6-26-2017

An Approach to Identify Mycobacteriophage Diversity Prior to DNA Sequencing

Charles Gregory
Western Kentucky University, charles.gregory940@topper.wku.edu

AN APPROACH TO IDENTIFY MYCOBACTERIOPHAGE DIVERSITY PRIOR TO DNA SEQUENCING

A Thesis Project

Presented in Partial Fulfillment of the Requirements for

the Degree Bachelor of Science with

Honors College Graduate Distinction at Western Kentucky University

By

Charles T. Gregory

May 2017

* * * * *

CE/T Committee:

Professor Claire Rinehart

Professor Rodney King

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Rinehart for his countless hours of guidance, the Bioinformatics and Information Science Center, and the High-Performance Computing Center here at WKU for supporting this project. I would also like to acknowledge Dex Wood and Jason Smith for their extensive programming advice throughout this project.

ABSTRACT

Over 6,869 Mycobacteriophages have been isolated and purified. Of these, 1,367 genomes have been sequenced at the DNA level and more are added each year through the SEA-PHAGES program. Sequenced mycobacteriophages are grouped into clusters based on a 50% or greater nucleotide identity. The number and breadth of these clusters represents the diversity present in the environment. Each year, as new phages are discovered by students in the SEA-PHAGES program, the question arises, "Which isolates should we sequence?" In order to sequence phages that represent the greatest possible diversity, and thus broaden under-represented clusters and identify new singletons, we need a rapid way to identify phage cluster membership or singleton status before selection for DNA sequencing. One approach is to identify unique short nucleotide sequences that are common across a cluster. Unique sequences could then be used as primers or probes to assign membership to a cluster or potential singleton group. A computer program called PhageUniqueSeq was written in *Go* language to identify all the oligonucleotides that are common to all members of a cluster but unique between clusters. The program generated millions of unique sequences that can be used as probes or in Polymerase Chain Reactions to determine sub-cluster assignment. Unique sequences will help us to target underrepresented phages for sequence analysis.

VITA

EDUCATION

Western Kentucky University, Bowling Green, KY — May 2017

B.S. in Biochemistry – Honors College Graduate

Graves County High School, Mayfield, KY — May 2013

PROFESSIONAL EXPERIENCE

Bioinformatics and Information Science Center, WKU — June 2016-

Research Assistant — Present

AWARDS & HONORS

Magna Cum Laude, WKU, May 2017

Ogden Research Scholarship, August 2015

PRESENTATIONS

Gregory, C., & Rinehart, C.A. (2015, March). An approach to identify mycobacteriophage diversity prior to DNA sequencing. Poster presented at the WKU Student Research Conference. Bowling Green, KY.

Gregory, C., & Rinehart, C.A. (2015, November). An approach to identify mycobacteriophage diversity prior to DNA sequencing. Poster presented at the 101st Kentucky Annual of Sciences Annual Meeting. Highland Heights, KY.

Gregory, C., & Rinehart, C.A. (2016, April). An approach to identify mycobacteriophage diversity prior to DNA sequencing. Poster presented at the KBRIN Bioinformatics Summit. Cadiz, KY.

CONTENTS

List of Figures

List of Tables

CHAPTER ONE:

INTRODUCTION

Viruses that infect bacteria are called bacteriophages or phages for short. Bacteriophages are the most numerous evolutionary particles on the planet (Wommack & Colwell, 2000). Bacteriophages are an important area of molecular biological research as many bacteria now have developed antibiotic resistance. Bacteriophages can act as new self-replicating target-specific antibiotics as they usually infect only one strain of bacteria. Their specificity makes them useful biocontrol agents to eliminate pathogens on food and attack antibiotic-resistant bacteria.

Bacteriophages that infect the soil bacteria, *Mycobacterium*, are called mycobacteriophages. As the genomes from mycobacteriophages have been sequenced, they have been grouped into clusters based upon a greater than fifty-percent identity in their DNA sequence to other phages within a cluster (Pope et al., 2011). There have been over 1,367 mycobacteriophages sequenced to date and they have been classified into 27 clusters (A-AA), with several additional mycobacteriophages existing as singletons or single members. Several of these clusters have been further divided into sub-clusters (Phagesdb).

Western Kentucky University (WKU) is a part of the Howard Hughes Medical Institute's SEA-PHAGES (Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science) program. This program provides undergraduate

students with authentic research experience early in their career, exploring the vast diversity of bacteriophages, and finding novel bacteriophages and bacteriophage-bacterium interactions. There are over 150 institutions exploring phage diversity within the Science Education Alliance. Exploring bacteriophage diversity is an important step towards furthering our understanding of the possibility of developing bacteriophages as alternative antibiotics.

In the Genome Discovery and Exploration course (GDEP is the SEA-PHAGES course at WKU) students isolate and purify mycobacteriophages. This course allows students to learn microbiology lab skills, explore their own research questions, and participate in individual and group research early on in their career. The course also serves to expand known bacteriophage diversity and phylogeny with novel bacteriophages being discovered by each student. During the second semester, students identify genes and annotate their location and function in the genomes of those bacteriophages that were sequenced between semesters. The annotations are reviewed for quality and submitted to GenBank with contributing students as authors.

Each year, as new phages are discovered by students, the question arises, "Which phage isolates should we sequence that will expand the diversity of the phage clusters and give us more members within underrepresented clusters?" Since HHMI usually provides funding to sequence only two bacteriophage isolates from a class of 20 students, it is of interest to choose phage isolates that will contribute the most to the diversity within our

current databases. Thus, there is a need to precisely classify new isolates into their cluster/subcluster group or into a new singleton classifications before selecting phages for DNA sequencing. As whole-genome sequencing gets cheaper and universities can choose to fund the sequencing of more bacteriophages, the sampling of diversity will certainly increase. However, there will still be a need to target novel phages or those that are underrepresented in the databases

CHAPTER TWO:

BACKGROUND

Bacteriophages

Bacteriophages are viruses that infect and replicate within bacteria. These viruses were first noticed in 1915 by Dr. Frederick Twort. He described these newfound viruses as nonpathogenic and filterable due to their extremely small size (Twort, F.W., LOND., L.R.C.P. , M.R.C.S., 1915). His results showed that these viruses, obtained from the environment, would turn micrococci cultures glassy and transparent. When these areas were viewed under a microscope, they showed nothing but minute granules. If a glassy section from one plate was transferred to a pure micrococcus culture, the effect was replicated, sometimes killing all the micrococcus. A solution containing the viruses could retain its destructive powers longer than six months. Even a one in a million dilution of a filtered phage lysate would render a surface unsuitable for growing micrococcus. Only after being heated to 60°C would the viruses become inert. Though Dr. Twort observed the effects of these viruses through their effects on micrococcus, the viruses' structure could not be viewed by available microscopes at the time. He did state that he could not discern whether they were viruses or ultra-microscopic bacteria that only grew on living material.

