CHARACTERIZATION OF VIRAL COMMUNITIES IN SOIL, ACTIVATED SLUDGE, AND INFLUENT

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisory committee. This thesis does not include proprietary or classified information

_____
Erin Jean Consuegra

Certificate of Approval:

_____	_____
Sharon Roberts	Mark Liles, Chair
Associate Professor	Assistant Professor
Biological Sciences	Biological Sciences
_____	_____
John F. Murphy	George T. Flowers
Professor	Dean
Plant Pathology	Graduate School

CHARACTERIZATION OF VIRAL COMMUNITIES IN SOIL, ACTIVATED SLUDGE, AND INFLUENT

Erin Jean Consuegra

A Thesis

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Auburn, Alabama
August 10, 2009

CHARACTERIZATION OF VIRAL COMMUNITIES IN SOIL, ACTIVATED SLUDGE, AND INFLUENT

Erin Jean Consuegra

_____
Signature of Author

_____
Date of Graduation

VITA

Erin Jean Consuegra, daughter of Dr. John L. Carroll and Dr. Lezlie D. Carroll, was born August 25, 1982, in Winter Haven, FL. In 2004 she received her Bachelors of Science degree in Biomedical Sciences at Auburn University in Auburn, Alabama. She married Erik M. Consuegra in March 2008. In 2005, she enrolled in Auburn University to pursue a Masters degree in Microbiology in the Department of Biological Sciences at Auburn University. During her enrollment as a graduate student she was supported by graduate teaching assistantships, graduate research assistantships, and an NSF GK-12 Fellowship.

THESIS ABSTRACT

CHARACTERIZATION OF VIRAL COMMUNITIES IN SOIL, ACTIVATED SLUDGE, AND INFLUENT

Erin Jean Consuegra

Master of Science, August 10, 2009
(B.S., Auburn University, 2004)

122 Typed Pages

Directed by Mark R. Liles

Viruses are the most abundant yet uncharacterized biological entities on the planet. This thesis is a survey of viral communities in soil, activated sludge (AS), and influent (IN). Both culture-dependent and culture-independent techniques were used to assess the viral communities. Linker-amplified shotgun subclone viral metagenomic libraries were constructed from all environments and a bacteriophage culture collection was constructed from the activated sludge sample for use in comparison of the two culture assessment methods.

The bioinformatics analysis of the metagenomic libraries revealed that the viral communities studied were not well characterized in the GenBank databases and supported the hypothesis that different environments harbor distinct populations of viruses.

ACKNOWLEDGEMENTS

The author would like to thank Mark R. Liles for making it possible to conduct research at Auburn University, and for his valuable lessons and guidance in the development and implementation of these projects. She would like to thank K. Erik Wommack and Jaysheel Bhavsar from the University of Delaware for their bioinformatic contributions and insight. The author would like to acknowledge Larissa Parsley for her contribution of the bacterial collection from activated sludge. She is grateful for the technical support, patience, and helpfulness of the members of the Liles lab. The author would like to thank her science instructors past and present who have given her knowledge and encouragement. She is also grateful for the support of her husband and family.

Style manual or journal used: Applied and Environmental Microbiology

Computer software used: Microsoft Word 2003, EndNote X1

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

Viruses are an integral part of every environment. They not only are of clinical importance to living organisms, but also play a major role in the recycling of nutrients and elements in the environment. The study of viruses and their role in the environments is a constantly expanding field today, but it wasn't that long ago that scientists were puzzled over "filterable toxins" not knowing they were about to discover the most diverse and abundant biological entities on the planet.

LITERATURE REVIEW

A brief history of virology

Before the advent of Germ Theory the causes of infectious disease were unknown. There were several different beliefs on what could cause disease, such as bad odors and witchcraft, but no real evidence to support these guesses. The bone formation of an ancient Egyptian skeleton is evidence of the oldest know viral caused infection, Polio, dating back to c. 3700 BC (27, 52). Around 1500 BC Egyptian hieroglyphs shows a leg deformity characteristic of polio disease symptoms (27, 82). Pharaoh Ramses V who ruled Egypt from 1149-1145 BC is thought to have suffered from smallpox (27). Although unsure of what was causing smallpox disease, the Chinese are said to have began practicing variolation as early as 1000 BC (27). This practice no doubt saved some of the Chinese people from the devastating effects of the variola virus and the disease it causes (27). Diseases in plants and animals were described as well. Aristotle described rabies in dogs, and tulip mosaic disease was first described in 1576 by Carolus Clusius (82). Before virology even existed as a science, the first vaccine was developed in 1796 for smallpox infections (82).

The microbial world became observable in the mid-1600's with the invention of the microscope by Leeuwenhoek. This, along with experiments from many scientists, led to the development of the germ theory. As the disciplines of bacterial pathogenesis and bacterial ecology were developing, the importance of viruses in disease and ecology remained unknown, waiting for their own discovery (82).

Aldolf Mayer, a German agricultural chemist, became the director of the Agricultural Experiment Station at Wageningen, Holland in 1876 (27). During his time in Holland, he investigated an infection of tobacco plants that gave infected leaves distinct coloring patterns. Mayer performed a number of experiments to show that the condition was infectious but was never able to find a bacterium or fungus that was associated with the disease. His conclusion was that the disease would be later revealed to be caused by a bacterium (27, 96). Following Mayer, a Russian scientist, Dmitrii Ivanovsky, also investigated tobacco diseases in the Crimea (27). Ivanovsky used a Chamberland filter-candle to filter the sap of infected leaves (27). By doing this he was able to discover that the infectious agent was smaller than anything known at that time. He made the assumption that the disease was caused by either a dissolvable toxin that remained in the filtered sap or by a bacterium so small it could pass through the filter pores (27). His discovery was presented to the St. Petersburg Academy of Science on February 12, 1892 (75). In 1898, Martinus W. Beijerinck began experiments with diseased tobacco plants in the Netherlands (27). Beijerinck also used a filter candle and showed that the filtered sap was still infectious and was confident that bacteria could not pass through the filter (27). He also ruled out bacterial infection by placing the sap on agar gel surfaces and observed there was measurable diffusion through the agar, whereas bacteria would have remained on the surface of the agar (27). Beijerinck was the first to apply the term "virus" (27). With this Beijerinck described a new class of disease agents that would later be discovered to be the most diverse biological entities on the planet. This began years of testing infectious agents for their filterability determining if the causative agents for a

number of unexplained diseases such as smallpox, cowpox, yellow fever, and rabies were indeed caused by viruses (27, 82, 147).

The discovery of the tobacco mosaic virus was soon followed by the discovery of viruses infecting humans and animals. It wasn't long until viruses that infected bacterial cells were discovered. Viruses that infect prokaryotes, called bacteriophage or phage, are the most numerous type of viruses present in many environments sampled and are very diverse. The discovery of bacteriophage did not occur without controversy. Perhaps the first description of phages was made by Hankin, a scientist doing research in the Jumna and Ganges Rivers in India. Hankin noted that these waters could kill diverse amounts of bacteria (67). However, Hankin is not credited with being the first to discover bacteriophages. This credit is shared between two scientists and caused years of controversy in what became known as the "Twort-d'Herelle Controversy" (136).

Frederick W. Twort was the first to publish on the subject in 1915 (143). He reported a phenomenon he called a "glassy transformation" that he observed in the growth of his bacterial culture (143). This discovery was made as a side observation on his major experiment dealing with the vaccinia virus (136, 143). He noticed the contaminating micrococci colonies appeared mucoid and watery. He described the colony morphology as glassy and then discovered that the glassy morphology could be induced in other regular morphological colonies by the introduction of some of the watery material (143). He also observed the glassy cultures under the microscope and observed the degenerations of the cells into small granules (143). He described the phenomenon as "an acute infectious disease of micrococci" (143). Two years later, in 1917, Félix d'Herelle independently described bacteriophages while at the Pasteur

Institute in Paris working with bacillary dysentery (43). He described "a microbe that was antagonistic to bacteria, lysed bacteria in liquid cultures, and killed bacteria in discrete patches, which he termed plaques, on the surface of agar spread with a film of bacteria" (43). d'Herelle was the first to coin the term bacteriophages as he saw them as "ultraviruses" that ate bacteria (136). His further research also verified several characteristics of phage infection: phage need living cells to multiply, phage lyse the host cell, phage infection and multiplication is cyclic, and phage can multiply indefinitely (136). d'Herelle went on to do important work in the fields of phage therapy and phage biology (136). Although both scientists independently discovered bacteriophage, it wasn't until another bacteriologist, Jules Bordet, dug up Twort's work that the controversy began (136). Bordet, a Nobel Prize winner, was also involved in phage research at the Pasteur Institue in Brussels. Due to disagreements about Boret's and d'Herelle's views of phages along with some personal issues, when Boret discovered Twort's paper on the "glassy transformation" he challenged Twort's claim to the discovery of bacteriophage (136). Although Twort didn't get fully involved in the controversy, Bordet and d'Herelle allowed the controversy to continue for about 10 years (136). It was finally put to rest by Paul-Christian Flue and E. Renaux, two independent scientists who reviewed the data and concluded that the discovery made by Twort in 1915 and the discovery made by d'Herelle in 1917 were identical (57, 136).

Twort did not pursue further research on bacteriophage, but d'Herelle did, and made discoveries that led to a field of therapeutic use known as phage therapy (136). Phage therapy is the use of phage to treat bacterial infections. d'Herelle noted that while the phage were multiplying the host cells were being lysed (136). He realized the

medical possibilities when it was noted that in dysentery patients phage titers increased in the stool samples (136). Based on his growing knowledge of phage biology, and his observations of phage titers in dysentery patients, d'Herelle and others began to realize the role phage could play in the treatment of infectious disease. Even though d'Herelle realized this, he was not the first to try it. The first scientists to use phage therapeutically were Bruynoghe and Masin from Louvain (24, 136). They injected phage prepared from a *Staphylococcus* culture into a cutaneous boil. They did see some improvement including a reduction in swelling and pain in the boil area and a reduction in fever (24).

d'Herelle's first work with phage therapy came in 1919 when we worked with a disease in chickens caused by *Bacillus gallinarum* (136). d'Herelle reported that the therapy was a success in that it offered a high degree of protection (41, 136). During his field trials he noticed that the phage treated flocks had shorter epidemics, lower death rates, and were not susceptible to recurrent rounds of infection (41). d'Herelle also tested phage therapy on water buffaloes (39). He discovered that the phage-treated buffaloes were protected against experimental inoculation of the bacterium *Pasteurella multocida* which causes barbone (39). Following his work with birds and buffaloes, d'Herelle wanted to see if phage therapy would be of therapeutic use in humans. d'Herelle, like other scientists in his time, decided the best way to do human trials was by self-administration, tests on his family members, and also some tests on his coworkers (39). He began by administering a Shiga-bacteriophage to himself and the selected parties both orally and subcutaneously (39). No adverse effects were observed and so he began applying the Shiga-bacteriophage to his patients with culture confirmed bacillary dysentery (39, 136).

d'Herelle also did phage therapy work on plague victims (42, 135). While he was in Alexandria, Egypt, d'Herelle came across 4 plague victims passing through on a ship. They were all culture-confirmed to have the bubonic plague and it wasn't long before d'Herelle was injecting their buboes with phage that infected phage that infected *Yersinia pestis*, the causative agent of the plague (40). Remarkably, all four patients had a speedy recovery (40). These results lead to an invitation to the Haffkine Institute in Bombay to work on phage therapy (136). After d'Herelle visited India, "The Bacteriophage Inquiry" began and was successful for the treatment of cholera. Consistent reports from India reported that the use of phage therapy was helping fight the infectious disease and suggested that it might be used as an established method of treatment for cholera (42, 135, 136).

The fast rise of interest in phage therapy was soon dissipated. A review on the use of phage therapy was conducted by The Council on Pharmacy and Chemistry in the 1930's (136). The Council concluded that due to the lack of a standardized method by which the phage were purified and processed it was difficult if not impossible to compare the published studies. They noted that there were inconsistencies on the reports of phage therapy published and were very concerned with the lack of knowledge on the general biology of phage (48, 136). With World War II and the discovery of antibiotics, research of phage therapy trials halted in the United States and most of Western Europe. However, after the war phage therapy research did continue in the Soviet Union and some other European countries. Since that time, phage therapy has been used in several studies from Pakistan, England, Romania, France, North America, and others. These studies range from human trials against cholera in Dacca to cattle trials against *E. coli*.

Newer studies using phage therapy to treat infections in aquatic environments appear to be promising (136).

During the excitement of the phage therapy era, phage biology was overlooked. It wasn't until the 1930's and 40's that scientists set out to learn more about the nature and biology of phage. In combination with the research on other viruses, such as tobacco mosaic virus and poliovirus, scientists began to observe the nature of viruses. Wendell Stanley was the first to obtain a virus in crystalline form and from there methods and criteria for viral purification and physical studies began (82). Martin Schlesinger, a German scientist, was the first to identify the chemical makeup of viruses as proteins and nucleic acids (128). When the electron microscope made its way onto the scene in 1940, many questions about the nature of viruses were answered. Scientists got their first glimpse at the character and different morphologies that viruses possess by the use of this newly invented microscope (128, 136).

Understanding viral multiplication and replication, however, was a different task and was studied by several scientists. An American, Emory Ellis, from Caltech began studying phage and was determined to understand the basic biology of phage and to see phages role as a model organism. Ellis was not a microbiologist, but instead was a physical chemist who appreciated the similarities in known viruses. Around this time, Max Delbrück from Germany, a trained physicist who studied gene structure, was recruited to come to Caltech and on an unrelated project. Delbrück, who had an interest in viruses, soon teamed up with Ellis (136). Basing their concept on d'Herelle's work, Ellis and Delbrück invented the "one-step growth experiment" to understand the

fundamental nature of viral replication (50). Delbrück recruited other scientists and soon led a team studying the T phages (136).

While Delbrück's group was focused on the lytic coliphages, another research group set out to explore the lysogenic phages described by Bordet. Beginning in the 1930's the Wollmans, a married couple who worked in Paris, determined that lysogenic phages appeared to be part of the "cellular hereditary apparatus" (157). Twenty years later, two Parisian scientists, Andŕe Lwoff and Antoinette Gutmann, coined the term "prophage" and through their experiments monitoring phage induction and release led to a better understanding of the nature of lysogeny (91). Lysogenic research found a model organism when Lederberg discovered the phage lambda (86, 136, 147). Lambda has greatly increased scientific knowledge of lysogeny as well as gene expression (136).

As technology improved and new techniques were invented, the field of virology advanced considerably. Viruses are used in many areas of scientific research from disease control to bioengineering. Phage research was once limited to culturable host cells, but today with the use of a culture independent "metagenomic" approach science is getting a glimpse of phage beyond the lab bench. Throughout the research presented in this thesis both classical techniques invented by the founding scientists in the field of microbiology as well as the new culture-independent techniques recently applied to study diverse viruses in natural environments are discussed. This research is aimed at assessing the diversity of viruses, especially bacteriophages, in both soil and wastewater microbial communities, as well as to compare the classical culture-dependent methods with the newer culture-independent methods for the first time in an evaluation of the diversity of viruses in a microbial community.

Viral biology and structure

Viruses are obligate intracellular parasites. They have no means of metabolism and are dormant when not infecting a suitable host. All viruses consist of two simple molecules: protein and nucleic acids (some eukaryotic viruses that infect animals can contain an envelope made out of phospholipids and proteins) (27). This section will explore the biology of the virus and will detail the structure, replication cycles, and genetic structures of viruses.

In 1959, Watson and Crick published a theoretical discussion on the structure of viruses that suggested the use of identical subunits to construct a protein coat arranged with cubic symmetry (38). This discussion soon shifted from theoretical to actual when the negative staining technique was published (72). For the first time scientists were able to visualize the shape of the virus by using the transmission electron microscope. It was noted early on that there were characteristic shapes and symmetry to viruses (73). It was Lwoff *et al.* as well as Casper *et al.* that went on to create the generally accepted terminology of capsid, capsomers, envelope, nucleocapsid, and virion (30, 92). The capsid describes the protein coat that surrounds the genome of the virus (92). The capsid is constructed by smaller identical proteins that self-assemble (92). These monomer proteins are referred to as capsomers (92). Together the capsid and the genome compose the virion (92). The virion is generally referred to as the infectious viral particle (30, 92). It was obvious that the capsids did contain the predicted symmetry. In fact, a majority of the capsids observed fell into two categories based on the type of symmetry: cubic or helical (71). Of course there were some viruses, primarily bacteriophage that

contained a more complex capsid. The cubic symmetrical capsids are icosahedral in shape containing a 5:3:2 symmetry. The symmetry of the icosahedral virus was further described by Casper and Klung (31). Helical capsids are tube-like in nature and resemble a "spiral staircase" wrapping around the viral genome. Some nucleocapsids may be surrounded by an envelope. The envelope is composed of a phospholipid bilayer, carbohydrates, and proteins. It is derived from a host cell wall (either external or internal) and may contain some virally encoded proteins. The envelope, found on animal viruses, serves many functions such as attachment, disguise, and protection. Many enveloped viruses, such as the HIV, depend on the envelope for viability. The more complex capsid morphologies seen in bacteriophage and pox viruses may include tails and baseplates. The structure of a bacteriophage is perhaps one of the most complex viral structures known. A typical bacteriophage is composed of an icosahedral shaped head which contains the genome, a collar, a sheath, a baseplate, spikes, and tail fibers. The tail fibers are responsible for the specific attachment of the phage to the host cell receptors. Some phages contain retractable sheaths which are used during infection.

Viruses have the highest degree of genomic structural diversity of any biological entity. Viral nucleic acid can either be DNA or RNA, but viruses do not contain both. The structure of the genome can either be circular, linear, or segmented. The viral genome can be single stranded, double stranded, or a combination of single and double stranded nucleic acids. The sense of the nucleic acid can either be positive sense, negative sense, or ambisense. The genome of the virus and the viral mechanism of mRNA production is used in the Baltimore classification system as a means of sorting and grouping viruses for classification and taxonomic purposes.

Viral replication cycles

Viruses are obligate intracellular parasites, and can only replicate with the aid of a host cell. As early virologists observed, viruses can exist in either a virulent (lytic) cycle or a temperate (lysogenic) cycle (3, 64). Since the majority of viruses discovered in the research presented in this thesis are bacteriophage, this discussion will focus on bacteriophage replication cycles.

Phages can exist in the lytic cycle if they produce progeny phage without integrating into the host genome, or in the lysogenic cycle which in most cases involved the integration of the phage genome as a prophage into the host cell chromosome or in some cases plasmids (3, 20, 25, 89, 118). If a phage exists in the prophage stage, it will be replicated along with the host cell genome or plasmid. Although little is known about the factors leading to lysogeny in nature, prophage will typically remain stably integrated or maintained in the host cell for many generations (20). Environmental stressors can induce the prophage to resume the lytic functions suppressed by the lysogenic phage (25, 97). Two well studied inducing factors of prophage include ultra violet light (UV-C, <300nm) and mitomycin C (2, 37, 77, 120). Freifelder reports that 90% of known phage are prophage, and Casjens states that 60% of sequenced bacterial genomes contain at least one prophage (29, 58, 77). Edwards and Rohwer report that 3% of genomic DNA content is composed of prophages and the 75% of the genes in prophage genomes do not have known functions (49).

Lytic phages are those phages that release their phage progeny by host cell lysis. Also called virulent phages, these phages are not lysogenic, but rather depend on lysing

their host cell in order to search for new hosts in which to replicate (20). The lytic phage replication cycle involves rapid production of intracellular phage particles (154). Due to the high amounts of host cell lysis due to virulent phage in natural environments, lytic phage play a role in the biogeochemical cycles which will be discussed in the Viruses in the Environment section. Virulent phages are also the most important phages used for phage therapy as they can decrease the numbers of their host cell population.

Bacteriophage have also been shown to exist in two alternative replication cycles: pseudolysogeny and chronic infection (98, 100). Romig *et al.* identified pseudolysogeny as "stalled lytic infections that resume when the dormant host begins vegetative growth" (126, 154). When host cells are abundant, some phage exist in the shedding replication cycle in which there is a constant production of phage particles (2).

Viral infection and replication

A virus can exist in three forms: mature virus existing outside the host cell, vegetative virus existing inside the host cell following adsorption, or as a provirus existing in a state of lysogeny with the host (3). There are several steps that occur during viral infection. The first step in viral infection is attachment and adsorption of the mature viral particle to the host cell. During this step the viral particle will bind specifically to a complementary host cell receptor. The second step is the penetration step involving the viral genome entering into the host cell. This can occur by a number of mechanisms including uncoating, endocytosis, and fusion. Some bacteriophage will release the enzyme lysozyme from the tail to break down a portion of the cell wall (23). Following this, the sheath may be used as a "drill" for penetration of host cell membrane.

