of the total number of phams, was shared with a phage in a *Streptomyces*-infecting cluster other than Cluster BE. Similarly, of the 63 total phams in Cluster BE that are shared with a phage belonging to another *Streptomyces*-infecting cluster, only eight phams (~ 2% of the total number of phams) were shared with a phage in a *Streptomyces*-infecting cluster other than Cluster BK. The relationship between these two clusters is also reflected in the Splitstree phylogeny of Figure 4.8, as these two clusters share a branch. A similar relationship is observed among the *Arthrobacter* phages discussed above, where *Arthrobacter*-infecting Clusters AM and AU, and Clusters AO and AR, share roughly 25% and 20% of their genes, respectively, in a landscape of phage clusters that largely show a high degree of cluster isolation [56].

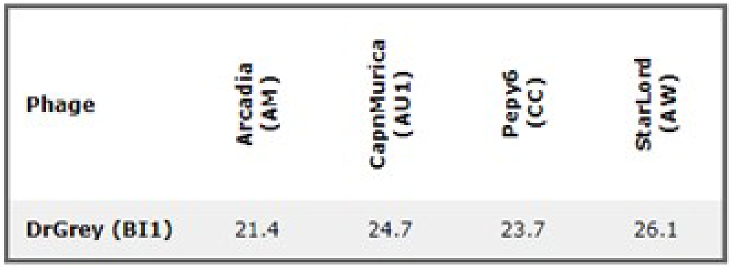
Another useful metric of diversity among these 45 phages is the extent to which the phages share gene phams with phages belonging to clusters that infect actinobacteria outside of the genus *Streptomyces*. Shown in Figure 4.19, the different clusters of *S. griseus* phages here have a wide range of proportions with respect to these cross-genus shared phams, ranging from around 6% of their total phams (Cluster BF and Raleigh) to around 31% of their total phams (Cluster BI). Interestingly, a majority of the clusters reported here (five of the seven clusters and phage Wentworth) appear to share genes with hosts outside of the genus *Streptomyces* with a greater frequency than they share genes with other *Streptomyces*-infecting clusters. Indeed, a majority of these clusters (Clusters BD, BG, BH, and BI) are at least twice as likely to share a pham with a phage in a cluster that infects another actinobacterial host as they are to share a pham with a phage in another *Streptomyces*-infecting cluster.

Nowhere in this data set is this phenomenon more pronounced than in the Cluster BI phages. Discussed above, the Cluster BI phages share no phams in common with other *Streptomyces*-infecting clusters. However, roughly 31% of their gene phams are shared with phages belonging to clusters that infect hosts outside of the genus *Streptomyces*. As illustrated in Table 4.1 and Figure 4.20, Cluster BI phages share a large number of phams with phages in clusters that infect hosts in the genera *Arthrobacter* and *Rhodococcus*. Of the 29 phams that are present in a Cluster BI phage as well as at least one other phage in cluster isolated on a host outside of the genus *Streptomyces*, 25 phams (~ 86%) are present in at least one *Arthrobacter*-infecting cluster. Further, of these 25 phams, 20 (~80%) of these phams are present in at least two different clusters of *Arthrobacter* phages and 18 (~72%) are present in three different clusters of *Arthrobacter* phages.

Similarly, 22 of the 29 shared phams (~ 76%) are present in at least one phage that infects a host in the genus *Rhodococcus*, although there is only evidence that these phams are present in a single *Rhodococcus*-infecting cluster, i.e. Cluster CC. Interestingly, the GCS values, shown in Figure 4.20, closely resemble those found between Clusters BE and BK of the *Streptomyces* phages.

Table 4.1: Cluster BI phams found in at least one other phage cluster isolated on a host outside of the genus *Streptomyces*. The 29 phams listed can be found in phages across 10 different clusters which comprise phages infecting hosts in six different genera.

