BIOCHEMICAL CHARACTERIZATION OF BACTERIOPHAGE IPHANE7 REPRESSOR PROTEINS

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Master of Chemistry

By

Vance Renaud, Erin Cafferty, Dr. Maria Gainey

Director: Dr. Maria Gainey

Assistant Professor of Chemistry

Chemistry Department

Committee Members: Dr. Amanda Storm, Biology

Dr. Jamie Wallen, Chemistry

July 2021

ACKNOWLEDGEDMENTS

I would like to thank my committee members and director for their assistance and encouragement. In particular, Dr. Gainey who has guided me through this project from the beginning. I also extend sincere thanks to the following people, without whom this thesis would not have been possible: The Hatfull laboratory, WCU Virus-Hunters, SEA-PHAGES program, Lori Neri, and Cecilia Baumgardner.

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

BIOCHEMICAL CHARACTERIZATION OF BACTERIOPHAGE IPHANE7 REPRESSOR PROTEINS

Vance Renaud, Master of Chemistry

Western Carolina University (07/2021)

Director: Dr. Maria Gainey

Bacteriophages are viruses that infect bacteria and can utilize two different pathways after infection; the lytic or lysogenic cycles. During the lytic cycle, the bacteriophage hijacks the bacterial machinery to replicate itself and destroy the bacterial cell. During the lysogenic cycle, the viral genome integrates into the bacterial genome, and the bacterium reproduces normally. The lysogenic cycle continues until a stimulus causes the virus to reenter the lytic replication cycle. Temperate bacteriophages use both the lytic and lysogenic cycles. Bacteriophage diversity is immense, and they are currently loosely classified into different genetic clusters based on sequence and gene content similarity. Cluster M bacteriophages are temperate and can be divided into 3 subclusters M1, M2, and M3. IPhane7 is a cluster M1 bacteriophage that was discovered at WCU in 2016. Temperate bacteriophages typically encode a repressor gene. The repressor protein binds to the viral genome during lysogeny and prevents expression of genes vital to lytic replication. The purpose of this project is to begin characterization of a novel cluster M repressor system. While cluster M bacteriophages have been shown to be temperate, bioinformatic analysis failed to provide a repressor gene candidate. A repressor gene candidate

(gene 1) was identified using an overexpression screen by undergraduate student Erin Cafferty. Interestingly Erin also determined that gene 2 may also be playing a role in repression of select cluster M bacteriophages but is toxic during overexpression assays.

Initial work using a single copy expression assay and a panel of cluster M viruses has established that gene 1 and gene 2 likely synergize during lysogeny to repress lytic gene replication. To begin biochemical characterization of IPhane7 gene 1, a solubility test has also been performed. The results of this test revealed that IPhane7 gene product 1 will likely be able to be expressed and purified and is soluble under conditions tested. Due to the Coronavirus pandemic, further biochemical characterization work will be a future direction of this project.

Repressor proteins typically bind to specific DNA sequences located near promoters. In order to bind specific DNA sequences, repressor proteins typically contain a helix-turn-helix DNA binding motif. A detailed bioinformatic analysis was performed on cluster M genomes, as well as gene products 1 and 2. Cluster M genomes were searched for predicted promoters, terminators, and repetitive elements. This analysis helped reveal potential areas of the genome that are likely subject to repressor regulation and binding. Structural prediction programs were used to predict the 3-dimensional structure of gene products 1 and 2. Detailed amino acid analysis were then performed to determine the conserved areas of these gene products that are likely essential to their function. Gene 1 is unique to cluster M bacteriophages. However, preliminary BLAST analysis of gene 2 has revealed gene 2 is found in other bacteriophages and in a wide range of Actinobacterial species and may have been taken from one of these hosts.

CHAPTER ONE: INTRODUCTION

Phages are viruses that replicate using a bacterial host. There are two replication pathways phages are capable of: the lytic cycle and the lysogenic cycle. During the lytic cycle, the phage utilizes bacterial replication machinery replicate itself and destroy the bacterial cell. Conversely, During the lysogenic cycle the viral genome forgoes immediate replication and instead integrates into the bacterial genome, the bacterium reproduces normally, reproducing with it, the viral genome. The lysogenic cycle can continue until an outside stimulus causes the virus to renter the lytic replication cycle. Temperate phages are phages that use both the lytic and lysogenic reproduction cycles.

The ability of temperate phages to differentiate between these two replication cycles stems whether or not the phage integrase is able to function. The phage's integrase is a site-specific enzyme which acts to recombine the DNA of the host bacterium and the phage DNA. During lysogeny this recombination results in the integration of the entire phage genome into the host chromosome, forming a lysogen. Inversely, if the integrase is prohibited from recombining the DNA strands, the lytic reproduction cycle takes over and the phage rapidly reproduces itself using the bacterium's cellular machinery. A highly researched siphoviridae phage infecting *Escherichia coli*, lambda phage, has been well studied. In lambda phage, the reproductive cycle is determined by the presence of two proteins, the Cl protein and the Cro protein. The cro protein initiates lysis and binds to the OR3 operator to stop Cl gene transcription. The cl protein is the repressor protein that self assembles into a dimer to bind to the OR1 operator to stop the transcription of the Cro protein and activates transcription of Cll, which then activates transcription from the associated promoters, which transcribe the integrase. When the cl protein

is present in sufficient quantity the integrase is activated and lysogeny occurs. When the cl protein is not present in sufficient quantity, cro is able to cease production of cl and initiate the lytic cycle (Johnson A. D. et. al, 1981).

Erin Cafferty, who began this project, tested 20 repressor gene candidates using the vector pSMEG in *Mycobacterium smegmatis* mc^2155. pSMEG is a high copy vector with an inducible promoter. Her results revealed that IPhane7 gene 1 over-expression prevents superinfection of cluster M viruses IPhane7 (M1) and Nanosmite (M3), suggesting that gene 1 is the repressor. Over-expression of gene 2 was also found to be toxic in this system. When Erin expressed gene 2 from the single copy vector pMH94 the expressed protein was no longer toxic to *M. smegmatis* cells, and inhibited infection of cluster M3 bacteriophage Nanosmite to a great degree ($1.82*10^{-5}$ efficiency of plating (EOP)), EOP is the ratio of the viral titer (PFU/mL) found on a plate containing construct of interest divided by the pMH94 viral titer. IPhane7 gene 2 only inhibit infection by IPhane7 roughly half ($4.45*10^{-1}$ EOP) Taken together, these results suggest that gene 2 is a prophage-mediated immune defense mechanism that could work synergistically with gene 1 during lysogeny to prevent superinfection of the lysogenic cell by other bacteriophages. A prophage-mediated defense mechanism is a mechanism by which the expression of a gene or group of genes from the prophage work to stop the infection of the lysogen by other phages. This mechanism keeps the lysogen from falling victim to lytic replication from a different phage and allows the lysogenic cell containing the prophage to survive. There are different types of prophage-mediated defense systems. In one example, when the phage Sbash, forms a lysogen with *M. smegmatis*, the infection of the lysogen by phage Crossroads activates Sbash genes 30 and 31, suspectedly causing membrane depolarization and a

loss of ATP in the cell, inhibiting the lytic growth of the infecting phage. Meanwhile, the prophage-mediated defense system of lambda phage uses a DNA binding protein RexAB to protect its lysogen from infecting phages. (Gentile, Gabrielle M et al, 2019). Our work sought to validate whether the gene 1 is indeed the cluster M repressor and begin its biochemical characterization and further explore the defensive role of gene 2.

Phages are classified according to clusters. The phage cluster system is based on nucleotide homology (genomes with a sequence similarity spanning >50% belonging to the same cluster), host similarity, and gene content similarity. Cluster M phages are mycobacteriophages, meaning they use the bacterium *M. smegmatis* mc^2155 as a host. All cluster M phages are likely temperate phages. These phages have siphoviral morphologies (long non-contractile tails, and double stranded DNA genomes), are homo-immune to each other, and have relatively large genomes (80.2 to 83.7 kbp) compared to other mycobacteriophages (shown in Figure 1 below) (Broussard, Gregory W et al, 2013; Hatfull, Graham F. et al, 2010).

Cluster M phages also have noncanonical genome architectures and several unusual sets of conserved repeated sequences. One such noncanonical genome feature is the position of the integrase gene (gene 129) which is located near the right end of the genome (Figure 1). This differs from the norm because typically the integrase in located near the center of the genome. Cluster M is divided into three subclusters according to sequence similarity (Hatfull, Graham F. et al, 2012): M1, M2, and M3. The M1 subcluster includes 7 phages, the M2 subcluster includes 6 phages, and the M3 subcluster contains only 1 phage. IPhane7 belongs to the M1 subcluster and was discovered through the Western Carolina University SEA-PHAGES program by Dylan Rood in Cullowhee, North Carolina in 2016. IPhane7 was isolated using the bacterial host *M. smegmatis* mc^2155.

A search for new phages to study is being performed by the SEA-PHAGES (Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science) program. This program is administered by Graham Hatfull's group at the University of Pittsburgh and the Howard Hughes Medical Institute's Science Education division (www.seaphges.org). The SEA-PHAGES program guides students through a two-semester long experiment starting with growing a stock of phage from a soil sample, isolating a single phage species, amplifying the population of the phage to a concentration high enough for electron microscope images to be taken, and the phage genome to be analyzed. The students then identify the family the phage belongs to and learn bioinformatic methods for manually confirming the locations of genes determined using Phamerator software (Cresawn, S.G. et al, 2011).

The repressor gene codes for a repressor protein that binds to specific sites in the integrated viral genome during lysogeny, stopping lytic gene transcription. Expression of a repressor is also known to cause a homo-immunity phenotype to closely related phages. Repressor mediated homo-immunity is a phenomenon by which a repressor protein from a prophage is similar enough to other, closely related phages that the repressor protein is able to bind to the other phage genomes during initial infection, halting their lytic replication cycles. This protects the cell from becoming productively infected by similar types of bacteriophages. The IPhane7 genome circularizes after infection, so while the repressor is designated as gene 1 (Figure 1), this is a somewhat arbitrary distinction. Generally, in phages the repressor gene is normally located near the integrase gene. Cluster M bacteriophages are of particular interest because although they all have an integrase gene, bioinformatics has failed to reveal any obvious repressor gene candidates.

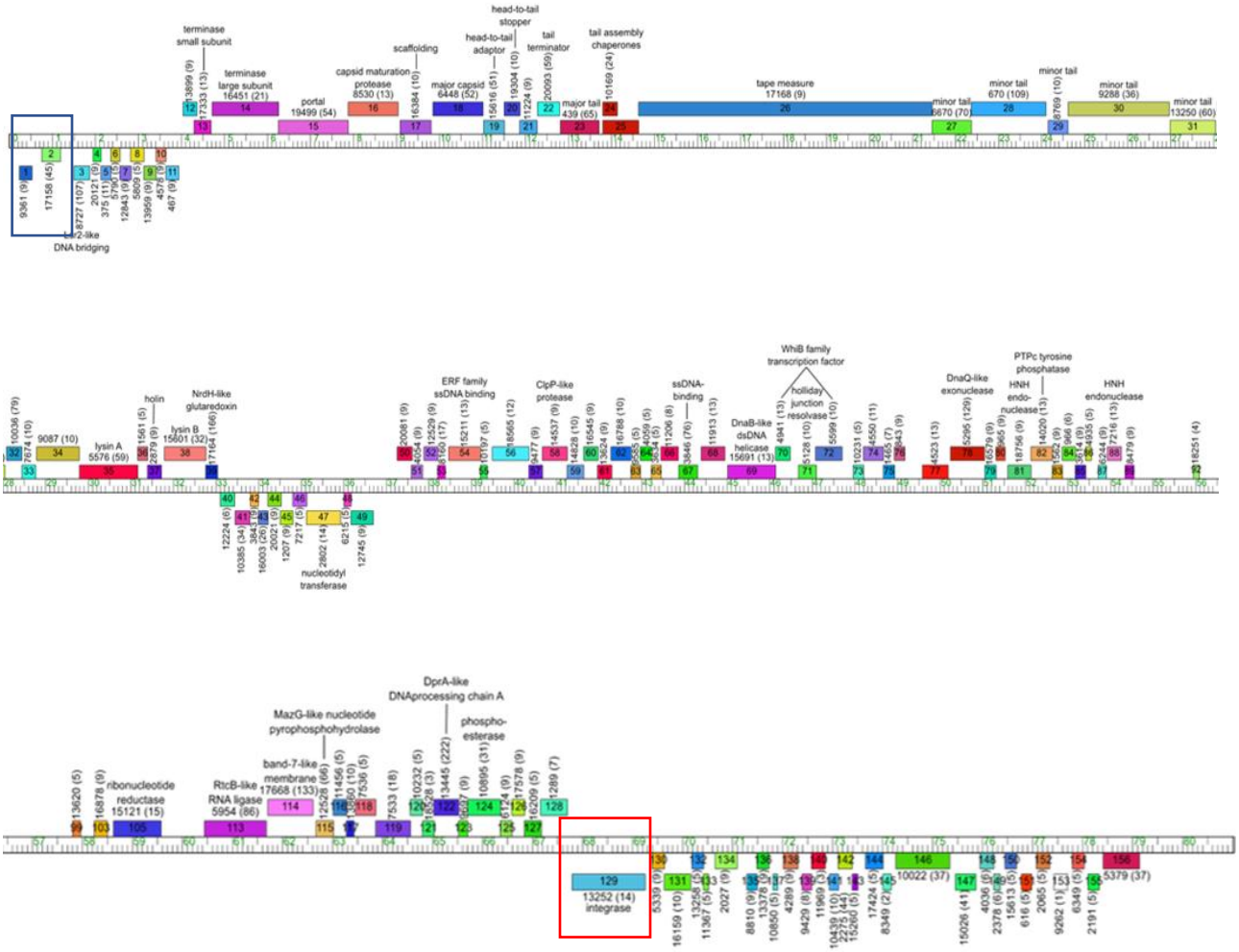


Figure 1. *IPhane7 genome map generated using Phamerator (Cresawn, S.G. et al, 2011). Each block represents a predicted gene, while the numbers on the ruler correspond to that area of the genome in kilobase pairs. Genes on the top and bottom of the ruler are predicted to be transcribed in the forward and reverse directions respectively. The five-digit number associated with each predicted gene indicates the family group of that gene. The blue box around genes 1 and 2 at the beginning of the genome illustrates the area of focus for the repressor gene and defense mechanism. The red box around gene 129 highlights the location of the integrase.*

CHAPTER TWO: METHODS

Vector Cloning

PCR Amplifications

Genes of interest were amplified by PCR from the IPhane7 genome using Q5 hot start polymerase and sequence specific primers. An EcoRI site (GAATTC) plus several guanines were added to the 5' ends of these primers. The primers were designed to have a total number of nucleotides between 20-35, a GC content over 60%. All primers were ordered from IDT. Primers stocks were resuspended to 100uM using sterile water and were further diluted to a 10uM working concentrations. PCR reactions were carried out using the following conditions: Initial denaturation 98° C, 30 seconds followed by 28 cycles of 98° C, 10 second denaturation, 66° C 30 second annealing, and 72° C, 60 second elongation.

Amplification of the correct size PCR products was confirmed by running 2.5uL of each PCR product mixed with 7.5uL 6x dye on a 1% agarose gel containing a 10uL/100mL concentration of ethidium bromide alongside NEB's 1KB standard DNA ladder.