Bacteriophage penetration may be aided by the presence of divalent ions such as magnesium or calcium (149). The third step occurs inside the host cell and involves biosynthesis and multiplication of viral components at the expense of the host cell processes. This process is highly variable depending on the construction of the phage genome. The fourth step involves viral maturation and virion assembly. Finally, the phage progeny will be released from the host cell by cell lysis or budding, mediated by the production of lysin and/or holin proteins that rupture the cell from within by inserting into and degrading the host cell membrane (3).

Bacteriophage classification

Presently there are three proposed methods for bacteriophage classification. The original and presently used method of classification was proposed by the International Committee on Taxonomy of Viruses (ICTV) over three decades ago. Many bacteriophage researchers agree that due to the increase in genomic knowledge, the ICTV methods need to be revised (104). In 1966 the International Congress of Microbiology met and created the ICTV with the purpose of developing a taxonomic system for viruses infecting plants, animals, fungi, and bacteria (later conventions included Archaea). Beginning in 1971, the committee has met regularly to update the taxonomic guidelines. The ICTV taxonomic system is based on a Linnaeus type hierarchal ranking based on shared characteristics and attributes (74, 146). Using this system, phage are categorized into families based on tail length and structure. The families are further broken down into genus and subgenus by criteria such as genome configuration, genome size, and host range (104). There are flaws and weaknesses with this system such as: the overlook of

genomic and proteomic information, the hierarchical system is based on vertical transmission of genetic characteristics whereas phages undergo large amounts of horizontal gene exchange (111), and as a system based on physical characteristics, classification is dependent on electron microscope images. Whether it is the problematic grouping of the *Salmonella* phage P22 with the *Podoviridae* instead of its genetically related cluster *Siphoviridae*, the discovery of six *Streptococcus pyogenes* prophage which cannot be recognized by the ITCV system, or the failure to classify half the phage for which entire genomic sequence is known, examples of the shortcoming of this classification system are seen regularly (15, 55, 104).

It is generally agreed that genomic information should be used to assign taxonomic identification to phage (104). In bacteria this is simpler due to the conservation of the 16S ribosomal RNA gene. However, in bacteriophage, the dilemma arises in the lack of a conserved gene or protein sequence common to all phage as well as biases due to overrepresentation of certain phage genes in the public databases (122). Some conserved genes have been identified for certain groups of phage such as DNA polymerases in the Phycodnaviridae and capsid protein in cyanophages, but no one conserved gene is found in all phage (34, 35, 61, 130). Even so, there are three proposed methods of phage classification based on the genomic and proteomic information.

Rohwer and Edwards proposed the "phage proteome tree" that groups phage relative to their nearest neighbors but also in the context of all other phage (104, 122). With this method a tree is constructed based on relationships between phage and proteins (104, 122). The entire predicted proteome for a phage was analyzed by the BLASTp or PROTDIST program and the results were used to construct a distance matrix (104, 122).

Data was then parsed based on an e-value of <0.1, <0.01, or <0.001 for the BLASTp data or penalty score of 5, 10, or 100 for the PROTDIST data (104, 122). The relationship between proteins allowed for the tree construction (104, 122). By using this relationship, Rohwer and Edwards were able to group the *Salmonella* phage P22 with the lamboid-like phages resolving one of the known shortcomings of the ITCV classification system (122).

The Pittsburg Bacteriophage Institute suggested that due to the mosaic genomes of phage due to horizontal gene transfer a strictly hierarchal system of classification could not be used (84, 104, 111). According to their proposed system of classification "domains" would be determined according to genome type (double stranded DNA, single stranded DNA, double stranded RNA, single stranded RNA), "divisions" would be determined based on defining characteristics such as whether the phage is tailed or filamentous (84, 104). Following these steps, these three tenets would further direct the classification process: (1) members of a group should exhibit similarity in one or more loosely defined cohesion mechanism, (2) significant sequence data should be available for evolutionary assignment to a taxonomic cluster, (3) grouping may be reticulate, in that phage may simultaneously belong to several groups based on multiple and/or differing criteria from the first two tenets (84, 104). Criticism of this method is based on the high complexity of the reticulate classification (104).

Praux *et al.* described a third method of classification based on comparative genomics of a structural gene module (104, 115). This method, based on dot plots of temperate lactococcal phage, was able to delineate and distinguish four phage species and two genera based on head morphology. Comparative genomics of nonstructural genes lumped all of the phage studied into one species, but based on the comparative genomics

of structural genes (which are the most conserved module in dairy phage) allowed for this distinction (104, 115). Based on the study, Praux *et al.* proposed five distinct genera for lactococcal phages (104, 115). The comparative genomics of structural gene approach was further supported by Chibani-Chennoufi *et al.* with their work on the SPO1-like genus (36).

As Nelson notes, all of the above mentioned genomic and proteomic approaches would ultimately cluster the majority of the bacteriophage in the same groups in which they are currently classified and would unlikely have an immediate significant impact (104). Nelson suggests that with the increase in environmental community genome sequencing of phage in multiple environments, and the fact that most of what these studies are finding are unknown, it would be best for scientists to wait a bit longer before reconstructing a phage phylogenetic tree (104).

Viruses and the environment

Viruses are the most abundant and diverse biological entity on the planet. Of the viral populations in natural environments it has been found that a majority are bacteriophage (13, 18, 19, 21). This is likely a consequence of prokaryotes being the most abundant potential hosts. It is estimated that there are approximately 10^{30} prokaryotes present on earth (151). Others have shown that there are approximately five to ten viral like particles (VLP) for every prokaryotic cell (121, 159). Together, these estimates suggest that there may be as many as 10^{31} free phage on the Earth. Viral counts by electron and epifluorescence microscopy show that there are approximately 10^{6}-10^{7} VLP per ml in the world's oceans and lakes and 10^{8}-10^{9} VLP per g of sediment and top

soil (9, 14, 44, 70, 93, 105, 147, 159). With viruses infecting all domains of life, the diversity of viral species exceeds that of living organisms. The diversity of phage species alone is estimated at around 100 million species (121). Viruses play a major role in the environment. They affect nutrient cycles, affect population dynamics, and are responsible for horizontal gene exchange.

Phages have been shown to be an important part of global biogeochemical cycles (17, 141, 142). The transfer of energy and nutrients to higher trophic levels in the ocean is greatly influenced by the microbial food web (12, 112). This microbial food web consists of a dissolved organic matter (DOM) cycle in which heterotrophic prokaryotes incorporate DOM and are lysed by viral infections (60, 137, 152). DOM may be consumed by a diverse population of marine heterotrophic bacteria and these bacteria are either lysed by phage or consumed by protists (60, 81). The lysis of bacteria results in an increase in the DOM which is then consumed again by the heterotrophic bacteria (137, 152). It is estimated that phage lyse around 50% of the bacterial population present in marine environments (60). Wilhelm *et al.* estimates that viral lysis is responsible for the recycling of 6-26% of the photosynthetically fixed organic carbon into (DOM) (152). This loop can be sped up or slowed down by the rate of viral lysis and this process may play a role in altering atmospheric CO_2 levels (59, 137).

Phages also affect the population dynamics of a given environment (98). In some environments phage may be the only predators to their bacterial hosts (22, 129). It has been estimated that phage are responsible for 20-50% of bacterial lysis, thereby greatly influencing the microbial food web and host cell populations (112, 137, 138). Examples

of the use of phage to control host cell populations can be applied in the form of phage therapy (reviewed previously).

Perhaps one of the more influential roles of viruses in the environment is their role in horizontal gene transfer. This process may be a driving force in both microbial and viral evolution (108, 150). Phages mediate genetic exchange through generalized and specialized transduction (1, 53, 76, 108, 109, 148). Jiang *et al.* estimates that there are 10^{14} transduction events per year in Tampa Bay Estuary, Florida (76). Paul *et al.* . estimated that every year in the world's oceans 10^{28} base pairs of DNA are transduced by bacteriophage (110). Phage have been found to contain virulence factors and other genes that are important in disease, in what is termed lysogenic conversion (16, 121). These genetic exchanges can alter the host cell phenotype (101, 116, 144).

Metagenomics

Even in the earliest studies of environmental microbiology scientists noted that there were more microorganisms present in the sample then they were able to cultivate in the lab. This phenomenon, known as the "great plate count anomaly", has been a driving impetus to develop culture-independent techniques, for example many studies now assess bacterial diversity and community structure based on 16S rRNA gene sequences directly PCR amplified from community genomic DNA. However, while rRNA sequence provide a means to observe the phylogenetic diversity of microbial communities without the cultivation bias, the functional diversity within a microbial community cannot be assessed via a single phylogenetic marker. To investigate the functional genetic diversity present in natural environments, investigators developed culture-independent methods for

cloning and sequencing genomic DNA isolated directly from natural environments (131). Metagenomics, a term coined by Jo Handelsman, is the use of molecular genomic techniques to study an entire community of microorganisms in a given habitat (66). The use of metagenomic studies, both sequence-based and function-based, has revolutionized the way scientist study microorganisms. The first metagenomic library was published in 1995, and since then numerous libraries have been made from varying environments and for varying target microbes including viruses (69). One of the largest metagenomic projects was the Sargasso Sea project which sequenced 1.045 billion base pairs of non-redundant sequences (145). Within this database of marine metagenomic sequences 148 previously unknown bacterial phylotypes and 1.2 million previously unknown genes were identified (145). Not only has metagenomics provided insight into a vast amount of unknown species and genetic heterogeneity present within natural environments, scientists have also been able identified novel enzymes and antibiotics that have had biotechnological applications (65, 132).

The use of metagenomics to study prokaryotic, eubacterial and archeal communities in diverse environments is been used with increasing frequency (127, 133). By studying 16S rRNA gene libraries, scientists have realized that prokaryotic diversity is far greater than previously expected, and the number of divisions within the domain Eubacteria has increased from only 12 in the year 1988 to over 40 recognized divisions currently, with many Eubacteria divisions lacking any cultured representatives (132). Metagenomic studies have recently been used to examine the genetic diversity of viral communities. Constructing and studying metagenomic libraries for viral communities has many challenges that must be overcome. There is no conserved gene, such as the 16S

rRNA gene in bacteria, found in all viruses. Without an evolutionarily conserved gene within all viral genomes, studying a viral population or viral metagenomic library for community structure is more difficult. Also, viruses contain many genes that would be toxic to an *E. coli* clone, which is a typical host used for analysis in the laboratory. These genes encoding for proteins such as lysozyme and holins must be interrupted before the cloning process (21). Rohwer's group was able to overcome the construction challenges by creating a linker amplified shotgun library based on technology developed by the Lucigen Corporation (Middleton, WI) (21, 124, 125). This library construction technique has been used several times to study the viral metagenome of human feces, surface seawater, marine sediments, viroplankton of the Chesapeake Bay, and soil (13, 18, 19, 21). Studies of these libraries indicate that 30%-75% of environmental viral sequences are completely novel (13, 18, 19, 21). Edwards and Rohwer reports that a comparison of 964,094 ORFs from the Sargasso Sea metagenome revealed that <1% had significant similarity (E-value $\leq 1 \times 10^{-5}$) to known phage genes (49, 145). With viruses having such ecological importance, this amount of unknown data is significant, but also not unexpected given the paucity of our knowledge of viral genetic diversity in natural environments.

Without an identifying conserved gene, taxonomic studies on these libraries use a variety of approaches for comparative genomic analysis. After being compared to GenBank non-redundant databases, a majority of the significant (e-value ≤ 0.001) hits in every viral metagenomic library analyzed to date are to bacterial proteins. These findings may appear to represent contamination of the libraries, but purification steps and additional studies have shown that this is not the case (19, 21). Hits to proteins of non-

viral origin are acceptable and expected due to the following reasons: (a) the viral purification process used to construct the libraries separates viruses from contaminating microorganisms and free DNA, (b) open reading frames in purified cultured phage genomes are often more similar to bacterial open reading frames than to those of other phages, (c) forward and reverse sequencing from a single clone can result in one viral and one bacterial read, (d) hits may represent uncharacterized prophages or sequences from transducing phages, (e) BLAST searches with the predicted ORFs from the genomes of cultured phages resulted in 50% and 30% of the significant hits being bacterial in origin, respectively (18, 19, 21, 111, 117), and (f) perhaps most significantly, the existing GenBank databases are largely populated by prokaryotic sequences and lack any significant inclusion of environmental viruses (at least currently). Edwards and Rohwer reports that "approximately 65% of phage genes have no homologues at all, even within other phage genomes or with sequenced phage genes" and they go on to estimate that "about 1% of the microbial metagenomes encode phage proteins" (49). A summary of the taxonomic breakdowns from each viral metagenomic library can be seen in Table 1. Of the viral hits, an overwhelming majority (~98%) hit to bacteriophage (13, 19, 21). This is consistent with the results of Wommack *et al.* (160).

Getting a glimpse at the taxonomic representation contained in the viral metagenomic library is useful, but understanding the richness and evenness of the community is important as well. Assessing this information from living microbial communities is done through tools such as a rarefaction curve which can plot the number of individuals counted in the sample verses the total number of species sampled (63). This method cannot be applied to the viral libraries due to there being no identifying gene

used for species determination. Previous studies have used contig assembly to assess the community diversity represented by the viral libraries. All of the sequence reads are run through an alignment process by which if they are 98% identical over a minimum of 20 base pairs they will form a contig. These contig parameters were determined by Brietbart *et al.* on the basis of stringency settings that were able to differentiate between the closely related T3 and T7 coliphages, indicating that when a contig forms it represents an alignment between sequences from the same phage or a very closely related phage (21). Sequences from species with higher relative abundance will form contigs, whereas species with lower relative abundance will be represented by single sequence reads. The statistical analysis program used to represent the contig data was created by Angly *et al.* (6). They developed a free online bioinformatics tool, Phage Communities from Contig Spectrum (PHACCS), which uses a modified Lander-Waterman algorithm to predict an underlying contig spectrum (6). The user inputs the experimentally determined contig spectrum, average genome size, average shotgun DNA sequence length, and the minimum overlap length used for the assembly (6). The PHACCS tool will output six possible community structure models (6). The user can then select the best descriptive model for a community structure based on the error value predicted by the PHACCS algorithm (6). The values of the error can be roughly interpreted as logarithms of the odds ratio of the observed contigs produced from the viral community distributions (6). Community structures using this tool for the viral metagenomic libraries presented can be seen in Table 2.

For several years the application of metagenomics to environmental studies was limited due to its high cost. In 2006 Margulies *et al.* published a sequencing-by-synthesis

technology that is less expensive to sequence per base pair and eliminated the need for the construction of a clone library (94). The downside is that is produces smaller sequence reads (100-200 bp) than the normal Sanger dideoxy sequencing method (94, 158). Although useful for whole genome sequencing approaches, its use for environmental metagenomic studies was questioned by Wommack *et al.* (62, 158). In their study, Wommack *et al.* compared BLAST and COG analysis for longer (~750 bp) and randomly created short reads for three metagenomic libraries. They showed that BLASTx searches against the GenBank nr database found far fewer homologs within the short-sequence libraries, especially in the viroplankton metagenomic library (158). They also were able to show that more distant homologs of microbial and viral genes are not detected by short-read sequences, thus coming to the conclusion that for environmental metagenomics, especially for viral communities, read length matters (158).

Viral ecology of soil

Although the microbiology of soil has been studied extensively, the ecology and diversity of viruses in the soil is poorly understood. Previous studies tend to focus on phage/host systems and population dynamics of cultured microorganisms (10, 11, 107). One study of phage in the rhizosphere of sugar beets showed that "temporal dynamics of phage and host populations were almost identical over three consecutive years" (10, 155). There is also extensive research to describe the interactions between viral particles and soil surfaces (153). Unfortunately, many of these studies were conducted with enteric bacteriophage or other phage which are exogenous to soils (153). From these studies it is known that the composition of soil influences viral adsorption. Factors such as pH (90),

dissolved organic matter (113), and soil content (such as water, clay, and organic matter content) (78, 99), as well as viral characteristics (32, 47) can affect viral interactions in the soil (153). Even with all of this work, it wasn't until recently that researchers examined the abundance of autochthonous soil viruses (9, 56, 153, 155).

The first report of direct counts of soil bacteriophage came in 2002 by Ashelford *et al.* In their study, viral direct counts were determined by TEM. Until this point, estimates of viral numbers in the soil had only been assessed by culture-dependent methods (8, 26, 28, 83, 95, 119) and, by these estimates, an average of 4×10^4 viruses g^{-1} of soil were reported. Due to potential biases due to nonspecific interactions of soil particles with DNA stains, and the desire to view phage morphology, Ashelford *et al.* only used TEM for their counts. They report a mean of 1.5×10^7 g^{-1} bacteriophages in an agricultural soil. Based on a second experiment where they spiked a known amount of phage in the soil, they estimate that this mean is underestimating the total viral counts by "40-fold for tailed phage and 8-fold for VLPs" (9). They claim that damage done to the viruses during the sample processing methods, along with some loss during storing and fixing was to account for this underestimation. Based on their data, they concluded that, "predation of bacteria by viruses will be an important factor in controlling and stimulating the growth of bacterial populations in soil" (9). They also concluded that bacteriophage are important mediators of gene transfer in soil (9).

A few years later, Williamson *et al.* looked at the abundance and diversity of viruses in six Delaware soils (153). They studied three types of soil: agricultural, coastal plain forest, and piedmont forest soils. They based their viral counts on epifluorescence microscopy and their diversity estimates on morphological differences seen by TEM.

Several observations were made: First, they found the highest VLP abundances in the wetland soil sample (~ 3 x 10^9 VLP g^{-1}), next highest in the forest upland soils (<3 x 10^9 VLP g^{-1}), and the lowest being in the agricultural soil (<1 x 10^9 VLP g^{-1}) (153). This correlated with data showing that the wetland soil sample also had the highest number of bacterial cells. The high number of VLPs observed in the wetland soil was explained by the high water content and the addition of viruses transported by runoff events from higher lands. The low VLP counts in the agricultural soil was explained by land practices affecting the bacterial counts (which were also very low) and viral counts in similar ways (153). Secondly, they found that the viral population of all six soils was dominated by tailed viruses assumed to be bacteriophage (153). They also noticed diversity differences among the viral communities of each soil based on observed morphologies. Varying capsid shapes were observed including rarely reported elongated capsids and filamentous viruses. A mean capsid diameter of around 50nm was observed to be consistent for all six soils (153). They concluded that based on morphology, the four forested soils had the highest diversity (153).

A study of bacteria, archaea, fungi and viruses in prairie, desert, and rainforest soil was conducted by Fierer *et al.* (56). For the viral diversity analyses, they constructed a viral metagenomic library. Using contig spectrum analysis (described above) they estimated the operational taxonomic unit (OTU) richness of the viral community to be greater than 10^{10} OTUs and the percent abundance of the most common OTU to be between 5% (rainforest) and <1% (desert) (56). Their data suggests that viral communities are both locally and globally diverse. Using a tBLASTx comparison of the metagenomic viral sequences to the GenBank nr/nt database, they found that the majority

of the sequences showed no significant (E-value <0.001) similarities. They observed several significant hits to phage with the most hits to *Actinoplanes* φAsp2, *Mycobacterium* φBxz1, *streptomyces venezuelae* φVWB, *Haloarcula hispanica* φSJ1, *Mycobacterium* φRosebush, and *Myxococcus xanthus* φMx8 (56). A comparison of the soil viral metagenomic libraries with other previously described metagenomic libraries suggested that "distinct habitats harbor distinct viral communities" (56).