After the invention of the electron microscope, more was discovered about these elusive viruses (Luria, S.E., & Anderson, T.F., 1942). Researchers from Columbia

University first viewed bacteriophages and characterized their infection of bacteria. They described these viruses as possessing an “extremely thin tail” and head of extremely dense internal structure. It was found that these viruses were readily absorbed by susceptible bacteria but not by non-susceptible bacteria, and upon lysis of the bacteria, the virus particles were released.

Applications of Bacteriophages

Bacteriophages have very useful applications as alternative antibiotics. Researchers from the Polish Academy of Sciences have fully summarized their effectiveness as antibiotics (Weber-Dąbrowska et al., 2016). They stated that since bacteriophages only target certain bacteria, they are good candidates for the growing issue of drug-resistant bacterium such as Methicillin-resistant *Staphylococcus aureus*, or MRSA. The researchers also found that bacteriophages have advantages over conventional antibiotics. For example, they multiply at the site of bacterial infection and only target the bacteria to which they can attach and infect. Due to this, they can be used to treat antibiotic-resistant strains without broad-spectrum selection for drug resistance in other species. Due to their abundance, bacteriophages can often be derived from most sources where their target bacteria exist. This could allow researchers to create new bacteriophage antibiotics for most bacteria, as in most cases a bacteriophage can be found that targets the host bacterium in question.

Bacteriophages are also useful to the food industry. Researchers from the University College Cork in Ireland and the Max Rubner Institute in Germany have employed bacteriophages to curb the spoilage of beer during commercial brewing (Deasy et al., 2011). Certain species of bacteria in the genus Lactobacillus can survive the natural disinfectants present in the brewing process, such as hops and ethanol, leading to the spoilage of beer (Deasy et al., 2011). Spoilage is caused by the lactic acid and acetic acids that the bacteria excrete. The researchers isolated a bacteriophage that was infectious to *Lactobacillus brevis* and assessed its capacity to prevent spoilage by monitoring the pH levels, which decrease as acids are secreted. Though they were using only a single bacteriophage, the results showed that the bacteriophage could control the bacterium's levels even when the beer's level of *L. brevis* contamination was quite high. Thus, the pH drop due to the bacterium was much less, 0.05pH, when contaminated beer was bacteriophage-treated. The FDA has also recognized bacteriophage applications as an antibiotic as well and has approved the use of bacteriophage P100 for the treatment of cheese as an antimicrobial against the pathogen *Listeria monocytogenes* (Food and Drug Administration).

Mycobacteriophages

Interest in mycobacteriophages gained speed in the 1990s as cases of opportunistic infections due to *Mycobacterium tuberculosis* surged during the Acquired Immunodeficiency Virus (AIDS) epidemic (Jacobs, 1992). Jacobs and other researchers

at the Albert Einstein College of Medicine studied mycobacteriophages and their uses as genetic vectors (Jacobs, W. R., Jr., Tuckman, M., & Bloom, B. R., 1987). They worked with mycobacteriophage DNA to create amplicons that would replicate as plasmids inside *E. coli* or as bacteriophages inside mycobacterium. This allowed for controlled gene transfer to occur between *M. smegmatis*, *M. bovis*, and *E. coli* for the first time. The work by Dr. Jacobs with mycobacteriophages was extended by research conducted at the University of Pittsburg by Dr. Graham F. Hatfull. Dr. Hatfull's research focused on bacteriophages that infected *Mycobacterium smegmatis*. *M. smegmatis* has a doubling time of 3 hours as opposed to a doubling time of 24 hours for *M. tuberculosis*, and *M. smegmatis* is not pathogenic in humans (Hatfull, 2012). Therefore, *M. smegmatis* is less expensive and less dangerous to work with, making it a prime host for mycobacteriophage research.

SEAPHAGES and GDEP

Dr. Hatfull saw the assessment of mycobacteriophage diversity as a unique opportunity to introduce high school and undergraduate students to microbiological research (Hatfull et al., 2006). Students could isolate a bacteriophage from the environmental samples and assist with genome annotation. Due to the sheer abundance of bacteriophages and the diversity contained therein, each student who isolated a bacteriophage would almost certainly isolate a novel, undiscovered bacteriophage (Pope et. al., 2015). This, combined with the prospect of finding new genes within each

bacteriophage, lets students experience scientific discovery. The small genome size of mycobacteriophages allows genome annotation to be managed by individual students. This independence and ownership of the bacteriophage can fuel their motivation for discovery. With support from the Howard Hughes Medical Institute's Science Education Alliance (SEA), Dr. Hatfull's work expanded into the Phage Hunters Advancing Genomics and Evolutionary Sciences (PHAGES) program that helps educators at universities across the globe advance science education and discovery. To date, the SEA-PHAGES program has isolated over 10,162 bacteriophages and 1,367 mycobacteriophages have been sequenced. Many of these genomes have been published in GenBank with students listed as contributing authors. The Genome Discovery and Exploration Program (GDEP) here at WKU is officially part of the larger SEA-PHAGES program.

Bacteriophage Diversity

Mycobacteriophages discovered through SEA-PHAGES can be described by clusters and sub-clusters. Mycobacteriophages are assigned to clusters based on a greater than 50% average nucleotide identity to one or more members of a current cluster (Hatfull, G. F., 2012). If the bacteriophage cannot meet a 50% average nucleotide identity with any known bacteriophage, it is designated a singleton. If it ever shares 50% nucleotide identity with another bacteriophage, it forms a new cluster with the other bacteriophage. It can sometimes be difficult to create local associations within clusters

using nucleotide identity. Sometimes bacteriophage genomes have a low level of similarity to one another over a significant portion of the genome. Sometimes they have a high level of similarity to one another over a small portion of their genomes. In these situations, DNA dotplots are useful to identify local associations within a cluster. Both scenarios usually result in lower than 50% nucleotide identity. Both scenarios also show different levels of diversity and divergence among bacteriophages within a cluster. The first scenario, broad low levels of similarity, describes two bacteriophages that long ago diverged evolutionarily but have residual similarity. The second scenario, high similarity over a short region of the genome, describes bacteriophages that are rather unrelated except for regions of horizontal gene transfer, in which both bacteriophages once infected the same host and had the opportunity to exchange genetic information. Associations within a cluster can lead to sub-clusters. Bacteriophages within a similar cluster but different sub-clusters generally share many genes but have overall lower nucleotide similarity.

Bacteriophage genes are also often associated, by amino acid sequence similarities, into groups of genes called “phamilies”, or gene families (Cresawn, S.G. et al., 2011). When bacteriophage genes are annotated and confirmed, their sequences are compared to current phams, and added to the pham if they present at least 32.5% amino acid identity to that pham’s other gene members. This analysis is done by a program called Phamerator (Cresawn, S.G. et al., 2011). Phams provide the scientific community with a model for horizontal gene transfer among viruses and by extension the bacteria

they infect. Genes within a pham often come from closely related phages, however genes within the pham may also come from more distantly related phages or the host itself.