Dephosphorylation, Ligation, and Transformation

1ug of the pMH94 vector was digested for 4 hours at 37 degrees, with EcoRI following NEB's standard 50uL digest protocol. After digestion 1uL CIP was added to mixture and incubated for 10 min at 37° C, then heat inactivated at 90°C for 30 seconds. Enzyme were removed using NEB's PCR and DNA clean up kit.

Ligation of digested PCR products and pMH94 vector was performed by adding 7.5uL PCR product to 2.5uL PHM94 vector, 10uL of ligation buffer and 1uL T4 DNA ligase for 5 minutes at room temperature.

Transformation of the pMH94 + DNA construct vectors into chemically competent c2981 *E. coli* cells was performed by adding 5uL of the ligation reaction to 50uL of c2981 *E. coli.* cells for 10 minutes, on ice, followed by heat shock at 42°C for 30 seconds, then placed back on ice for 3 minutes. 300uL of SOC media was added to each transformation, and the tubes were placed in a shaker at 37° C for1hr. 60 uL of each transformation was plated onto agar plates containing LB plus 50 ug/mL kanamycin. Colonies were allowed to incubate overnight at 37 degrees.

Colony Screening/Vector Proliferation

Colony PCR was used to screen for correctly ligated clones. The forward and reverse primers were designed to sit ~100bp before and after the pMH94 EcoRI site. PCR reactions were carried out using the following conditions: Initial denaturation 94° C, 300 seconds followed by 30 cycles of 94° C, 30 second denaturation, 58° C 30 second annealing, 68° C, 120 second elongation, and a 68°C, 300 second final extension. OneTaq DNA polymerase was used for all screening PCR reactions.

Amplification of the correct size PCR products was confirmed by running 7.5uL of each PCR product mixed with 2.5uL 6x dye on a 1% agarose gel containing a 10uL/100mL concentration of ethidium bromide alongside NEB's 1KB standard DNA ladder.

Colonies that were screened were archived on an LB plate. A pipette tip was touched to each colony then was gently scraped across a fresh LB agar plate containing 50ug/ml concentration of kanamycin, the pipette tip was then shaken into each PCR reaction. This was left to incubate at 37°C overnight.

Sanger sequencing was used to confirm positive clones. Positive colonies were transferred in to 5mL LB plus 50ug/mL Kanamycin and grown to saturation at 37° C overnight in the shaking

incubator. The culture was miniprepped using the monarch miniprep kit to extract the pMH94 + construct vector from the culture prior to Sanger Sequencing.

Electroporation

~200 ng of the pMH94 vector or pMH94 vector plus genes of interest was then electroporated into 50uL of electrocompetent *M. smegmatis* cells that had been thawed on ice. 1uL of vector was added to 50uL of cells and incubated for 10 minutes on ice. Cells were then transferred into ice cold 1mm cuvettes. Cells were then shocked using standard *E. coli* settings. The cells were then immediately transferred into 1mL of 7H9 media (9mL 7H9 neat, 1 mL AD supplement, 100uL 100mM $CaCl_2$. Cells were shaken at 37° C for 2hrs. 60uL of cells were plated onto 7H9 agar plates (1.44 g 7H9 broth base, 4.5 g agar, 267 mL deionized water, 30 mL AD supplement, and 1.5 mL 40 % glycerol) then placed at 37 degrees for 4 days. Colonies formed on the plates were scraped with a pipette tip and transferred to a test tube with 5mL of 7H9 media with Tween 80 (4.5 mL 7H9 neat, 0.5 mL AD supplement, 50uL CaCl2, 12.5uL tween 0.5uL kanamycin (50mg/mL)). These reaction tubes incubated in a 37°C shaker for 48 hours.

In some experiments, Tween 80 was not used. In these experiments, the *M. smegmatis* bacteria went through this initial 48-hour incubation cycle in the 7H9 media containing tween, then ~100uL of the culture was taken and placed in a reaction tube with 5 mL of 7H9 media without tween. These reaction tubes the incubated in a 37° C shaker for 48 hours.

Protein Expression

Protein Expression Vector Cloning

Protein expression vector cloning utilized the same cloning techniques used in cloning genes into pMH94, (primer design, PCR amplification of gene 1 with primers, digestion of

vector, dephosphorylation, ligation, transformation) the only differences being, the primers for gene 1 were designed for the vector PLM303, due to the necessity for the gene to clone into the vector in the correct direction the forward and reverse primers were designed with different restriction enzyme sites the forward primer utilized the Kpnl (GGTACC) whereas the reverse primer utilized the EcoRI restriction enzyme (GAATTC). PCR reactions were carried out using the following conditions: Initial denaturation 98° C, 30 seconds followed by 30 cycles of 98° C, 10 second denaturation, 62° C 30 second annealing, 72° C, 30 second elongation, and a 72°C, 120 second final elongation. Amplification of the correct size PCR products was confirmed by running 2.5uL of each PCR product mixed with 7.5uL 6x dye on a 1% agarose gel containing a 10uL/100mL concentration of ethidium bromide alongside NEB's 1KB standard DNA ladder.

In order for the gene to clone in phase with the promoter a nucleotide had to be added in front of the gene 1 ATG start codon. PLM303 has an asparagine tag that is designed to be translated as a part of the protein during DNA translation. This helps with protein purification as the asparagine tag has high affinity to a nickel column. This asparagine tag is designed to be cleaved off at the protease site LEVLFQGP-VP, due to the use of the Kpnl site only the VP amino acids would remain linked to protein 1 after the purification process.

Gene 1 Protein Expression and Solubility Testing

Once the genes were cloned into the PLM303 vector and transformed into *E. coli* BL21 cells, the cells were selected for kanamycin resistance using the same methods used with the pMH94 vector. The *E. coli* colony found to contain the PLM303 vector containing IPhane7 gene 1 was placed in a 5 ml culture and incubated overnight to saturation. The culture was then diluted to 1:100 in a LB broth with a 50ug/ml concentration of kanamycin. This broth was incubated for 2 hours 40 minutes to an optical density (OD) 600 A of E. coli cells. 0.5 ml of this broth was

pelleted and resuspended in nickel buffer. 50 ul of isopropylthio—galactoside (IPTG), an inducing agent was added to induce expression of the MBP + gene 1 section of the vector. The broth culture was incubated for 3 more hours, the OD was recorded (1.511 A), 0.5 ml of the culture was pelleted and resuspended in 250 ul nickel buffer, the remaining culture was sonicated in 1-minute increments with a 2-minute cool down period with the sonicator set to setting 5. Sonication was repeated 5 times. The freshly sonicated solution was evenly distributed between 4 Oakridge tubes and centrifuged at 12,000 rpm for 20 minutes to spin down the cellular debris. 20 ul of the supernatant was added to 20 ul 2x dye. The T=0 solution and T=3hr solutions were racked and 20 ul of their supernatants were also added to 20 ul 2x dye and these were put on a 15% SDS-page gel and run at 150 v for 120 minutes.

Homo-immunity assay

Viral Titter

The viral titers were determined using 10-fold serial dilution of viral stocks into phage buffer. 3uL of each dilution was pipetted onto the top agar mixed with *M. smegmatis* using a multi-channel micro-pipette. The plates were put in a 37° C incubator for 48 hours.

At lower concentrations of virus, individual plaques can be counted. From the number of individual plaques formed in a single 3uL droplet of diluted viral stock, the concentration of the viral stock can be calculated in plaque forming units (PFU)/mL. The number of plaques formed in a 3uL droplet of serial diluted viral stock is multiplied by 10^n where n is the number of dilutions away from the stock the in the serial dilution the droplet came from, to calculate the number of plaque forming units in 3 uL of the stock solution. This was then divided by 0.003mL

to give the number of plaque forming units in 1mL of stock solution. This titer was done in duplicate and the average concentration was taken.

Viral stocks were then diluted to $5*10^8$ PFU/mL with phage buffer for future experiments.

Homo-immunity Assay

The homo-immunity assays were performed using a top agar containing the *M. smegmatis* cultures on 7H9 agar plates with kanamycin. 7H9 plates were made with 1.44g 7H9 base, 4.4g agar, and 1.3mL 40% glycerol. The solution was sterilized via autoclave then 30mL AD supplement and 30uL Kanamycin (50mg/mL) were added using sterile technique.

The top agar was prepared by combining 25 mL warm 2x top agar, 25mL 7H9 neat, 0.5 mL 100mM $CaCl_2$, and 5uL kanamycin. While still warm (55°C) 5mL top agar was mixed with 5mL *M. Smegmatis* in 7H9 media culture and pipetted onto 2 7H9 agar plates and allowed to solidify. Efficiency of plating was calculated by dividing the viral titer for each plate by the vector only plate titer.

Bioinformatic methods

Blast and PSI-BLAST

BLAST and PSI-BLAST were performed using NCBI (NCBI, 2004) and PhagesDB.org (Russel, Daniel A. et al, 2017) databases. When searching specifically for phage-based results while trying to determine if the presence of genes similar to IPhane7 Gene 1 and 2 were due to the presence of a prophage or if the gene likely originated in a bacterium, PhagesDB.org was used. all other BLASTs and PSI-BLASTs were performed using NCBI.gov.

Table 1. *PhagesDB nucleotide BLAST were performed using the following settings*:

Program:	blastn
Database:	Actinobacteriophages
Expect:	10
Matrix:	Blosum62
Alignment view:	Pairwise

Table 2. *PhagesDB Protein BLAST were performed using the following settings*:

Program:	Blastp
Database:	Actinobacteriophages
Expect:	10
Matrix:	Blosum62
Alignment view:	Pairwise

Table 3. *The NCBI.gov blastn BLASTs were performed using the following settings*:

Database:	Nucleotide collection (nr/nt)
Organism:	none
Exclude:	None
Limit to:	None
Optimize for:	Highly similar sequences (megablast)

Table 4. *The NCBI.gov blastp BLASTs were performed using the following settings*:

Database:	Non-redundant protein sequences (nr)
Exclude:	None
Algorithm:	blastp (protein-protein BLAST)

Table 5. *The NCBI.gov PSI-BLASTs were performed in blastp using the following settings:*

Database:	Non-redundant protein sequences (nr)
Exclude:	None
Algorithm:	PSI-BLAST (Position-Specific Iterated BLAST)

Conserved repeats

Table 6. *Conserved repeats were found using MEME Suite (Timothy L. Bailey et al, 1994). The MEME Suite settings were set as:*

motif discovery mode:	Classic mode
sequence alphabet:	DNA, RNA, or Protein
Site distribution:	Any number of Repetitions (anr)
Number of motifs:	5

Terminators

Table 7. Terminators were found using Softberry findterm program (V. Solovyev et al, 2011) with default settings:

energy threshold value:	-11

Promoters

Promoters were found using DNA Master by inserting the sequence into the program and running the promoter prediction function.

Table 8. *The promoter search settings were set as follows:*

Site:	Sigma 70
-10 Weight:	1.0
-35 Weight:	1.0
Scores to keep:	50
Site method:	Geometric
Forward Strand:	Yes
Reverse Strand:	Yes
Merge Method:	Geometric
Spacing Weight:	0.1

Structural Predictions

Protein structural predictions were found using I-TASSER (B. Webb et al, 2016) and Modeller (J Yang et al, 2015). In I-TASSER the amino acid sequence responsible for the protein was

submitted to I-TASSER, which created a prediction for the structure of the protein from a database containing structural consequences of amino acid combinations. Modeller is a python-based program that allows for the curation of data to compare to the amino acid sequence in question. The structurally known repressor gene for the T7 phage was selected as the comparative data, and the program was run to give the Modeller protein structural prediction.

Table 9. *ClustalW (Thompson JD, et al, 1994). was used to perform all amino acid alignment comparisons with settings set as*:
General setting Parameters: Clustal

Pairwise alignment:	Slow/Accurate
Sequences:	Sequences: Protein
Pairwise Alignment Parameters for SLOW/ACCURATE	
Gap open penalty:	10.0
Gap Extension Penalty:	0.1
Weight Matrix:	BLOSUM (for PROTEIN)

Table 10. *Multiple Alignment Parameters*:

Gap open Penalty:	10
Gap Extension Penalty:	0.05
Weight Transition:	No
Hydrophilic residues for Proteins:	GPSNDQERK
Hydrophilic Gaps:	YES
Weight Matrix:	BLOSUM (for PROTEIN)

.

Table 11. *BoxShade parameters. BoxShade (Falquet L, et al. 2003) was used to shade in amino acid alignments made using ClustalW for easier visualization of alignment similarities.*

Output Format	RTF_New
Font Size	10
Consensus Line	No consensus line
Fraction of Sequence	0.5
Input sequence format	ALN

CHAPTER THREE: CONFIRMATION OF IPHANE7 REPRESSOR GENE USING A SINGLE -COPY VECTOR WITH ENDOGENOUS PROMOTERS RESULTS AND DISCUSSION

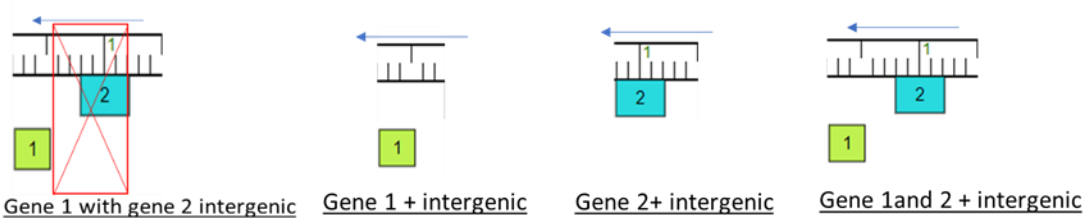


Figure 2**.** *Image depicting the regions of the IPhane7 genome that were cloned into pMH94.*

When expressed from the multi-copy vector pSMEG, IPhane7 gene 1 was able to prevent infection by other cluster M viruses. While this is a promising indicator that gene 1 is the repressor, it is not conclusive evidence, as this phenotype could have been caused by overexpression of gene 1 rather than a genuine homo-immune response. To rule out the latter possibility, the homo-immunity test was performed with a single copy vector.

The vector pMH94 does not have promoters that will promote the expression of genes cloned into the EcoRI site of the vector, therefore genes cloned into the EcoRI site must rely on endogenous promoters to be expressed. This lack of promoters in pMH94 required the cloning in of the intergenic regions upstream of the gene for the gene to be expressed, which allowed for the presence of a promoter upstream of both gene 1 and gene 2. Furthermore, the pMH94 vector integrates into the *M. smegmatis* genome to create a system closely mimicking repressor expression during lysogeny. Due to Erin's preliminary results with gene 2 expressed from the

pMH94 vector, we tested the four constructs shown in Figure 2 above. Gene 1 plus 1-2 intergenic region, gene 2 plus gene 2-3 intergenic region, gene 1 and gene 2 plus 1-2 and 2-3 intergenic regions, and gene 1 plus gene 2-3 intergenic regions were cloned into pMH94. Gene 1 and gene 2 were cloned in together to determine if these two genes have a synergistic effect. The Gene2 intergenic region was placed in front of gene 1 to determine if the gene 2-3 intergenic region's promoter had a stronger effect with gene 1 than the gene 1-2 intergenic region's promoter.

Results

The vector pMH94 alone, as well as pMH94 with the constructs shown in Figure 2 were electroporated into *M. smegmatis*. Liquid cultures of the transformed *M. smegmatis* were grown and mixed with top agar to create a lawn onto which the indicated viruses were spotted during immunity testing. The viruses used in the homo-immunity test were all diluted to $5*10^8$ PFU/mL before the immunity experiments took place.