Viral ecology of wastewater

There have been many studies which have looked at the role viruses play in activated sludge. Most of the studies have examined coliphages or other specific host/phage relationships and the role of these relationships in the wastewater system (7, 46, 85, 87, 140). Only a handful of studies have looked at total viral communities in activated sludge (51, 68, 79, 80, 88, 106). Out of these studies, all of them were culture-dependent studies. The exception to this was Otawa *et al.* who used direct counts by epifluorescent microscopy and pulse field gel electrophoresis (PFGE) to study viral communities (106). From these studies several observations have been made. Ewert and Paynter observed an increase in viral numbers in activated sludge compared with influent based on their direct counts by electron microscopy (51). This observation was supported by other studies, and the conclusion was made that the activated sludge process is involved in the promotion of viral production (68, 106). Total viral counts in activated sludge reported have varied based on isolation techniques and counting methods, but highest counts were observed by Otawa *et al.* who observed $10^8 – 10^9$ viral like particles (VLPs) per ml of activated sludge (106). Culture-dependent studies that have looked for

phage infecting bacterial cultures isolated from activated sludge have found that phage could be isolated from around 30%-60% of the cultured bacterial isolates (79, 80, 88). Hantula *et al.* found phage that infected approximately 18% of their cultured bacterial isolates from activated sludge and that the phages exhibited a broad host range (68). They also noted that approximately 16% of those phage were thermosensitive (68). It has been noted by several studies that coliphages are seen in high titers in influent, but seem to decline in activated sludge (64, 161). Another unexplained phenomenon observed in a culture collection from activated sludge was the disappearance and reappearance of a cultured phage (68).

The only culture-independent study of viruses in activated sludge was conducted by Otawa *et al.* They looked at total viral communities using epifluorescence microscopy and PFGE from 14 full-scale waste water treatment plants (WWTP) (106). From their studies they noted the range of size of viral genomes in activated sludge to be between 40 and >200kb with 40-70 kb sized genomes the most frequent (106). They also observed similar PFGE band patterns which indicated that activated sludge processes could contain common viral communities (106). Varying temporal behaviors in the laboratory-scale reactor were observed with two distinct bands emerging and disappearing within a very short time period (106).

The study of viral communities in activated sludge has many practical applications. Withey *et al.* suggests many ways in which bacteriophages could be used to improve effluent and sludge emissions. They suggest that the application of phage therapy to activated sludge could be used for pathogen control, improving dewaterablility and digestibility, and the control of various activated sludge bacterial communities such

as those that produce foam and the non-phosphate accumulating bacteria (156). Although it is known that phages exist in high numbers in activated sludge, the exact role the phages play in this environment is poorly understood (106).

Through the development of newer non-culture based techniques science has been given a glimpse into the previously unseen world of total microbial communities. The research presented in this thesis uses a combination of the classical culture based techniques as well as the newer metagenomic techniques to give a better understanding of the total viral communities in soil, activated sludge, and influent. It also presents a comparative study of the two techniques to determine the precision of each technique relative to each other.

Table1. A comparison of taxonomic breakdown of significant hits from four metagenomic libraries (13, 18, 19, 21). All percentages are approximates.

	Human Feces	Scripps Pier	Mission Bay	Mission Bay Sediments	Chesapeake Bay Viroplankton
Viral	27%	38%	31%	47%	44%
Eubacterial	50%	32%	23%	31%	47%
Archaeal	6%	2%	<1%	3%	1%
Eukaryotic	4%	5%	12%	9%	7%
Other[a]	13%	28%	33%	11%	N/A

[a] Breitbart et. al also categorized hits into repeat and mobile genetic element categories, Bench et. al. did not include these categories

Table 2 – Comparison of PHACCS analysis between four metagenomic libraries

	Feces	Scripps Pier	Mission Bay	Mission Bay Sediments	Viroplankton in Cheseapeak Bay[a]
Rank-abundance form	Power law	Power law	Power law	Power law	Power law
Error	9.79	1.84	2.15	0.0104	Not reported
Richness (genotypes)	2,390	3,350	7,180	7,340	4,110
Evenness	0.873	0.932	0.9	1	0.999
Most abundant genotype	4.8%	2.03%	2.63%	0.0136	0.065%
Shannon-Wiener index	6.8 nats	7.57 nats	7.99 nats	8.9 nats	8.31 nats

[a]Based on all sequences with a 50kb average genome size

METAGENOMIC ANALYSIS OF THE VIRAL COMMUNITY IN A WISCONSIN SOIL AND A COMPARATIVE ANALYSIS OF SOIL VIRAL METAGENOMES

Abstract

A linker-amplified shotgun subclone viral library was constructed from wetland soil at the Curtiss Prairie at the University of Wisconsin's Arboretum. From the clones, 1,399 (1.2MB) sequence reads were analyzed. Sequences were compared against the GenBank databases as well as other viral metagenomes from soil and other environments. Based on a tBLASTx comparison of the 1,399 sequences fragments to the GenBank nr/nt and env_nt databases, 50% were known, 17% were unknown, and 32% were novel. Using a 0.001 cutoff E-value, 17% of the significant hits were to viral genomes with bacteriophages representing 93% of these viral hits. Siphoviridae were the most commonly found significant hit. Contig assembly indicated a high frequency of hits with similarity to an *Actinoplanes* Asp2 phage. A comparison of viral metagenomes by G+C content and BLAST comparisons further supports the conclusion that soils harbor a distinct community of viruses that differs from the viral communities in other environments, and that the soil viral community reflects the phylogenetic diversity of their prokaryotic hosts, with a high relative abundance of Actinophages present in soil viral metagenomic libraries.

Introduction

The microbiology of soil has been studied extensively, but the ecology and diversity of viruses in the soil is poorly understood. Historically, soil viral research has involved studies on phage-host systems and population dynamics of the cultured soil microorganisms (10, 11, 107) and interactions of phage and soil particles (mostly using exogenous enteric bacteriophage) (153). Recently there have been studies conducted to look at the abundance of autochthonous soil viruses (9, 56, 153, 155).

Few studies have been conducted to examine the total viral communities in soil, but these have demonstrated that viruses are abundant and maintain an important niche in the soil environment. Ashelford *et al.* a reported grand mean of 1.5×10^{7} VLP g^{-1} bacteriophage in soil, but clamed this was an eight-fold underestimation of the VLPs in the soil environment (9). Williamson *et al.* followed this with a study that reported ~ 3×10^{9} VLPs g^{-1} in a wetland soil sample, $<3 \times 10^{9}$ VLPs g^{-1} in a forest upland soil, and $<1 \times 10^{9}$ VLPs g^{-1} in an agricultural soil (153). This study suggested that water content, topography, and land practices can affect the abundance and types of viruses found (153). However, all of their samples were dominated by tailed bacteriophages (9, 153). Not only did these studies report that soil viruses are abundant, they also suggested ecological roles that these viruses might play in the soil environment. Ashelford *et al.* reported that "predation of bacteria by viruses will be an important factor in controlling and stimulating the growth of bacterial populations in soil" (9), and also concluded that bacteriophage are important mediators of gene transfer in soil (9).

At the onset of this study, no previous culture-independent analysis of soil viral communities had been published. Currently, besides the library described in this paper,

two other soil viral metagenomic libraries have been constructed and analyzed to our knowledge. One of those libraries, constructed by Wommack *et al.*, is reported in this study (personal communication). The other library was constructed by Fierer *et al.* (56). Using contig spectrum analysis (described above) the operational taxonomic unit (OTU) richness of the viral community was estimated to be greater than 10^{10} OTUs and the percent abundance of the most common OUT was estimated to be between 5% (rainforest) and <1% (desert) (56). These data suggested that viral communities are both locally and globally diverse. Using a tBLASTx comparison of the metagenomic viral sequences to GenBank nr/nt, the majority of the sequences showed no significant (E-value <0.001) similarities. The most frequently observed significant hits to phage genes were to *Actinoplanes* φAsp2, *Mycobacterium* φBxz1, *Streptomyces venezuelae* φVWB, *Haloarcula hispanica* φSJ1, *Mycobacterium* φRosebush, and *Myxococcus xanthus* φMx8 (56). A comparison of the soil viral metagenomic libraries with other previously described metagenomic libraries suggested that each habitat harbors its own distinct community of viruses, as the soil viral communities are distinct from viral communities in other natural environments and more similar to one another than to any other database of viral community metagenomes (56).

The work in this chapter supports and extends the conclusions of previously published analyses of soil viral communities. The soil chosen for this study was a wetland soil at the Curtiss Prairie at the University of Wisconsin's Arboretum, which is described as the world's oldest restored prairie in existence. Planted by the Works Progress Administration in the 1930's during the Great Depression, the Curtiss Prairie was the only restored prairie in the UW-Arboretum to have native plants planted

according to a water saturation gradient (Dr. Joy Zedler, UW-Madison, personal communication) and has been the subject of numerous ecological restoration studies. Despite not being the first published study to examine soil viral communities via a culture-independent methodology, this study will add to the wealth of new information regarding viral communities in natural habitats and has laid the foundation for research efforts in later chapters.

Materials and Methods

Sample collection and processing

Approximately 500 g of soil were collected from the Curtiss Prairie, UW-Arboretum on July 14, 2004. The soil was homogenized in a beef extract buffer for 4 hours (155). Soil debris and microbial cells were removed by differential centrifugation and the viral concentrate was filtered through a 0.45 micron PVDF membrane.

DNA isolation

The filtered viral sample was treated with Benzonase (250 units/µl final concentration) at 37˚C overnight to eliminate any contaminating DNA. Following benzonase inactivation (addition of 10 mM EDTA and heating at 70°C for 10 min) proteinase K (1mg ml^{-1} final) and 1% sodium dodecyl sulphate were added and incubated at 37˚C for 2 hours to degrade viral protein coats. Proteins were removed by phenol:cholorfom extraction and DNA was recovered by an isopropanol precipitation. The purified DNA was sent to the Lucigen Corp. for library construction.

Library construction and sequencing

Random shotgun genomic libraries of the viral communities were prepared at the Lucigen Corp. (21, 124). Viral DNA was fragmented to approximately 1 to 3 kb fragments and end-repaired to blunt-end each dsDNA molecule. The dsDNA was then ligated to an adaptor (20 bp dsDNA sequence), and then primers specific to the adaptor sequence were used to PCR amplify the dsDNA fragments, using only 15 rounds of amplification. The viral amplicons were then purified over a column and ligated into the pSmartLC vector, a low-copy cloning vector that has transcriptional terminators flanking the cloning site to prevent transcription of viral gene (21, 124). Since many viral-encoded gene products may be toxic to *E. coli*, this decreases cloning bias associated with underrepresentation of clones containing genes toxic to the bacterial host. The ligation was transformed into an *E. coli* strain DH10B and a glycerol stock was prepared of the transformation mixture. The transformation was shipped to Auburn University, where the number of transformants per microliter was assessed. Transformants were plated onto LB agar containing Kan (30 μg/ml) and were picked using the Genetix QPix2 colony picking robot into 96-well format. After growth, sterile 50% glycerol was added to each well (final concentration, 15%) and plates were frozen at -80°C. Eighteen plates were selected and sent to the Lucigen Corp. for BigDye sequencing.

Sequence analysis

Sequences were screened for vector sequence, linker sequence, and minimum base quality using Sequencher 4.8 program. Sequences were then manually monitored for sequence quality. Following removal of poor sequences there were 1,399 quality sequences (1.2MB). For taxonomy assignments sequences were compared against the

GenBank nr/nt and env databases using a tBLASTx comparison. The top 5 hits were used for analysis and only matches with an E-value of ≤0.001 were considered significant. Significant hits were categorized as either viral, bacterial, archaeal, eukaryotic, or a mobile genetic element. In cases where multiple significant hits were observed for a single query sequence, the sequence was preferentially classified as viral if the hit occurred within the top five hits (21). Mobile elements consisted of plasmids and synthetic sequences. Significant hits to phages were further classified into phage families according to The International Committee on Taxonomy of Viruses (ICTV) classification (74).

Contig assembly

Contig assembly was conducted using the Sequencher 4.8 program using varied stringency settings. The lowest stringency setting used was 80% identity with 20bp overlap, the medium stringency setting used was 85% identity with 20bp overlap, and the highest stringency setting was 98% identity with 20bp overlap. Contig assembly at 80% and 85% identities resulted in large overlapping contigs with the largest contig containing 41 and 18 sequences, respectively. The contig spectra used for diversity estimates was the most stringent assembly (98% ID, 20bp overlap) and the resulting spectra is [1339 25 2 1 0 0 0 0 0 0 0 0 0 0 0]. Diversity estimates were constructed using the free online PHACCS tool (6) found at http://biome.sdsu.edu/phaccs/. The best fitting rank-abundance form was the logarithmic.

Community comparison methods

G+C distribution plots and heat plots were generated using a bioinformatics script constructed at the University of Delaware (unpublished data and personal

communication). The average G+C content of each sequence was calculated and then plotted to produce a distribution of G+C content for each viral metagenomic library. Viral metagenomic library sequences from the Chesapeake Bay, Delaware soil, Wisconsin soil, two Yellowstone National Park hot springs (Octopus Spring, Little Hot Creek), a hydrothermal vent sample, and a deep sea sample were compared against sequences from the GenBank nr/nt and env_nr databases, a viral database, and cross compared against the Chesapeak Bay and Delaware soil viromes using tBLASTx. A heat plot was constructed using low E-values to represent higher similarity. Collective E-values for each library comparison were then based on a color scale to indicate similarity.

Results and Discussion

Metagenomic library breakdown

A soil viral metagenomic library was constructed from a prairie wetland soil at the Univ. Wisconsin Arboretum, Madison, Wisconsin. A total of 1,728 viral genome fragments were subcloned and sequenced. Following sequence clean up 1,399 unique sequence fragments were analyzed. The average read length was 806 base pairs and the G+C content of each sequence was normally distributed between 30 and 69%, with a mean of 56%.

Bioinformatic analysis

Following sequence generation, the quality sequence reads from the soil library were compared to multiple databases in GenBank. Sequence fragments were categorized as either 1) known, 2) unknown, or 3) novel based on comparison to the non-redundant nucleotide and the environmental nucleotide databases by tBLASTx analysis. A

sequence with a significant hit to the nr/nt database was considered known. A sequence with no significant hit to the nr/nt database and a significant hit to the env/nt database was considered as unknown. A sequence with no significant hit to either database was considered novel. Of the 1,399 sequence fragments, 50% were known, 17% were unknown, and 32% were novel (Figure 1A). The known hits were further broken down into taxonomic groups. Taxonomic assignments were made based on significant hits to the nr/nt database using a tBLASTx comparison. The top 5 significant hits for each sequence were considered. If a hit to a viral genome was in the top 5 the sequence was preferentially called as viral. Hits to mobile genetic elements were classified in a separate category and included plasmids and synthetic sequences. Of the 703 known hits, 69% hit to bacterial genomes, 17% hit to viral genomes, 4% hit to MGE, and 0.8% hit to archaeal genomes (Figure 1B). The taxonomic groups for both libraries were further categorized.

The high number of non-viral hits in the taxonomic distribution may appear to be counterintuitive if this library represents solely viral genome sequences. However, after reviewing previous viral metagenomic analyses from natural environments and considering the high frequency of lysogenic viruses, the high number of non-viral hits is to be expected. First, the methodology used to isolate and purify the viruses takes measures to prevent contaminating prokaryotic cells (i.e., differential centrifugation and membrane filtration) and contaminating DNA (i.e., exonuclease digestion) from interfering with the sample. The CsCl gradient step will free the viral sample of not only contaminating cell, but also contaminating free floating DNA that might be present (21, 134). An exonuclease step was also included prior to viral lysis to further eliminate any

contaminating prokaryotic DNA. Secondly, since viruses require such an intimate relation with their host, it is reported that many viruses mimic their host cell characteristics (such as G+C content and tetranucleotide usage frequencies) leading to an expected "background of homology" between virus and host (13, 114). Thirdly, studies on purified phage cultures show that that 30%-50% of phage ORFs will give significant BLAST hits to bacterial homologs (19, 33, 111, 123). Pedulla *et al.* showed that purified phage genomes are more similar to bacterial ORFs than to other phage ORFS (19, 111). Fourth, non-viral hits could represent uncharacterized prophage or transducing phages. Finally, with only a handful of viral metagenomic studies conducted, there is little information on viral communities represented in the GenBank database. This accounts for the low percentage of viral hits to the GenBank database.

Eleven different phyla as well as some unclassified hits were represented in the 488 bacterial hits. The phylogenetic breakdown of the AS bacterial hits to the representing phylum are as follows: Proteobacteria (n=239, 49%), Actinobacteria (n=173, 35.5%), Firmicutes (n=26, 5.3%), Cyanobacteria (n=8, 1.6%), Deinococcus-Thermus (n=10, 2%), Chloroflexi (n=8, 1.6%), Chlorobi (n=3, 0.6%), Planctomycetes (n=2, 0.4%), Candidate TG1 (n=1, 0.2%), Bacteroidetes (n=5, 1%), Acidobacteria (n=7, 1.4%), and Unclassified (n=6, 1.2%) (Figure1C). The alpha-Proteobacteria were the most frequently represented bacterial phylum consisting of 18% of the bacterial hits.

Similar to other studies which show prokaryotic viruses are more numerous than eukaryotic viruses (13, 19, 21), we found 93% of the viral hits to be hits to bacteriophage, 6% to eukaryotic viruses, and 1% to viral-unclassified. The tailed phage were the most represented viral hits with the Siphoviridae being the most abundant viral family (n=65,

53.7%). Three different families of Eukaryotic viruses were found: Herpesviridae (n=5, 4.1%), Anelloviridae (n=1, 0.8%), and Togaviridae (n=1, 0.8%) (Figure 1D). A hit to a ssRNA virus is interesting considering this is a library constructed from dsDNA. The sequence fragment that hit to the Togaviridae genome did not give a significant protein hit. Although a majority of significant hits from a BLASTx comparison to the GenBank nr protein database revealed unknown viral proteins, a number of different functional viral proteins were also observed including: structural, terminases, integrases, portal, primases, lysozyme, helicases, recombinases, and DNA polymerases.

Based on the taxonomic breakdown the soil sample appears to be a very diverse community. The high relative abundance of Siphoviridae may indicate a large amount of prophage or transducing phage within the library. In comparison with the Fierer *et al.* study, similar phage were observed in both phage studies. Similar phage include: *Actinoplanes* Asp2, *Mycobacterium* Bxz1, *Mycobacterium* Rosebush, *Myxococcus xanthus* Mx8, and *Burkholderia cepacia* BcepC6B (56). Our results would further support their observation that the soil harbors a distinct set of viruses and that viral communities in different soils have commonalities.

Contig assembly

Since there is no universally conserved set of genetic loci among viruses, assessing the diversity and viral species abundance from a metagenomic library analysis is challenging. One method of estimating viral genetic diversity is by contig assembly. Contig assembly aligns sequences that share homology to form a contig containing multiple sequences. The specificity of the assembly (% identity and number of overlapping base pairs) can be adjusted to obtain contigs containing sequences from very

closely related phages. Breitbart *et al.* determined that using the contig assembly parameters of 98% identity with a 20bp overlap enabled differentiation of T3 and T7 coliphages (21). Contig assembly was conducted on the WI soil viral library using these highest stringency settings. A contig spectra of [1339 25 2 1 0 0 0 0 0 0 0 0 0 0] was obtained at this highest stringency of contig assembly and used for submission to the PHACCS online tool. The logorithmic rank-abundance form was the best fit with an error of 2.8. This algorithm estimated 1,700 different viral genotypes with the most abundant virus representing 2.3% of the community.

Since a population of phage that infect a single host species may contain significant genetic diversity between individuals it is hypothesized that lowering the stringency settings will allow for the grouping of sequences from related phage. Varying assemblies were run at different stringencies by changing the % identity from 80% to 85% to 98%. Each contig assembly produced unique contig spectra. As Fierer *et al.* reported, the most stringent assembly of 98% identity and 20 bp overlap gave very few (n=28) overlapping contigs with the highest number of sequences in a contig being 4. Contig assembly at 85% ID gave 77 overlapping contigs with the highest number of sequences in a contig being 18. Contig assembly at 80% ID gave the greatest amount of overlapping contigs with the highest number of sequences in a contig being 41. Regardless of stringency settings, multiple contigs were observed to have strong homology with *Actinoplanes* Asp2 phage which is known to infect an Actinobacteria host (Figure 2). The largest overlapping contig obtained was one containing 41 sequence fragments. When compared against the GenBank nr/nt database, by tBLASTx comparison, the top hit of the contig was to Actinoplanes phage phiAsp2 (Figure 3). The

high frequency of contiguous sequence hits to the same phage indicates the high relative abundance of this phage or closely related phages that are presumably specific to high G+C Gram-positive Actinobacteria.

Comparison of two soil metagenomes

A soil viral metagenomic library was also constructed from Delaware soil by Wommack *et al.* and the first comparative soil viral metagenomic analysis was conducted. The Delaware soil viral metagenome contained 63.9% non-significant hits to the GeneBank nr/nt database whereas the Wisconsin soil had a much lower percentage of non-significant hits (50%) to this database. Both libraries contained a large percentage of hits to bacterial proteins, although these are likely phage-derived.