Methods for Identification of Clusters before DNA Sequencing

To identify the greatest possible bacteriophage diversity, it is imperative to discover a bacteriophage's cluster membership before sequencing. To determine bacteriophage cluster membership before sequencing, an experimental approach is needed.

Currently, bacteriophage plaque morphology is used alongside restriction enzyme digests to predict cluster. Plaque morphology is a characteristic of bacteriophages seen when they are grown on a bacterial lawn. A plaque is an area on the dish where the bacteria have died from phage infection. These plaques are often characteristic of the bacteriophage that created them. They can differ in size, shape, turbidity (translucence), and may even have extra concentric rings around the main plaque. Bacteriophages within the same cluster often have similar plaque morphologies. Researchers from North Carolina Central University used these methods to predict the cluster of bacteriophages found within the Neuse River Basin in Durham, NC (Leslie et al., 2014). They showed that the bacteriophages they isolated, which were from the A cluster, all had the same HindIII restriction site locations. HindIII is a restriction enzyme and cuts the DNA at specific sites (A^AGCTT). Their work shows that one can use a combination of restriction enzymes and plaque morphology to predict cluster membership.

Another approach to bacteriophage cluster prediction prior to DNA sequencing has been addressed previously by Chris R. Gissendanner et al. (2014). They developed a web-based program that assesses probable cluster placement from restriction enzyme patterns. This program allows users to enter fragment patterns, from a restriction enzyme assay of bacteriophage DNA, to the Phage Enzyme Tool and get a prediction on cluster assignment of the bacteriophage. This tool provides one pathway for cluster prediction.

During a normal GDEP course, bacteriophage DNA is isolated and can be utilized to identify a bacteriophage's cluster, if unique sequences can be identified that are characteristic of a cluster. We can find sequences that are unique to a cluster by defining all possible subsequences of a genome, that is N nucleotides long, and then finding a set of candidates common to all phage sets within a cluster. To find sequences unique to a cluster, the set of common candidates for a cluster are compared to all possible phages outside the cluster. Only those common set candidates, that are not found in any other cluster, are considered to unique sequences for the cluster.

DNA Probes

One approach to using unique DNA sequences to identify bacteriophage cluster membership is to use the unique sequences as probes. DNA probes are DNA sequences that are tagged in some way (e.g. a fluorescent tag). When the probes bind to DNA and are irradiated, the tag will fluoresce and indicate cluster membership specific to the unique sequence probe.

Figure 1 demonstrates how DNA probes can be applied to bacteriophage screening. Early in bacteriophage isolation, many plaques are present on a host culture and often represent many different bacteriophages. The plaque lift binds bacteriophage DNA from individual plaques onto a filter. The DNA on the filter is then denatured and the single-stranded DNA can bind to the labeled DNA probes. The bacteriophage plaques can be screened for certain clusters by performing a plaque lift and then hybridizing a cluster-specific probe to the plaques. By using a mixture of probes from different underrepresented clusters, screening for several rare phages could be accomplished at once.

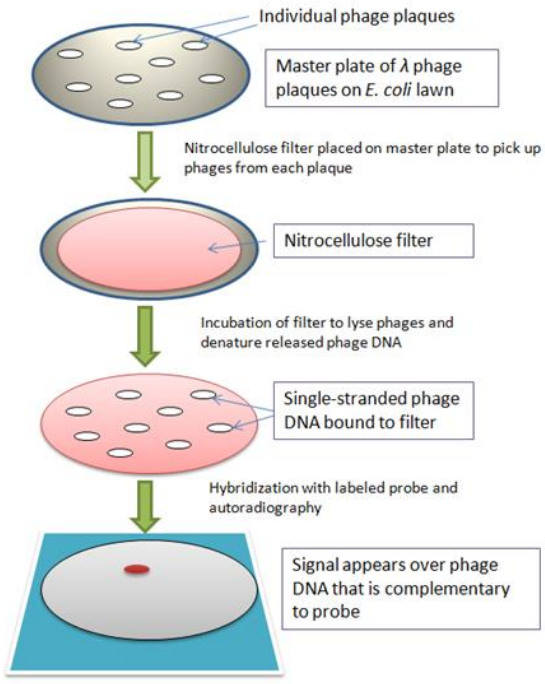

Figure 1: Bacteriophage Plaque Lift. A bacteriophage plaque lift can be used in conjunction with DNA probes to indicate cluster very early in the bacteriophage selection process (NPTEL).

Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) is another method that can be used to discriminate between clusters by using unique sequences in the design of PCR primer pairs. PCR is a technique in which one can amplify a specific section of DNA from a template.

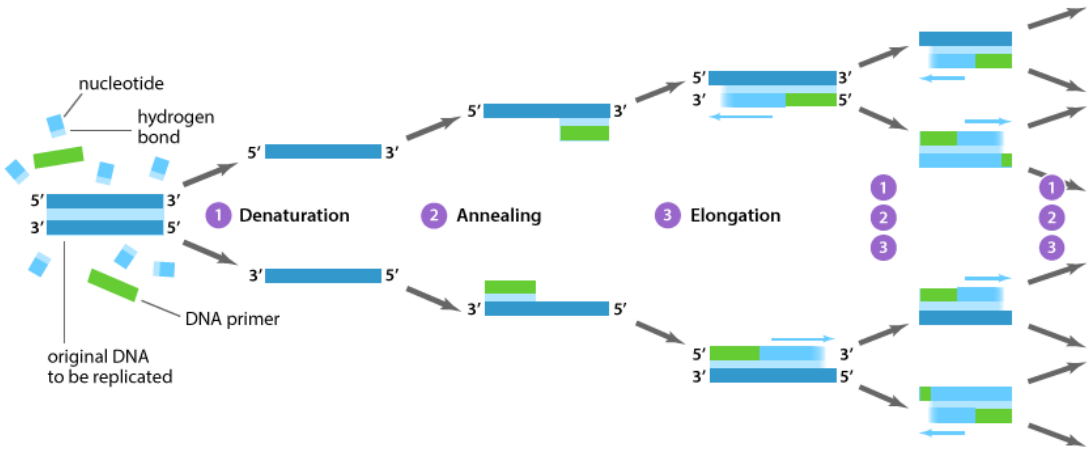

Figure 2: Polymerase Chain Reaction. PCR involves the exponential amplification of a target sequence using DNA primers as guides (ABMGood).