Due to the reduced propagation of the viruses IPhane7 and Bongo (M1) in the first 4 homo-immunity experiments (as seen in Figure 3 by the failure to infect the empty vector), the results were not accurate; the problem was found to be the $CaCl_2$ used in the bacterial cultures. When this was replaced for experiment 5 the issues ceased, therefore the results from experiment 5 will be reported. The results labeled experiment 1 in Figure 3 are shown are representative of the first 4 experiments. The results of the accurate experiment (experiment 5) are reported as both viral titer and Efficiency of Plating (EOP). EOP is the ratio of the viral titer (PFU/mL) found on a plate containing construct of interest divided by the pMH94 viral titer. The smaller this ratio, the greater the effect the of the particular construct on virus replication. The results of experiment 5 have shown that gene 1 does indeed repress the propagation of IPhane7 with an

(EOP) of $2.05*10^{-4}$. IPhane7's close relative Bongo has an EOP of $2.08*10^{-4}$ indicating it to was repressed. Reindeer (M1) has an EOP of $2.00*10^{-2}$, this is much less inhibited than either IPhane7 or Bongo. Pegleg (M1) has an EOP of $3.03*10^{-1}$, this is the least inhibited of the M1 phages. Nanosmite, the only subcluster M3 in the panel is unaffected or lowly effected with an EOP $4.45*10^{-1}$. The relatively high EOP values for Pegleg and Nanosmite could be caused by variation in the assay rather than repression per se.

Gene 2 also had a repressive effect on the propagation of the cluster M viruses; however, the repressive effect was most notable on viruses less closely related to IPhane7. IPhane7 had an EOP of $4.41*10^{-1}$, Bongo had an EOP of $3.33*10^{-1}$, and Pegleg had an EOP of $1.36*10^{-1}$. Reindeer was not repressed, however, Nanosmite, the only M3, had an EOP of $1.82*10^{-5}$. Genes 1 + 2 together had the largest repressive effect, all viruses had an EOP of $3.03*10^{-5}$ or less except for the virus Kumao, a singleton being used as a control to determine plate to plate variation, which was not affected. In Figure 4 it appears that gene 1 affected Kumao, but this is because of a spotting error. The virus labeled MrMagoo was replaced with Kumao because MrMagoo may have been mislabeled and in fact not been MrMagoo or could potentially be a mutated form of MrMagoo. Only one plaque was recovered from the MrMagoo viral stock sent from the Hatfull laboratory, which is suspicious. However, since it is the only M2 tested thus far so we cannot rule out this result yet because it could be that M2s are not repressed by the IPhane7 genes. We will not know until an M2 immunity test panel can be performed. Gene 1 with the gene 2 intergenic region acting as the promoter region had a repressive effect, however one that was not as intense as gene one with the gene 1-2 intergenic region.

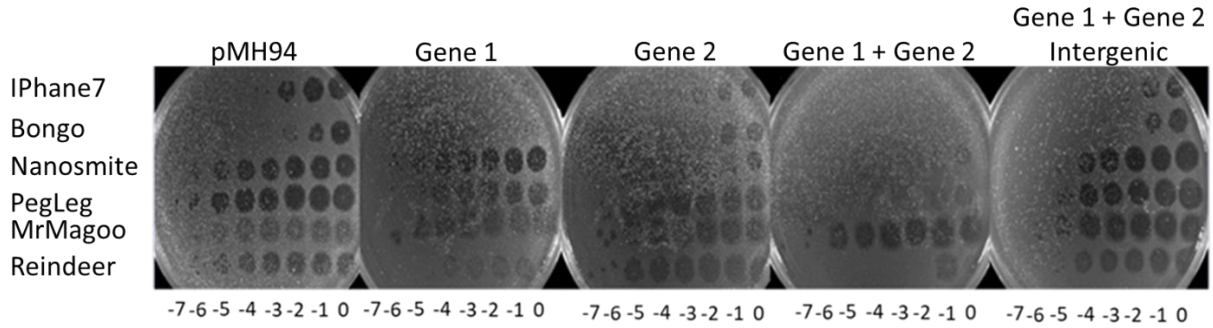


Figure 3. *Representative plates from each condition used in experiment 1. Each test was performed in triplicate. This experiment was performed using serial dilutions from $5*10^{8}$ PFU to $5*10^{1.}$*

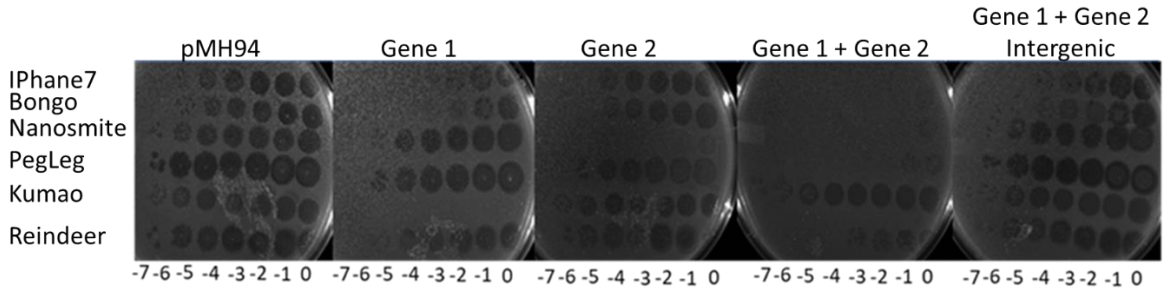


Figure 4. Pictures of a representative plate from each condition used in experiment 5. Each test was performed in triplicate. This experiment was performed using serial dilutions from 10^{-8} PFU to 10^{-1}

Table 12. *Table recording the average viral titer in PFU/mL and efficiency of plating (EOP) for each virus and experimental condition used in the fifth homo-immunity test.*

Phage	pMH94	Gene 1	Gene 2	G1 + G2	G1 + G2 Int	G1 EOP	G2 EOP	G1 + G2 EOP	G1 + G2 Int EOP
IPhane7 (M1)	$5.67E^{+08}$	$1.17E^{+05}$	$2.50E^{+08}$	$0.00E^{+00}$	$5.00E^{+07}$	$2.06E^{-04}$	$4.41E^{-01}$	$0.00E^{+00}$	$8.82E^{-02}$
Bongo (M1)	$4.00E^{+08}$	$8.33E^{+04}$	$1.33E^{+08}$	$1.00E^{+08}$	$2.00E^{+08}$	$2.08E^{-04}$	$3.33E^{-01}$	$2.50E^{-05}$	$5.00E^{-01}$
Nanosmite (M3)	$1.83E^{+09}$	$1.00E^{+09}$	$3.33E^{+04}$	$1.00E^{+04}$	$2.17E^{+09}$	$5.45E^{-04}$	$1.82E^{-05}$	$5.45E^{-06}$	$1.18E^{+00}$
Pegleg (M1)	$2.20E^{+09}$	$6.67E^{+08}$	$3.00E^{+08}$	$6.67E^{+04}$	$2.00E^{+09}$	$3.03E^{-01}$	$1.36E^{-01}$	$3.03E^{-05}$	$9.09E^{-01}$
Kumao (Singleton)	$5.00E^{+09}$	$4.00E^{+09}$	$3.00E^{+09}$	$3.00E^{+09}$	$3.33E^{+08}$	$8.00E^{-01}$	$6.00E^{-01}$	$6.00E^{-01}$	$6.67E^{-01}$
Reindeer (M1)	$1.33E^{+09}$	$2.67E^{+08}$	$2.00E^{+09}$	$2.17E^{+06}$	$3.50E^{+06}$	$2.00E^{-01}$	$1.50E^{+00}$	$1.63E^{-03}$	$2.63E^{-01}$

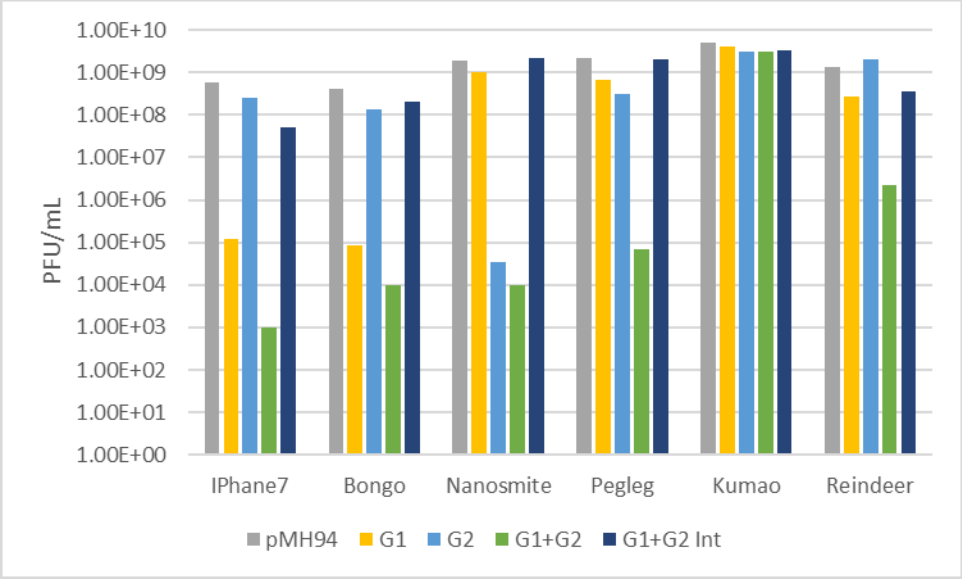


Figure 5A

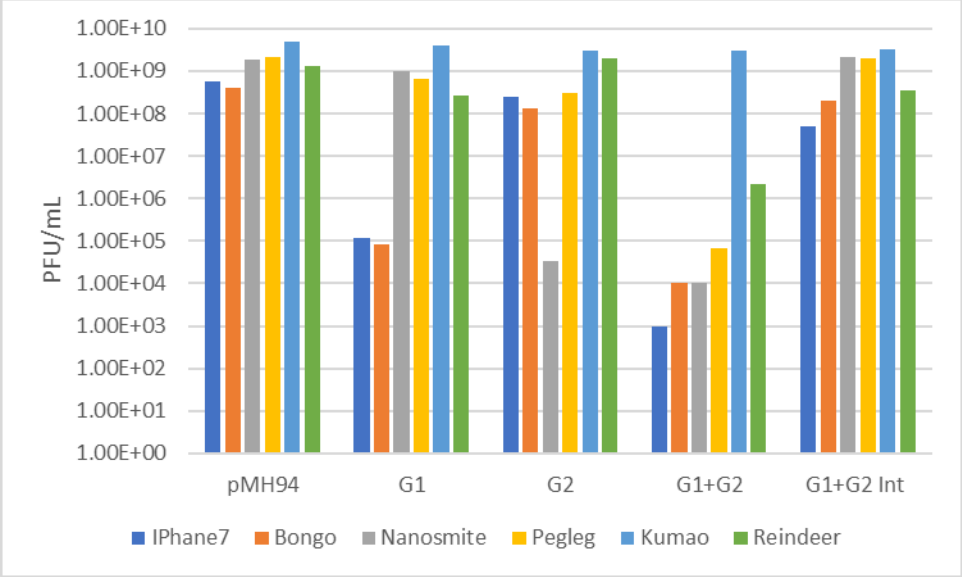


Figure 5B

Figure 5. *Average viral titer for each experimental condition used in experiment 5 is shown A. grouped by phage and B. grouped by gene. pMH94 and Kumao acts as controls. IPhane7 and Bongo are heavily repressed by the presence of gene 1. Nanosmite is heavily repressed in the presence of Gene 2. IPhane7, Bongo, Nanosmite, and Pegleg are all heavily repressed by the presence of both gene 1 and gene 2. Gene 1 with the gene 2 intergenic region is less effective than gene 1 with its indigenous intergenic region. Reindeer is not repressed much by anything, with the exception of gene 1 and gene 2 together.*

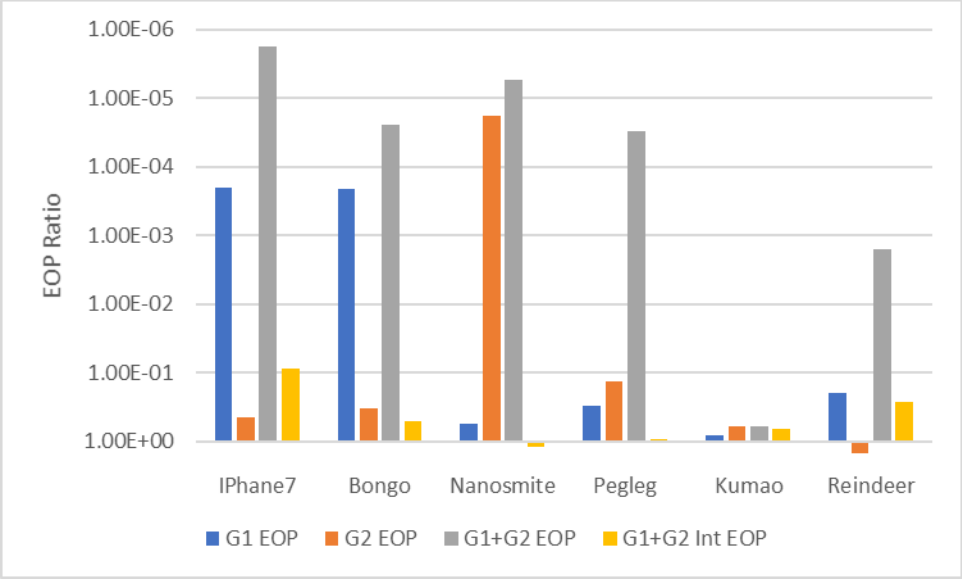


Figure 6. Average Efficiency of Plating for each experimental condition used in experiment 5 is shown. The EOP ratio is represented in a negative logarithm, the smaller the ratio, the larger the bar on the chart. IPhane7 lytic capabilities were repressed most significantly by gene 1 and gene 2 together, significantly by gene 1, and less significantly by gene 1 with the gene 2 intergenic region. Bongo lytic capabilities were repressed most significantly by gene 1 and gene 2 together, and significantly by gene 1. Nanosmite lytic capabilities were repressed very significantly by gene 2 alone and gene 1 + gene 2. Pegleg lytic capabilities were repressed significantly by gene 1 + gene 2. Reindeer lytic capabilities were repressed by gene 1 + gene 2 but not as significantly as the other phages. Kumao served as a control and was unaffected by the presence of any IPhane7 genes.

Discussion

The results of the single copy homo-immunity testing helped to confirm that gene 1 is the repressor gene due to its repression of IPhane7. Also, gene 1's repression of the virus most similar to IPhane7 (Bongo) is an indicator of the homo-immunity phenotype characteristic of a repressor gene. The lack of effect of this gene on other M1 phages Pegleg and Reindeer discourages this idea somewhat, however possible reasons for this discrepancy are explored further in bioinformatics. Gene 2 has the effect of minor repression of M1's but completely represses the M3 Nanosmite. This phenotype is not currently explained. Gene 1 + gene 2

repressed all M1's and the M3 more effectively than either gene 1 or gene 2 alone. This phenotype could be explained by several different things and further experimentation is needed to assign a cause for this phenotype. One hypothesis for the gene 1 + gene 2 phenotype was that gene 1 is getting expressed twice, once by the promoter in the gene 1-2 intergenic region and once in the gene 2-3 intergenic region. If the strength of the effect on gene 1 and gene 1 with gene 2 intergenic are equal to the effect of gene 1 + gene 2 there could be a strong likelihood this hypothesis is correct.