Since a majority of the sequence fragments from both libraries did not have significant similarity to sequences in the GenBank database, community comparisons were made in two other ways. First, a G+C distribution was constructed based on the individual G+C content of each sequence fragment in each library. The %G+C profiles for both libraries were very similar (Figure 4). Both libraries showed a mean %G+C distribution of 56%. The DE soil had a larger G+C percentage range of between 24% and 74% and the WI soil had a G+C% range of 30% to 69%. The similar G+C profiles indicated that the soils share similar viral communities dominated by high G+C genome content Gram-positive microorganisms. To determine if the soils shared a distinct viral community from marine viral communities, the soil G+C profile was compared to a G+C profile from a Chesapeake Bay viral metagenomic library (13). The Chesapeake Bay viral metagenome has a very different %G+C profile from the two soil libraries, with a mean of 46% G+C. This data suggests that the two soil viral metagenomes are more

similar to each other than they are to the Chesapeake Bay viral metagenome. This also further supports the hypothesis of Fierer *et al.* that distinct natural environments harbor distinct viral communities (56).

Another method to determine similarities between the two libraries is by BLAST comparison. A heat plot of library BLAST comparisons was produced that included the two soil viral metagenomes as well as other viral metagenomes from the Chesapeake Bay, Deep-sea, Hypothermal vents, and Yellowstone Hot Springs (Figure 5). As expected, the two soil viral metagenomes were more closely related to each other than to the Chesapeake Bay. The DE soil was compared to all of the viral metagenomes and it was more similar to the WI soil than it was to any other environment although it did share a noticeable similarity to the Little Hot Creek hot spring, the deep-sea, and the Hypothermal vent viral metagenomes. This similarity between soil viral metagenomic libraries is striking, with 46% of the WI sequence fragments having a significant homolog to a DE soil sequence fragment, whereas only 12% of the WI sequence fragments had a significant homolog to a Chesapeake Bay sequence fragment.

A comparative metagenomic analysis shows us that the soil viral communities are dominated by phage that presumably infect high %G+C Gram-positive Actinobacteria. None of the culture-independent surveys of soil viruses have included a culture-dependent analysis. It would be interesting to determine if a culture-dependent viral analysis from the same soil sample would also be dominated by phage infecting Gram-positive Actinobacteria. Unfortunately, no bacterial culture collection was prepared from the Wisconsin or Delaware soil samples, although soil was stored for further analysis from the WI study. A significant difficulty in pursuing a comparison of culture-

dependent and –independent viral community analyses is the considerable diversity of the soil bacterial community, considered to be the most phylogenetically diverse microbial community on Earth. Future studies are needed to examine how differing methodologies (based on culture-dependent and/or culture-independent analysis) reflect the viral communities present in natural environments.

Figure 1A – Classification of BLASTX hits against GenBank databases. Known hits are defined by significant (E-value ≤0.001) hits to the GenBank nr/nt database, unknown hits are defined as hits with a significant hit only to the GenBank env, and novel hits are defined by non-significant hits.

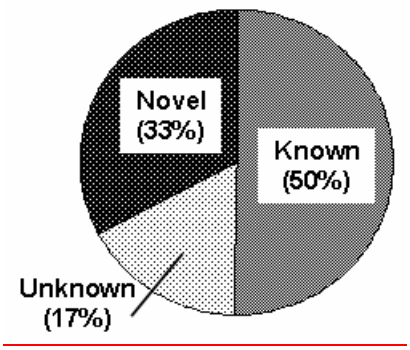


Figure 1B – Biological group breakdown of the WI soil library based on significant hits.

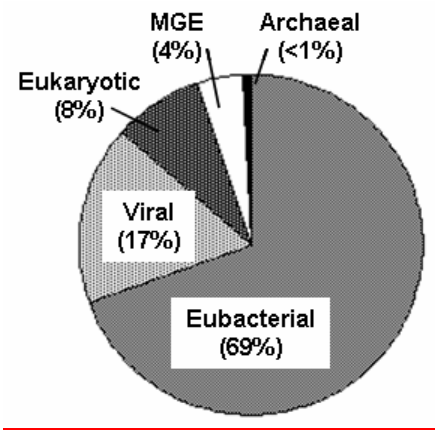


Figure 1C – Significant hits to Eubacteria divisions.

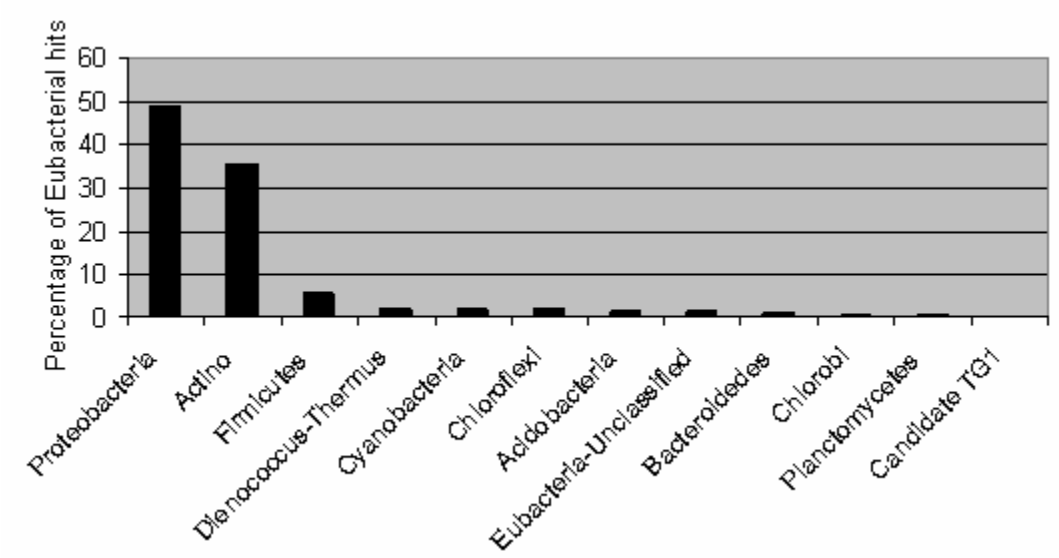


Figure 1D – Viral groups according to ICTV classification.

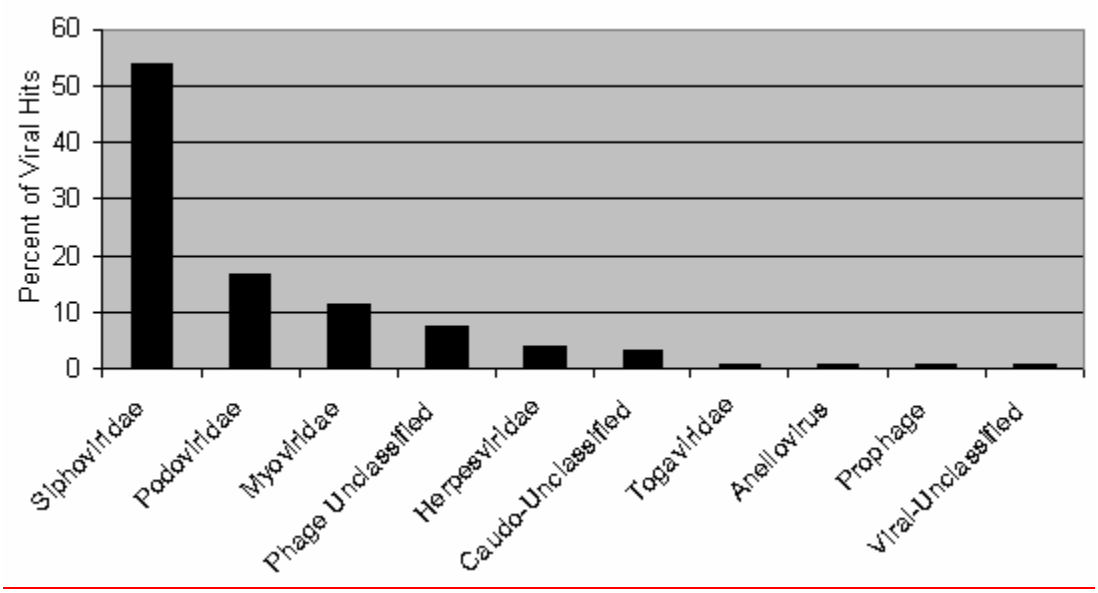


Figure 2 – Contig Assembly plots to show abundance of hits to Acinoplanes phage phiAsp2 (denoted in black).

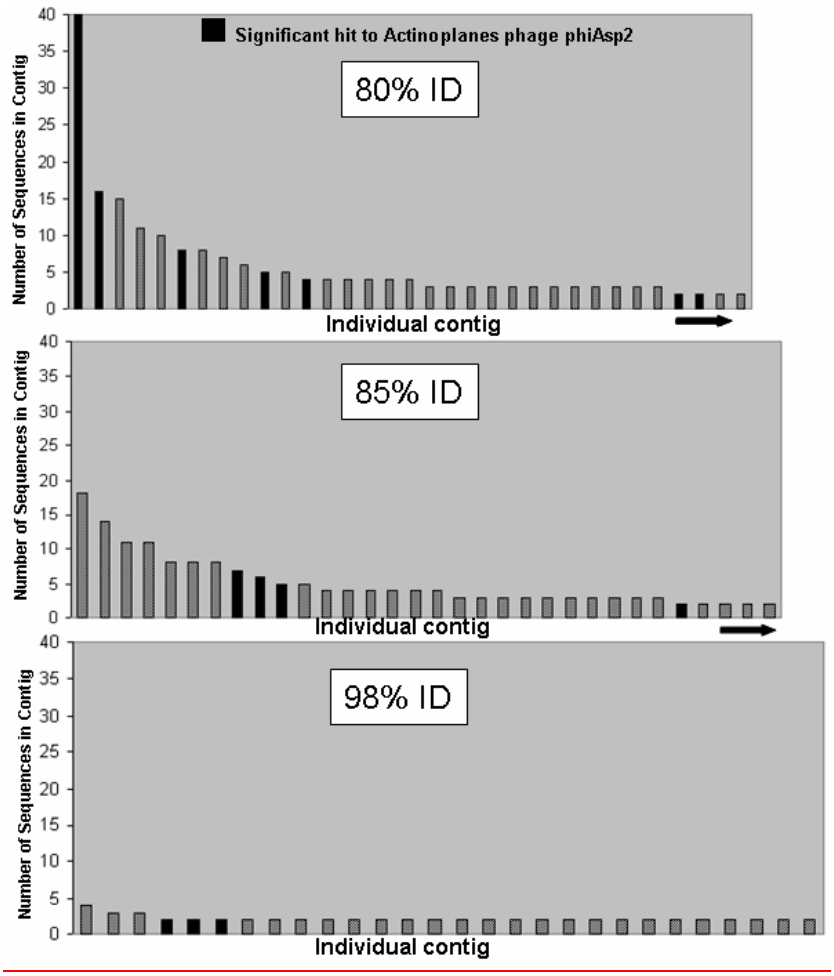
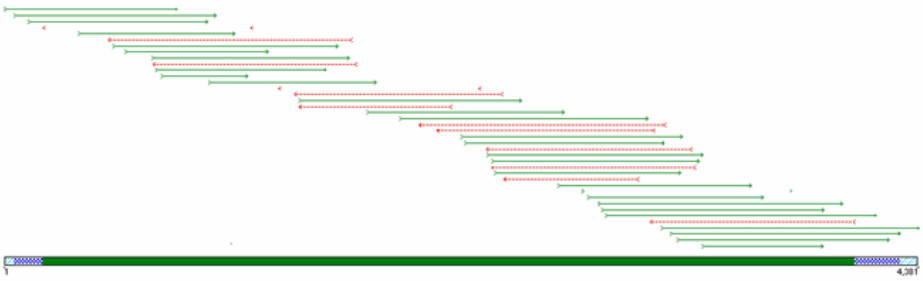


Figure 3 – Alignment of a contig with 41 sequences, and the top 5 BLASTx results.



Sequences producing significant alignments:	(bits)	Value
gb|AY576796.1| Actinoplanes phage phiAsp2, complete genome	91	6e-39
gb|AF547430.1| Mycobacteriophage PG1, complete genome	73	1e-29
gb|DQ398046.1| Mycobacteriophage Orion, complete genome	73	1e-29
emb|CT573213.2| Frankia alni str. ACN14A chromosome, complete se...	91	6e-26
gb|DQ398044.1| Mycobacteriophage Cooper, complete genome	64	9e-22

Figure 4 – G+C profiles of the DE soil, WI soil, and Chesapeake Bay viral metagenomic libraries. The DE soil and WI soil library show a similar G+C profile and both contain a mean G+C distribution of 56% compared to the Chesapeake Bay profile which contained a mean G+C distribution of 46%.

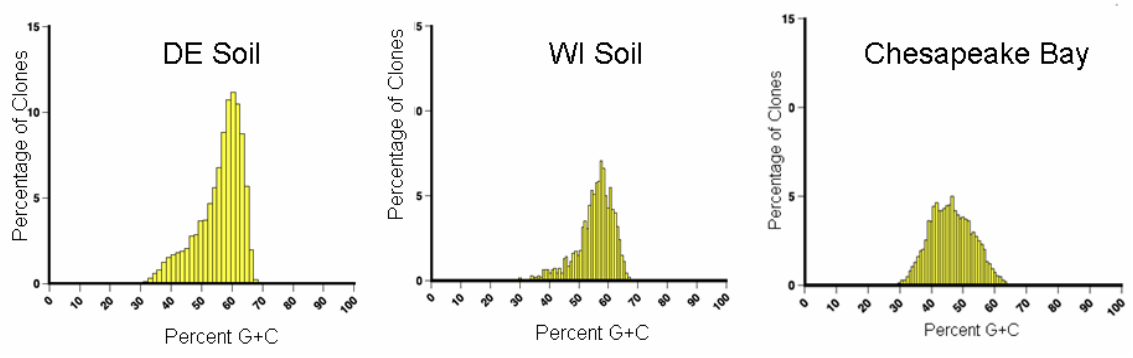
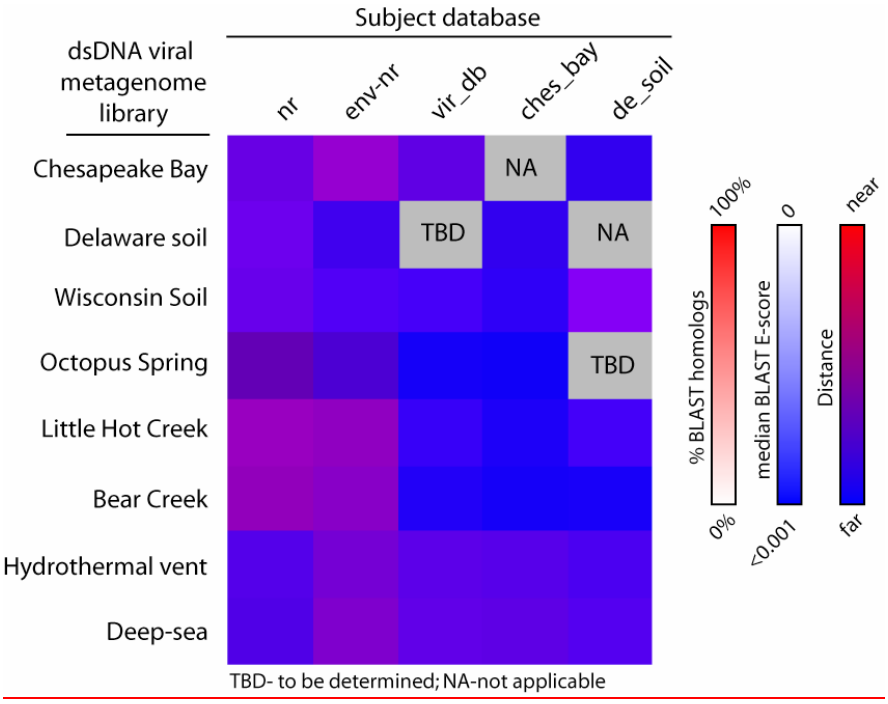


Figure 5 - A heat plot showing similarities between dsDNA viral metagenomic libaries.



VIRAL COMMUNITIES IN ACTIVATED SLUDGE AND INFLUENT

Abstract

This study is the first viral metagenomic analysis of activated sludge and influent samples as well as the first comparison between culture-dependent and –independent analyses of viral communities. A phage culture collection containing 28 plaque-positive (13 purified) phage was obtained. From the viral metagenomic libraries 1,161 sequences from AS (activated sludge) (0.76 MB) and 618 sequences from IN (influent) (0.41MB) were analyzed. Based on BLAST comparisons against the GenBank databases, it was determined that the AS sample was much more diverse (10,001 genotypes) and less characterized (35% novel) compared to the IN sample (160 genotypes and 18% novel). DNA from cultured AS viruses was amplified using primers derived from sequences that were abundant in the AS viral metagenomic library. A phylogenetic analysis of these viral sequences from cultured phage revealed that the cultured phage were more similar to the sequences obtained from a metagenomic library than to any sequence within the GenBank database. These data reveal the striking degree of functional genetic diversity within viruses in wastewater treatment samples, and the paucity of viral sequences within existing GenBank databases.

Introduction

Viruses are the most diverse and abundant biological entity in the biosphere, and many studies show that a majority of viruses are bacteriophage (13, 18, 19, 21). Although some specific virus/host systems are well studied with a very strong bias towards medically important viruses, total environmental viral communities in natural environments are just beginning to be studied in terms of their functional genetic diversity and contributions to microbial ecology. Due to the necessity of a cultured host, total viral communities are impossible to study *in toto* solely by using culture-dependent methods. With the emergence of new culture-independent methods and metagenomic tools, science is just now getting a glimpse at the abundance of viruses in natural environments.

Bacteriophage play an active role in the ecology of the environments in which they exist. Phages can affect biogeochemical cycles (17, 141, 142), are predators to most organisms present (22, 129), affect population dynamics (98), and are responsible for significant amounts of lateral gene transfer between microorganisms (76). Culture-independent metagenomic studies on viral communities have been conducted for several environments and have shown that the diversity and abundance of viruses present exceeds the estimate given by the classical culture-dependent studies (13, 19, 21).

Historically the focus of wastewater viral studies has been on specific host/viral interactions and the use of phages as indicators of the presence of specific hosts. More recently, culture-independent methods have been applied to study the viral communities in activated sludge (AS) (106). Total viral counts in AS have varied based on isolation techniques and counting methods, but highest counts were observed by Otawa *et al.* who

observed $10^8 – 10^9$ viral like particles (VLPs) per ml of AS (106). From culture-dependent studies on AS, phage have been isolated from 30%-60% of the cultured bacterial isolates (68, 79, 80, 88). Hantula *et al.* found phage that infected approximately 18% of their cultured bacterial isolates from AS and that the phages exhibited a broad host range (68). They also noted that approximately 16% of those phage were thermosensitive as well as the appearance and disappearance of cultured phage (68). It has been observed in several studies that coliphages are seen in high titers in influent, but seem to decline in activated sludge (64, 161). Otawa *et al.* used pulse field gel electrophoresis to examine viral communities from 14 full-scale wastewater treatment plants. They noted the range of size of viral genomes in AS to be predominately between 40 and 200kb with 40-70 kb sized genomes the most frequent, and similar banding patterns were observed between samples suggesting that the viral communities had many viruses in common among AS communities (106). Varying temporal patterns in viral populations in a laboratory-scale reactor were observed with two distinct bands emerging and disappearing within a very short time period (106).

Recently there has been discussion of using phages to improve the efficiency of the wastewater treatment process, including pathogen control and foam control in the AS community (see source (156) for a review). With the idea of using phage as a means of bacterial population control in AS, it is even more important to understand the viral community that exists naturally in AS. Although it is known that phages exist in high numbers in AS, the exact role the phages play in this environment is poorly understood (106), and a survey of the total viral community has yet to be done. This study presents the results of two viral metagenomic libraries constructed from AS and influent (IN) from

the Southside Wastewater Treatment Plant in Auburn, Alabama. Also, a culture-dependent survey of this same AS phage community and a sequence batch reactor was conducted. None of the previous metagenomic surveys of viral communities have attempted to link culture-dependent with culture-independent analyses; therefore, our study is not only novel from the point of view of examining activated sludge viral communities, but also in linking disparate methods of viral identification (i.e., culture-dependent and –independent) to appreciate the abundance and diversity of viruses in natural environments as well as the biases inherent in the methodology employed for their study.