Pham Identity	Predicted Function	Clusters	Host genera
26743	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
6711	HNH endonuclease	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
4131	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
1601	terminase, large subunit	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
21494	-	AU, BI	*Arthrobacter, Streptomyces*
5056	-	AM, AU, AW, BI, CC, DJ	*Arthrobacter, Gordonia, Rhodococcus, Streptomyces*
3273	portal protein	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
1952	capsid maturation protease	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
2671	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
36813	-	AW, BI, CC, DJ	*Arthrobacter, Gordonia, Rhodococcus, Streptomyces*
7626	major tail protein	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
4669	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
2340	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
5705	tape measure protein	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
5344	minor tail protein	AU, AW, BI	*Arthrobacter, Streptomyces*
32248	minor tail protein	AU, AW, BI	*Arthrobacter, Streptomyces*
2794	-	AU, BI	*Arthrobacter, Streptomyces*
36449	-	BI, DJ	*Gordonia, Streptomyces*
20255	-	BI, CC	*Rhodococcus, Streptomyces*
46374	-	AM, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
2337	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
6612	DNA polymerase/primase	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
3272	helicase	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
50073	-	BI, DJ	*Gordonia, Streptomyces*
32606	hydrolase	AW, BI, CC, DJ	*Arthrobacter, Gordonia, Rhodococcus, Streptomyces*
4051	ATP-dependent helicase	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
4419	HTH DNA-binding domain protein	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
712	-	AM, AU, AW, BI, CC	*Arthrobacter, Rhodococcus, Streptomyces*
51119	-	A12, A20, BI, E, EC	*Microbacterium, Mycobacterium, Streptomyces*



Phage	Arcadia (AM)	CapnMurica (AU1)	Pepy6 (CC)	StarLord (AW)
DrGrey (BI1)	21.4	24.7	23.7	26.1

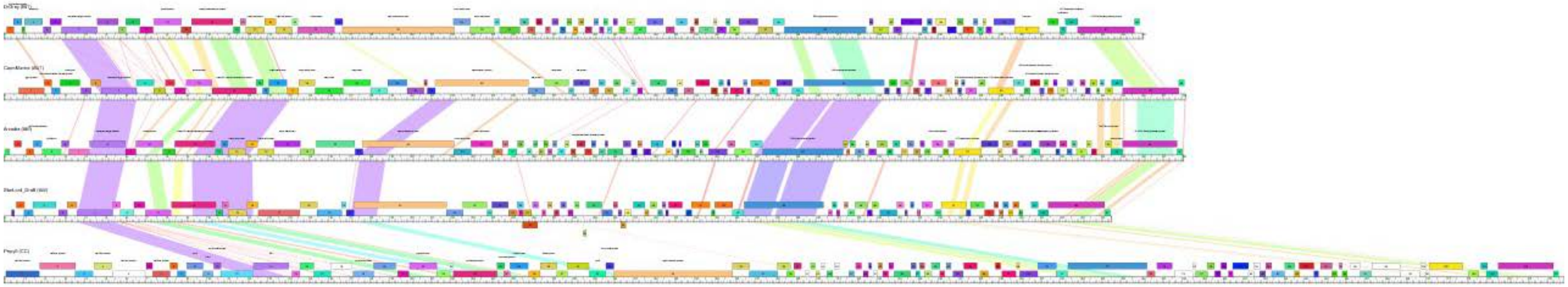


Figure 4.20: GCS values and pairwise alignment for Cluster BI phage DrGrey and phages (1) Arcadia, (2) CapnMurica, (3) Pepy6, and (4) StarLord. From top to bottom, the pairwise alignment illustrates phages DrGrey, Arcadia, CapnMurica, StarLord, and Pepy6.

Further, four phams (5056, 36813, 36449, and 32606) are present in phage clusters that infect a *Gordonia* host, although, like the phams present in *Rhodococcus*-infecting phages, their presence appears to be limited to a single cluster, i.e. Cluster DJ. A single pham, 51119 (gp 30 in both OlympicHelado and Spectropatronm) is present in phages belonging to clusters that infect hosts in the genera *Mycobacterium* (Clusters A12, A20, and E) and *Microbacterium* (Cluster EC).

4.3 Conclusions

Analysis of Gene Content Similarity (GCS) and Splitstree-generated phylogenetic representation of the same largely reinforced the cluster assignments discussed in Chapter 3. Using Phamerator, a program designed specifically for parsing out the complex relationships between the mosaic genomes of bacteriophages, each predicted gene of the 45 phages included here was sorted into one of the 1300 total gene "phamilies," or "phams," present in this collection of *Streptomyces*-infecting phages. Each cluster of phages was analyzed for pham content and similarities and differences are highlighted above.