Another possibility is that gene 2 and gene 1 products form a complex that is more effective than either one by themselves. If this is the case further testing with a florescent polarization assay could show that a complex is being formed by measuring the speed of the spin of fluorescently labeled DNA molecules in the presence of the gene 1 protein, gene 2 protein, and both gene 1 and gene 2 proteins. If the fluorescently labeled DNA in the presence of both gene 1 and gene 2 had a slower spin than gene 1 alone it would indicate a complex has been formed between proteins 1 and 2.

Several issues came up in these experiments. In the first four experiments IPhane7 and Bongo were not propagating to the $5*10^8$ PFU/ml on pMH94 only plates. This issue influenced the results and made it difficult to find an accurate EOP. The $CaCl_2$ used in the first 4 experiments had been autoclaved. Once this $CaCl_2$ was replaced with $CaCl_2$ that was filter sterilized this issue was resolved.

The phage labeled MrMagoo did not behave as expected in that it was not affected by gene 1 or gene 2 in any significant way. This could be due to the phage being mislabeled and in fact not being MrMagoo. This could also be due to the phage being a mutated version of MrMagoo, one that is not affected by gene 1 or gene 2. Another possibility is that M2s such as

MrMagoo are not similar enough to IPhane7 for IPhane7's repressor gene to have an immunity effect on it. MrMagoo has both gene 1 and gene 2, however the Gene 1 sequences are less similar than the gene 1 sequences between subcluster M1 phages. The gene 2 similarity between M1 phages and M2 phages is somewhat lacking, while they are similar enough to be grouped into the same family, significant differences can be found, particularly in the C-terminus.

CHAPTER FOUR: BIOINFORMATIC ANALYSIS RESULTS AND DISCUSSION

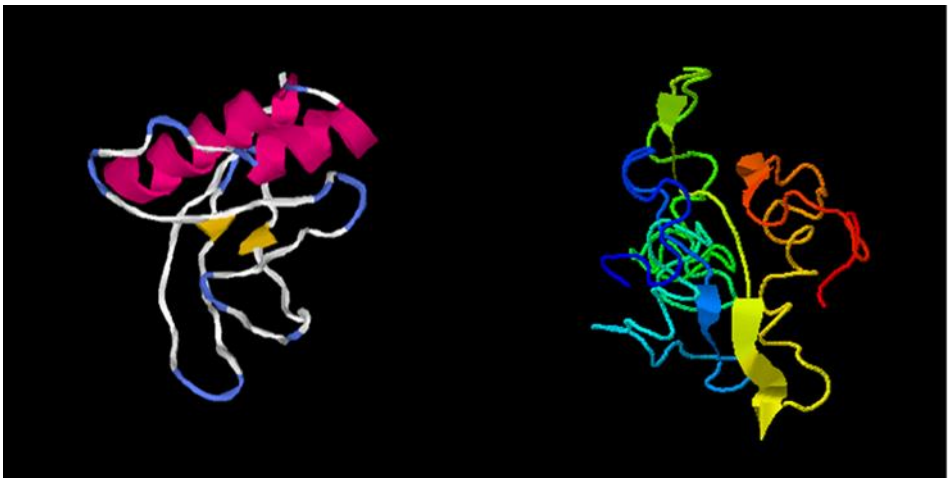
Bioinformatic analysis was used to find information about IPhane7s suspected repressor gene through the use of protein prediction software, amino acid alignments and overlays with known repressor genes. This was done to add further confidence to the identification of gene 1 being the repressor gene as well as to understand why the homo-immunity results varied from phage to phage within the M1 subsluster. Information about possible binding site for the repressor gene were also searched for using terminator, promoter, and conserved repeat searches. This was done to narrow down possible binding sites to decrease the time and resources needed to positively identify this site by biochemical methods.

Bioinformatic analysis is performed using the DNA sequence of the IPhane7 genome and accessing databases storing thousands of other genomes (Zheng Zhang et al, 2000). Bioinformatics is used to compare the genome in question, or genes within that genome to those that may have been studied previously, or specific genes found in other genomes that may have been identified previously. Bioinformatic tools such as I-TASSER and Modeller allow for the prediction of protein structure (B. Webb et al, 2016; J Yang et al, 2015). The predicted folds can be searched for in similar proteins in related genomes to determine if the structures are conserved across genomes. Through the bioinformatic tools BLAST and PSI-BLAST the nucleotide and amino acid sequences of IPhane7 genes 1 and 2 were run in the NCBI database to determine the most similar sequences found in other genomes (Zheng Zhang et al, 2000; NCBI, 2004). The similar sequences were then compared to determine which residues are conserved between the different genomes using ClustalW (Thompson JD, et al, 1994). Softberry was used to determine transcriptional terminators in the genome (V. Solovyev et al, 2011). The IPhane7 genome can

also be searched for repeat sequences that could be used in regulation of gene expression using the MEME Suite (Timothy L. Bailey et al, 1994). A bioinformatic tool within the program DNA Master was also used to search the IPhane7 genome for promoter regions. Promoter locations, terminator locations, and conserved repeat locations were used in tandem to determine likely areas the repressor binds in order to repress lytic replication (Pope, Welkin H. et al, 2014).

Results

I-TASSER predicted a helix-turn-helix motif in the gene 1 product. (Indicated in pink in Figure 8A by the two distinct alpha helices in near parallel orientation), which is indictive of a DNA binding protein. I-TASSER predicting a DNA binding motif in gene 1 further indicates that protein 1 is the repressor protein. I-TASSER did not predict any meaningful motifs in IPhane7 protein 2. Modeller is a different protein structure prediction software which also predicted a helix turn helix in IPhane7 gene 1 (Figure 8). The Modeller results come from a one-on-one comparison to the T7 phage repressor gene which has an experimentally confirmed helix-turn-helix motif. The experimentally confirmed Lambda phage repressor protein (figure 9) has a clear Helix-Turn-Helix motif that has a helix construct (helix 2 – helix 3) that closely resembles the motif predicted by I-TASSER and Modeller. This Lambda phage repressor protein helix 2 – helix 3 construct was isolated and superimposed onto the Modeller Helix-Turn-Helix prediction for comparison (Figure 10). The size of the helices as well as the spacing of the helices are remarkably similar to the Lambda Phage helices. The orientation is slightly off, but as the Modeller image is a prediction this is acceptable.



A B

Figure 7A. *Image of g1 structure prediction using I-TASSER. This prediction software predicted a helix-turn-helix structure (indicated in pink). The helix-turn-helix is high confidence (8/10); however, the rest of the protein is low confidence (4/10 or lower). B. Image of g2 prediction using I-TASSER. This prediction had several high confidence structures in a prediction with an overall low confidence, however these structures were not of consequence to DNA binding or protein ligand interactions.*

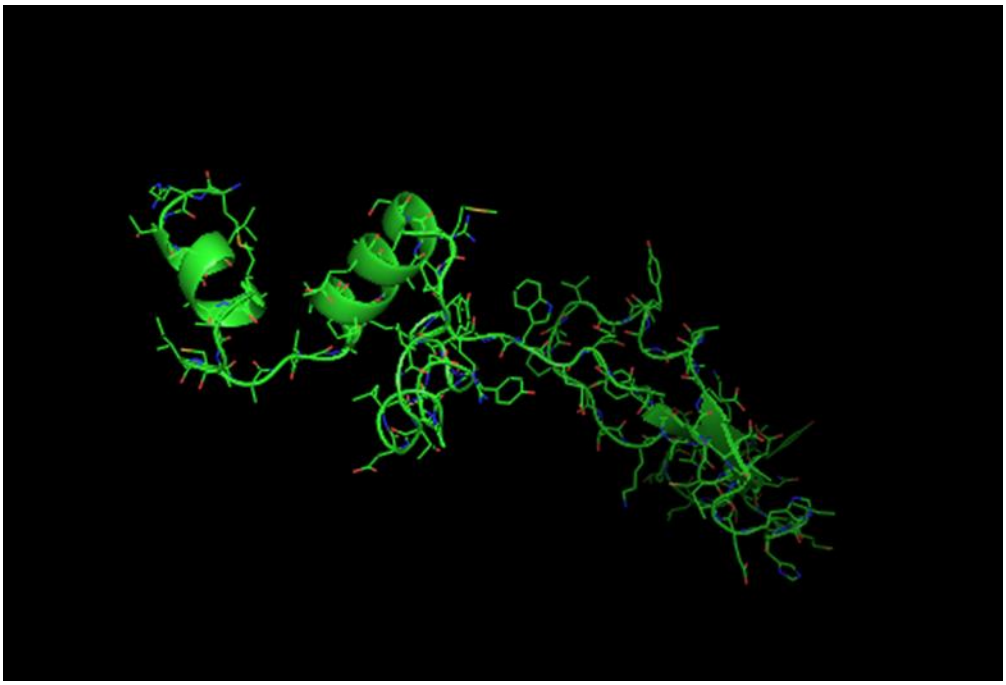


Figure 8. *gene 1 product structure prediction of Modeller software using a known T7 repressor protein as structure for comparison. This independently predicted structure gives credence to the accuracy of the I-TASSER predicted helical structures separated by a small amino acid strand without a secondary structure which appears to be a helix-turn-helix motif.*

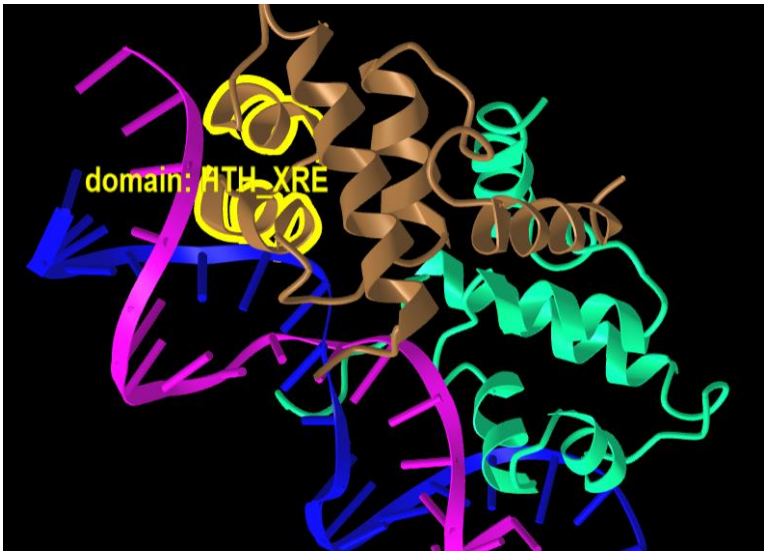


Figure 9. *Illustrated protein structure of the lambda phage repressor protein binding to DNA. This structure was experimentally found using x-ray crystallography (Beamer LJ. et al, 1992). The brown structure is one repressor protein, the green structure is a second repressor protein, the pink and blue structures are DNA strands. The helix 2 and helix 3 regions are highlighted as these two helices correspond to the helices predicted by I-TASSER and Modeller. 3D model from the Molecular Modeling Database (Madej T. et al, 2014)*

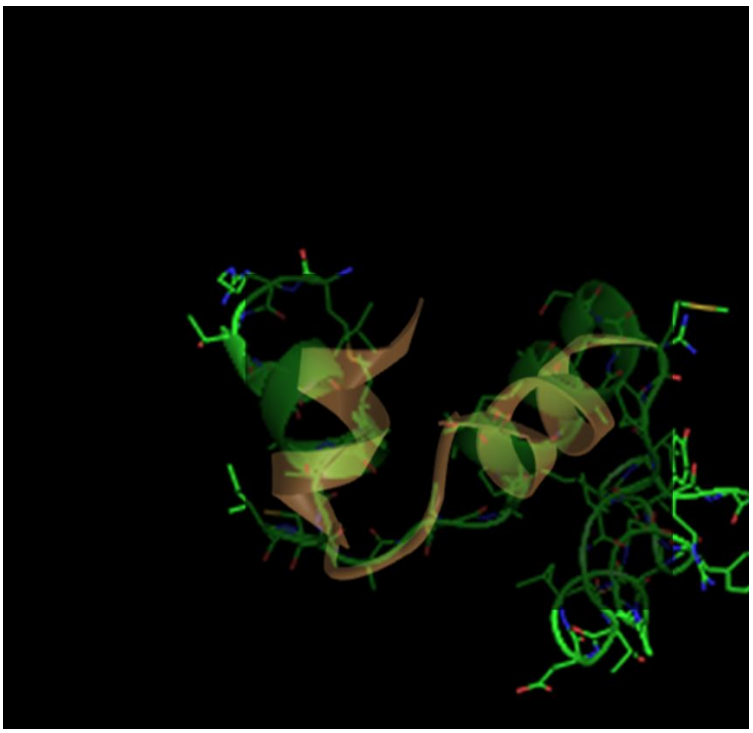


Figure 10. *superimposition of helices 2-3 from Lambda phage repressor over the Modeller predicted helices. The helices are of similar size and distance from each other however the orientation is slightly off.*

A helix-turn-helix motif is conserved in the gene 1 protein across cluster M phages according to I-TASSER predictions of a panel of cluster M phages representative of each subcluster, though the amino acid sequence of the helix-turn-helix seems to be highly conserved only within each individual subcluster (figure 10). While the gene 1 amino acid sequence is highly conserved throughout M1 phages, it is not completely conserved in Reindeer, where there is a slight difference in amino acid sequence within the helix-turn-helix motif that changes the makeup of the motif. This indicates that while gene 1 likely has a similar function throughout the cluster M phages, the DNA binding site of the protein could differ for each of the subclusters M1, M2, and M3. Analysis of amino acid sequences has indicated that IPhane7 gene 1 is remarkably similar to other M1 phages. IPhane7 gene 1 has an analogous gene in the M2 and M3 genomes; however, subcluster M2 and M3 phage's gene 1s are not as similar to IPhane7 gene 1 as the other M1 phage's gene 1's. Using BLAST, the IPhane7 protein 1 has been determined to only have a 42.27% identity to MrMagoo (M2) protein 1, whereas IPhane7 gene 1 had a 100% identity to the corresponding gene in Bongo (M1).

Figure 11. *Amino acid alignment of Cluster M panel. Exact amino acid matches are shaded black, similarities are shaded grey, and dissimilar amino acids are not shaded. Alignment created using ClustalW. Similarities shaded using BoxShade.*

Figure 12. *Cluster M panel of gene 1 amino acid alignment with predicted Helix-Turn-Helix motif highlighted. IPhane7, Bongo, Pegleg, and Reindeer are subcluster M1 phages. The M1 Helix-Turn-Helix predictions are highlighted red. MrMagoo, Rey, and Genevab15 are subcluster M2 phages. The M2 Helix-Turn-Helix predictions are highlighted green. Nanosmite is a subcluster M3 phage. The M3 Helix-Turn-Helix prediction is highlighted blue. The Helix-Turn-Helix motifs followed a general pattern of matching up very well by subcluster, however subcluster M1 phage Reindeer does have significant differences compared to the other M1 phages. The amino acid alignment was created using ClustalW. The Helix-Turn-Helix motifs were predicted using I-TASSER.*

The Helix-Turn-Helix motif predicted by I-TASSER is broken down into 3 distinct roles based on amino acid characteristics, the helix, the turn, and the recognition helix (Table 13A.) this predicted breakdown, however, is not deemed particularly accurate because the turn section

of a Helix-Turn-Helix motif is characterized by a 3 amino acid long sequence of non-polar amino acids, but I-TASSER predicted turns spanning from two to seven amino acids long. The amino acid sequences of the I-TASSER predicted Helix-Turn-Helix motifs were compared to the known Helix-Turn-Helix from the lambda phage repressor gene (Table 13B.) to determine a more accurate breakdown of the helix turn helix components.