Materials and Methods

Sample collection and isolation of total viral community

Two hundred milliliter samples were collected from the IN and AS aeration basin at the Southside Wastewater Treatment Plant (Auburn, AL) on July 10, 2007. Samples were taken at 3:30 pm and were processed that afternoon.

To determine the best way to extract viruses from the activated sludge flock particles, varying extracting methods were previously tried before the library sampling day. Based on previous studies it was shown that beef extract was an effective elluent of viruses from activated sludge and solid sludge samples (4, 5, 102). Viral extraction trials were performed using varying ratios of beef extract to sludge and varying percentage of beef extract to determine the best extraction method. After mixing AS with beef extract at different volume ratios, the samples were agitated at room temperature for 30 min then

the cellular matter was spun down by differential centrifugation. The viral-containing supernatant was filtered through a 0.45 micron PVDF filter. The purified viral sample was stained with SYBR Gold (Molecular Probes, Inc.) and viewed with an epifluorescent microscope. Based on microscopic counts it was determined that 10% beef extract at equal (or greater) volume ratio with AS provided the highest viral like particle (VLP) yield (data not shown).

AS samples for library construction were combined with equal volume of 10x beef extract. Samples were then agitated for 30 min at room temperature, and subjected to centrifugation at 10,000 x g for 30 min. The supernatant was polyethylene glycol (PEG) precipitated at 4˚C overnight. The overnight PEG solution was subjected to centrifugation at 7,000 x g for 45 min. The viral pellet was resuspended in SM buffer [100 mM NaCl, 8 mM MgSO4, 50 mM Tris-HCl (pH 7.5)]. The resuspended solution was subjected to centrifugation (12,000 x g for 10 min) and the supernatant was loaded onto a cesium chloride (CsCl) step gradient and subjected to ultracentrifugation at 100,000 x g for 12 hours. The 1.35-1.5 mg ml^{-1} fraction was collected and used for DNA extraction as well as viral direct counts.

Isolation of total viral community DNA

The purified viral sample was treated with Benzonase (250 units μl^{-1} final concentration) at 37˚C overnight to eliminate any contaminating prokaryotic DNA. Following benzonase inactivation (addition of 10 mM EDTA and heating at 70°C for 10 min), proteinase K (1mg ml^{-1} final) and 1% sodium dodecyl sulphate were added and incubated at 37˚C for 2 hours to degrade viral protein coats. Proteins were removed by phenol:cholorfom extraction and DNA was recovered by isopropanol precipitation. The

DNA was used for pulsed field gel electrophoresis and sent to the Lucigen Corp. (Middleton, WI) for library construction.

Direct viral counts

The CsCl purified viral solution was stained with SYBR Gold (Molecular Probes, Inc.) and the VLPs were directly counted by epifluorescent microscopy using the Zeiss Axiovert 200 Inverted Fluorescence Microscope. VLPs from 10 fields of view were counted. No contaminating cells were observed. Following repetitive sample views, the average counts were taken and it was determined that there were $6x10^{6}$ VLP ml^{-1} in the AS sample.

Bacterial culture collection

AS samples were serially diluted and plated onto LB and SWWA media and incubated at room temperature for at least one week. Unique bacterial mophotypes were restreaked for isolation and genomic DNA was isolated by bead beat extraction method. The 16S rRNA gene from each culture was PCR amplified using universal bacterial primer sets (27F and 1492R) and the resulting PCR products were restriction digested with RsaI and restriction fragments were resolved on a 1.5% agarose gel. Unique bacterial cultures were stored in glycerol at -80˚C (Parsley *et al.*, manuscript in preparation). In many cases multiple bacterial phylotypes were co-isolated but eventually could be separated by sequential restreaking.

Viral culture collection and viral purification

Bacterial cultures from the activated sludge and bioreactor were grown in both LB and LB supplemented with 1mM CaCl and 1mM MgCl at room temperature on a shaker until log phase growth was observed. The cultures were then incubated with

approximately $1.09x10^6$ purified VLPs isolated from AS and the incubation was continued under the same conditions overnight. Following incubation the solutions were evenly divided and half the culture was chloroform lysed and the other half of the culture was stored as a 7% DMSO stock. Spot tests with both chloroform lysed and non lysed enrichment cultures as well as with $2.7x10^5$ purified VLP from AS were performed on all of the cultures to observe plaque formation. Plaque positive samples were then purified by soft agar isolation (3). Phages that were plaque purified were collected for TEM and for DNA extraction. DNA was extracted from viruses as described previously

Library construction and sequencing

Random shotgun genomic libraries of the viral communities were prepared at the Lucigen Corp. (21, 124). Viral DNA was fragmented to approximately 1 to 3 kb fragments and end-repaired to blunt-end each dsDNA molecule. The dsDNA was then ligated to an adaptor (20 bp dsDNA sequence), and then primers specific to the adaptor sequence were used to PCR amplify the dsDNA fragments, using only 15 rounds of amplification. The viral amplicons were then purified over a silica column and ligated into the pSmartLC vector, a low-copy cloning vector that has transcriptional terminators flanking the cloning site to prevent transcription of viral genes (21, 124). Since many viral-encoded gene products may be toxic to *E. coli*, this decreases cloning bias associated with underrepresentation of clones containing genes toxic to the bacterial host. The ligation was transformed into an *E. coli* strain DH10B and a glycerol stock was prepared of the transformation mixture. The transformation was shipped to Auburn University, where the number of transformants per microliter was assessed. Transformants were plated onto LB agar containing Kan (30 μg/ml) and were picked

using the Genetix QPix2 colony picking robot into 96-well format. After growth, sterile 50% glycerol was added to each well (final concentration, 15%) and plates were frozen at -80°C. Seven 96-well plates (672 clones) from IN and 13 96-well plates (1,248 clones) from AS were sent to the Lucigen Corp. for sequencing.

Pulse field gel electrophoresis

DNA from the same sample each library was constructed from was analyzed by pulse field gel electrophoresis (PFGE). DNA was electrophoresed through a 1% agarose gel for 14 hours at 6 V per cm with a 30 sec switch time, while being cooled to 4^{o}C. Following electrophoresis the gel was stained in ethidium bromide for 30 min, then destained in deionized water for 15 min and visualized with an Alpha imager® HP gel documentation system (Alpha Innotech Corporation, San Leandro, CA, USA).

Sequence analysis

Sequences were screened for vector sequence, linker sequence, and minimum base quality using Sequencher 4.8. Sequences were then manually monitored for sequence quality. Following removal of poor sequences there were 1161 quality sequences from AS (0.76 MB) and 618 quality sequences from IN (0.41MB). For taxonomy assignments, sequences were compared against the GenBank nr/nt and env databases using a tBLASTx comparison. The top 5 hits were used for analysis and only matches with an E-value of ≤0.001 were considered significant. Significant hits were categorized as either viral, bacterial, archeal, eukaryotic, or a mobile genetic element. In cases where multiple significant hits were observed for a single query sequence, the sequence was preferentially classified as viral if the hit occurred within the top five hits (21). Mobile elements consisted of plasmids, synthetic sequences, transposons,

transposases, pathogenecity island, and insertion sequences. Significant hits to phages were further classified into phage families according to The International Committee on Taxonomy of Viruses (ICTV) classification (74).

Open reading frames (ORFs) were identified using Chromas Pro v. 1.41. Each ORF was compared to the GenBank non-redundant protein database by a BLASTx comparison. Hits to bacterial genomes were further compared to a list of known prophages. Proteins identified as being located on the genome of a known prophage were categorized in the viral category. This constitutes 2.4% of the phage hits in the AS library and 0.7% of the phage hits in the IN library.

Estimates of viral diversity and community structures

It is possible to predict viral population structure based on contig assembly analysis of metagenomic libraries. A viral community with low genomic diversity would have a larger number of sequences per each contig compared to a high diversity community, based on the contig spectrums produced by assembly with Sequencher 4.8 as described by Breitbart *et al.* (21). The online PHACCS tool was used for assessing viral community diversity (6). The contig spectra were as follows: AS: 892 113 32 5 2 0 0 0 0 0 0 and IN: 522 47 13 2 1 0 0 0 0 0 0 0. The exponential rank abundance form was predicted to be the best fit for the AS viral community and the lognormal rank abundance form was predicted by PHACCS to be the best fit for IN.

Culture dependent and culture independent comparison method

I. Southern Blot with viral metagenomic DNA probe against cultured phage

DNAs isolated from all cultured viral isolates were denatured by boiling then 5 microliters of DNA was spotted onto a nitrocellulose membrane. A digoxigenin (DIG;

Boehringer Mannheim) labeled probe was created using pooled metagenomic clone plasmids as template for PCR with a pSmart vector-specific primer set. A number of hybridizations with varying stringencies were performed, and probe hybridization was detected by Anti-Digoxigenin-AP Fab fragments (Roche Molecular Biochemicals) with NBT/BCIP (Roche Molecular Biochemicals) used for colorimetric detection.

II. PCR of cultured phage DNA with primers based on contig sequences

Contig assembly of viral metagenomic sequences was performed using assembly criteria of 80% identity with at least a 20 bp overlap in order to assemble sequences from the most abundant viral types (21). These contigs were then analyzed by BLAST comparisons to GenBank databases and contigs with significant hits to bacterial taxa that were present in the AS bacterial culture collection were selected. Primers were designed to target regions of the contigs that were 100% identical with homologous genes (Table 1). Touchdown PCR was performed using each primer set on all cultured phage DNAs. PCR amplicons were analyzed by gel electrophoresis, and amplicons chosen for sequencing were column purified (PCR clean-up kit, Promega). PCR products were sequenced at Auburn University using forward and reverse primer reactions. A consensus sequence with 100% agreement between forward and reverse sequences was identified using Chromas Pro sequence analysis software.

Phylogenetic analysis

The evolutionary history was inferred using the Maximum Parsimony method (45). The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed (54). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of

replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (54). The MP tree was obtained using the Close-Neighbor-Interchange algorithm [(103), pg. 128] with search level 3 (54, 103) in which the initial trees were obtained with the random addition of sequences (10 replicates). The tree is drawn to scale, with branch lengths calculated using the average pathway method [see pg. 132 in ref. (103)] and are in the units of the number of changes over the whole sequence . All positions containing gaps and missing data were eliminated from the dataset (Complete Deletion option). For contig110 there were a total of 110 total positions in the final dataset, out of which 58 were parsimony informative. For contig30 there were a total of 88 positions in the final dataset, out of which 65 were parsimony informative. Phylogenetic analyses were conducted using MEGA4 (139).

Results and Discussion

Isolation of bacteriophages infecting cultured bacteria

The first objective of this study was to isolate bacteriophages that infect the collection of unique bacterial isolates previously identified from the AS community (Parsley *et al.*, manuscript in preparation). Viral enrichment was performed in the presence and absence of Ca and Mg. Following plaque assays, the supplemented and non-supplemented media produced similar yields for phage enrichment and identification with the supplemented media producing two additional plaque positive cultures. Chloroform lysis of the bacterial cells following enrichment led to the identification of 14 and 8 additional plaque-forming enrichments from the supplemented and non-supplemented media, respectively. Of the 56 cultured isolates from AS, 24 cultures

showed plaque formation following a spot lysis test of the viral enrichments onto a lawn of the pure bacterial culture (Table 2). Of the 16 sequencing batch reactor-derived bacterial cultures, four were plaque positive (Table 2). Following lytic tests, soft agar overlays were done to obtain phage in pure culture. Of 28 plaque positive phage/host combinations, only 13 were able to be purified (11 from AS hosts and 2 from SBR hosts). Out of the 15 that were not able to be purified, it is probable that many of these were lysogenic phage. It is also possible that some phage are thermosensitive as suggested by Hantula *et al.* (68). In one case it was not necessary to use a viral enrichment to identify a phage, as a pseudolysogenic phage was identified in a *Brevibacterium sanguinis* isolate (Figure1). To our knowledge, this is the first report of a pseudolysogenic phage observed to infect *Brevibacterium sanguinis*. The inability to purify some phage isolates resembles the results of Hantula *et al.*, who observed several phage that produced isolated plaques during the first round of infection in high titers, but no plaques were observed during subsequent infection.

Pulsed field gel electrophoresis analysis of viral communities

The viral DNA samples were electrophoresed using PFGE, revealing a range of viral genome sizes predominantly in the 50 kb size range (Figure 2). The DNA from sample date July 2006 was too concentrated to observe individual bands. There is a noticeable difference in bands between June and July 2007 sample dates with the July sample containing more intense bands at a lower genome size. The influent and the activated sludge samples from July 2007 show similar average genome size and DNA intensities at the expected 50 kb range. In each of these samples it was difficult to visualize individual phage bands, and therefore comparison of phage communities via

PFGE suffers from low resolution making this technique useful in determining the approximate average genome size of the most abundant viral types and lacking in the ability to track distinct phage populations or in gathering information regarding viral genetic diversity.

Metagenomic libraries

Viral metagenomic libraries were constructed from both IN and AS samples. A total of 1,161 and 618 sequencing reads were determined for the AS and IN libraries, respectively. For the AS virome the average read length was 644 bp. PCR amplification of a pooled library template, using vector-specific primers, revealed a range of insert sizes from 2.5 kb to <250 bp, with an approximate average insert size of 1kb (Figure 3).

Bioinformatic analysis

Following sequence generation, the quality sequence reads from AS and IN were compared to multiple databases in GenBank. Fragments were categorized as either 1) known, 2) unknown, or 3) novel based on comparison to the non-redundant nucleotide and the environmental nucleotide databases by tBLASTx analysis. A sequence with a significant hit to the nr/nt database was considered known. A sequence with no significant hit to the nr/nt database and a significant hit to the env/nt database was considered as unknown. A sequence with no significant hit to either database was considered novel. For the AS library, 55% of the sequences were known, 13% were unknown, and 32% were novel. For the IN library 74% of the sequences were known, 8% were unknown, and 18% were novel (Figure 4A). The known hits were further broken down into taxonomic groups. Taxonomic assignments were made based on significant hits to the nr/nt database using a tBLASTx comparison. The top 5 significant

hits for each sequence were considered. If a hit to a viral genome was in the top 5 the sequence was preferentially called as viral. Hits to mobile genetic elements were classified in a separate category and included plasmids, insertion sequences, integrases, transposons, transposases, synthetic sequences, and vector sequences. Of the 641 known hits for the AS library, 389 (61%) were to bacterial genomes (mean E-value: 1.04E-4), 140 (22%) were to viral genomes (mean E-value: 4.72E-5), 57 (9%) were to MGE (mean e-value: 5.21E-5), 53 (8%) were to eukaryotic genomes (mean e-value: 9.76E-5), and 2 (<1%) were to archaeal genomes (mean e-value:6.5E-4). Of the 457 known hits for the IN library, 309 (68%) were to bacterial genomes, 76 (17%) were to MGE, 66 (14%) were to viral genomes, 6 (1%) were to eukaryotic genomes (Figure 4B). The taxonomic groups for both libraries were further categorized.

The high number of non-viral hits in the taxonomic distribution may be counterintuitive if these libraries represent solely viral genome sequences. However, after reviewing previous viral metagenomic analyses from natural environments and considering the high frequency of lysogenic viruses, the high number of non-viral hits is to be expected. First, the methodology used to isolate and purify the viruses takes measures to prevent contaminating prokaryotic cells (i.e., differential centrifugation and membrane filtration) and contaminating DNA (i.e., exonuclease digestion) from interfering with the sample. The CsCl gradient step will free the viral sample of not only contaminating cell, but also contaminating free floating DNA that might be present (21, 134). An exonuclease step was also included prior to viral lysis to further eliminate any contaminating prokaryotic DNA. Secondly, since viruses require such an intimate relation with their host, it is reported that many viruses mimic their host cell

characteristics (such as G+C content and tetranucleotide usage frequencies) leading to an expected "background of homology" between virus and host (13, 114). Thirdly, studies on purified phage cultures show that that 30%-50% of phage ORFs will give significant BLAST hits to bacterial homologs (19, 33, 111, 123). Pedulla *et al.* showed that purified phage genomes are more similar to bacterial ORFs than to other phage ORFS (19, 111). Fourth, non-viral hits could represent uncharacterized prophage or transducing phages. Finally, with only a handful of viral metagenomic studies conducted, there is little information on viral communities represented in the GenBank database. This accounts for the low percentage of viral hits to the GenBank database. Since no metagenomic study has been conducted on activated sludge or influent, we would expect a large percentage of the sample sequences to be uncharacterized.

Eleven different phyla as well as some unclassified hits were represented in the 389 AS bacterial sequences. The phylogenetic breakdown of the AS bacterial hits to the representing phylum are as follows: Proteobacteria (n=281,72.2%), Actinobacteria (n=73, 18.7%), Firmicute (n=14, 3.6%), Unclassified (n=7, 1.2%), Spirochetes (n=3, 0.8%), Acidobacteria (n=2, 0.5%), Deinococcus-Thurmus (n=2, 0.5%), Bacteroidetes (n=2, 0.5%), Cyanobacteria (n=2, 0.5%), Chloroflexi (n=1, 0.25%), Thermotogae (n=1, 0.25%), Chlorobi (n=1, 0.25%). The gamma-Proteobacteria were the most commonly found significant hit representing 26% of the total bacterial sequences (Figure 4C)

Ninety-five percent of the viral hits found in the AS library were to bacteriophage. There were hits to the eukaryotic viruses in Herpesviridae (n=4, 2.9%), Poxviridae (n=2, 1.4%), and Phycodnaviridae (n=1, 0.7%). Of the bacteriophage hits, 10 sequences gave hits to unclassified bacteriophages. Of the known tailed bacteriophage hits there were 46

hits to Myoviridae (32.9%), 44 hits to Siphoviridae (31.4%), 28 hits to Podoviridae (20%) and 5 hits to unclassified Caudovirales (3.8%). Ten of the viral hits were to unclassified phage (7.1%) (Figure 4D). The viral sequence hits were then further analyzed for functional categorization (Table 5).

Although the IN library contained a slightly higher percentage of hits to bacterial genomes, there were only 7 different phyla represented as well as an unclassified bacterial hit. The bacterial phylogenetic breakdown of the IN library is as follows: Proteobacteria (277, 89.6%), Firmicute (10, 3.2%), Actinobacteria (8, 2.6%), Bacteroidetes (7, 2.3%), Spirochetes (3, 1.0%), Cyanobacteria (2, 0.6%), Acidobacteria (1, 0.3%), and unclassified bacteria (1, 0.3%). In this library the alpha-Proteobacteria were the most common phyla representing 58% of the bacterial hits (Figure 4C).

Similar to the AS library, 94% of the IN viral sequences were to bacteriophage. An interesting observation was a hit to the ssRNA viral family Noroviridae. The other eukaryotic viral families represented were Herpesviridae (2, 3%) and Baculoviridae (1, 1.5%). All of the bacteriophage hits were to the Caudovirales and were further broken down as follows Myoviridae (29, 43.9%), Podoviridae (22, 33.3%), Siphoviridae (9, 13.6%), and unclassified Caudovirales (2, 3.0%) (Figure 4D). The viral sequence hits were then further analyzed for homology based functional categorization (Table 5).

From these data, several comparisons between the two samples can be made. First, the genetic diversity of viruses within the AS sample appears to be much less similar to GenBank database sequences compared to the IN sample, as only 55% of the AS library contained a known sequence fragment compared to 74% of the influent data. Also, the AS library contained a higher percentage (32%) of novel sequences than the IN

library (18%). Secondly, based on the taxonomic data, the AS library appears to be more diverse than the IN library. There are 13 different bacterial phyla represented in the AS library compared to only 7 in the IN library. With AS being less similar to known sequences and more phylogenetically diverse, it is hypothesized that the AS viral community has a significantly greater genomic heterogeneity compared to the IN viral community. This is supported by analysis of viral community structures by PHACCS, which estimated the AS viral community to include 10,001 viral genomes whereas the IN viral community was estimated to have 160 viral genomes. Third, it is interesting to note the higher frequency of hits to eukaryotes in the AS sample. This finding is not surprising considering the diverse species of protozoa and algae present in AS. It is also interesting to note the higher frequency of sequences with homology to MGEs in the influent library. It was originally hypothesized that the AS would be a more suitable environment for horizontal gene exchange due to the high cell density and close cell contact. However, the lower incidence of MGE in the activated sludge sample could be a consequence of the higher incidence of Siphoviridae, and/or reflect a higher frequency of MGE-like sequences within mammalian gut flora that contribute to the IN sample. Since the AS community has been evolving as a stable community for many years as opposed to the influent community which is a more transient community, this community stability is reflected in the greater diversity of prokaryotes and viruses within the AS sample.