Short DNA sequences called primers are used to flank the specific region to be amplified. The primers are complementary to small sections of the template and can bind to the template DNA when it is separated into individual strands. The region between the

bound primers on the template DNA is the target sequence to be amplified. To begin the reaction, the mixture of primers and template DNA is heated so that the DNA is melted into individual strands. The temperature is then reduced so that the excess of primers have an opportunity to bind to their complementary targets on the DNA. Once the primers are bound, the temperature is raised slightly and a DNA polymerase in the solution can now extend the primer sequence through the region between the two primers. This generates a double-stranded product. When it is melted again, the product can bind to complementary primers and be amplified as the temperature cycling continues. Repeating the temperature cycling leads to the exponential amplification of the product DNA between the primer pairs.

Since these primers would be unique to a certain cluster, they serve as the test discriminator and identify the bacteriophage's cluster. A PCR, however, needs two primers. So, two unique sequences must be used as primers, but not just any two unique sequences. Randomly choosing two unique sequences as a primer pair could create false negatives due to the following: the primers binding being too far from or too close to one another (predicted product length), the primers binding to one another (complementarity), the primers binding to themselves (self-complementarity), or too high of a melting temperature difference between primers. Random primers might also produce multiple bands if they have multiple binding sites on a genome template. Unique primers must be paired, based on their projected product length, their complementarity, their self-

complementarity, uniqueness of their binding sites, and their melting temperature difference.

It is necessary to view the length of the PCR products in order to confirm that the product length matches the predicted distance between the unique primer pairs. If there are no products, or there are multiple products, or the product is the incorrect size, then it can be concluded that the target does not belong to the cluster that the primer pair was designed to identify. To view the size of the amplified DNA fragments, agarose gel electrophoresis is used.

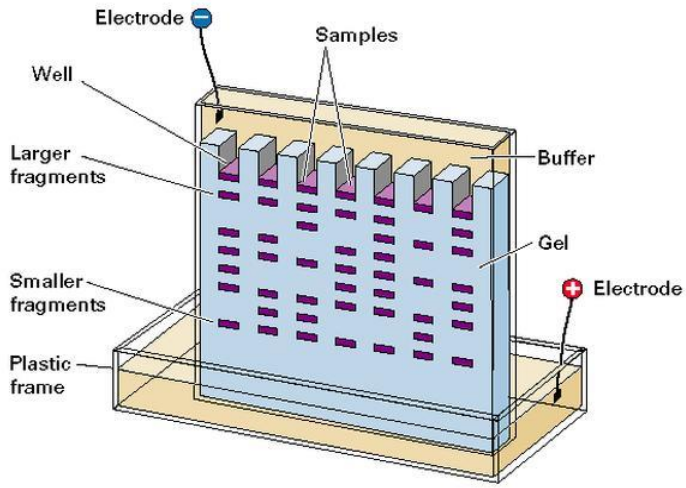

Figure 3: Gel Electrophoresis. Gel Electrophoresis can sort biological molecules by size. This can be used to identify if a PCR product is the proper size, compared to a standard (RegentsGeneticTechnology).

Gels that are made of agarose can separate DNA fragments based on their size. The resolution of the fragments is dependent upon the agarose concentration in the gel. Due to

the negatively charged phosphate backbone, DNA molecules can be pulled through an agarose gel via an electric current towards the positive electrode. Due to the restrictive nature of the "pores" in the agarose gel, smaller molecules are pulled farther than larger molecules and thus DNA fragments of differing lengths can be resolved from each other.

If the projected product size from PCR is smaller than 100 base pairs (bp), it may not be viewable on an electrophoresis gel, and if it is larger than 10,000 bp, it may not be possible for the DNA Polymerase in the PCR to fully amplify it. Therefore, primer pairs must be selected with binding site locations that differ between 100 and 10,000 bp. Fragment resolution is also dependent on the gel's agarose concentration. A 1% gel can resolve fragments from 500bp to 2000bp. A 2% agarose gel, however, can resolve fragments smaller than 500bp, but is not ideal for resolving fragments near 2000bp. Therefore, selecting unique sequence locations within 500 to 2000 bp of each other would allow the resolution of PCR fragments and the estimation of the size, relative to a standard DNA with known lengths, on a 1% agarose gel.

When performing gel electrophoresis, knowing the projected product size is important for determining a successful or a failed PCR. If primers bind to off-target sites, a product will be produced; however, it will often be of a different size than the intended target. Knowing projected product sizes allows one to identify such false-positives. Projected product size for the primer matches will not be a number, but a range of values. Though the unique sequences are conserved between bacteriophages of the same cluster,

the product lengths between two unique sequences is not necessarily conserved from bacteriophage to bacteriophage within a cluster. It is important to choose primer pairs that have a low product size variability from bacteriophage to bacteriophage, within a cluster, to lessen the chance of a false positive falling within the projected product size range. Selected primer pairs from each cluster should be tested for all of the bacteriophages within the cluster to ensure that product sizes are falling within a narrow expected range.

qPCR

Another cluster identification approach that uses the unique sequences in conjunction with PCR, is real-time quantitative PCR (qPCR) with TaqMan probes (ThermoFisher-TaqMan). qPCR quantitates the PCR amplified product using a fluorescent signal from a DNA probe, such as a TaqMan probe (ThermoFisher). The TaqMan probe is a unique sequence specific to the amplified region that contains a fluorophore and a quencher attached to opposite ends of the DNA probe. The quencher ensures that the fluorophore doesn't fluoresce unless separated from the quencher.

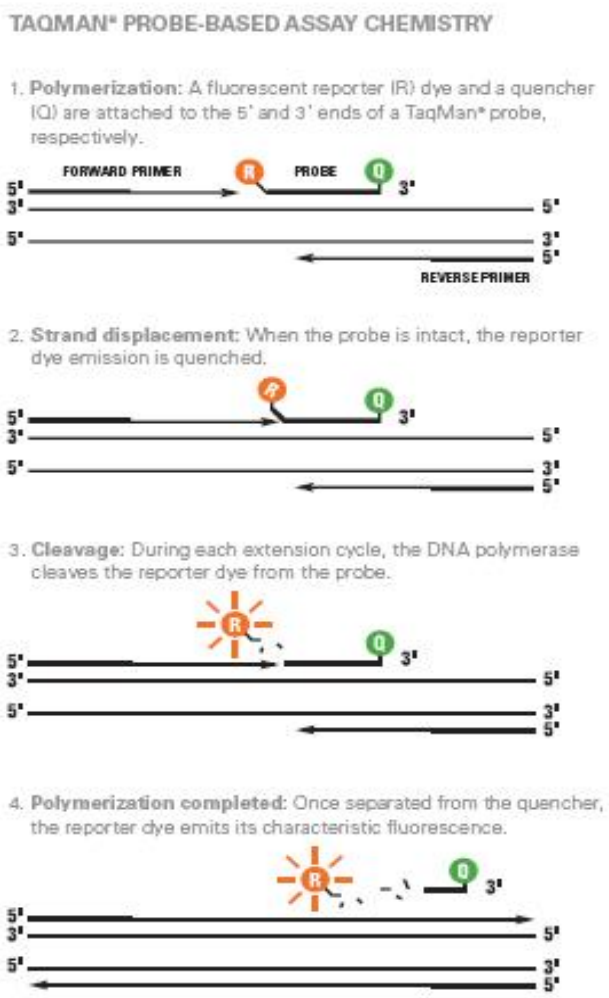

Figure 4: TaqMan q-PCR. TaqMan probes can provide an indication of cluster membership even before the PCR is completed without the need for an electrophoresis gel (BioSynthesis).