Metrics of diversity used both within and among phage clusters here include the proportion of total phams present in a cluster that are (1) cluster-identifier phams; (2) orphams; (3) phams present in at least one phage classified in another *Streptomyces*-infecting cluster; and (4) phams present in at least one phage classified in another cluster comprised of phages infecting a host outside of the genus *Streptomyces*. Notably, pham placement in the first two classifications (cluster-identifiers and orphams) are, by definition, mutually exclusive. Placement in the third and fourth classifications are not, and a single pham may be present in both groups, e.g. pham 50477 is present in all of the Cluster BF phages included in this study (HaugeAnator, Immanuel3, Percastrophe, Romero, ToriToki, and ZooBear) and all of the Cluster BD phages, (another *Streptomyces*-infecting cluster), but also present in phages belonging to clusters A2 – A19, which infect hosts in the genera *Mycobacterium* and *Gordonia* .

The above metrics are useful for eliciting relationships and measuring diversity among groups of phages that are (1) largely related, (2) moderately related, or (3) not seemingly related, at least at the nucleotide level. The numbers of cluster-identifier phams and orphams are useful metrics to determine the relatedness (and divergence) of phages for which a relationship is likely evident at the nucleotide level. The higher the percentage of cluster-identifier phams, the less diverse the cluster appears with respect to its individual members. Conversely, the higher the percentage of orphams, i.e. genes with no close relatives in the database, the higher the measure of diversity within the cluster.

Above, Clusters BF and BG illustrate this point. Cluster BF has GCS values ranging from 93.7 (Percastrophe and Immanuel3) to 98.4 (ZooBear and Romero and ToriToki), meaning that the phages in that cluster share roughly between 93.7 and 98.4% of their total phams in common. Meanwhile, the Cluster BG phages share GCS values between 65.4 (Xkcd426 and BabyGotBac) and 98.6 (BayC, TP1604, and Salete). Likewise, the two clusters of phages have cluster-identifier pham proportions of 84% and 41%, respectively. It is clear that the phages comprising Cluster BF are more alike (and potentially isolated, see discussion in Chapter 5) and less diverse from one another, than those phages comprising Cluster BG. A look at the proportion of orphams among these clusters reinforces this assertion. The total number of phams in Cluster BF which are identified as orphams comprises about 3%, while this same measure in Cluster BG is 26%. In essence, one fourth of the total gene phams of Cluster BG have no homologues in the cluster (or the database), while one thirty-third of the phams of Cluster BF fit the same description. Based on these criteria, the phages of Cluster BF are clearly less diverse, with respect to one another, than those belonging to Cluster BG.

Care should be taken when using these metrics, however. As discussed in the introduction and in Chapter 3, clustering is itself, an inexact science and is fraught with complications. This is particularly evident in the Cluster BG phages, where six of the seven phages actually share a high degree of similarity and a single phage, Xkcd426, is remarkably different, at the pham level, than its cluster-mates. In fact, when Xckd426 is removed from the cluster and the metrics are recalculated, the proportion of cluster-identifier phams jumps to 60% (a roughly 46% increase) and the proportion of orphams drops to 4% (a decrease of roughly 85%). However, for the time being, Xkcd426 remains within the threshold genometric values for classification in Cluster BG and, as there is no other phage in the database that is both (1) similar enough to Xkcd416 and (2) dissimilar enough the rest of the Cluster BG phages, does not qualify for subclustering. As such, the metrics calculated for Cluster BG remain valid and the phages of Cluster BG is, as asserted above, more diverse than their counterparts in Cluster BF. This point is simply made to be illustrative of the care that must be taken when making sweeping statements about the phages clustered through this classification scheme.

Overall, diversity within clusters ranges greatly from cluster to cluster. As a whole, it may be said that (of the clusters here and excluding Raleigh and Wentworth) the clusters can be arranged from those with the most intra-cluster diversity to the least intra-cluster diversity as BK > BE > BG > BD > BH > BI > BF.

The above metrics are also useful in determining a cluster's level of genetic communication (through horizontal gene transfer, etc.) with phages in other clusters (1) infecting closely related or identical hosts, or (2) infecting hosts that are not related at the genus level. Overall, the level of genetic communication between clusters presented here varies greatly, with Cluster BF sharing only about 9% of its phams with phages in other clusters, up to Clusters BI and BE, sharing about 31% of their phams each. Clusters BG, BH, BK, BD, and phages Raleigh and Wentworth, share about 10%, 12%, 29%, 30%, 13%, and 17%, of their phams, respectively, with phages in other clusters. The relative likelihoods that each cluster (or individual phage) is in genetic communication with another cluster of phages is, from highest to lowest, BE and BI > BD > BK > Wentworth > Raleigh > BH > BG > BF.