Combined phylogeny

Both the AS and the IN libraries were compared to the nr/nt database using BLASTx analysis. The top 5 significant hits were reviewed for each sequence and if a viral protein was in the top 5 hits then the sequence was identified as viral in origin.

Following this analysis, the tBLASTx data and BLASTx data were compared for each sequence to determine a combined taxonomic grouping for the two libraries. If a viral hit was found in either the tBLASTx comparison or the BLASTx comparison the sequence was categorized as viral. If there was a discrepancy between the taxonomic classification by both BLASTx and tBLASTx analyses then ORFs for the sequence were predicted using Chromas Pro and a BLASTx analysis was performed for each ORF against the nr/nt database. All ORF BLASTx results were then compared to the tBLASTx data to look for a consistent taxonomic result. Most of the unresolved taxonomies were resolved by this method, but for a few sequences the combined taxonomic classification was based on the BLAST hit with the lowest E-value, thus the most significant homolog was used in taxonomic classification.

For the AS library a taxonomic breakdown based on a combined tBLASTx and BLASTx analyses is as follows: Nonsignificant (n=473, 40.7%), bacterial (n=365, 31.4%), Viral (n=223, 19.2%), Eukaryotic (n=59, 5.1%), MGE (n=37, 3.2%), Archaeal (n=4, 0.3%). For the IN library: Nonsignificant (n=138, 22.3%), Bacterial (n=298, 48%), Viral (n=111, 18%), MGE (n=63, 10.1%), Eukaryotic (n=7, 1.1%), Archaeal (n=1, 0.2%) (Figure 4E).

The combined phylogenetic classification based on BLASTx and tBLASTx analyses gives us a more complete determination of the phylogenetic classification of the viral libraries, but it is important to consider some potential biases in the analysis. Since ORFs were not called for all of the sequences it could be argued that some small percentage of the sequences might be classified in an incorrect taxonomic category. For example, Breitbart *et al.* noticed several of their clones would give a viral hit when

sequenced at one end, but a bacterial or eukaryotic hit when sequenced at the opposite end of the insert (21). Using BLASTx data without calling ORFs will produce a result for a single protein per sequence whether the sequence contains information for one or more than one protein. The BLASTx result for each sequence will be to the protein on the sequence that has the lower E-value. Therefore, some of the sequences fragments categorized in the bacterial, archaeal, and eukaryotic groups could contain a viral protein with a higher E-value on another location on the fragment. This would make the combined phylogeny an underestimate of viral homology. However, as the average length of each sequence read was 644 bp and 662 bp for the AS and IN libraries respectively, and the probability is low that multiple ORFs with differing taxonomic classification would exist on a single sequence read, the estimate of taxonomic distribution of the viral libraries should be reasonably close to what would be achieved by calculating taxonomic distributions for each ORF on every sequence. As was previously discussed, the percentage of hits to bacterial (or other host genomes) is likely a vast overestimation of the contribution of bacterial-derived genes to the viral genomes, and is most likely a function of the paucity of environmental viral genomes within GenBank databases. Due to this severe database bias, the exact taxonomic distribution within the viral metagenomic libraries is also severely biased in favor of prokaryotes, so that the small effect that may be seen from incorporating BLAST analysis of each ORF would be inconsequential to the actual taxonomic distribution. Fortunately, by submitting the sequence from this study into the GenBank databases, future studies of viral communities will be less biased in favor of prokaryotic taxa.

Library diversity estimates

Since there is no universally conserved set of genetic loci among viruses, assessing the diversity and viral species abundance from a metagenomic library analysis is challenging. One method of accessing this information is by contig assembly. Contig assembly aligns sequences that share homology to form a contig containing multiple sequences. You can adjust the specificity of the alignment (% identity and number of overlapping base pairs) to obtain contigs containing sequences from very closely related phages. Breitbart *et al.* determined that using the contig assembly parameters of 98% identity with a 20bp overlap enabled differentiation of T3 and T7 coliphages (21). A contig assembly was performed using Sequencher 4.8 on both the AS and IN library with assembly criteria of 98% identity and 20bp overlap. The contig spectra obtained from both libraries were then processed the PHACCS tool (http://biome.sdsu.edu/phaccs/) (6). The exponential and the niche-preemption rank-abundance forms were the best fit for the AS library (Table 4). Based on the exponential form, the AS library contained 10,001 different viral genotypes with the most abundant virus representing 2.27% of the community. The lognormal rank-abundance form was the best fit for the IN library (Table 5). Based on the lognormal form the IN sample contained 160 different viral genotypes with the most abundant virus representing 7% of the community (for direct comparison purposes, the exponential form applied to the IN contig spectrum yielded a 9 point increase in error over the lognormal form, but predicted 10,001 different genotypes, with the most abundant virus representing 3.4% of the community). While these may be conservative estimates, they are consistent with the other data suggesting that AS is more diverse than IN. The high relative abundance of the most abundant virus in the IN library

may suggest that there was a particular lytic virus that was dominant during the time of sampling.

Viral proteins

Following the tBLASTx and BLASTx analysis, sequences in the combined phylogeny viral category were further examined to determine viral protein functional assignment. ORFs were called on each sequence and a BLASTx analysis was conducted against the nr/nt database. ORFs with significant hits were assigned into functional categories. There were 33 different groups of functional protein hits for the AS library and 22 for the IN library (Table 5). This is consistent with the taxonomic results suggesting that the activated sludge sample was more functionally diverse than the influent sample.

Culture-independent vs. culture-dependent results

To determine the similarities between the culture-dependent and culture-independent community assessments, two methods were utilized. A Southern blot (dot-blot) DNA hybridization using a DIG-labeled probe constructed from the pooled AS library DNA was performed against all of the cultured viral isolates. No hybridization was observed between the pooled metagenomic library and the cultured isolates, despite successful hybridization to positive control samples (data not shown). This could be due to the cultured viruses existing at low relative abundance within the activated sludge community and therefore were not represented in the viral metagenomic library. This demonstrates the extreme bias in the cultivation of environmental viruses, as they may not represent the most relatively abundant viruses within the community.

The second method to compare culture-dependent and –independent analyses of viral communities was based on the contig assembly data from the AS metagenomic library. Contig assembly at 80% mismatch and 20 base pair overlap resulted in 152 different contigs with at least 2 sequences per contig. Contigs were compared to the GenBank nr database by BLASTx and sorted based on similarities to the cultured isolate phylogeny. From the data there were four contigs that showed similar phylogeny matches to those represented in the culture collection. The four phylogenies represented by the contigs and cultures are: *Citrobacter sp.*, *Aeromonas sp.*, *Acinetobacter sp., and Klebsiella sp*. Oligonucleotide primers were designed from the four contigs. Contig 134 was over 600 bp long so two primers were developed specific to that contig sequence (Table 1). DNA from the purified phages was extracted and amplified in a PCR reaction using the metagenomic sequence-specific primers. Out of the 65 different reactions, only three gave products. The three viral PCR products, one from virus 47 (contig 30-specific primers) and two from viruses 5 and 53 (contig 110-specific primers) were sequenced in both the forward and reverse directions. These sequences were compared to the metagenomic sequences within the specific contigs used for primer design as well as other top matches from the GenBank database. A phylogenetic tree was constructed to determine the phylogenetic relationships between the cultured viral sequences, the metagenomic sequences, and related sequences in GenBank (Figures 5 and 6). In every case the viral metagenomic sequences were more closely related to the cultured phage sequence than to any other sequence in the GenBank database. The phylogenetic affiliation between the culture-independent viral sequences and the culture-dependent viral sequences had high bootstrap support, with > 50% support for each clade.

TABLE 1 – Designed primers from metagenome library used for amplification of viral isolate DNA.

Contig ID	Forward Primer	Reverse Primer
C19	5'-CAC AAC GTC GTC GTG TAG – 3'	5' – GAC GTA AGG AGC ACT TGG – 3'
C30	5' – TGC TGT TGG TCA TGG AAG – 3'	5' – TCT ATC GAT CTG GCT TGG – 3'
C110	5' – GTA TCA GAT ACA AGC GGC – 3'	5' ATC GCC ATT TGA ACA TAG – 3'
C134A	5' – ACG AGA GAG GAT GCT CAC – 3'	5' GCT CAC TCA TTA GGC ACC – 3'
C134B	5' – TGA GAG AGT TGC AGC AAG – 3'	5' – TGG TGT CGA TGG TAG AAC – 3'

Table 2. – Phylogenetic breakdown of the bacterial host to the cultured phage. The phylogeny of the bacterial hosts on the x-axis and the number of isolates that tested plaque positive.

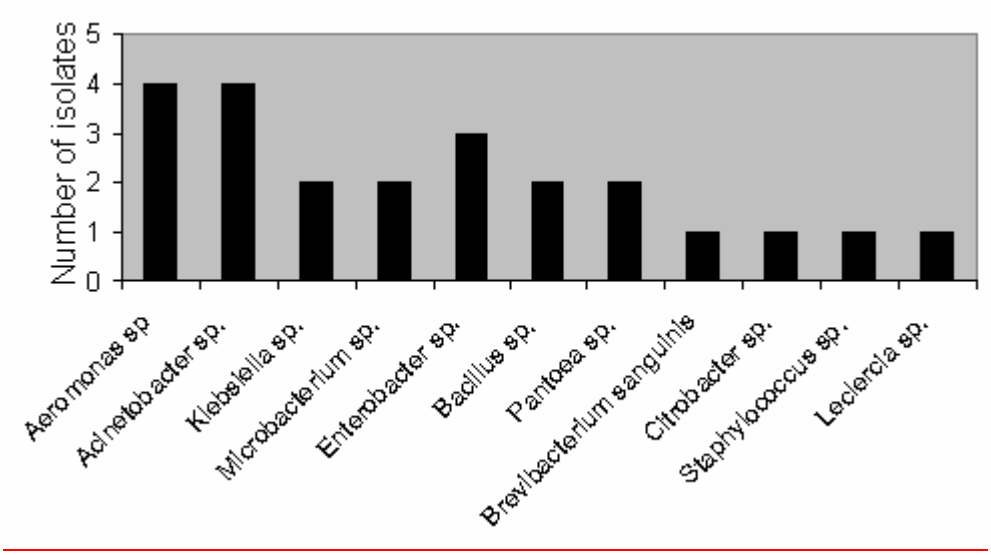


Figure1 – Plaques from a pseudolysogenic bacteriophage on *Brevibacterium sanguinis*. Consecutive streaks of isolated colonies with no plaque formation yielded plaque production.

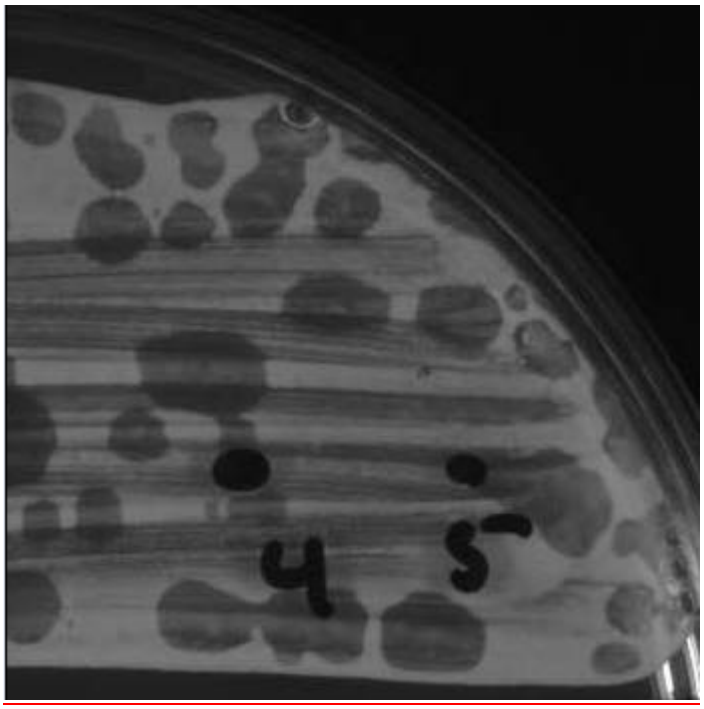


Figure 2. Pulse Field Gel Electrophoresis: Image of stained viral community DNA. Lane one represents the molecular weight marker. Lane 2-4 are activated sludge samples from the given dates. Lane 5 is a sequencing batch reactor sample from the given date. Lane 7 is the influent sample from given date. Libraries were constructed from the same samples as lane 4 and 7.

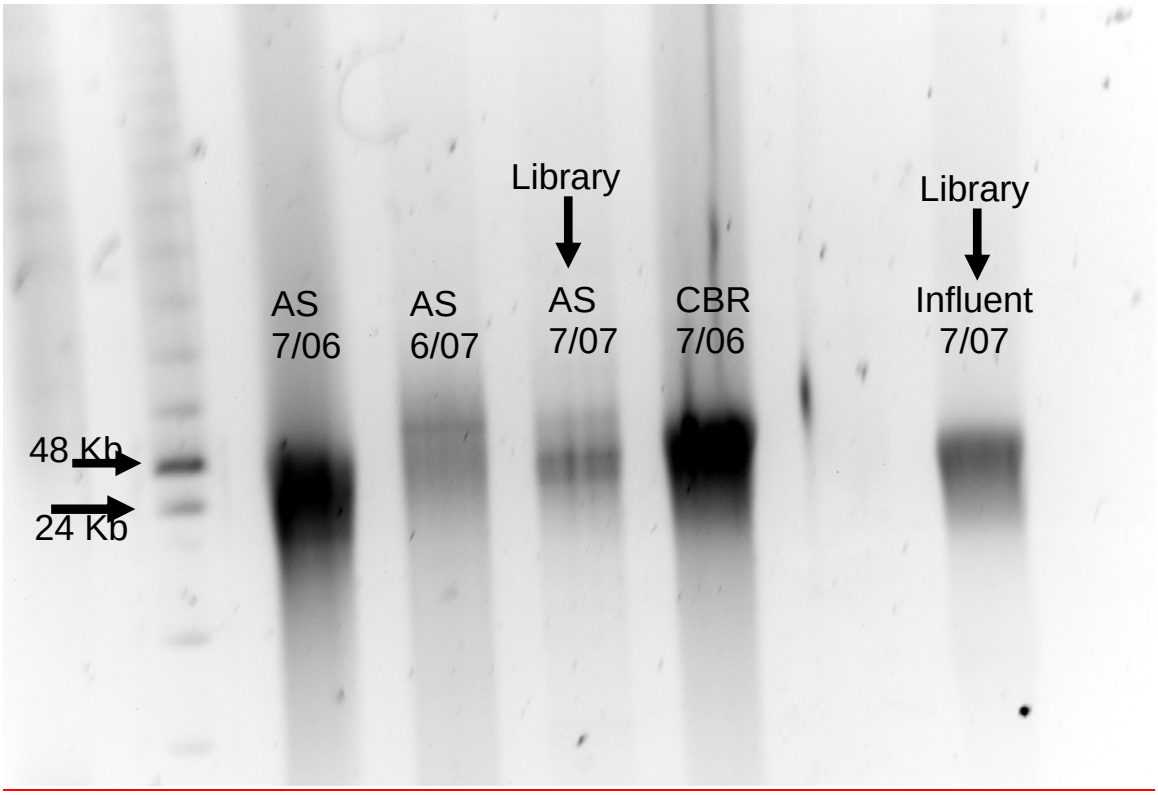


Figure 3– Inserts from a plasmid preparation on the AS viral metagenomic library were amplified and run on a 1% agarose gel to determine the average insert size.

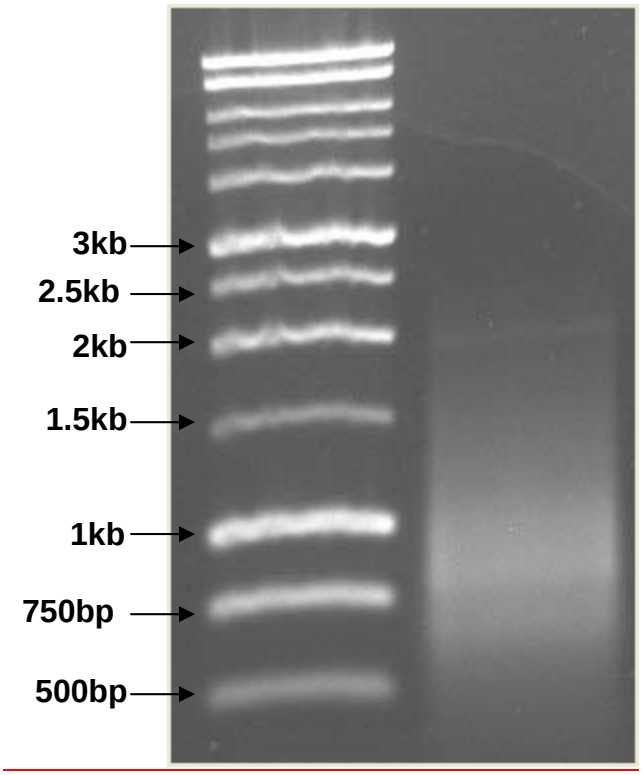


Figure 4A – Comparison of metagenomic analysis between the AS and IN metgenomic libraries. tBLASTx comparisons between the GenBank nr/nt and env/nt databases were conducted. Sequences with significant (≤ 0.001 E-value) hits to the nr/nt database were considered known, sequences with significant hit to the env/nt database but a nonsignificant hit to the nr/nt database was considered unknown, and sequences with no significant hit to either nucleotide database was considered novel.

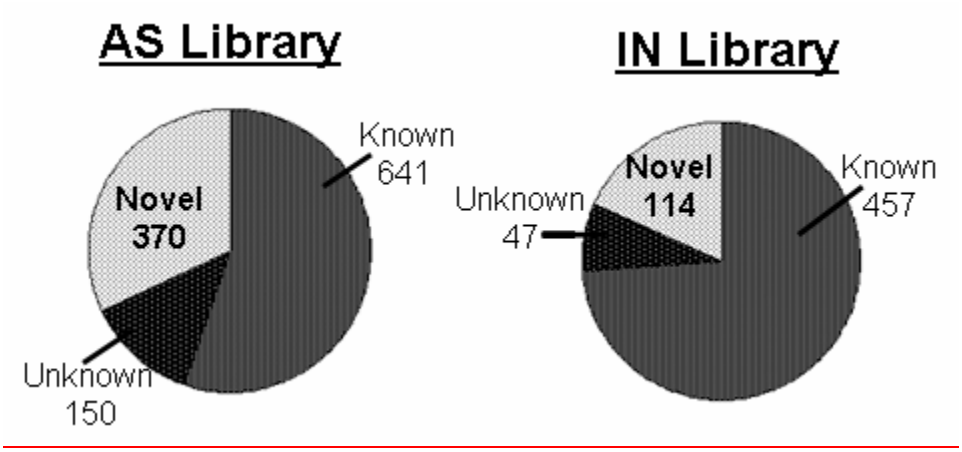


Figure 4B – Phylogenetic breakdown of the significant hits from the AS and IN metagenomic libraries based on a tBLASTx comparison against the Genebank nr-nt database.

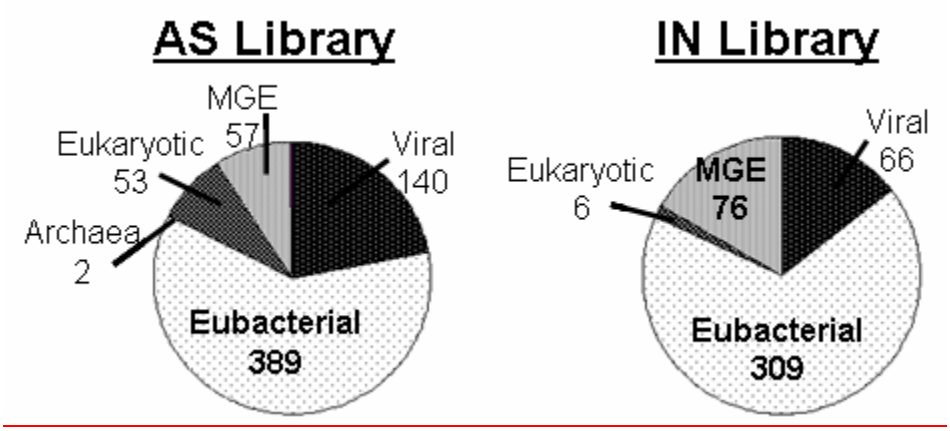


Figure 4C – Phylogenetic breakdown of the eubacterial hits from the AS and IN metagenomic libraries based on a tBLASTx comparison against the GenBank nr/nt database.