Reindeer stands out as having a significantly different helix compared to the other M1 phages. The M2 phages are all similar to each other but differ from the M1 phages in both the helix and the recognition helix. The M3 phage Nanosmite is different from both the M1 and M2 phages in the recognition helix while hiving a similarity to M2 phages in the helix breakdown.

Table 13. *Breakdown of the helix-turn-helix motifs predicted in the cluster M gene 1 proteins. Table 13A is a breakdown of how I-TASSER predicted the Helix-Turn-helix to be arranged however, experimentally confirmed Helix-Turn-Helix motifs are known to have a three amino acid long turn. Table 13B is a more realistic hypothesized Helix-Turn-Helix breakdown based on the experimentally confirmed lambda phage repressor protein's Helix-Turn-helix.*

Phage	Helix	Turn	Recognition Helix
Bongo(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
Pegleg(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
IPhane7(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
Reindeer(M1)	KLFALIAAPIAATAAIAL	AP	TAEARTAMYR
MrMagoo(M2)	VGIAAAALIC	APVAQAD	DAMFR
Rey(M2)	LGIAAAALFF	APVAQAD	DAMFR
GenevaB15(M2)	VGIAAAALIC	APAAQAD	QSMYR
Nanosmite(M3)	VSVAAAALA	LAP	MAQ

A

Phage	Helix	Turn	Recognition Helix
Lambda	QESVADKM	GMG	MGQSGVGALFN
Bongo(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
Pegleg(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
IPhane7(M1)	TTLTALICAAIA	LAP	AAEARSAMYR
Reindeer(M1)	KLFALIAAPIAATAAIA	LAP	TAEARTAMYR
MrMagoo(M2)	VGIAAAALI	CAP	VAQADDAMFR
Rey(M2)	LGIAAAALF	FAP	VAQADDAMFR
GenevaB15(M2)	VGIAAAALI	CAP	AAQADQSMYR
Nanosmite(M3)	VSVAAAALA	LAP	MAQ

B

IPhane7 gene 2 is also remarkably similar to the second gene in the other M1 phages. IPhane7 Gene 2 has an analogous gene present in each of the M2 and M3 phages, however, the entire gene is not conserved in the M2 or M3 genomes. This gene also has strong genetic similarities to a gene found in cluster L genomes. This gene in the L genomes has a truncated C-terminus and is non-functional. The cluster M genes were aligned with these luster L genes to determine if the truncation of M2 and M3 genes would affect the functionality of the gene as it does cluster L genes (Figure 13), Iphane7 gene 2 also has similarities to a domain-containing

protein found in a host of Mycobacterial species though this genes presence in bacterial species could be due to a lysogenic relationship this is explored in Table 14.

Using an amino acid alignment of a panel of phages including IPhane7 (M1), MrMagoo (M2), Nanosmite (M3), JoeDirt(L1), Snenia(L3), Faith(L2), DyoEdafos(L4), and Kumao (singleton), the portions of the IPhane7 gene 2 sequence that is conserved was discovered. It was discovered that JoeDirt, Snenia, Faith, DyoEdafos, and Kumao all have truncated C-termi (Erin previously discovered that in the L2s the C-terminus is truncated, and when this C-terminal extension of IPhane7 gene 2 is removed, then the gene is no longer functional). MrMagoo and Nanosmite both have very slightly truncated C-termi, however further experimentation is needed to determine if these proteins have lost functionality.

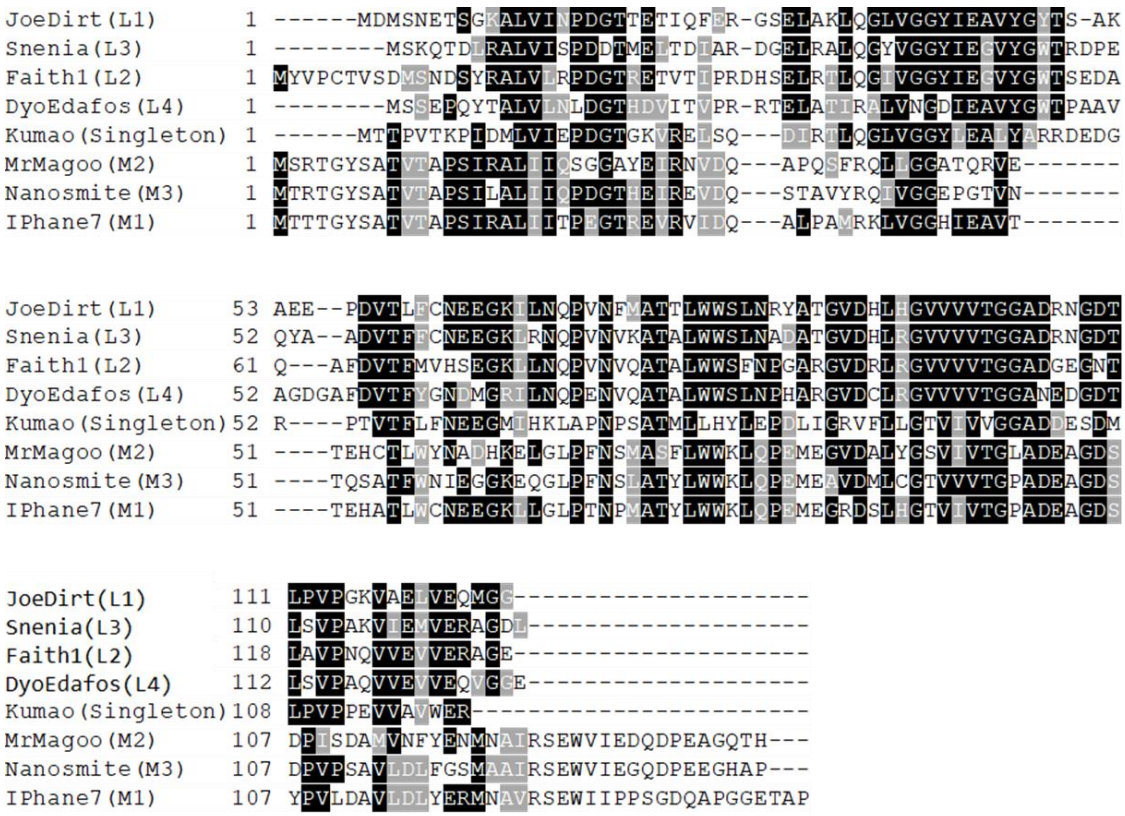


Figure 13. *IPhane7 gene 2 comparison to similar genes. Erin Cafferty found that the cluster L phages genes in the same family as IPhane7 gene 2 were rendered non-functional by the truncation in the C-terminus. This alignment compares the truncation seen in the cluster L phages to the smaller truncation seen in the M2 and M3 subclusters. While there is an amino acid elimination in M2 and M3 gene 2 it is not as substantial as the deletion seen in the L phages, it is unclear if this smaller truncation would render the protein non-functional. The Alignment was created using ClustalW. The shading of similar genes was done using BoxShade.*

An amino acid alignment of the all the cluster M phages was necessary to determine which phages have gene 2 genes different enough from Iphane7 gene 2 to have a different reaction than IPhane7 to the presence of IPhane7 gene 2. To condense this cluster M alignment into a more palatable size, an amino acid alignment of each subcluster was performed to then group the phages into representative groups and pick one phage from each group to compare in the broader cluster M amino acid alignment. From the M1 alignment (Figure 14A) 2 groups were chosen, represented by Iphane7 and Reindeer. From the M2 alignment (Figure 14B) 2 groups were chosen represented by MrMagoo and Ray. IPhane7, Reindeer, MrMagoo, and Ray were put

in an alignment alongside the lone M3 phage Nanosmite for a broad cluster M amino acid alignment (Figure 14C). this broad cluster M alignment highlighted the differences in sequence between the different subclusters.

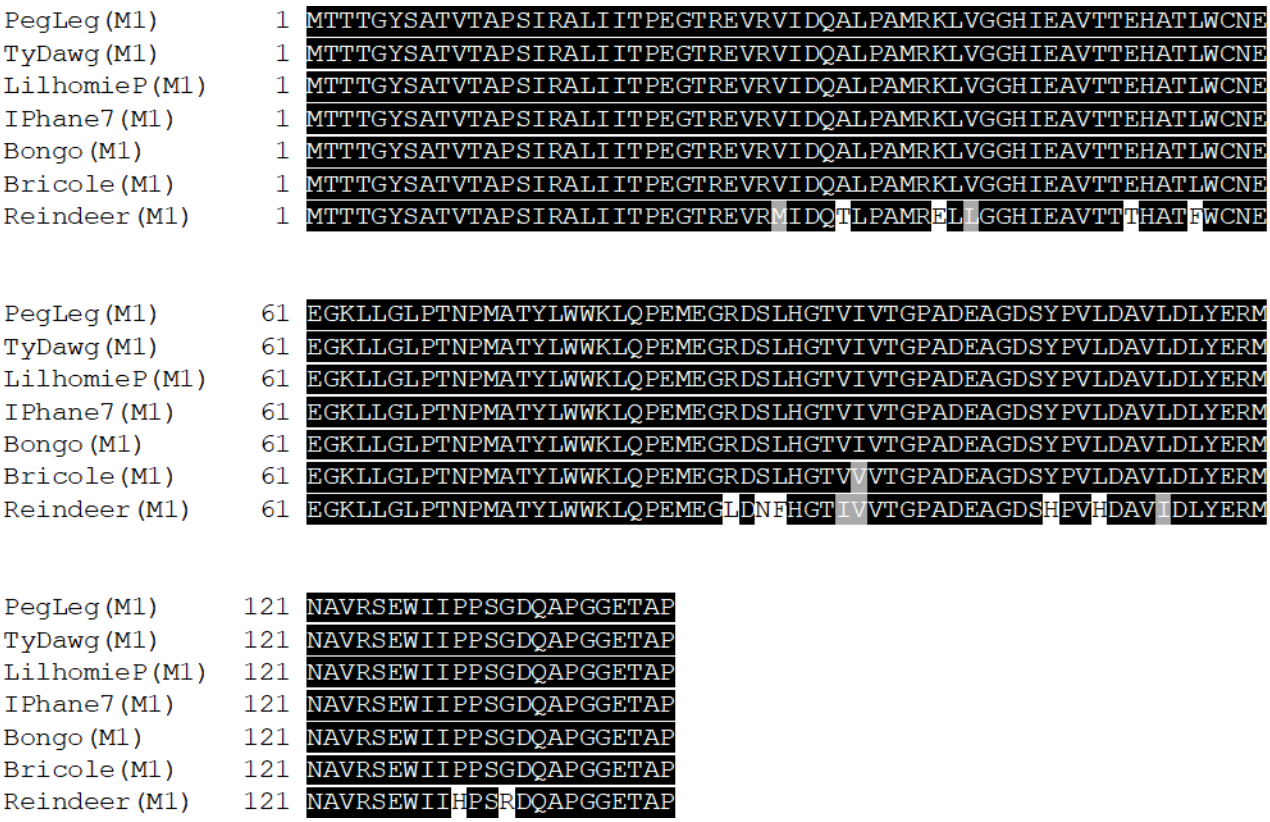


Figure 14A. *Subcluster M1 alignment of gene 2. All M1 phages have identical gene 2 sequences to IPhane7 with the exception of Bricole and Reindeer. Bricole is identical with the exception of 1 amino acid which is similar rather than identical to the corresponding amino acid in IPhane7. Reindeer has significant differences to IPhane7. This amino acid alignment is used to justify using only Iphane7 and Reindeer to represent subcluster M1 in a broader cluster M panel. Alignment created using ClustalW. Similarities shaded using BoxShade.*

Figure 14B. *Subcluster M2 alignment of gene 2. Of the M2 phages MrMagoo, GardenSalsa, GenevaB15, and Aziz all have identical gene 2 sequences. Rey has some significant changes to the gene 2 sequence when compared to the other M2 sequences. This alignment is used to justify using only MrMagoo and Rey in a broader cluster M alignment. Alignment created using ClustalW. Similarities shaded using BoxShade.*

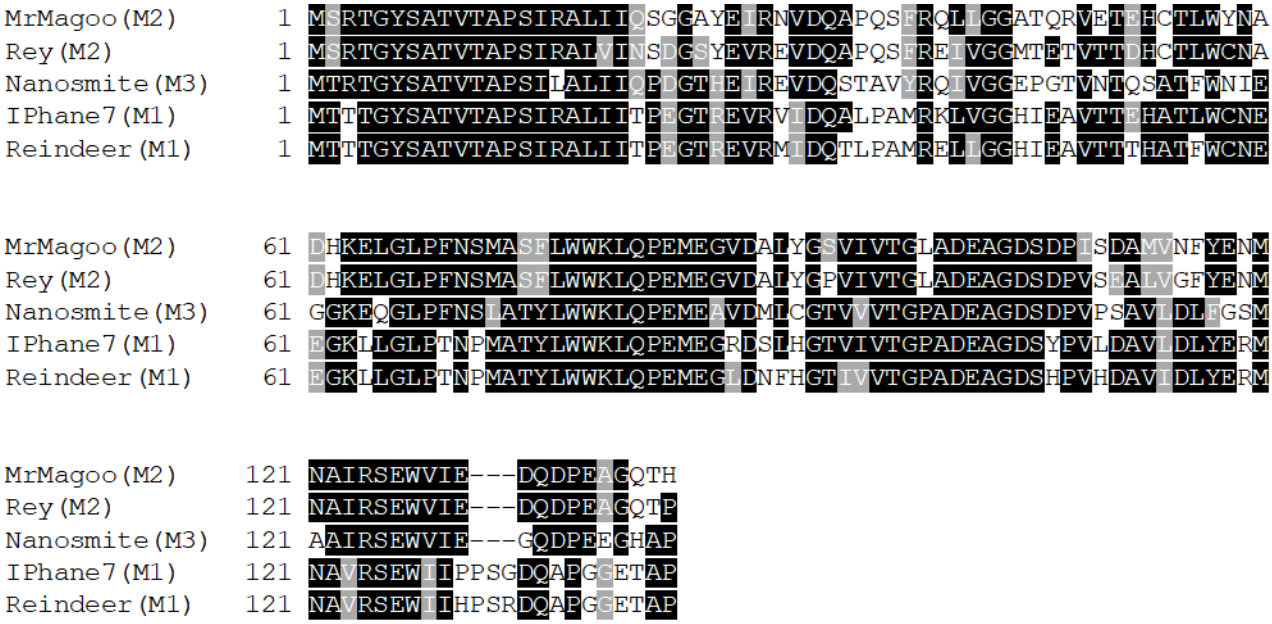


Figure 14C *Cluster M alignment of gene 2. This panel is representative of all cluster M phages because of the similarities within the M1 and M2 subclusters. This alignment highlights how the similarities between sequences become less frequent when comparing across subclusters. These genetic differences could indicate why the M3 phage Nanosmite is affected by lysogens containing IPhane7 gene 2. Alignment created using ClustalW. Similarities shaded using BoxShade.*

IPhane7 genes 1 and 2 had similarity to genes found in Mycobacterium species. This raised the question of whether these genes originated in a Mycobacterium and were then stolen and utilized by an ancestor of IPhane7, or if the genes were found in the bacterial species due to a prophage being present in the sequenced bacterium. To determine whether or not the genes in question originated from a bacterium or a prophage, the gene map of the bacterium species was studied and the genes on either side of the gene in question were BLASTed in the PhagesDB database. Because the PhagesDB database only contains phage genomes, it is probable that if the genes found in the mycobacterium genome on either side of the gene in question are found in the PhagesDB database, the gene with similarity to IPhane7 gene 1 would be present in the bacterium because a prophage containing the gene was present in the bacterium. Vice versa, if the genes on either side of the gene were not present in the PhagesDB database, it becomes more likely the gene originated in a bacterium and was stolen by an ancestor of IPhane7.