The story shifts when discussing the genetic communication above within the framework of "whom" the phages in a cluster are likely to be in communication with. Two metrics were used (presented above) and highlight differences in the clusters. The first, communication with other phages known to infect hosts in the genus *Streptomyces*, is particularly interesting with respect to the clusters presented here. It is a rational presumption that phages who are known to infect a single host or very closely related hosts, e.g. hosts that are in the same genus, would have an increased access to a common gene pool than phages that infect hosts of different genera. However, considering the relatively moderate-to-high levels of genetic communication discussed above (10% - 31%), the proportions of phams within each cluster that are shared between phages of other *Streptomyces*-infecting clusters is relatively low. The proportions of phams shared between these clusters and other *Streptomyces*-infecting clusters range from 0% (Cluster BI, discussed above) to 24% and 26% (Clusters BK and BE, respectively, also discussed above). Indeed, most of the clusters have proportions of phams shared with other *Streptomyces*-infecting clusters that fall under 10%, a value that is unexpectedly low considering the close proximity granted by the ability to infect a closely-related (if not identical) host. Overall, clusters and individual phages most likely to be in direct genetic communication with phages belonging to other *Streptomyces*-infecting clusters are (in descending order): BE > BK > Raleigh > Wentworth > BD and BF > BG and BH > BI. Conversely, this metric speaks to the relative isolation of these clusters from both one another and from other *Streptomyces*-infecting clusters, a phenomenon discussed more thoroughly below.

Lastly, a view is taken to the levels of communication that these phages share with phages infecting hosts outside of the genus *Streptomyces*. With the exception of Clusters BE and BK and phage Raleigh, all clusters presented here are more likely to share phams with phages infecting hosts outside of the genus *Streptomyces* than they are to share phams with phages in other *Streptomyces*-infecting clusters. Overall, the likelihood that a phage in one of the clusters or an individual phage presented here is in direct genetic communication with another phage infecting hosts outside of the genus *Streptomyces* is (in descending order): BI > BD > BE > BK > Wentworth > BH > BG > BF and Raleigh.

Two major themes have emerged from this study of 45 phages that infect the soil bacterium *S. griseus*. The first is that the degree of genetic diversity among these phages is both (1) high in degree, and (2) not uniform. The second theme to emerge is that different clusters of *Streptomyces*-infecting phages appear to be isolated from one another, a relatively surprising notion considering they all infect a single host species and are assumed to be in direct genetic contact with one another.

The degree of genetic diversity among these phages is high. Unlike the phages of the *Propionibacterium* [65], the *Streptomyces* phages do not appear to be variations on just a couple of sequences and instead appear to be comprised of a rich array of genetic sequences, presumably acquired through genetic contact with a diverse gene pool. Indeed, in this regard the *Streptomyces* phages presented here appear to more closely mirror those phages isolated using hosts in the genera *Mycobacterium* [55]*, Gordonia* [57], and *Arthrobacter* [56].

The genetic diversity of the phages in study is also not uniform. As discussed in previous chapters, when a closer look is taken at the genetic composition of the phages included in this study, clusters begin to emerge. The abundance of genomic information presented here may be sorted into seven clusters, two subclusters, and two phages standing alone. This is comparable to the *Mycobacterium* phages (n = 60) and the *Arthrobacter* phages (n = 46), which were sorted into (1) nine clusters and five singletons, and (2) ten clusters and three singletons, respectively, when a similar number of phage isolates were sequenced and characterized. The *Streptomyces* phages appear to be less diverse than the phages of *Gordonia*, which were sorted into 14 clusters and 14 singletons when a similar number (n = 65) were sequenced and characterized. It is unclear at this time, however, if the breadth of diversity observed in this set of 45 phages is truly representative of the overall *Streptomyces*-infecting phage population or if what we are observing is merely representative of the inherent limitations in isolation procedures, e.g.

enrichment conditions, etc. If the latter is the case, then perhaps new proportions will emerge as isolation conditions, e.g. incubation temperature, growth media, incubation time, isolation methodology (such as enrichment vs. direct plating), etc., are varied.