During qPCR, in the annealing phase, a TaqMan probe will bind along with the normal PCR primers somewhere between the primers in the product sequence. As extension occurs, the probe is cleaved and the fluorophore is released from the probe that also harbors the quencher. Once cleaved, fluoresces from the fluorophore is no longer

quenched and can be detected by the qPCR machine, thus indicating the presence and quantity of the product from the intensity of the fluorescence. By using unique cluster sequences for both primers and the TaqMan probe, it would be possible to identify a bacteriophage cluster without the need for gel electrophoresis. The added unique TaqMan sequence would also provide more specificity than just using two unique sequences for normal PCR primers or the one sequence for the probes in a plaque lift. However, the addition of a fluorophore and a quencher makes Taq-Man probes more expensive.

Using one of the three unique sequence approaches avoids issues with the Phage Enzyme Tool and using plaque morphologies or restriction enzymes. PCR, qPCR, and the use of probes can all use low quantities of DNA in a mixture and can show a strong specific signal from a mixture of DNA targets. Probes or PCR could be used on bacteriophage plaque samples without even the need for a DNA isolation step due to the free-floating bacteriophage DNA within a plaque. An electrophoretic gel would then be used to resolve any PCR products, and the lengths could be used to identify cluster membership or contamination. Unique sequences created at the sub-cluster level, can be more precise than restriction enzyme and plaque methods. Any one of these experimental designs that use unique sequences could provide WKU's GDEP program and the Science Education Alliance with tools to aid in their decision process for selecting new mycobacteriophages for sequencing. Using plaque lifts, PCR, or q-PCR, universities could select for bacteriophages from smaller, low-diversity clusters before sequencing bacteriophage genomic DNA. With plaque lifts, selection could begin even prior to phage

purification. This should be extremely useful for identifying bacteriophage diversity in under-represented bacteriophage clusters.

CHAPTER THREE:

PROGRAM DESIGN AND DEVELOPMENT

PhageUniqueSeq is a computer program that was developed to identify sequences that are unique to a single sub-cluster but are common to all phages in the sub-cluster. PhageUniqueSeq works at the sub-cluster level because some clusters are so large that it may be hard to identify sequences common to all bacteriophages within the cluster. Also, some of the subclusters are only sparsely represented and it is desirable to find more members. The program first downloads JavaScript object notation (JSON) data for all sequenced bacteriophages provided by phagesdb.org using their application program interface (API). JSON is a standard format to provide data through the JavaScript programming language from a web server. Phagesdb's API is a simple interface to present server data. This JSON data was used to parse out each bacteriophage genome with its associated name, sub-cluster, and host strain genus. For each bacterial strain, all genomes were split into all possible sub-sequences of a determined size, usually between 18 and 25 bp long. These sub-sequences were encoded from character strings, which are typically 40 bytes long for a 20bp sequence, into 4 byte (64-bit) long integers. This allowed for a 10-fold reduction in memory consumption per sub-sequence string. Without this step, the computation of these sequences became difficult even for high-end servers. The sub-sequences were then screened to determine their sub-cluster, bacteriophage, and strain membership. If for a specific host strain, a sub-sequence was found with

membership in only one sub-cluster and it had membership in all bacteriophages within that sub-cluster, then it was considered common to every phage within the sub-cluster and unique to that sub-cluster. These sub-sequences are referred to as "unique sequences". These results were then saved into a database managed by the program.

To provide a method of testing the unique sequences for compatibility to work in PCR, PhageUniqueSeq paired the unique sequences together and the pairs were filtered. The program's filtering criteria included complementarity, melting temperature, GC content, product length limits, off-site targets, and other properties that help produce good PCR results. The specific criteria in the program limited product length from 500bp to 2000bp, allowed no off-site targets, no self-annealing primers, and limited the maximum melting temperature difference between each primer to 5° C. The Wallace formula for determining primer melting temperature was chosen for speed and simplicity (Wallace et al., 1979). Complementarity and GC content information was recorded but not used for selection of primer pairs. This was done to leave more choices of primer pair candidates for the user. The pairs were generated for each sub-cluster by finding the locations of the unique primers on the forward and reverse strands for each bacteriophage. The locations were compared and a positive pair status was assigned if the primer on the reverse strand was within the accepted product size range. The product size range of 500bp and 2000bp was decided to be ideal for the polymerase to generate a product and for resolving PCR products of this size on a 1% agarose electrophoresis gel.

The PhageUniqueSeq program was first written in Python using the Biopython package and was run on a Dell Inspiron laptop. High system memory (RAM) usage and long program runtimes quickly became an issue. Therefore, the project was rewritten in Java using the BioJava and QuestDB packages. To further remove the RAM usage barrier, the project's testing machine became an experimental computing node at WKU's High Performance Computing Center. This machine had 96 GB of RAM as compared to the 8GB of RAM on the laptop. Using the QuestDB package, PhageUniqueSeq could generate millions of unique primers and primer-primer matches. QuestDB is useful in that it is a database written in Java and it is faster than other Java databases such as HyperSQL and H2, allowing for writing rates at 2 million rows a second and reading rates at 15 million rows a second. QuestDB achieved this with very low RAM usage, allowing more RAM to be devoted to PhageUniqueSeq's algorithms (QuestDB.org). The versatility of the database allowed PhageUniqueSeq to save all the generated primers in a relatively short amount of time and made them accessible to those familiar with Standard Query Language (SQL), a common standard to accessing databases. The primer matching algorithm was also developed in Java using the QuestDB database. However, RAM usage and runtimes were still higher than ideal, and using the project from WKU's HPCC was not practical for public access. Therefore, the project was rewritten once more in Google's Go language which offers faster runtimes and more efficient memory usage than Java or Python. This allowed PhageUniqueSeq to once again be run on a laptop or normal server. Though QuestDB was not compatible with the Go language, writing the

data to comma delimited files proved to be just as fast as QuestDB. Though this sacrificed the SQL accessibility of QuestDB, comma delimited file parsers are present in most programming languages and comma delimited files can be read by programs like Microsoft Excel.