It was determined that because the mycobacterium species containing a gene similar to IPhane7 gene 1, also contained multiple other genes similar to phage genes in a row, this strand contained a prophage when sequenced. The mycobacterium species containing the gene similar to IPhane7 gene 2 (DUF3846) did not contain other genes known in the phage databank. Genes did show up in the BLAST results, but the E-value was so high on these results (1.5 and 3.2 respectively) this was likely due to random chance rather than an evolutionary relationship (an E-value of 1.0 indicates that one might expect to see 1 match with a similar score simply by chance (BLAST, NCBI, 2004)). It was determined that IPhane7 gene 2 was likely a bacterial gene incorporated into the genome by an ancestor of IPhane7 (Hatfull, Graham F. et al, 2010). The genes on either side of the DUF3846 gene found in the bacterial genome were then searched in InterPro to determine the function of these genes in other organisms. The gene upstream of

DUF3846 was Gamma-glutamylcyclotransferase and the gene downstream of DUF3846 was methionine adenosyltransferase. Both of these genes code for enzymes. These two genes are the only identified genes in the operon DUF3846 is found in.

Table 14. *genes found upstream and downstream of the gene-1-like gene found in the bacterial species all genes downstream of the gene New54473.1 are identified phage genes with good E-Values. The gene upstream of New54473.1 is unknown and has a poor E-Value in the phage database PhagesDB.*

Relation to New54473.1	Function in Virus	E-Value of Best BLAST Hit
1 gene upstream	Unknown	0.87
New54473.1	Unknown	$1E^{-9}$
1 Genes downstream	Anti-Repressor	$2E^{-69}$
2 Genes downstream	Excise	$1E^{-6}$
3 Genes downstream	HTH DNA binding protein	$1E^{-21}$
4 Genes downstream	Immunity Repressor	$6E^{-20}$
5 Genes downstream	Integrase	$9E^{-70}$

Table 15. *genes found upstream and downstream of the gene-2-like gene found in the bacterial species the genes both upstream and downstream of DUF3846 are unknown and have poor E-Values in the phage database PhagesDB.*

Relation to DUF 3846	Function in Viruses	E-Value of Best BLAST Hit
1 gene upstream	Unknown	1.5
DUF3846	Unknown	$9E^{-6}$
1 gene downstream	Unknown	3.2

Table 16. *genes found upstream and downstream of the gene-2-like gene DUF3846 found in bacterial species, searched for familial relationships in InterPro (Blum M. et al, 2020)*

Relation to DUF 3846	Function in Bacteria
1 gene upstream	Gamma-Glutamylcyclotransferase
DUF3846	Unknown
1 gene downstream	Methionine Adenosyltransferase

Terminators, promoters, and conserved repeats

Terminators function to stop the transcription of an operon. Softberry's findterm terminator search was used to find potential terminators throughout the genome in both the forward and reverse directions and the locations of each potential promoter was recorded. Softberry's findterm software searches only for rho-independent terminators because rho-dependent terminators are not accurately found bioinformatically. However, rho-dependent terminators are not defined in mycobacterial phages, so this does not affect the search.

DNA Master promoter search was used to find promoters throughout the entire IPhane7 genome in both the forward and reverse directions with a low score threshold of 0.500 to eliminate low probability promoters from occurring in the search and the locations of each potential promoter were recorded. Only sigma-70 promoters were searched for in DNA master because other promoter classifications are ill-defined and cannot be searched for bioinformatically with any confidence. This limitation has led to experimentally located promoters being left out of the promoter search results. It is not known what percentage of promoters are sigma-70 however this promoter search does find potential promoters upstream of the majority of operons in the IPhane7 genome. This lends to the software's usefulness. Even

though not all promoters are found, a large enough percentage of promoters are found to narrow down potential repressor protein binding sites.

Conserved repeats are a regulatory tool used by phages and thought to help ribosomes recognize binding sites and assist in initiating translation (Pope Welkin H. et al, 2014). It is thought that a repressor designed to bind to this same binding site would utilize these conserved repeats in the same way. MEME Suite conserved repeat search was used to find the top 5 repeats best conserved throughout the IPhane7 genome, then used to find the top 5 best conserved repeats for the cluster M phages Bongo(M1), Pegleg(M1), Rey(M2), and Reindeer(M1) (Table 19). The MEME Suite conserved repeat search found that conserved repeats were largely conserved between different genomes within the M1 subcluster, with the exception of Reindeer, which only had 2 matching conserved repeats and 3 similar conserved repeats containing the cores of the conserved repeat found in IPhane7 out of the 5 compared to IPhane7s conserved repeats and Pegleg conserved repeat 4, which contained a core of the IPhane7 conserved repeat 4. Reindeer being the exception matches up to the EOP data from the homo-immunity assay. Rey, the only subcluster M2, had no matching conserved repeats, and 5 similar conserved repeats out of the 5 compared to IPhane7s conserved repeats (Table 20). The E-Values of conserved repeats were recorded from Meme suite, and it was found that only the first conserved repeat of each phage is considered significant.

Using Softberry findterm search data, DNA Master promoter search data, and MEME Suite conserved repeats search data. A potential DNA binding spot for the repressor protein is at the promoter found at BP 60531. This promoter follows a predicted terminator (BP 60277) and has potentially regulatory conserved repeat 1 (BP 60329) present between the terminator and promoter. Another potential DNA binding location could be the promoter at BP 57607 which

has similar characteristics to the promoter at BP 60277 with the exception being the conserved repeat associated with the promoter. The promoter at BP 57607 is associated with conserved repeat 4.

Table 17. *IPhane7 terminator search using Softberry. The transcriptional direction is indicated by the chain direction, the starting base pair is indicated by the start column, the length of the terminator is indicated by the length column, the confidence in the terminator is indicated by the score. A higher score is better, all scores present are acceptable.*

chain	start	Length	score
Forward	31740	37	-17.9
Forward	33753	42	-13.1
Forward	44329	42	-26.9
Forward	54582	45	-29.9
Forward	60277	41	-25.8
Forward	62623	44	-30.1
Forward	69455	44	-13.0

chain	start	Length	score
Reverse	167	58	-3.5
Reverse	1778	42	-23.8
Reverse	33053	35	-16.3
Reverse	52017	34	-13.3
Reverse	67718	51	-26.3
Reverse	69333	55	-35.5
Reverse	70608	33	-19.3
Reverse	74216	35	-18.3
Reverse	78603	27	-15.0

Table 18A. Table 18B.

Promoters found using DNA Master. The transcriptional direction is indicated by the chain direction, the starting base pair is indicated by the start column, the length of the terminator is indicated by the length column, the confidence in the terminator is indicated by the score. A higher score is better, all scores present are acceptable.

Chain	Start	Space	Score
Forward	80755	18	0.678
Forward	57607	17	0.676
Forward	71228	17	0.656
Forward	34046	16	0.640
Forward	70633	17	0.637
Forward	28833	16	0.635
Forward	13952	17	0.634
Forward	46965	16	0.632
Forward	3719	17	0.632
Forward	16379	19	0.623
Forward	66399	15	0.621
Forward	76744	17	0.616
Forward	4060	18	0.609
Forward	80267	18	0.605
Forward	54340	18	0.605
Forward	43025	15	0.605
Forward	80752	17	0.602
Forward	11098	17	0.599
Forward	35195	16	0.597
Forward	59976	17	0.595
Forward	60531	17	0.595
Forward	74866	15	0.594
Forward	6193	16	0.593
Forward	42355	18	0.591
Forward	63462	17	0.589
Forward	28833	17	0.589

Chain	Start	Space	Score
Reverse	80663	16	0.705
Reverse	70208	15	0.658
Reverse	14540	18	0.642
Reverse	58533	17	0.626
Reverse	9893	17	0.625
Reverse	73514	17	0.624
Reverse	46863	15	0.624
Reverse	1184	17	0.619
Reverse	38225	16	0.617
Reverse	5163	17	0.611
Reverse	61743	18	0.609
Reverse	80276	17	0.609
Reverse	23011	18	0.609
Reverse	62628	18	0.609
Reverse	80661	18	0.607
Reverse	71213	19	0.605
Reverse	32748	17	0.601
Reverse	37907	16	0.600
Reverse	19625	17	0.599
Reverse	58668	16	0.598
Reverse	22944	17	0.597
Reverse	4559	17	0.594
Reverse	47263	17	0.593
Reverse	58719	18	0.590

Table 19A. *Conserved repeats found in IPhane7 using MEME suite suite alongside the conserved repeats E-Value determined by MEME suite. In order to be statistically significant a conserved repeat should have an E-value of 0.5 of less. CR 1 is statistically significant. CR2-CR5 are not statistically significant.*

CR#	Sequence	E-value
CR1	CTGACCTGCGATTACAG(A/G)A	$4.2e^{-6}$
CR2	TCCTGTAATCACCC	$1.1e^{1}$
CR3	CACACG(GA)AGAAGGGA	$6.2e^{1}$
CR4	C(T/A)GGTTCGAATCC(A/T)G	$5.3e^{2}$
CR5	GGTTCGAATCC	$1.2e^{6}$

Table 19B. *Conserved repeats found in Bongo using MEME suite alongside the conserved repeats E-Value determined by MEME suite. In order to be statistically significant a conserved repeat should have an E-value of 0.5 of less. CR 1 is statistically significant. CR2-CR5 are not statistically significant.*

CR#	Sequence	E-value
CR1	CTGACCTGCGATTACAG(G/A)A	$4.0e^{-6}$
CR2	TCCTGTAATCACCC	$1.1e^{1}$
CR3	CACACGGAGAAGGGA	$7.1e^{1}$
CR4	C(T/A)GGTTCGAATCC(A/T)G	$5.0e^{2}$
CR5	GGTTCGAATCC	$1.2e^{6}$

Table 19C. *Conserved repeats found in Pegleg using MEME suite alongside the conserved repeats E-Value determined by MEME suite. In order to be statistically significant a conserved repeat should have an E-value of 0.5 of less. CR 1 is statistically significant. CR2-CR5 are not statistically significant.*

CR#	Sequence	E-value
CR1	CTGACCTGCGATTACAG(G/A)A	$4.1e^{-6}$
CR2	TCCTGTAATCACCC	$1.1e^{1}$
CR3	CACACGGAGAAGGGA	$7.2e^{1}$
CR4	TGCTGGTTCGAATCC	$3.8e^{3}$
CR5	A(G/A)ATTGG(T/C)(G/A)A(G/T)(C/G)(T/C)CGC(C/T)TG(G/A)(C/T)TC A(A/T/C/G)AA(C/T)CA(G/A)GAG(G/C)(C/T)(C/T)(G/C)C(G/C)(G/C)GT TC(G/A)A(A/C)TC	$3.1e^{4}$

Table 19D. *Conserved repeats found in Rey using MEME suite alongside the conserved repeats E-Value determined by MEME suite. In order to be statistically significant a conserved repeat should have an E-value of 0.5 of less. CR 1 is statistically significant. CR2-CR5 are not statistically significant.*

CR#	Sequence	E-value
CR1	CTGACCTG(C/T)GATTACAGGA(T/A)(C/G)(G/C)(T/C)GTGA	$1.4e^{-1}$
CR2	TCCTGTAATCACTC(T/C)G	$9.3e^{-2}$
CR3	A(A/C)(T/C)(A/T)CACACGGAGAAGGGAA	$1.7e^{2}$
CR4	(T/C)GCAGGTTCGAATCCTG	$1.5e^{2}$
CR5	AAGAAGAT	$4.0e^{7}$

Table 19E. *Conserved repeats found in Reindeer using MEME suite alongside the conserved repeats E-Value determined by MEME suite. In order to be statistically significant a conserved repeat should have an E-value of 0.5 of less. CR 1 is statistically significant. CR2-CR5 are not statistically significant.*

CR#	Sequence	E-value
CR1	CTGACCTGCGA(T/A)TACAG(G/A)AC	$6.4e^{-5}$
CR2	TCCTGTAATCAC(C/T)CT	$1.9e^{1}$
CR3	ACACACGGAGAAGGA(A/T)A(G/A)A(A/T/C/G)CATG(A/T/C/G)C(A/T)(T/A)A	$3.7e^{2}$
CR4	TGC(A/T)GGTTCGAATCC(T/A)G(A/T/G/C)C	$2.9e^{3}$
CR5	TG(T/G)TTGATGCGCTG(C/A)ACC(G/A)TAAGGC(T/G)GT	$3.1e^{2}$

Table 20A. *Comparison of IPhane7 conserved repeat 1 to conserved repeat 1 of each phage. Differences from IPhane7 are highlighted in red.*

IPhane7	CTGACCTGCGATTACAG(A/G)A
Bongo	CTGACCTGCGATTACAG(G/A)A
Pegleg	CTGACCTGCGATTACAG(G/A)A
Rey	CTGACCTG(C/T)GATTACAGGA(T/A)(C/G)(G/C)(T/C)GTGA
Reindeer	CTGACCTGCGA(T/A)TACAG(G/A)AC

Table 20B. *Comparison of IPhane7 conserved repeat 2 to conserved repeat 2 of each phage. Differences from IPhane7 are highlighted in red.*

IPhane7	TCCTGTAATCACCC
Bongo	TCCTGTAATCACCC
Pegleg	TCCTGTAATCACCC
Rey	TCCTGTAATCACTC(T/C)G
Reindeer	TCCTGTAATCAC(C/T)CT

Table 20C. *Comparison of IPhane7 conserved repeat 3 to conserved repeat 3 of each phage Differences from IPhane7 are highlighted in red.*