Another theme that emerges is that different clusters of *Streptomyces*-infecting phages appear to be more isolated that other clusters. As discussed above, the proportions of related genes shared between the different clusters of *Streptomyces*-infecting phages is remarkably low, in most cases below 10%, and in several of those, below 5%. Considering that all of these phages are known to infect a single host (here, *S. griseus*), there is a presumption that they are in direct genetic contact with one another through a common gene pool. However, the fact that the phages in this dataset share so few genes in common with one another is contradictory to that assertion. While it is unclear at this time why this is the case, one possibility is that a number of the phages in this study, while they may have to ability to infect species in the genus *Streptomyces*, in reality have different, but overlapping, host ranges in nature. This hypothesis is supported by the broad diversity in genometric values, particularly the large range, between ~40 and ~70%, of G+C content observed in the dataset. The hypothesis is further supported by the observation that, as a whole, the phages in this study appear to be almost twice as likely to share genes with phages infecting hosts in a genus other than *Streptomyces* than they are with each other.

4.4 Looking Ahead

The 3,314,409 nucleotides of sequence generated through the sequencing of these 45 phages is a rich amount of data, alive with opportunity. While this study has provided insights into the diversity of the phages isolated and used to generate this data, the data itself is truly at the beginning of its useful life. It is a sincere hope that the data will be further analyzed in the future, as it is almost assured to provide additional insight into the wildly diverse and interesting world of bacteriophages.

It is also important to note that 45 phages is a small data in the grand scheme of the phage population, especially considering that the global estimate of phage particles is around 10^{31}. Studies such as this one, databases such as the Actinobacteriophage Database, and ultimately the scientific community as a whole, greatly benefit from network effects, i.e. as usership increases, the value to each individual user increases on a greater-than-linear scale. As the number of *Streptomyces*-infecting phage isolates increases, so will the larger understanding of the truest natures of their diversity.

REFERENCE LIST

1. Weaver W. Molecular Biology: Origin of the Term. Science 1970;170: 581-582.

2. Summers WC. From culture as organism to organism as cell: historical origins of bacterial genetics. J. Hist. Biol. 1993;24: 171-190.

3. Twort FW. An investigation on the nature of the ultra-microscopic viruses. Lancet 1915;2: 1241-1243.

4. Duckworth DH. Who discovered bacteriophage? Bacteriol. Rev. 1976;40: 793-802.

5. d'Hérelle F. Bacterial mutations. Yale J. Biol. Med. 1931;4: 455-461.

6. Muller HJ. Variation Due to Change in the Individual Gene. Am. Nat. 1922;56: 32-50.

7. Demerec M, Fano U. Bacteriophage-resistant mutants in *Escherichia coli*. Genetics. 1945;30: 199-136.

8. Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 1943;28: 491-511.

9. Lederberg J, Lederberg EM. Replica plating and indirect selection of bacterial mutants. J. Bacteriol. 1952;63: 399-406.

10. Homes FL. Reconceiving the Gene: Seymour Benzer's Adventures in Phage Genetics. New Haven: Yale University Press; 2006.

11. Lederberg EM, Lederberg J. Genetic studies of lysogenicity in *Escherichia coli*. Genetics. 1953;30: 51-64.

12. Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biology. 2017 Sep 18. doi: 10.1371/journal.pbio.2003243

13. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227: 561-563.

14. Volkin E. The function of RNA in T2-infected bacteria. Proc. N. A. S. 1960;46: 1336-1349.

15. Chibani-Chennoufi S, Bruttin A, Dillmann M, Brüssow H. Phage-Host Interaction: An Ecological Perspective. J. Bacteriol. 2004;186: 3677-3686.

16. Yoshida-Takashima Y, Takaki Y, Shimamura S, Nunoura T, Takai K. Genome sequence of a novel deep-sea vent epsilonproteobacterial phage provides new insight into the co-evolution of *Epsilonproteobacteria* and their phages. Extremophiles 2013;17: 405-419.

17. Ashelford KE, Day MJ, Fry JC. Elevated Abundance of Bacteriophage Infecting Bacteria in Soil. App. Environ. Microbiol. 2002;69: 285-289.