CHAPTER FOUR:

PROGRAM TESTING ALGORITHM

The PhageUniqueSeq program's algorithms have undergone extensive testing to ensure that 1) the primer sequences for each sub-cluster truly are unique to their sub-cluster and are present in every bacteriophage within their sub-cluster, and 2) the primer pairs are indeed found within those bacteriophages and would produce fragments of expected size. This testing was necessary since the program's algorithms were designed for speed and efficiency, not simplicity. When writing complex algorithms, it can be easy to make mistakes. Their correctness was ensured by testing algorithms that were simple but were not optimized and required much more time and memory resources. The testing algorithms' simplicity was necessary to ensure there were no errors when testing.

To test the uniqueness of the primer sequences, each genome sequence and its reverse complement were searched for the location of the primer. If a single location was not found in every bacteriophage within the sequence's designated sub-cluster, then the algorithm was considered to be incorrect and was revised. Next, the testing algorithm searched for the location of each primer in the forward sequence and the reverse complement in every other bacteriophage outside the cluster but within the same host strain. If a location was present in any other sub-cluster, then the algorithm was considered to be incorrect and revisions to the program were made. To test the correctness of the matched primers for use in PCR, the primer pair's difference in

location was tested to ensure that they fell within 500 to 2000 bp of each other and generated the same product sizes as they did when they were generated via the primer matching algorithm. These testing algorithms were run at major milestones during development to ensure that following major changes, no mistakes were made in the core algorithms.

CHAPTER FIVE:

RESULTS

Currently, the program generates about 12 million unique sequences of base pair sizes 18 to 25 for all strains and sub-clusters of bacteriophage found on phagesdb.org, but that does not mean that all sub-clusters have unique sequences. Table 1 shows a summary of the number of unique sequences and the number of phage members found in each subcluster. Subcluster A2 was found to have no unique sequences but has a high number of members (73). To check the correlation of the number of unique sequences, in a subcluster, vs the number of phages in the subcluster, Figure 5 was generated. It indeed shows a decrease in the number of unique sequences as the number of phage members increase within a sub-cluster, but there was a wide range of variability even within a cluster (see K1-K6 and A1-A18 in Table 1).

After the primer-matching algorithm was completed, selected primer pairs from the A1, A4, A6, and K5 sub-clusters (shown in table 2) were tested on the known bacteriophages Badger (A4), TheloniusMonk (A1), Achebe (A4), Gruunaga (A6), AlleyCat (K5), and Ruin (A4) via PCR and gel electrophoresis. These bacteriophages were chosen since their DNA was readily available at the time. The results of the gel are shown in Figure 6.

Cluster	Phage Count	Unique Sequence Count
A1	124	716
A10	11	1867
A11	12	22555
A12	3	1512
A13	1	45751
A14	1	40419
A16	1	42525
A17	1	33311
A18	1	45424
A2	73	0
A3	81	82
A4	92	1088
A5	28	1177
A6	24	2424
A7	3	788
A8	8	14928
A9	17	193
AA	2	133525
B1	162	16096
B2	24	31212
B3	22	27705
B4	12	650
B5	7	2899
B6	5	14933
B7	1	63732
C1	93	14522
C2	2	30501
Cuke	1	68683
D1	14	29641
D2	1	65635
Dori	1	56324
DS6A	1	59301
E	83	22324
F1	118	3
F2	6	10715
F3	1	13545
F4	1	28477
F5	1	24518
G1	35	2280
G2	3	5119
G3	2	16296
G4	1	39722
H1	4	1296
H2	1	70364
I1	3	13635
I2	2	10560
J	30	1053
K1	55	9
K2	8	9158
K3	6	3663
K4	8	28037
K5	11	899
K6	11	7
L1	8	39248
L2	18	6599
L3	10	14100
L4	1	69325
M1	5	59996
M2	4	11111
M3	1	73764
MooMoo	1	44681
Muddy	1	47935
N	23	1045
O	8	38512
P1	22	91
P2	1	33223
Q	7	47943
R	5	40899
S	7	43569
Sparky	1	51258
T	5	4660
U	2	52945
V	3	59662
W	3	33881
X	2	71212
Y	2	1908
Z	2	18784

Table 1: 18bp Mycobacteriophage Cluster Summary. A summary of the number of phages per sub-cluster and the number of unique 18bp sequence per cluster for the mycobacteriophages.

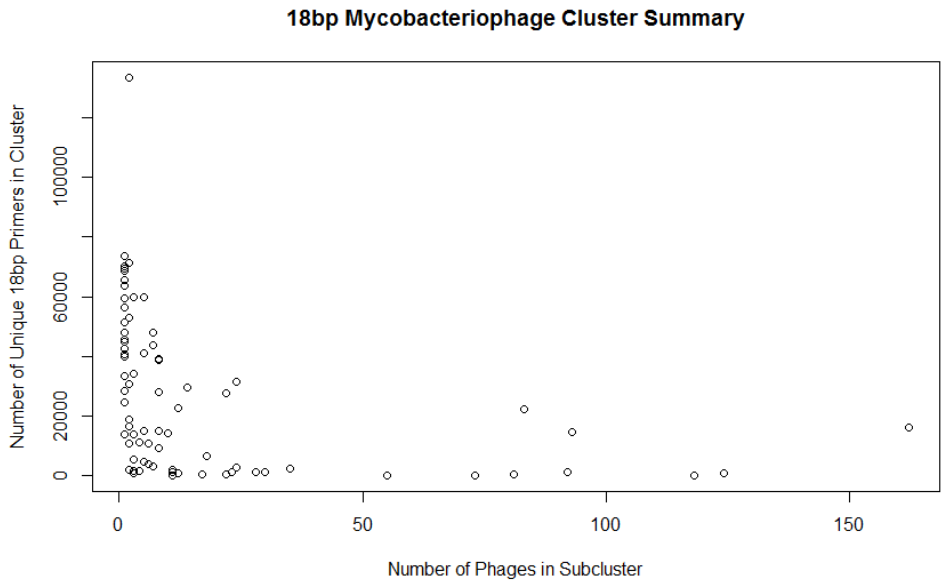

Figure 5: 18bp Mycobacteriophage Cluster Summary. In general, as the phage number increases within a sub-cluster there tends to be a lower unique primer count, though a few sub-clusters with high phage numbers still have a relatively high number of unique primers.

Cluster	Primer 1	Primer 2	Expected product length (bp)
A6	CCGATGGTTGCAGGAGTAGGGG	CGAAGAGAACATGCGCGAGCAGA	514
K5	GGGGATGATGACGGCGATTTCC	GTTCGCGCCCATCGCGGTA	606
A1	GGACATGACCGAGGACATCGCC	CAGCAAGAGCAGCAAGCCCA	875
A4	CGTCGACCCATGTTTTCCTCCACTT	TACTGCCCCCGGACGATC	1014

Table 2: Test Primers. Primers that were tested using the bacteriophages Badger (A4), TheloniusMonk (A1), Achebe (A4), Gruunaga (A6), AlleyCat (K5), and Ruin (A4).