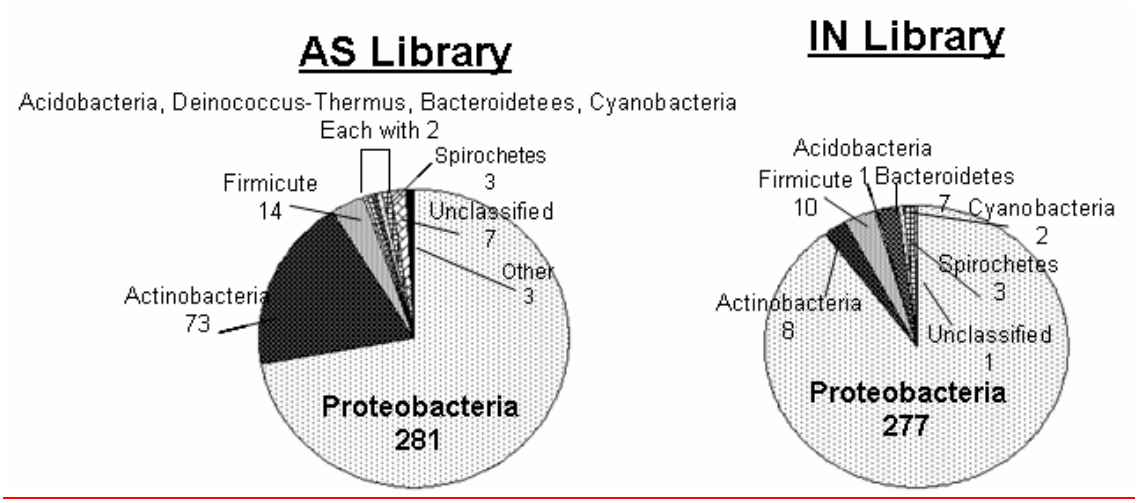


Figure 4D – Viral breakdown from the AS and IN metagenomic libraries based on a tblastx comparison against the GenBank nr/nt database.

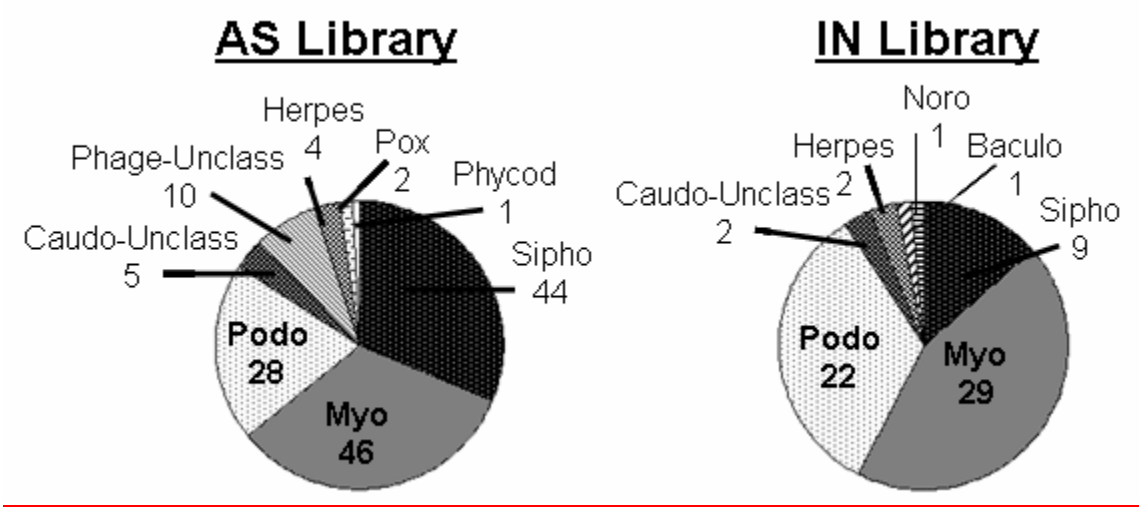


Figure 4E - Taxonomic groupings based on a combination of the tBLASTx and BLASTx comparisons giving significant hits..

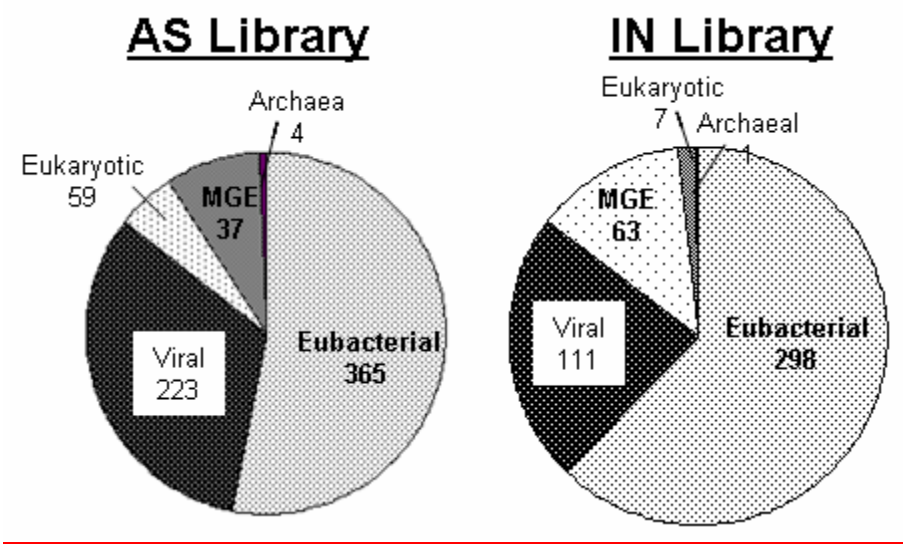


Table 3 - PHACCS results for the AS library contig spectra based on a 98% similarity and a 20bp overlap.

Rank-abundance form	Error	Richness (genotypes)	Evenness	Most abundant genotype (percentage of community)	Shannon-Wiener Index
Power	22.5	113	0.976	3.88%	4.61 nats
Exponential	**15.5**	**10001**	**0.518**	**2.27%**	**4.77 nats**
Logarithmic	26.1	104	0.985	4.11%	4.57 nats
Lognormal	19.6	133	0.955	3.8%	4.67 nats
Niche-preemption	**15.5**	**2220**	**0.619**	**2.27%**	**4.77**
Broken stick	16.9	170	0.919	3.36%	4.72

Table 4 - PHACCS results for the IN library contig spectra based on a 98% similarity and a 20bp overlap.

Rank-abundance form	Error	Richness (genotypes)	Evenness	Most abundant genotype (percentage of community)	Shannon-Wiener Index
Power	11	120	0.94	7.1%	4.5 nats
Exponential	19	10001	0.47	3.4%	4.4 nats
Logarithmic	14	92	0.96	7.3%	4.4 nats
Lognormal	**10**	**160**	**0.89**	**7%**	**4.5 nats**
Niche-preemption	19	10001	0.47	3.4%	4.4 nats
Broken stick	17	120	0.91	4.5%	4.4 nats

Table 5 – Breakdown of functional proteins found in the AS and IN metagenomic libraries based on sequence homology and blastx comparison to the GenBank nr protein database.

Protein	AS Library	IN Library
Unknown	113	40
DNA Methylase	15	2
DNA Replication	2	0
DNA Packaging	1	0
DNA Binding	2	0
DNA Transposition	0	1
DNA Polymerase	6	0
RNA Polymerase	0	1
RNA Ligase	1	0
Reverse Transcriptase	2	12
Ribonucleoside Reductase	1	0
Endonuclease	5	2
Exonuclease	0	1
Recombinase	2	0
Primase	2	0
Terminase	11	7
Transposase	1	0
Integrase	6	6
Helicase	8	1
Protease	1	0
Carboxylase	0	1
Portal	12	0
Lysozyme	3	0
Prophage antirepressor	0	5
Host specificity	0	2
Prohead Protease	2	0
Head Completion	2	0
Head Morphogenesis	2	0
Head Protein	0	1
Phage Coat	1	0
Capsid	2	0
Virion Structure	4	5
Internal Virion	1	0
Capsid Scaffolding	0	9
Encapsidation	0	2
Tegument	0	1
Tail Fiber	7	8
Tail Tape	4	0
Tail Assembly	4	0
Tail Component	1	0
Tail Sheath	1	6
Tail Tube	0	3
Tape Measure	0	1
Type III Restriction	1	0
Integral Membrane	1	0
TerL	1	0
Late Promotor	0	1
Totals	228	118

Figure 5 - Evolutionary relationships of phage sequences derived from culture-dependent and –independent sources and related IS4 transposases.

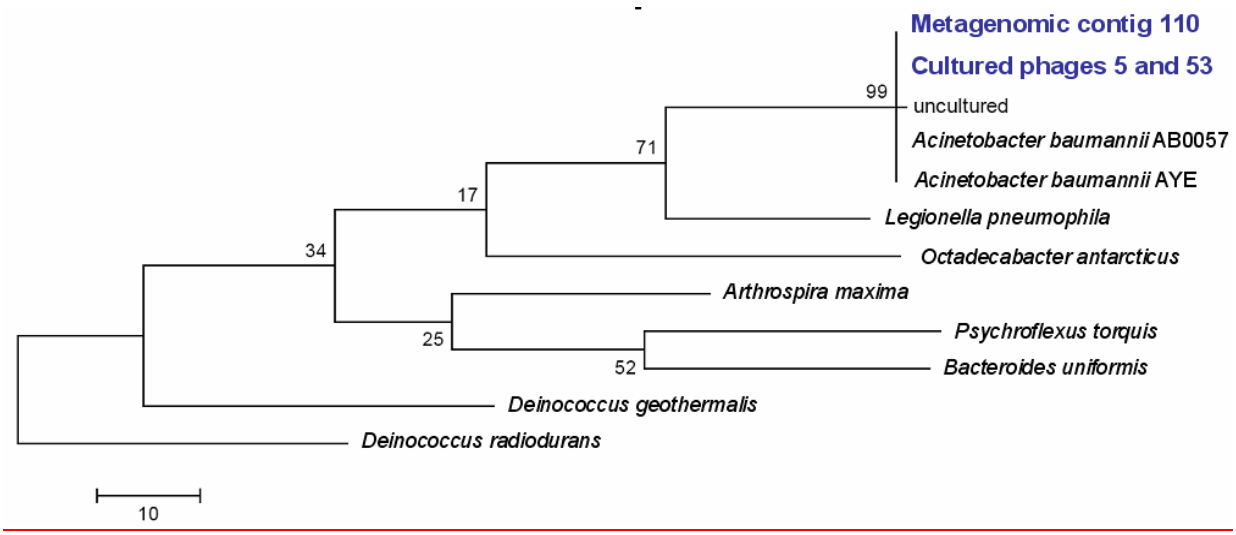
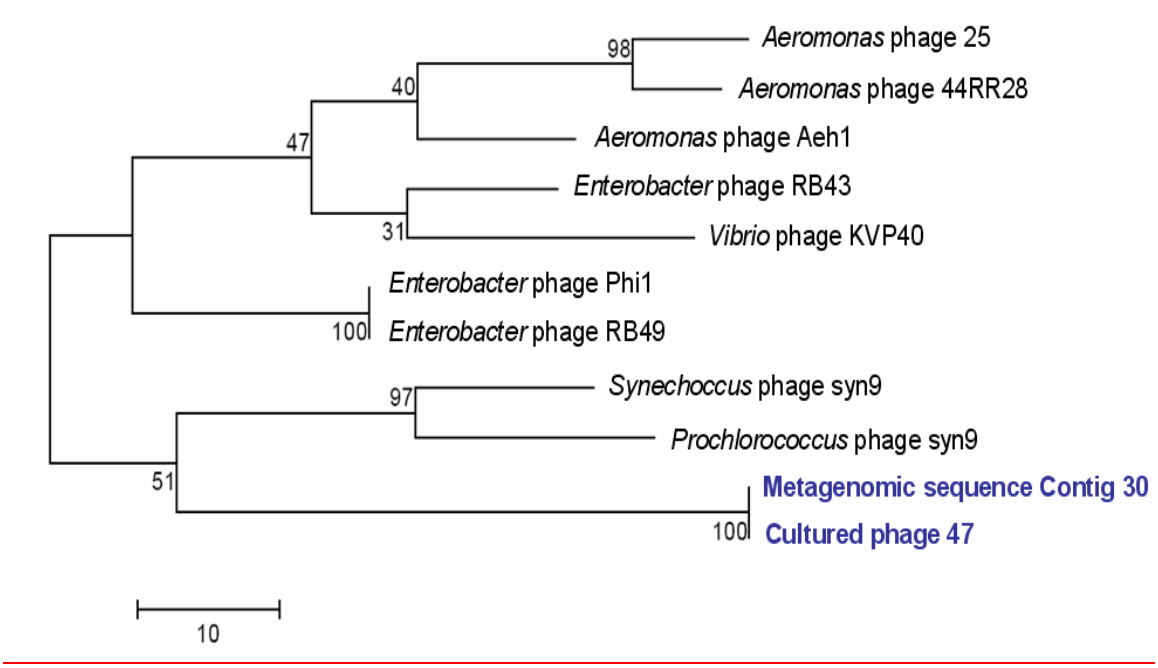


Figure 6 - Evolutionary relationships of phage sequences derived from culture-dependent and –independent sources and related phage gp4 head completion proteins



COMPREHENSIVE BIBLIOGRAPHY

1. **Acheson, D. W. K., J. Reidl, X. Zhang, G. T. Keusch, J. J. Mekalanos, and M. K. Waldor.** 1998. In Vivo Transduction with Shiga Toxin 1-Encoding Phage. Infection and Immunity **66:**4496.
2. **Ackermann HW, D., M.S.** 1987. Viruses of Prokaryotes, vol. 1. CRC Press, Boca Raton.
3. **Adams, M. H.** 1959. Bacteriophages.
4. **Ahmed, A. U., and D. L. Sorensen.** 1995. Kinetics of pathogen destruction during storage of dewatered biosolids. Water Environment Research **67:**143-150.
5. **Alouini, S., and M. D. Sobsey.** 1995. Evaluation of an extraction-precipitation method for recovering Hepatitis A virus and poliovirus from hardshell clams(Mercenaria mercenaria). Water Science & Technology **31:**465-469.
6. **Angly, F., B. Rodriguez-Brito, D. Bangor, P. McNairnie, M. Breitbart, P. Salamon, B. Felts, J. Nulton, J. Mahaffy, and F. Rohwer.** 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics **6:**41.
7. **Araujo, R. M., A. Puig, J. Lasobras, F. Lucena, and J. Jofre.** 1997. Phages of enteric bacteria in fresh water with different levels of faecal pollution. J Appl Microbiol **82:**281-6.

8. **Ashelford, K. E., M. J. Day, M. J. Bailey, A. K. Lilley, and J. C. Fry.** 1999. In Situ Population Dynamics of Bacterial Viruses in a Terrestrial Environment. Applied and Environmental Microbiology **65:**169-174.

9. **Ashelford, K. E., M. J. Day, and J. C. Fry.** 2003. Elevated Abundance of Bacteriophage Infecting Bacteria in Soil. Applied and Environmental Microbiology **69:**285-289.

10. **Ashelford, K. E., J. C. Fry, M. J. Bailey, A. R. Jeffries, and M. J. Day.** 1999. Characterization of Six Bacteriophages of Serratia liquefaciens CP6 Isolated from the Sugar Beet Phytosphere. Applied and Environmental Microbiology **65:**1959-1965.

11. **Ashelford, K. E., S. J. Norris, J. C. Fry, M. J. Bailey, and M. J. Day.** 2000. Seasonal Population Dynamics and Interactions of Competing Bacteriophages and Their Host in the Rhizosphere. Applied and Environmental Microbiology **66:**4193-4199.

12. **Azam, F., T. Fenchel, J. G. Field, J. S. Gray, L. A. Meyer-Reil, and F. Thingstad.** 1983. The ecological role of water-column microbes in the sea. Marine ecology progress series. Oldendorf **10:**257-263.

13. **Bench, S. R., T. E. Hanson, K. E. Williamson, D. Ghosh, M. Radosovich, K. Wang, and K. E. Wommack.** 2007. Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol **73:**7629-41.

14. **Bergh, O., K. Y. Boersheim, G. Bratbak, and M. Heldal.** 1989. High abundance of viruses found in aquatic environments. Nature **340:**467-468.

15. **Botstein, D., and I. Herskowitz.** 1974. Properties of hybrids between Salmonella phage P22 and coliphage lambda. Nature **251:**584-9.

16. **Boyd, E. F., and H. Brüssow.** 2002. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends in Microbiology **10:**521-529.

17. **Bratbak, G., m. Heldal, T.F. Thingstad, B. Riemann, and O.H. Haslund.** 1992. Incorporation of viruses into the budget of microbial C-transfer. A first approach. Mar. Ecol. Prog. Ser **83:**273-280.

18. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proceedings of the Royal Society B: Biological Sciences **271:**565-574.

19. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol **185:**6220-3.

20. **Breitbart, M., Rohwer, F., Abedon, S.T.** 2005. Phage ecology and bacterial pathogenesis. *In* M. K. Waldor, Friedman, D.I., Adhya, S. L. (ed.), Phages: their role in bacterial pathogenesis and biotechnology. ASM Press, Washington D. C.

21. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A **99:**14250-5.

22. **Breitbart, M., L. Wegley, S. Leeds, T. Schoenfeld, and F. Rohwer.** 2004. Phage community dynamics in hot springs. Appl Environ Microbiol **70:**1633-40.

23. **Brown, D. T., J. M. MacKenzie, and M. E. Bayer.** 1971. Mode of host cell penetration by bacteriophage phi X174. J Virol **7:**836-46.

24. **Bruynoghe, R., and J. Masin.** 1921. Essais de therapeutique au moyen de bacteriophage. C.R. Soc. Biol. Paris **84:**1120-1121.

25. **Campbell, A.** 2006. General aspects of lysogeny, p. 66-73. *In* R. Calendar (ed.), The Bacteriophages, 2nd ed. Oxford University Press, New York, N.Y.

26. **Campbell, J. I. A., M. Albrechtsen, and J. Sorensen.** 1995. Large Pseudomonas phages isolated from barley rhizosphere. FEMS Microbiology Ecology **18:**63-74.

27. **Cann, A.** 2005. Principles of molecular virology, 4th ed. Elsevier Academic Press, Amsterdam ; Boston.

28. **Casida, L. E., and K. C. Liu.** 1974. Arthrobacter globiformis and Its Bacteriophage in Soil 1. Applied and Environmental Microbiology **28:**951-959.

29. **Casjens, S.** 2003. Prophages and bacterial genomics: what have we learned so far? Molecular Microbiology **49:**277-300.

30. **Caspar, D. L., R. Dulbecco, A. Klug, A. Lwoff, M. G. Stoker, P. Tournier, and P. Wildy.** 1962. Proposals. Cold Spring Harb Symp Quant Biol **27:**49-50.

31. **Caspar, D. L., and A. Klug.** 1962. Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol **27:**1-24.

32. **Chattopadhyay, S., and R. W. Puls.** 2000. Forces dictating colloidal interactions between viruses and soil. Chemosphere **41:**1279-1286.

33. **Chen, F., and J. Lu.** 2002. Genomic Sequence and Evolution of Marine Cyanophage P60: a New Insight on Lytic and Lysogenic Phages. Applied and Environmental Microbiology **68:**2589.

34. **Chen, F., and C. A. Suttle.** 1996. Evolutionary Relationships among Large Double-Stranded DNA Viruses That Infect Microalgae and Other Organisms as Inferred from DNA Polymerase Genes. Virology **219:**170-178.

35. **Chen, F., C. A. Suttle, and S. M. Short.** 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Applied and Environmental Microbiology **62:**2869-2874.

36. **Chibani-Chennoufi, S., M. L. Dillmann, L. Marvin-Guy, S. Rami-Shojaei, and H. Brussow.** 2004. Lactobacillus plantarum bacteriophage LP65: a new member of the SPO1-like genus of the family Myoviridae. J Bacteriol **186:**7069-83.

37. **Cochran, P. K., C. A. Kellogg, and J. H. Paul.** 1998. Prophage induction of indigenous marine lysogenic bacteria by environmental pollutants. Marine Ecology Progress Series **164:**125-133.

38. **Crick, F. H., and J. D. Watson.** 1956. Structure of small viruses. Nature **177:**473-5.

39. **d'Herelle, F.** 1926. The Bacteriophages and Its Behavior Williams and Wilkins, Baltimore, Md.