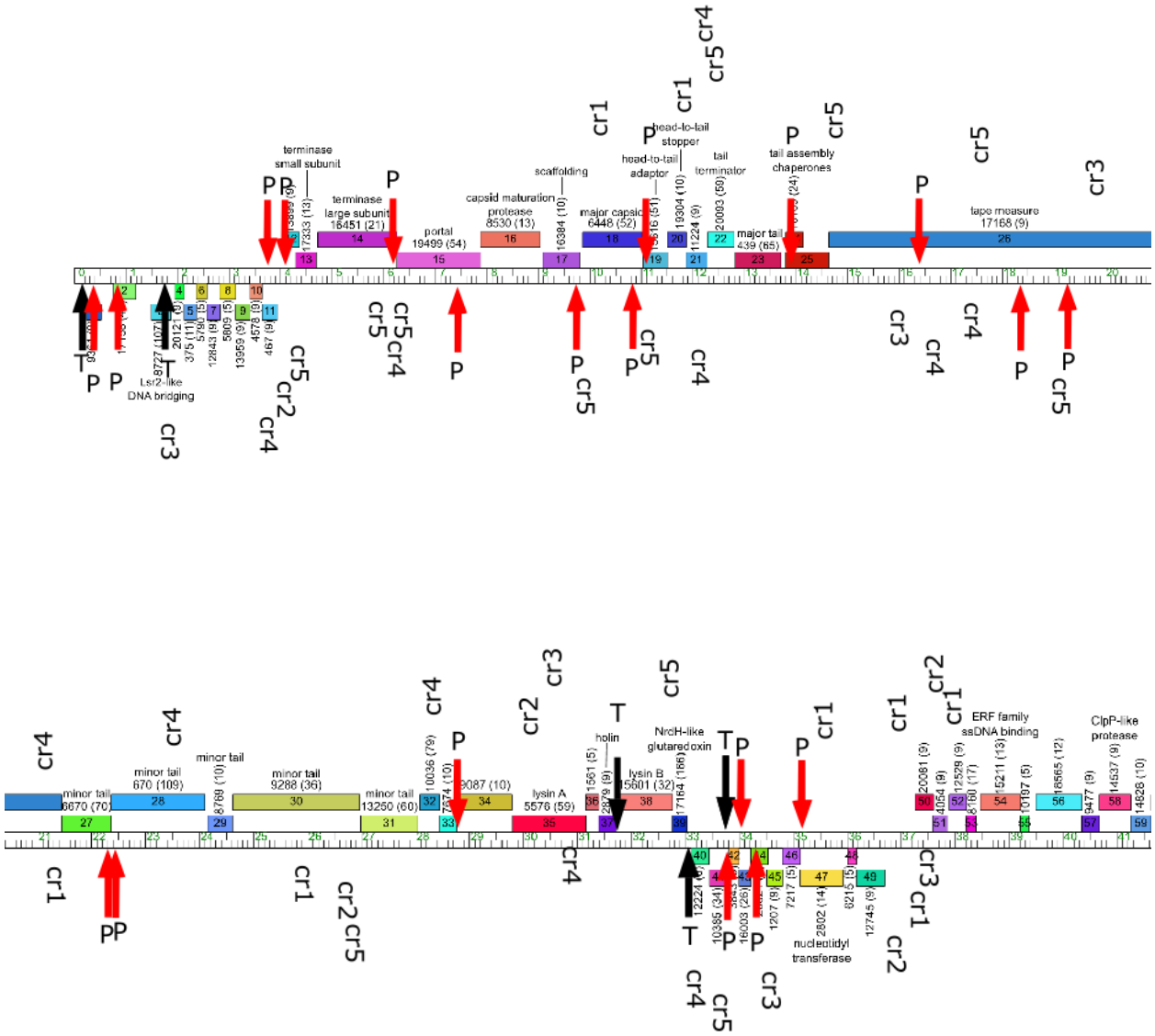
IPhane7	CACACG(G/A)AGAAGGGA
Bongo	CACACGGAGAAGGGA
Pegleg	CACACGGAGAAGGGA
Rey	A(A/C)(T/C)(A/T)CACACGGAGAAGGGAA
Reindeer	ACACACGGAGAAGGA(A/T)A(G/A)A(A/T/C/G)CATG(A/T/C/G)C(A/T)(T/A)A

Table 20D. *Comparison of IPhane7 conserved repeat 4 to conserved repeat 4 of each phage. Differences from IPhane7 are highlighted in red.*

IPhane7	C(T/A)GGTTCGAATCC(A/T)G
Bongo	C(T/A)GGTTCGAATCC(A/T)G
Pegleg	TGCTGGTTCGAATCC--
Rey	CAGGATTCGAA-CCTGC(G/A)
Reindeer	TGC(A/T)GGTTCGAATCC(T/A)G(A/T/G/C)C

Table 20E. *Comparison of IPhane7 conserved repeat 5 to conserved repeat 5 of each phage. Differences from IPhane7 are highlighted in red.*

IPhane7	GGTTCGAATCC
Bongo	GGTTCGAATCC
Pegleg	A(G/A)ATTGG(T/C)(G/A)A(G/T)(C/G)(T/C)CGC(C/T)TG(G/A)(C/T)TCA(A/T/C/G)AA(C/T)CA(G/A)GAG(G/C)(C/T)(C/T)(G/C)C(G/C)(G/C)GTTC(G/A)A(A/C)TC-
Rey	ATCTTCTT
Reindeer	TG(T/G)TT-GA-TGCGCTG(C/A)ACC(G/A)TAAGGC(T/G)GT


terminase small subunit
terminase large subunit 16451 (21)
portal 19499 (54)
capsid maturation protease 8530 (13)
scaffolding
major capsid 6448 (52)
head-to-tail adaptor
head-to-tail stopper
tail terminator
tail assembly chaperones
major tail 439 (65)
tape measure 17168 (9)
Lsr2-like DNA bridging
holin
lysin A 5576 (59)
lysin B 15601 (32)
NrdH-like glutaredoxin
minor tail
minor tail 670 (109)
minor tail 9288 (36)
minor tail 13250 (60)
minor tail 6670 (70)
ERF family ssDNA binding
ClpP-like protease
nucleotidyl transferase

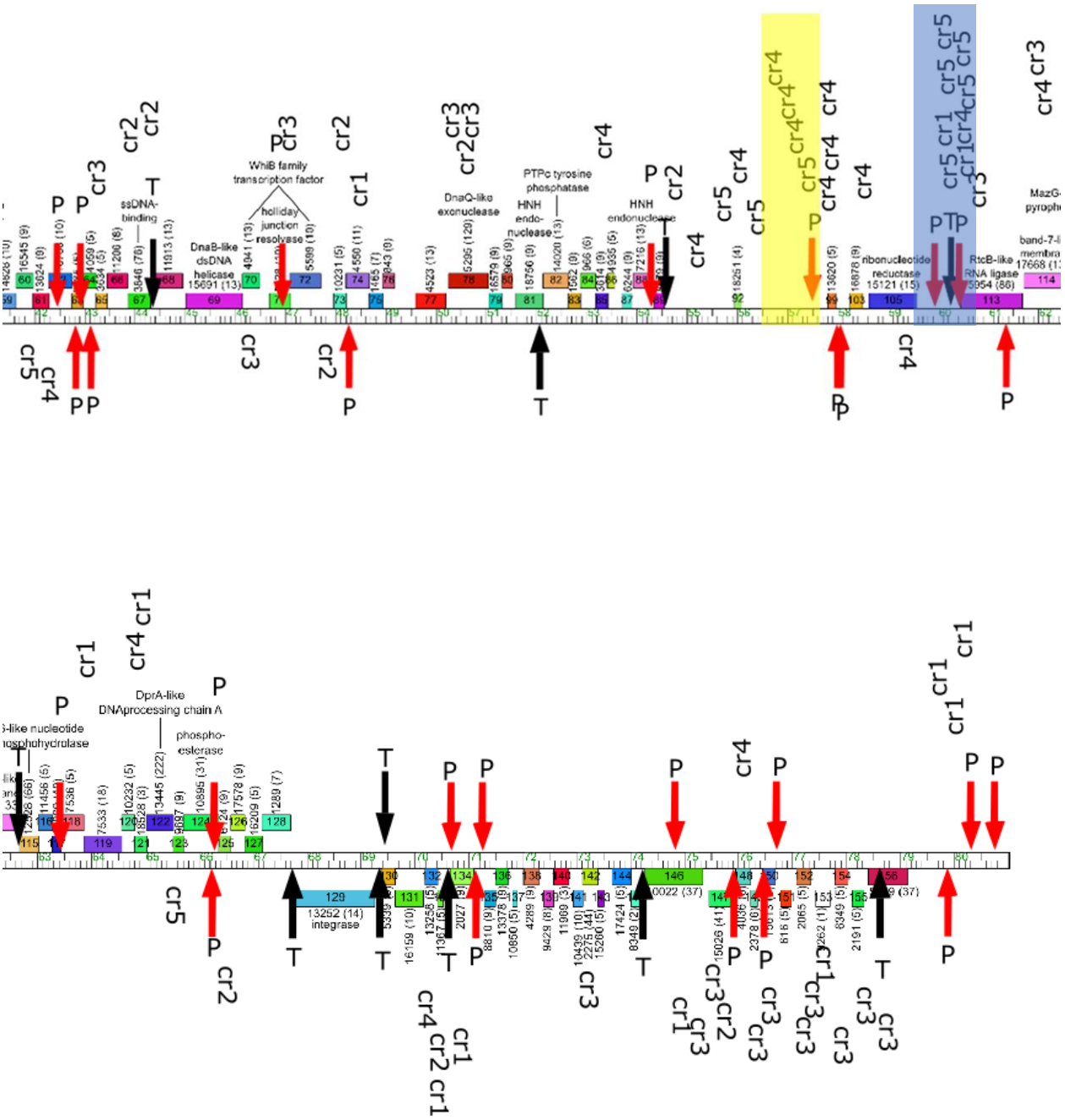


Figure 15. *Gene map of IPhane7 with predicted promoters, predicted terminators, and predicted conserved repeats 1,2,3,4,5 with predicted conserved repeats and predicted binding promoter highlighted in blue. The secondary binding promoter is highlighted in yellow.*

Discussion

Using I-TASSER, the structures of gene 1 and gene 2 were predicted. Interestingly, I-TASSER predicted a helix-turn-helix site on gene 1 protein. This helix-turn-helix motif has been shown to be a DNA binding site in other proteins. Repressors must bind DNA to function, and the fact that I-TASSER predicted a helix-turn-helix motif supports that gene 1 could be the repressor. I-TASSER was used rather than Phyre2 because Phyre2 was not able to predict a structure at all when the amino acid sequence was submitted. The helix-turn-helix motif was a high confidence prediction in I-TASSER. I-TASSER has a 0-9 confidence score on each individual amino acid, all amino acids associated with the 2 helices predicted had a confidence score of 7-9. However, because only one of the two prediction programs predicted a helix-turn-helix, a third prediction program, Modeller, was used to add confidence to the helix-turn-helix prediction by I-TASSER. Modeller uses the data from specific known protein structures to predict the structure of the target protein. When Modeller used the repressor gene structure from the T7 phage as a reference to predict the structure of IPhane7 protein 1, a clear helix-turn-helix was predicted. The T7 phage's repressor protein was used as a reference for Modeller because the T7 repressor protein has had its structure confirmed via x-ray crystallography and is therefore an extremely high confidence model. I-TASSER and Phyre2 did not predict any meaningful motifs related to DNA binding in the gene 2 protein. A structural prediction of IPhane7 gene 2 was not attempted using Modeller because Modeller uses known structures to form its predictions, and there were no relatives of gene 2 that have had the structure identified and uploaded to the database that were found. The helix2-helix3 construct of the Lambda phage repressor has a close structural similarity to the I-TASSER and Modeller predictions of IPhane7 protein 1 Helix-Turn-Helix. The superimposed image created using this helix2-helix3 construct

and the Modeller prediction adds credence to the hypothesis that the helical construct predicted by Modeller is indeed a Helix-Turn-Helix which in turn adds credence to the hypothesis that IPhane7 gene 1 codes for the repressor.

With a helix-turn-helix motif predicted in IPhane7 gene 1, the protein prediction software was used to determine if the helix-turn-helix was predicted in the other cluster M phages. I-TASSER found that a helix-turn-helix is predicted in each of the cluster M gene 1 proteins, however while the helix-turn-helix structure is conserved, the amino acid sequence responsible for the helix-turn-helix is different for each subcluster. Reindeer is also shown to have a slightly different helix-turn-helix site from other M1 gene 1 proteins, and this difference is seen phenotypically in the lack of homo-immunity to reindeer in the homo-immunity assay discussed previously. This suggests that while protein 1 in each cluster M phage likely has the same role of repressor, the DNA binding site may be different for the different subclusters.

IPhane7 gene 2 has similarities to the second gene in the other M1 phages. IPhane7 Gene 2 has an analogous gene present in each of the M2 and M3 phages; however, the genetic sequence is not perfectly conserved in the M2 or M3 genomes. This gene also has strong genetic similarities to a gene found in cluster L genomes, as well as a domain-containing protein DUF3846 found in a host of Mycobacterial species. The operon in which the gene was found had 2 identified genes coding for Gamma-glutamylcyclotransferase and Methionine adenosyltransferase as well as a number of hypothetical proteins of unknown function. Both of these known proteins are transferases and catalyze specific chemical reactions. The proximity of these enzymes to DUF3846 could potentially indicate that this gene has a related function, but it is not clear how this type of function would induce an immune response in lysogens. The M2 representative MrMagoo's gene 2 has a small 3 amino acid truncation at the C terminus, as does

the M3 representative Nanosmite. It remains to be seen if this 3 amino acid truncation has an effect on the functionality of the protein. The similar genes in cluster L were found to have a much larger 20-21 amino acid truncation to the C terminus. The truncation in the cluster L genes have rendered the associated proteins non-functional according to Erin Cafferty's data. The singleton Kumao also has a similar amino acid sequence to IPhane7 gene 2, and similarly to the cluster L genes, Kumao has a 24 amino acid truncation. If this 3 amino acid truncation is found to render Nanosmite gene 2 non-functional this could explain through trait loss, why the presence of IPhane7 protein 2 has such a drastic effect on the lytic activity of Nanosmite in the homo-immunity assay. If IPhane7 gene 2 is, as suggested by PSI-BLAST data, a gene borrowed from a bacterium, but also, as suggested by homo-immunity assay data, has a roll in lysogen defense, it could be that cluster M phages have developed a resistance to the effects of this protein, and Nanosmite, having rendered the gene non-functional has lost its resistance to the protein.

Similar genes to both IPhane7 gene 1 and IPhane7 gene 2 are found in other actinobacterial species. To determine if either of these genes originated in bacteria, the genomes of bacterial strands containing the genes in question were reviewed. The sequences of the genes on either side of the gene in question were run through a phage database BLAST. For the mycobacterium gene similar to gene 1, multiple genes with high similarity to known phage genes were found back-to-back directly downstream of the gene in question. This indicates that the gene 1 analog found in the mycobacterium is due to the bacterium having a phage genome integrated into the bacterial genome. The gene 2 analog found in a mycobacterium was found to be surrounded by genes unknown to the phage archive. This indicates that the IPhane7 gene 2 analog is not found in the mycobacterium as a result of the mycobacterium becoming a

prophage. Rather, this indicates that IPhane7 gene 2 originated in a bacterial species, and somewhere in the evolutionary path of IPhane7, an ancestor integrated the bacterial gene into its own genome.

A promoter search using Softberry, a terminator search using DNA Master, and a conserved repeat search using MEME suite has been conducted on the IPhane7 genome and the findings were plotted on a gene map. Promoters are the areas of the genome where proteins bind to initiate transcription. In order to stop transcription of the DNA necessary to start lytic reproduction the repressor binds to a promoter. Bioinformatically finding terminators also helps narrow down the possible locations of DNA binding by the repressor gene because the promoter search is imperfect, and several promoters will not be found in the search. However, when an operon ends with a terminator, a promoter must be present prior to the next gene. Conserved repeats have regulatory roles within the genome (Pope, Welkin H. et al, 2014), it is therefore important to use the locations of these conserved repeats in determining the repressor binding site. Conserved repeats 1, 2, and 3 are conserved throughout other cluster M phages. Conserved repeat 2 is conserved between gene 72 and 73 in IPhane7 and Bongo, this conserved repeat is also conserved in this location in Pegleg between genes 71 and 72, and genes 80 and 81 in Rey. MEME suite E-Values indicate that conserved repeats 2-4 were insignificant findings, because of this conserved repeat regulatory roles were only evaluated using conserved repeat 1. Even so, the other discovered conserved repeats remain in the results because these results match up to the conserved repeats found by the Pope team in *Cluster M Mycobacteriophages Bongo, PegLeg, and Rey with Unusually Large Repertoires of tRNA Isotypes*, in which the team found the conserved repeats 1-3 to be significant findings.