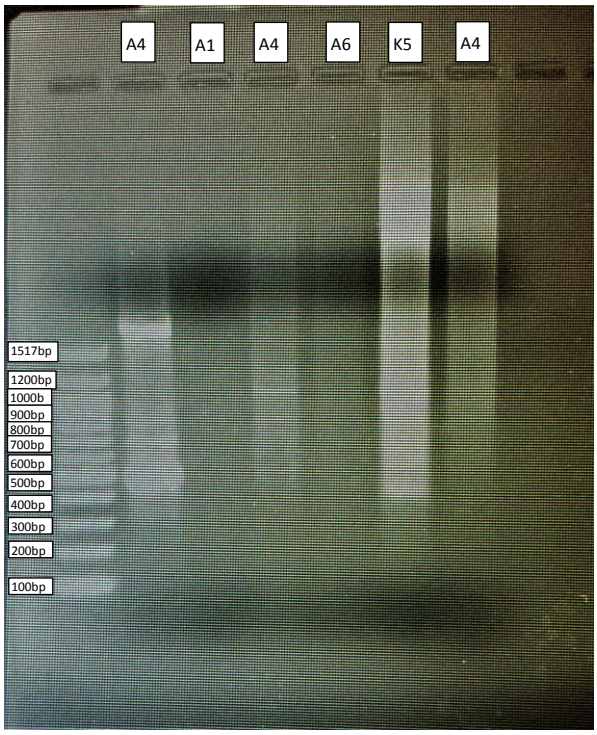

Figure 6: Primer Test Gel. This gel shows the PCR results of the above primers with bacteriophages from their clusters. From left to right: New England BioLabs 100bp DNA ladder, Badger (A4), TheloniusMonk (A1), Achebe (A4), Gruunaga (A6), AlleyCat (K5), and Ruin (A4).

CHAPTER SIX:

DISCUSSION

Figure 4 shows the relationship between mycobacteriophage sub-cluster member count and the number of unique 18bp sequences per sub-cluster. Generally, as the sub-cluster size increases, the number of unique sequences decreases. This is most likely a result of the increase in the genetic diversity of the subcluster as new bacteriophages are added to the subcluster. The increase in genetic diversity can cause sequences that were once unique and common to all bacteriophages in a sub-cluster to become not common to all bacteriophages within the sub-cluster, as the new member of the sub-cluster may not contain them. Cluster A2 has 73 phage members but does not have any 18-25 bp long sequences that are common and unique to all its members. However, Cluster B1 has 162 bacteriophages and 16,096 18bp unique sequences. It seems certain sub-clusters can accept increases in genetic diversity with minimal loss to the number of unique sequences, while others cannot. It is possible that these more resistant sub-clusters have common highly conserved regions that can handle new bacteriophage additions without incurring a loss of unique sequences. Clusters, like A2, on the other hand may have lower common conservation and with the addition of new diverse phages all unique sequences were lost.

This relationship also brings to light some of the potential problems with the primer pairs developed by PhageUniqueSeq for PCR. For smaller sub-clusters, the

incidence of false-negatives will be higher, because when new bacteriophages are added to a cluster, diversity rises and the number of unique sequences drops. It is possible that the bacteriophage in question does belong to a certain sub-cluster but the primers used (from that sub-cluster) are not found in this new bacteriophage. For larger sub-clusters, there is less chance of this happening, as they have fewer primers and those that remain have been consistently conserved across many phages. There is a point, however, when these large sub-clusters may lose all unique sequences, due to the low inherent conservation, as has happened in the A2 sub-cluster.

Using the PhageUniqueSeq's unique sequences as probes is a better route than PCR to alleviate the issue with false negatives, if several unique sequences of a single sub-cluster are used. Using more sequences increases the chances that at least one of the differentiating probes will bind and signal cluster membership. Using probes along with the plaque lift also allows for screening of several under-represented clusters at once, making it much more versatile. PCR can be used to test for more than one primer pair at once, but to avoid the issue of all the different primers binding to one another, many PCR reactions would have to be run or all of the primer pairs would need to pre-screened to ensure that there was minimal complementarity and defined cluster-specific product sizes. Probes can also be stripped from blots, and the blots can be recycled and hybridized to additional probes. In PCR, primers become part of the product sequence and are consumed by the reaction. Using the unique sequences in qPCR offers a better route than PCR to alleviate the issue with false positives since it increases the number of

unique sequences used, from two to three, two sequences for the primers and one for the TaqMan probe. The added specificity of the third sequence would increase the chance that the signal produced is a true positive indication of sub-cluster membership. There would, however, be a higher chance of a false negatives, if one of the sequences doesn't bind, therefore producing no fluorescence.

During development of PhageUniqueSeq, primer matches were tested on currently available bacteriophages to assess the accuracy of the sequences as primers. Figure 5 shows the results from the PCR and gel electrophoresis. Though the initial interest was to confirm the accuracy of the projected product size within the expected product size range, it was quickly discovered, from the gel results, that more profound issues existed . For lane 2, multiple bands can be seen, and for lanes 3 and 4, no bands were found at all. This test brought to light issues within the program design that have been corrected in the current iteration. It also prompted the design of testing programs to verify the PhageUniqueSeq output. Corrections were made to the primer-matching algorithm to ensure that any primer pair can only have one product for each bacteriophage within the cluster. This ensures that there aren't multiple products which would give ambiguous results to the researcher. Corrections were also made to ensure that primers are in the correct orientation. Before, primers could have been on different strands, which would have resulted in only one primer on the strand being extended. Without a second primer on the extension strand, the product cannot be predicted.

Though PhageUniqueSeq does check for exact-match off-targets when selecting primer pairs, it doesn't consider off-targets that include mismatches. This is important when considering PCR, since off-targets that are one to a few base pairs different could potentially bind to an off site, even though they are not exact matches. This would create the situation of a unique primer binding to bacteriophage DNA that doesn't belong to that primer's cluster, giving a false-binding positive that would most likely not give a predictive length product. This would not necessarily ruin the PCR, but could lead one to consider the aberrant product size as being derived from contamination.

PhageUniqueSeq's approach can alleviate issues with the current methods described earlier. The method described by Leslie et al. (2014), using plaque morphologies and restriction digests, can sometimes be too ambiguous. Plaque morphologies can sometimes also be so similar among bacteriophages in different clusters as to not be useful in determining cluster. Plaque morphologies are also often characteristic of cluster but not sub-cluster. Bacteriophages that differ by cluster can also show the same restriction enzyme fragment patterns, which is unhelpful and can be misleading. The creators of the Phage Enzyme Tool stated that high quality restriction digests are needed to obtain accurate results. High quality restriction digests require large amounts of purified DNA. Being a participant of the GDEP program, I know how difficult it is to obtain large amounts of purified DNA. I was not able to get enough bacteriophage from my plates due to the lysogeny of my bacteriophage. My phage plaques would disappear from the plates from week to week. Though others may not

have this problem, it is not a trivial process to obtain a high bacteriophage concentration needed for DNA purification. The creators of the Phage Enzyme Tool had problems with this as well since some of their test samples could not be used due to either contamination or low DNA quality. Consequently, there is sometimes not enough DNA to run a restriction digest if one wants to also sequence the bacteriophage genome at a later time. Additionally, the program also doesn't always give a straightforward prediction and can predict more than one cluster for a set of restriction enzyme patterns.