18. Van Regenmortel MHV, Ackermann H, Calisher CH, Dietzgen RG, Horzinek MC, Keil GM, et al. Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. Arch. Virol. 2013;158: 1115-1119.

19. Maniloff J, Ackermann HW. Taxonomy of bacterial viruses: establishment of tailed virus genera and the order *Caudovirales*. Arch. Virol. 1998;10: 2051-2063.

20. Arnold HP, Zillig W, Zeise U, Holz I, Crosby M, Utterback TT, et al. A Novel Lipothrixvirus, SIFV, of the Extremely Thermophilic Crenarchaeon *Sulfolobus*. Virology. 2000;267(2): 252-266.

21. Prangishvili D, Arnold HP, Gotz D, Zeise U, Holz I, Kristjansson J, et al. A Novel Virus Family, the *Rudiviridaea*: Structure, Virus-Host Interactions and Genome Variability of the *Sulfolobus* Viruses SIRV1 and SIRV2. Genetics. 1999;152: 1387-1396.

22. Haring M, Peng X, Brugger K, Rachel R, Stetter KO, Garrett RA, et al. Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the *Globuloviridae*. Virology. 2004;323: 233-242.

23. Pietila MK, Atanasova NS, Oksanen HM, Bamford DH. Modified coat protein forms the flexible spindle-shaped virion of haloarchael virus His1. Environmental Microbiology. 2013;15(6): 1674-1686.

24. Cabilly S. The Basic Structure of Filamentous Phage and its Use in the Display of Combinatorial Peptide Libraries. Molecular Biotechnology. 1999;12: 143-148.

25. Roux S, Krupovic M, Poulet A, Debroas D, Enoult F. Evolution and Diversity of the *Microviridae* Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. PLoS ONE. 2012;7(7): 1-12.

26. Prangishvili D, Forterre P, and Garrett RA. Viruses of the Archaea: a unifying view. Nature Reviews Microbiology. 2006;4: 837-848.

27. Kivela HM, Kalkkinen N, Bamford DH. Bacteriophage PM2 Has a Protein Capsid Surrounding a Spherical Proteinaceous Lipid Core. Journal of Virology. 2002;76: 8169-8178.

28. Suttle CA. Marine Viruses – major players in the global ecosystem. Nat Rev Microbiol. 2007;5: 801-812.

29. Hendrix RW. Bacteriophage genomics. Curr Opin Microbiol. 2003;6: 506-511.

30. Stern A, Sorek R. The phage-host arms race: Shaping the evolution of microbes. Bioessays. 2010;33: 43-51.

31. Hyman P, Abedon S. Bacteriophage Host Range and Bacterial Resistance. Advances in Applied Microbiology. 2010;70: 217-248.

32. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nature Reviews Microbiology. 2010;8: 317-327.

33. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Current Opinion in Microbiology. 2011;14: 321-327.

34. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage phiX174 DNA. Nature. 1977;265: 687-695.

35. Hatfull G. Bacteriophage Genomics. Curr Opin Microbiol. 2008;11(5): 447-453.

36. Juhala RJ, Ford ME, Duda RL, Youlon A, Hatfull GF, Hendrix RW. Genomic Sequences of Bacteriophages HK97 and HK022: Pervasive Genetic Mosaicism in the Lambdoid Bacteriophages. J. Mol. Biol. 2000;299: 27-51.

37. Casjens SR. Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microb. 2005;8(4): 451-458.

38. Grose, JH, Casjens SR. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014 Nov. doi: 10.1016/j.virol.2014.08.024.

39. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Scientific Reports. 2013;3: 2471.

40. Smith, MCM, Hendrix RW, Dedrick R, Mitchell K, Ko CC, Russell D, et al. Evolutionary Relationships among Actinophages and a Putative Adaptation for Growth in *Streptomyces*. Journal of Bacteriol. 2013;195(21): 4924 – 4935.

41. Gomez-Escribano JP, Bibb MJ. *Streptomyces* coelicolor as an expression host for heterologous gene clusters. Methods Enzymol. 2012;517: 279-300.

42. Anné J, Van Mellart L. *Streptomyces lividans* as host for heterologous protein production. FEMS Microbiol Lett. 1993;114(2): 121-128.

43. Bekiesch P, Basitta P, Apel A. Challenges in the Heterologous Production of Antibiotics in *Streptomyces*. Arch Pharm. 2016;349: 594-601.