40. **d'Herelle, F.** 1925. Essai de traitement de la pest buboniqe par le bacteriophage. Presse Med. **33:**1393-1394.

41. **d'Herelle, F.** 1921. Le Bacteriophages: Son Role dans I'Immunite, Masson, Paris, France.

42. **d'Herelle, F.** 1930. Studies on Asiatic cholera. Indian Med. Res. Memoirs **14**.

43. **d'Herelle, F.** 1917. Sur un microbe invisible antagoniste des bacilles dysenteriques. C.R. Acad. Sci. Paris **165**.

44. **Danovaro, R., and M. Serresi.** 2000. Viral Density and Virus-to-Bacterium Ratio in Deep-Sea Sediments of the Eastern Mediterranean. Applied and Environmental Microbiology **66:**1857-1861.

45. **Dayhoff, M. O.** 1978. Atlas of protein sequence and structure. Vol. 5. Suppl. 3. National biomedical research foundation.

46. **Dias, F. F., and J. V. Bhat.** 1965. Microbial Ecology of Activated Sludge. Ii. Bacteriophages, Bdellovibrio, Coliforms, and Other Organisms. Appl Microbiol **13:**257-61.

47. **Dowd, S. E., S. D. Pillai, S. Wang, and M. Y. Corapcioglu.** 1998. Delineating the Specific Influence of Virus Isoelectric Point and Size on Virus Adsorption and

Transport through Sandy Soils. Applied and Environmental Microbiology **64:**405-410.

48. **Eaton, M. D., and S. Bayne-Jones.** 1934. Bacteriophage therapy. JAMA **103**.

49. **Edwards, R. A., and F. Rohwer.** 2005. Viral metagenomics. Nat Rev Microbiol **3:**504-10.

50. **Ellis, E. L., and M. Delbruck.** 1939. THE GROWTH OF BACTERIOPHAGE. The Journal of General Physiology **22:**365-384.

51. **Ewert, D. L., and M. J. Paynter.** 1980. Enumeration of bacteriophages and host bacteria in sewage and the activated-sludge treatment process. Appl Environ Microbiol **39:**576-83.

52. **Faraj, A. A.** 2006. Poliomyelitis: Orthopaedic management. Current Orthopaedics **20:**41-46.

53. **Faruque, S. M.** 2000. Sunlight-Induced Propagation of the Lysogenic Phage Encoding Cholera Toxin. Infection and Immunity **68:**4795.

54. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783-791.

55. **Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A. N. Suvorov, S. Kenton, H. S. Lai, S. P. Lin, Y. Qian, H. G. Jia, F. Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S. W. Clifton, B. A. Roe, and R. McLaughlin.** 2001. Complete genome sequence of an M1 strain of Streptococcus pyogenes. Proc Natl Acad Sci U S A **98:**4658-63.

56. **Fierer, N., M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R. A. Edwards, B. Felts, and S. Rayhawk.** 2007. Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil? Applied and Environmental Microbiology **73:**7059-7066.

57. **Flu, P.-C., and E. Renaux.** 1932. Le phenomene de Twort et la bacteriophage. Ann Inst Pasteur **48:**15-18.

58. **Freifelder, D.** 1987. Molecular biology. Jones & Bartlett Inc., Boston.

59. **Fuhrman, J. A.** 1999. Marine viruses and their biogeochemical and ecological effects. Nature **399:**541-548.

60. **Fuhrman, J. A., and R. T. Noble.** 1995. Viruses and protists cause similar bacterial mortality in coastal seawater. Limnology and Oceanography **40:**1236-1242.

61. **Fuller, N. J., W. H. Wilson, I. R. Joint, and N. H. Mann.** 1998. Occurrence of a Sequence in Marine Cyanophages Similar to That of T4 g20 and Its Application to PCR-Based Detection and Quantification Techniques. Applied and Environmental Microbiology **64:**2051-2060.

62. **Goldberg, S., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, and F. M. Lauro.** 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft

assemblies of marine microbial genomes. Proceedings of the National Academy of Sciences **103:**11240.

63. **Gotelli, N. J., and R. K. Colwell.** 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4:**379-391.

64. **Goyal, S. M., C. P. Gerba, and G. Bitton.** 1987. Phage ecology. Wiley, New York.

65. **Handelsman, J.** 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev **68:**669-85.

66. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol **5:**R245-9.

67. **Hankin, E. H.** 1896. L'action bactericide des eaux de la Jumna et du Gange sur le microe du cholera. Ann Inst Pasteur **10:**511-523.

68. **Hantula, J., A. Kurki, P. Vuoriranta, and D. H. Bamford.** 1991. Ecology of bacteriophages infecting activated sludge bacteria. Appl Environ Microbiol **57:**2147-51.

69. **Healy, F. G., R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam.** 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Applied Microbiology and Biotechnology **43:**667-674.

70. **Hewson, I., J. M. O'Neil, J. A. Fuhrman, and W. C. Dennison.** 2001. Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. Limnology and Oceanography **46:**1734-1746.

71. **Horne, R. W., and P. Wildy.** 1961. Symmetry in virus architecture. Virology **15:**348-73.

72. **Horne, R. W., and P. Wildy.** 1963. Virus Structure Revealed by Negative Staining. Adv Virus Res **10:**101-70.

73. **Huxley, H. E., and G. Zubay.** 1960. The structure of the protein shell of turnip yellow mosaic virus. J. Mol. Biol **2:**189-196.

74. **ICTV** 2007, posting date. International Committee on Taxonomy of Viruses [online] [Online.]

75. **Ivanovsky, D.** 1882. Concerning the mosaic disease of the tobacco plant. St. Petsb. Acad. Imp. Sci. Bul. **35:**67-70.

76. **Jiang, S. C., and J. H. Paul.** 1998. Gene Transfer by Transduction in the Marine Environment. Applied and Environmental Microbiology **64:**2780-2787.

77. **Jiang, S. C., and J. H. Paul.** 1994. Seasonal and diel abundance of viruses and occurrence of lysogeny/bacteriocinogeny in the marine environment. Marine ecology progress series. Oldendorf **104:**163-172.

78. **Jin, Y., and M. Flury.** 2002. FATE AND TRANSPORT OF VIRUSES IN POROUS MEDIA. ADVANCES IN AGRONOMY **77:**40-103.

79. **Khan, M. A., H. Satoh, H. Katayama, F. Kurisu, and T. Mino.** 2002. Bacteriophages isolated from activated sludge processes and their polyvalency. Water Research **36:**3364-3370.

80. **Khan, M. A., H. Satoh, T. Mino, H. Katayama, F. Kurisu, and T. Matsuo.** 2002. Bacteriophage-host interaction in the enhanced biological phosphate removing activated sludge system. Water Science & Technology **46:**39-43.

81. **Kirchman, D. L.** 1994. The uptake of inorganic nutrients by heterotrophic bacteria. Microbial Ecology **28:**255-271.

82. **Knight, C. A.** 1974. Molecular virology. McGraw Hill, New York.

83. **Lanning, S., and S. T. Williams.** 1982. Methods for the direct isolation and enumeration of actinophages in soil. J. Gen. Microbiol **128:**2063-2071.

84. **Lawrence, J. G., G. F. Hatfull, and R. W. Hendrix.** 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol **184:**4891-905.

85. **Leclerc, H., S. Edberg, V. Pierzo, and J. M. Delattre.** 2000. Bacteriophages as indicators of enteric viruses and public health risk in groundwaters. J Appl Microbiol **88:**5-21.

86. **Lederberg, E. M., and J. Lederberg.** 1953. Genetic Studies of Lysogenicity in Escherichia Coli. Genetics **38:**51-64.

87. **Lee, S. H., M. Onuki, H. Satoh, and T. Mino.** 2006. Isolation, characterization of bacteriophages specific to Microlunatus phosphovorus and their application for rapid host detection. Letters in Applied Microbiology **42:**259-264.

88. **Lee, S. H., H. Satoh, H. Katayama, and T. Mino.** 2004. Isolation, physiological characterization of bacteriophages from enhanced biological phosphorus removal activated sludge and their putative role. Journal of microbiology and biotechnology **14:**730-736.

89. **Lehnherr, H.** Bacteriophage P1. *In* R. Calendar (ed.), The Bacteriophages. Oxford University Press, New York, N.Y.

90. **Loveland, J. P., J. N. Ryan, G. L. Amy, and R. W. Harvey.** 1996. The reversibility of virus attachment to mineral surfaces. Colloids and Surfaces A: Physicochemical and Engineering Aspects **107:**205-221.

91. **Lwoff, A., and A. Gutmann.** 1950. Recherches sur un Bacilus megatherium lysogene. Ann Inst Pasteur **78**.

92. **Lwoff, A., T. F. Anderson, and F. Jacob.** 1959. Remarks on the characteristics of the infectious viral particle. Ann Inst Pasteur (Paris) **97:**281-9.

93. **Maranger, R., and D. F. Bird.** 1996. High concentrations of viruses in the sediments of Lac Gilbert, Québec. Microbial Ecology **31:**141-151.

94. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, and Z. Chen.** 2005. Genome

sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-380.

95. **Marsh, P., and E. M. H. Wellington.** 1994. Phage-host interactions in soil. FEMS Microbiology Ecology **15:**99-107.

96. **Mayar, A.** 1886. Concerning the mosaic disease of tobacco. Landivertschaftlichen Versuchs-Stationen **32:**451-567.

97. **McDaniel, L., and J. H. Paul.** 2005. Effect of Nutrient Addition and Environmental Factors on Prophage Induction in Natural Populations of Marine Synechococcus Species. Applied and Environmental Microbiology **71:**842-850.

98. **Mei, M. L., and R. Danovaro.** 2004. Virus production and life strategies in aquatic sediments. Limnology and Oceanography **49:**459-470.

99. **Meschke, J. S., and M. D. Sobsey.** 2003. Comparative reduction of Norwalk virus, poliovirus type 1, F+ RNA coliphage MS 2 and Escherichia coli in miniature soil columns. Health-related Water Microbiology **47:**85-90.

100. **Miller, R. V.** 2006. Marine phages, p. 534-544. *In* R. Calendar (ed.), The Bacteriophages, 2nd ed. Oxford University Press, New York, N.Y.

101. **Mirold, S., W. Rabsch, H. Tschäpe, and W. D. Hardt.** 2001. Transfer of the Salmonella type III effector sopE between unrelated phage families. Journal of Molecular Biology **312:**7-16.

102. **Monpoeho, S., A. Maul, B. Mignotte-Cadiergues, L. Schwartzbrod, S. Billaudel, and V. Ferre.** 2001. Best Viral Elution Method Available for

Quantification of Enteroviruses in Sludge by Both Cell Culture and Reverse Transcription-PCR. Applied and Environmental Microbiology **67:**2484-2488.

103. **Nei, M., and S. Kumar.** 2000. Molecular Evolution and Phylogenetics. Oxford University Press, USA.

104. **Nelson, D.** 2004. Phage taxonomy: we agree to disagree. J Bacteriol **186:**7029-31.

105. **Ogunseitan, O. A., G. S. Sayler, and R. V. Miller.** 1990. Dynamic interactions ofPseudomonas aeruginosa and bacteriophages in lake water. Microbial Ecology **19:**171-185.

106. **Otawa, K., S. H. Lee, A. Yamazoe, M. Onuki, H. Satoh, and T. Mino.** 2007. Abundance, diversity, and dynamics of viruses on microorganisms in activated sludge processes. Microb Ecol **53:**143-52.

107. **Pantastico-Caldas, M., K. E. Duncan, C. A. Istock, and J. A. Bell.** 1992. Population Dynamics of Bacteriophage and Bacillus Subtilis in Soil. Ecology **73:**1888-1902.

108. **Paul, J. H.** 1999. Microbial Gene Transfer: An Ecological Perspective. Journal of Molecular Microbiology and Biotechnology **1:**45-50.

109. **Paul, J. H., M. E. Frischer, and J. M. Thurmond.** 1991. Gene Transfer in Marine Water Column and Sediment Microcosms by Natural Plasmid Transformation. Applied and Environmental Microbiology **57:**1509-1515.

110. **Paul, J. H., M. B. Sullivan, A. M. Segall, and F. Rohwer.** 2002. Marine phage genomics. Comp Biochem Physiol B Biochem Mol Biol **133:**463-76.

111. **Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix, and G. F. Hatfull.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171-82.

112. **Pomeroy, L. R.** 1974. The ocean's food web, a changing paradigm. Bioscience **24:**499-504.

113. **Powelson, D. K., J. R. Simpson, and C. P. Gerba.** 1991. Effects of organic matter on virus transport in unsaturated flow. Applied and Environmental Microbiology **57:**2192-2196.

114. **Pride, D. T., T. M. Wassenaar, C. Ghose, and M. J. Blaser.** 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics **7**.

115. **Proux, C., D. van Sinderen, J. Suarez, P. Garcia, V. Ladero, G. F. Fitzgerald, F. Desiere, and H. Brussow.** 2002. The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. J Bacteriol **184:**6026-36.

116. **Pruzzo, C., and G. Satta.** 1988. Capsular antigenic variations by lysogenic conversion inKlebsiella pneumoniae: Relationship with virulence. Current Microbiology **16:**259-263.

117. **Puig, A., N. Queralt, J. Jofre, and R. Araujo.** 1999. Diversity of Bacteroides fragilis Strains in Their Capacity To Recover Phages from Human and Animal Wastes and from Fecally Polluted Wastewater. Applied and Environmental Microbiology **65:**1772-1776.

118. **Ravin, N. V.** N15: the linear plasmid prophage. *In* R. Calendar (ed.), The Bacteriophages. Oxford University Press, New York, N.Y.

119. **Reanney, D. C., and S. C. N. Marsh.** 1973. The ecology of viruses attacking Bacillus stearothermophilus in soil. Soil Biol. Biochem **5:**399-408.

120. **Roberts, J. W., Brodetsky A.M.** 1983. Lysogenic induction, p. 123-144. *In* R. W. Hendrix (ed.), Lambda II. Cold Spring Harbor Laboratroy, Cold Spring Harbor, N.Y.

121. **Rohwer, F.** 2003. Global phage diversity. Cell **113:**141.

122. **Rohwer, F., and R. Edwards.** 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol **184:**4529-35.

123. **Rohwer, F., A. Segall, G. Steward, V. Seguritan, M. Breitbart, F. Wolven, and F. Azam.** 2000. The complete genomic sequence of the marine phage Roseophage SI01 shares homology with nonmarine phages. Limnol. Oceanogr **45:**408-418.

124. **Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam.** 2001. Production of shotgun libraries using random amplification. Biotechniques **31:**108-12, 114-6, 118.

125. **Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam.** 2001. Production of Shotgun Libraries Using Random Amplification. BIOTECHNIQUES **31:**108-119.

126. **Romig, W. R., and A. M. Brodetsky.** 1961. ISOLATION AND PRELIMINARY CHARACTERIZATION OF BACTERIOPHAGES FOR BACILLUS SUBTILIS. Journal of Bacteriology **82:**135-141.

127. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, and C. Minor.** 2000. Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. Applied and Environmental Microbiology **66:**2541-2547.

128. **Schlesinger, M.** 1936. The Feulgen Reaction of the Bacteriophage Substance. Nature **138:**508.

129. **Schoenfeld, T., M. Patterson, P. M. Richardson, K. E. Wommack, M. Young, and D. Mead.** 2008. Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol **74:**4164-74.

130. **Short, C. M., and C. A. Suttle.** 2005. Nearly Identical Bacteriophage Structural Gene Sequences Are Widely Distributed in both Marine and Freshwater Environments. Applied and Environmental Microbiology **71:**480-486.

131. **Staley, J. T., and A. Konopka.** 1985. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annual Reviews in Microbiology **39:**321-346.

132. **Steele, H. L., and W. R. Streit.** 2005. Metagenomics: advances in ecology and biotechnology. FEMS Microbiol Lett **247:**105-11.

133. **Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong.** 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology **178:**591-599.

134. **Steward, G. F., J. L. Montiel, and F. Azam.** 2000. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnology and Oceanography **45:**1697-1706.

135. **Summers, W. C.** 1993. Cholera and plague in India: the bacteriophage inquiry of 1927-1936. Journal of the history of medicine and allied sciences **48:**275-301.

136. **Summers, W. C.** 2005. History of Phage Reserach and Phage Therapy, p. 3-17. *In* M. K. Waldor, Friedman, D. I, Adhya, S. L. (ed.), Phages their role in bacterial pathogenesis and biotechnology, 1 ed. ASM Press, Washington DC.

137. **Suttle, C. A.** 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol **5:**801-12.

138. **Suttle, C. A.** 1994. The significance of viruses to mortality in aquatic microbial communities. Microbial Ecology **28:**237-243.

139. **Tamura, K., J. Dudley, M. Nei, and S. Kumar.** 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Molecular Biology and Evolution **24:**1596.

140. **Tanji, Y., K. Mizoguchi, M. Yoichi, M. Morita, K. Hori, and H. Unno.** 2002. Fate of Coliphage in a Wastewater Treatment Process. Journal of Bioscience and Bioengineering **94:**172-174.

141. **Thingstad, T. F., M. Heldal, G. Bratbak, and I. Dundas.** 1993. Are viruses important partners in pelagic food webs? Trends in Ecology & Evolution **8:**209-212.

142. **Thingstad, T. F., and R. Lignell.** 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity, and carbon demand Aquatic microbial ecology **13:**19-27.

143. **Twort, F. W.** 1915. An investigation on the nature of ultramicroscopic viruses. Lancet **2:**1241-1243.

144. **Vaca-Pacheco, S.** 1999. The Clinically Isolated FIZ15 Bacteriophage Causes Lysogenic Conversion in Pseudomonas aeruginosa PAO1. Current Microbiology **38:**239-243.

145. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, and W. Nelson.** 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science **304:**66-74.

146. **Viruses, I. C. o. T. o.** 2008, posting date. International Committee on Taxonomy of Viruses. [Online.]

147. **Waldor, M. K., D. I. Friedman, and S. L. Adhya.** 2005. Phages : their role in bacterial pathogenesis and biotechnology. ASM Press, Washington, D.C.

148. **Waldor, M. K., and J. J. Mekalanos.** 1996. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science **272:**1910.

149. **Watanabe, K., and S. Takesue.** 1972. The requirement for calcium in infection with Lactobacillus phage. J Gen Virol **17:**19-30.

150. **Weitz, J. S., H. Hartman, and S. A. Levin.** 2005. Coevolutionary arms races between bacteria and bacteriophage. Proceedings of the National Academy of Sciences **102:**9535-9540.

151. **Whitman, W. B., D. C. Coleman, and W. J. Wiebe.** 1998. Prokaryotes: The unseen majority, p. 6578-6583, vol. 95. National Acad Sciences.

152. **Wilhelm, S. W., and C. A. Suttle.** 1999. Viruses and Nutrient Cycles in the Sea. BIOSCIENCE-WASHINGTON- **49:**781-788.

153. **Williamson, K. E., M. Radosevich, and K. E. Wommack.** 2005. Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol **71:**3119-25.

154. **Williamson, K. E., J. B. Schnitker, M. Radosevich, D. W. Smith, and K. E. Wommack.** 2008. Cultivation-Based Assessment of Lysogeny Among Soil Bacteria. Microb Ecol.

155. **Williamson, K. E., K. E. Wommack, and M. Radosevich.** 2003. Sampling natural viral communities from soil for culture-independent analyses. Appl Environ Microbiol **69:**6628-33.

156. **Withey, S., E. Cartmell, L. M. Avery, and T. Stephenson.** 2005. Bacteriophages—potential for application in wastewater treatment processes. Science of the Total Environment, The **339:**1-18.

157. **Wollman, E., and E. Wollman.** 1937. Les phases de bacteriophages (facteurs lysogenes). C. R. Soc. Biol Paris **124:**931-934.

158. **Wommack, K. E., J. Bhavsar, and J. Ravel.** 2008. Metagenomics: read length matters. Appl Environ Microbiol **74:**1453-63.

159. **Wommack, K. E., and R. R. Colwell.** 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev **64:**69-114.

160. **Wommack, K. E., R. T. Hill, M. Kessel, E. Russek-Cohen, and R. R. Colwell.** 1992. Distribution of viruses in the Chesapeake Bay. Appl Environ Microbiol **58:**2965-70.

161. **Zhang, K., and K. Farahbakhsh.** 2007. Removal of native coliphages and coliform bacteria from municipal wastewater by various wastewater treatment processes: implications to water reuse. Water Res **41:**2816-24.