Using the PCR approach with PhageUniqueSeq as the lone discriminator to identify the cluster membership of every phage would require roughly 89 PCRs with 178 primers and enough DNA for all the reactions. However, if the goal is to identify more diversity in underrepresented clusters or to identify new members of potentially new clusters, then only a subset of primers or probes would be needed. A quick scan of Table 1 identifies 37 subclusters with a membership less than 5. If primer pairs, or probes, were designed to work together in a single reaction and still identify the subcluster membership then they could be multiplexed together. If 4 subcluster reactions could be multiplexed together then only 10 reactions would be needed to screen for underrepresented cluster membership. This is a research area of practical application that still needs to be developed, but the unique sequences identified by PhageUniqueSeq will be the starting place to use in these combinatorial designs.

Bacteriophages are fascinating viruses that have many practical uses in the industrial and medical communities. Their specificity for bacteria make them perfect candidates for alternative antibiotics at a time when our traditional antibiotics are failing. They also provide new avenues of control over bacteria in food products. For these reasons, bacteriophage research to discover new diversity in bacteriophages is occurring at over 100 institutions, including WKU. My program, PhageUniqueSeq, will help universities that employ the SEA PHAGES program, to make better decisions in determining which newly isolated bacteriophages to sequence. PhageUniqueSeq potentially provides an inexpensive and accurate method to determine a bacteriophage's cluster before sequencing, and therefore provides a crucial step in increasing the identification of diversity of bacteriophages. Further work will be necessary to 1) design multiplexed probe or primer sets from the unique sequences, 2) test the selected unique sequences in experimental probe or PCR settings, and 3) to make the data generated by this program readily available to other researchers. The ideal approach to present this data to the public would be a web server or API interface that can present the data in a usable manner, and allow it to be filtered. User input should be used to determine the filter parameters for selection of the primer pairs that the user is specifically interested in, based on sub-cluster, primer complementarity, melting temperature, and product size range.

References

ABMGood. Polymerase Chain Reaction (PCR) - An introduction. Retrieved March 27, 2017, from https://www.abmgood.com/marketing/knowledge_base/polymerase_chain_reaction_introduction.php

BioSynthesis. TaqMan® vs. SYBR® Green Chemistries. Retrieved April 30, 2017, from http://www.biosyn.com/tew/taqman-vs-sybr-green-chemistries.aspx

Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R.W., Hatfull, G.F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics, 12*(395). doi: 10.1186/1471-2105-12-395.

Deasy, T., Mahony, J., Neve, H., Heller, K. J., & Sinderen, D. V. (2011). Isolation of a virulent Lactobacillus brevis phage and its application in the control of beer spoilage. *Journal of Food Protection, 74*(12), 2157–2161.

Food and Drug Administration. Agency Response Letter GRAS Notice No. GRN 000218. Retrieved December 01, 2016, from http://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/NoticeInventory/ucm153865.htm

Gissendanner, C. R., Wiedemeier, A. M. D., Wiedemeier, P. D., Minton, R. L., Bhuiyan, S., Harmson, J. S., Findley, A. M. (2014). A web-based restriction endonuclease

tool for mycobacteriophage cluster prediction. *Journal of Basic Microbiology, 54*(10), 1140-1145. doi: 10.1002/jobm.201300860

Hatfull, G. F. (2012). The secret lives of mycobacteriophages. *Adv Virus Res, 82*,179-288. doi: 10.1016/B978-0-12-394621-8.00015-7.

Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M. E., et al. (2006) Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genet 2*(6), e92. doi:10.1371/journal.pgen.0020092

Jacobs, W. R., Jr. (1992). Advances in mycobacterial genetics: New promises for old diseases. *Immunobiology, 184*(2–3),147–156.

Jacobs, W. R., Jr., Tuckman, M., and Bloom, B. R. (1987). Introduction of foreign DNA into mycobacteria using a shuttle phasmid. *Nature, 327*(6122), 532–535.

Leslie, F. H., Gerald-Goins, T., Williams-Devane, C., (2014). Plaque morphology and restriction analysis patterns of bacteriophage genomic DNA as a preliminary screening method for bacteriophage cluster/sub-cluster assignments. *ProQuest Dissertations and Theses.*

Luria, S. E., & Anderson, T. F. (1942). The identification and characterization of bacteriophages with the electron microscope. *Proc Natl Acad Sci USA, 28*,127-130.1. PMID:16588529; http://dx.doi.org/10.1073/pnas.28.4.127

NPTEL. Module 4 : CONSTRUCTION OF DNA LIBRARIES. Retrieved April 30, 2017, from http://nptel.ac.in/courses/102103013/module4/problems/4.html

PhagesDB. Host genus: mycobacterium. Retrieved December 01, 2016, from http://phagesdb.org/hosts/genera/1/

Pope, W. H., Jacobs-Sera, D., Russell, D. A., Peebles, C. L., Al-Atrache, Z., et al. (2011) Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. *PLoS ONE 6*(1). doi:10.1371/journal.pone.0016329

Pope, W. H., Bowman, C. A., Russell, D. A., Jacobs-Sera, D., Hendrix, R. W., Lawrence, J. G., & ... Jacobs Jr., W. R. (2015). Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Elife, 1-65. doi:10.7554/eLife.06416.001

RegentsGeneticTechnology. Gel Electrophoresis 3-4. Retrieved March 27, 2017, from https://regentsgenetictechnology.wikispaces.com/Gel+Electrophoresis+3-4

ThermoFisher. Real-time PCR (qPCR) defined. Retrieved April 30, 2017, from https://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/qpcr-education.html

ThermoFisher-TaqMan. TaqMan Assays for all your qPCR needs. Retrieved April 30, 2017, from https://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays.html

Twort, F.W., LOND., L.R.C.P. , M.R.C.S. (1915). An investigation on the nature of ultra-microscopic viruses. *Bacteriophage, 186*(4814), 1241–1243.

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T., Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res, 6*(11), 3543-57.

Weber-Dąbrowska, B., Jończyk-Matysiak, E., Żaczek, M., Łobocka, M., Łusiak-Szelachowska, M., Górski1, A. (2016). Bacteriophage procurement for therapeutic purposes. *Front Microbiol, 7*, 1177.

Wommack, K. E., Colwell, R. R.,.(2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev, 64*(1), 69-114.