44. Diaz LA., Hardisson C, Rodicio R. Isolation and Characterization of Actinophages Infecting *Streptomyces* species and Their Interaction with Host Restriction-Modification Systems. J. of Gen. Microbiol. 1989;135: 1847-1856.

45. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, and Hatfull GF. Phamerator: a bioinformatics tool for comparative bacteriophage genomics. BMC Bioinformatics. 2011;12: 395.

46. Gordon D, Abajian C, Green P. Consed: A Graphical Tool for Sequence Finishing. Genome Research. 1998;8: 195-202.

47. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Research. 1998;8: 186-194.

48. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Research. 1998;8: 175-185.

49. Russell DA. Sequencing, Assembling, and Finishing Complete Bacteriophage Genomes. In: Clokie M., Kropinski A., Lavigne R. (eds) Bacteriophages. Methods in Molecular Biology, vol 1681. New York, NY: Humana Press; 2018.

50. Ross MG, Russ C, Costello M, A. Hollinger, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013 May 29. doi: 10.1186/gb-2013-14-5-r51

51. Delcher AL, Harmon D, Kasif S, White O, and Salzburg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Research. 1999;27(23): 4636-4641.

52. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes, and viruses. Nucleic Acids Research. 2005 Jun 27. doi: 10.1093/nar/gki487.

53. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Research. 2013;41: 36-42.

54. Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007;23(8): 1026-1028

55. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko CC, et al. Comparative genomic analysis of sixty mycobacteriophage genomes: Genome clustering, gene acquisition, and gene size. J Mol. Biol. 2010;397(1): 119-143.

56. Klyczek KK, Bonilla JA, Jacobs-Sera D, Adair TL, Afram P, Allen KG, et al. Tales of Diversity: Genomic and morphological characteristics of forty-six *Arthrobacter* phages. PLoS ONE. 2017 Jul 17. doi: 10.1371/journal.pone.0180517.

57. Pope WH, Mavrich TN, Garlena RA, Guerrero-Bustamante CA, Jacobs-Sera D, Montgomery MT, et al. Bacteriophages of *Gordonia* spp. Display a Spectrum of Diversity and Genetic Relationships. mBio. 2017 Aug 15. doi: 10.1128/mBio.01069-17.

58. Laslett D, Canback B. ARAGORN, a program for the detection of transfer RNA and transfer-messenger RNA genes in nucleotide sequences. Nucleic Acids Research. 2004;32: 11-16.

59. Liu M, Gill JJ, Young R, Summer EJ. Bacteriophages of wastewater foaming-associated filamentous *Gordonia* reduce host levels in raw activated sludge. Scientific Reports. 2015;5: 13754.

60. Dyson ZA., Tucci J, Seviour RJ, Petrovski S. Lysis to Kill: Evaluation of the Lytic Abilities, and Genomics of Nine Bacteriophages Infective for *Gordonia* spp. And Their Potential Use in Activated Sludge Foam Biocontrol. PLoS ONE. 2015 Aug 4. doi: 10.1371/journal.pone.0134512.

61. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 2006;23(2): 254-267.

62. Layton SR, Hemenway RM, Munyoki CM, Barnes EB, Barnett SE, Bond AM, et al. 2016. Genome Sequences of *Streptomyces* Phages Amela and Verse. Genome Announc. 2016 Feb 18. doi: 10.1128/genomeA.01589-15.

63. Hughes LE, Shaffer CD, Ware VC, Aguayo I, Aziz R, Bhuiyan S, et al. Eight Genome Sequences of Cluster BE1 Phages That Infect *Streptomyces* species. Genome Announc. 2018 Jan 11. doi: 10.1128/genomeA.01146-17.

64. Donegan-Quick R, Gibbs ZA, Amaku PO, Bernal JT, Boyd DAM, Burr AR, et al. Genome Sequences of Five *Streptomyces* Bacteriophages Forming Cluster BG. Genome Announc. 2017 Jul 13. doi: 10.1128/genomeA.00502-17.

65. Marinelli LJ, Fitz-Gibbon S, Hayes C, Bowman C, Inkles M, Loncaric A, et al. *Propionibacterium acnes* Bacteriophages Display Limited Genetic Diversity and Broad Killing Activity against Bacterial Skin Isolates. mBio. 2012 Sep 25. doi: 10.1128/mBio.00279-12.