Conserved repeats have regulatory roles, so the fact that the best conserved repeat is conserved throughout and has a low E-Value indicates that the protein 1 helix-turn-helices of each cluster M phage could use the same conserved repeat to find its binding site. BLAST data of the gene 1 sequence alongside I-TASSER protein predictions indicate that IPhane7, Bongo, and Pegleg have identical gene 1 sequences and identical helix-turn-helix motifs, while Reindeer has a similar but not identical gene 1 sequence and helix-turn-helix motif. Rey, belonging to the M2 subcluster, has the least similar gene 1 sequence and helix-turn-helix motif. The homo-immunity assay shows that while IPhane7 and Bongo have similar reactions to the presence of Iphane7 protein 1, Pegleg and Reindeer were not affected by the presence of protein 1 to anywhere near the same degree.

Based on the gene map created mapping out the promoters, terminators, genes, and conserved repeats, the most likely location for the repressor gene to bind was chosen as the promoter found at bp 60277. This promoter is found downstream of the previous terminator and has a conserved repeat 1 upstream and a conserved repeat 1 downstream of the promoter site before the next gene. The operon this promoter potentially regulates contains RNA ligase and a DNA Processing Chain. Repressors must bind to operators coding for gene necessary early in the lytic cycle; because these two proteins would be vital to the early lytic replication cycle of the virus if no bacterial proteins were available to hijack, thus indicating that a repressor binding to this promoter would be a viable location for successful lytic repression. This promoter being the binding site of the repressor gene could explain why phage Pegleg, which has an identical repressor gene would not be greatly affected by the presence of IPhane7 gene 1. This stretch of the genome is dissimilar to the parallel stretch of genome in Pegleg, which has a HNH endonuclease inserted into the corresponding location of the predicted promoter.

Though this method of using bioinformatically found conserved repeats, terminators, and promoters to predict binding locations can indicate a strong starting position for testing the DNA binding capabilities of IPhane7 protein 1 it is imperfect, so other possible binding locations have been identified and further biochemical testing will be needed to definitively determine a DNA binding site for protein 1.

CHAPTER FIVE: BIOCHEMICAL CHARACTERIZATION OF THE IPHANE7 GENE 1

RESULTS AND DISCUSSION

Biochemical characterization of the IPhane7 repressor would give insight into how the repressor functions to repress lytic transcription. To perform biochemical characterization, large amounts of gene 1 and gene 2 proteins must be successfully expressed and purified. Characterization of the IPhane7 repressor begins with the expression of IPhane7 protein 1 and protein 2 because protein 1 is believed to be the repressor and protein 2 appears to have a superinfection defense response that improves the effect of protein 1. Expression of Protein 1 occurred in BL21(DE3) *E. coli* cells using the maltose binding protein (MBP) fusion vector PLM303. A solubility test was performed to determine if the protein was expressed and soluble in the nickel buffer solution that would be used in the nickel column during purification. A maltose binding protein is used because at 97 amino acids long, the gene 1 protein is very small, and therefore would likely be difficult to purify using a nickel column. This vector also utilizes a His-6 tag which will be used in binding to the nickel column in the purification process. Once the MBP-target protein is purified, the protein of interest (IPhane7 protein 1 or 2) is then cleaved from MBP with a precision protease along the proteolytic cleavage site. The target protein can then be separated from MBP. The expression of gene 2 must be done using the same technique as was used on gene 1 and then both proteins must be expressed and purified on a large scale.

Results

The preliminary protein 1 solubility test was successful. As seen in Figure 16 a band not present in the sample from before the introduction of the inducing agent is present in the T=3

hours after introduction of inducing agent and the post sonication/centrifugation lanes. These bands have traveled the correct distance from the injection well of the SDS-page gel to indicate a protein of 50-55 kDa. IPhane7 protein 1 fused to MBP has a calculated weight of 52 kDa.

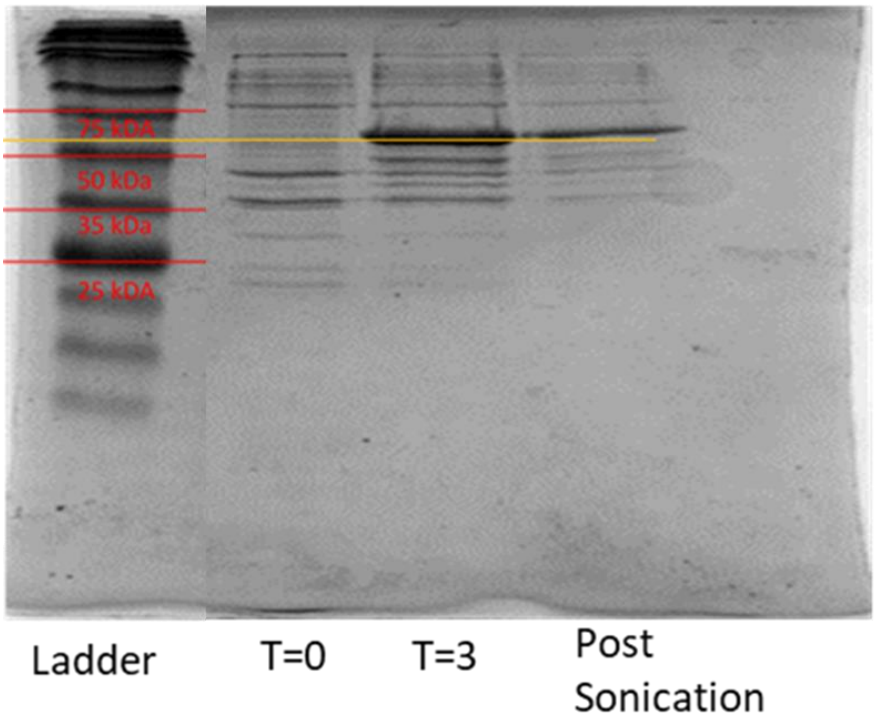


Figure 16. *Gel image of protein 1 with maltose binding protein solubility test. The lane closest to the ladder is loaded with a sample taken at the time 0 of inducing protein production, the lane in the middle is loaded with a sample taken at time 3 hours after inducing protein production, and the rightmost lane is loaded with a sample taken 3 hours after inducing protein production, and after sonication/centrifugation of cells. The band present in the middle and rightmost lanes represent the gene 1 protein. A 15% SDS protein gel was used. The ladder on the left has four weights superimposed by red lines. In ascending order 25kDa, 35kDa, 50kDa, and 75kDa the bands superimposed by the yellow line between the 50kDa and 75kDa ladder markers are at approx. 50-55 kDa.*

Discussion

The gene 1 protein has been successfully expressed and found to be soluble, meaning the protein will likely be able to be purified, However the solubility of the protein when protease

cleaved from the MBP has yet to be tested. The primers to anneal the gene 2 construct to the pLM303 vector have been designed but the IPhane7 gene 2 has not yet been tested for solubility in the nickel buffer.

CHAPTER SIX: FUTURE WORK

If it is found that IPhane7 gene 1 and gene 2 are soluble in solution when the proteins are cleaved from the MBP, biochemical characterization can be used to definitively answer several questions still remaining, the role gene 2 plays in the immune response of its lysogen, the binding site of IPhane7 can be confirmed, the stoichiometry of the binding repressor can be discovered, and the crystalline structure of the repressor protein can be observed.

The role gene 2 has in the immune response can be found with purified gene 2 product using a *M. smegmatis* 2-hybrid system. This involves splitting a transcription factor into two domains, a binding domain and an activating domain. One of these domains is fused to IPhane7 protein 2, and one of these domains is fused to a potential binding partner of IPhane7 protein 2. In close proximity these two domains can function together. If the transcription of a reporter gene also added to the system occurs it is known that protein 2 and the potential binding partner were bound.

The binding site location of IPhane7 protein 1 could be experimentally narrowed down to a specific location using Electrophoretic Mobility Shift Assay (EMSA) one protein 1 is purified. EMSA involves cutting sections of IPhane7 genome, mixing the cut sections with gene 1 protein, and running the solution on an electrophoresis gel. DNA bound to protein will run down the gel slower than free DNA. This means if the IPhane7 genome is evenly divided into sections, the section IPhane7 gene 1 is binding to will clearly have moved a lesser distance than the DNA without protein binding. The section with DNA binding can then be divided into sections and the process repeated. This could be repeated until a specific binding site is discovered.

Once the DNA binding site of the repressor protein is discovered through EMSA the binding stoichiometry can be determined by small angle x-ray scattering. this process does not require crystallization so is not as difficult to perform as x-ray crystallography yet still gives useful information about how the protein binds to DNA. The process of crystalizing IPhane7 protein 1 could be started once the binding site of the protein is discovered through EMSA as well. Upon successful crystallization of the protein, x-ray crystallography could be performed to discover the crystalline structure of the protein when bound to DNA, if the crystals are ordered enough to refract x-rays.

REFERENCES

1. Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

2. J Yang, Y Zhang. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Research, 43: W174-W181 (2015). (PDF and supplementary).

3. Hatfull, Graham F. (2012) The Secret Life of Mycobacteriophages. Adv. Virus Res 82, 179-288, DOI: 10.1016/B978-0-12-394621-8.00015-7

4. Broussard, Gregory W., Oldfield, Lauren M., Villanueva, Valerie M., Lunt, Bryce L., Shine, Emilee E., and Hatfull, Graham F. (2013) Integration-dependent bacteriophage immunity provides insights into the evolution of genetic switches. Mol. Cell 49, 237-248, DOI: 10.1016/j.molcel.2012.11.012

5. Pope, Welkin H., Anders, Kirk R., Baird, Madison, et. al (2014) Cluster M Mycobacteriophages Bongo, PegLeg, and Rey with Unusually Large Repertoires of tRNA Isotypes. J. Virol. 88, 2461-2480, DOI: 10.1128/JVI.03363-13

6. Hatfull Graham F., Jacobs-Sera D, Lawrence JG, et al. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol. 2010;397(1):119-143. doi:10.1016/j.jmb.2010.01.011

7. V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

8. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994. [pdf]

9. Daniel A Russell, Graham F Hatfull, PhagesDB: the actinobacteriophage database, Bioinformatics, Volume 33, Issue 5, 1 March 2017, Pages 784–786, https://doi.org/10.1093/bioinformatics/btw711

10. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.

11. BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2021 03 20]. Available from: https://www.ncbi.nlm.nih.gov/blast/

12. Cresawn, S.G., Bogel, M., Day, N. et al. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics 12, 395 (2011). https://doi.org/10.1186/1471-2105-12-395

13. Gentile, Gabrielle M et al. “More Evidence of Collusion: a New Prophage-Mediated Viral Defense System Encoded by Mycobacteriophage Sbash.” mBio vol. 10,2 e00196-19. 19 Mar. 2019, doi:10.1128/mBio.00196-19

14. Johnson A. D., Poteete A. R., Lauer G., Sauer R. T., Ackers G.K., Ptashne M. lambda Repressor and cro--components of an efficient molecular switch. Nature. 1981 Nov 19;294(5838):217-23. doi: 10.1038/294217a0. PMID: 6457992.

15. Thompson JD, et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res., 1994, vol. 22 (pg. 4673-4680)

16. Falquet L, Bordoli L, Ioannidis V. Pagni, M. Jongeneel C.V. Nucleic Acids Research, 2003 PMID: 12824417

17. Beamer LJ, Pabo CO. Refined 1.8 A crystal structure of the lambda repressor-operator complex. J Mol Biol. 1992 Sep 5;227(1):177-96. doi: 10.1016/0022-2836(92)90690-l. PMID: 1387915.

18. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. " MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res. 2014 Jan; 42(Database issue):D297-303

19. Blum M, Chang H, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, RichardsonL, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A and Finn RD The InterPro protein families and domains database: 20 years on. Nucleic Acids Research, Nov 2020, (doi: 10.1093/nar/gkaa977)

Supplemental Material

Table A1. *conserved repeat 1 in IPhane7* using MEME suite.

Strand	Position	P-value
positive	10036	$8.28e^{-05}$
positive	11797	$4.79e^{-05}$
negative	21192	$4.79e^{-05}$
negative	25927	$1.30e^{-05}$
positive	35524	$1.52e^{-05}$
positive	36958	$4.42e^{-12}$
negative	37008	$3.88e^{-05}$
positive	37848	$1.70e^{-12}$
positive	48309	$4.79e^{-05}$
positive	60329	$1.70e^{-12}$
positive	63830	$1.70e^{-12}$
positive	64993	$4.79e^{-05}$
negative	70187	$7.71e^{-05}$
negative	70568	$4.42e^{-12}$
negative	74778	$1.27e^{-05}$
negative	77377	$2.87e^{-05}$
positive	79936	$1.27e^{-05}$
positive	80108	$2.87e^{-05}$
positive	80682	$7.71e^{-05}$

Table A2. *Conserved repeat 2 in IPhane7 using MEME suite*

Strand	Position	P-value
negative	3969	$2.53e^{-09}$
negative	6013	$1.14e^{-05}$
positive	9876	$1.14e^{-05}$
negative	26575	$2.62e^{-05}$
positive	28498	$6.46e^{-05}$
positive	29800	$2.62e^{-05}$
negative	36719	$3.85e^{-07}$
positive	37218	$2.53e^{-09}$
positive	43778	$1.17e^{-07}$
positive	43880	$2.62e^{-05}$
negative	47645	$1.14e^{-05}$
positive	47918	$2.53e^{-09}$
positive	50196	$2.53e^{-09}$
positive	54627	$4.23e^{-08}$
negative	54866	$6.46e^{-05}$
negative	62140	$6.46e^{-05}$
negative	66156	$2.62e^{-05}$
negative	70415	$2.62e^{-05}$
negative	72884	$6.46e^{-05}$
negative	75399	$6.46e^{-05}$
negative	75422	$2.53e^{-09}$

Table A3. *Conserved repeat 3 in IPhane7 using MEME suite*

Strand	Position	P-value
positive	30359	$5.81e^{-08}$
negative	34541	$5.57e^{-05}$
positive	42944	$5.57e^{-05}$
positive	50232	$2.84e^{-05}$
positive	60578	$1.54e^{-05}$
positive	61289	$5.57e^{-05}$
negative	70433	$2.37e^{-05}$
negative	72833	$6.54e^{-05}$
negative	75354	$2.03e^{-09}$
negative	75876	$2.03e^{-09}$
negative	76364	$1.25e^{-09}$
negative	76436	$1.54e^{-05}$
negative	76993	$1.25e^{-09}$
negative	77060	$1.54e^{-05}$
negative	77352	$9.23e^{-07}$
negative	77712	$1.25e^{-09}$
negative	78043	$1.25e^{-09}$
negative	78343	$3.57e^{-05}$

Table A4. *Conserved repeat 4 in IPhane7 using MEME suite*

Strand	Position	P-value
negative	3465	$2.37e^{-05}$
negative	9461	$7.34e^{-05}$
positive	11848	$6.30e^{-05}$
negative	11933	$2.37e^{-05}$
negative	15529	$7.34e^{-05}$
negative	16594	$1.80e^{-05}$
negative	17233	$2.70e^{-06}$
positive	20966	$6.30e^{-05}$
positive	23009	$1.80e^{-05}$
positive	27772	$1.80e^{-05}$
negative	30209	$9.11e^{-05}$
negative	30497	$1.80e^{-05}$
negative	32884	$1.86e^{-06}$
negative	41842	$3.17e^{-05}$
positive	52969	$1.80e^{-05}$
positive	53855	$9.11e^{-05}$
positive	54828	$3.12e^{-09}$
positive	55309	$4.92e^{-06}$
positive	55748	$7.79e^{-10}$
positive	56579	$3.12e^{-09}$
positive	56661	$2.29e^{-06}$
positive	56956	$4.54e^{-05}$
positive	57414	$4.54e^{-05}$
positive	57540	$7.79e^{-10}$
positive	57699	$1.43e^{-07}$
positive	58051	$7.79e^{-10}$
negative	58131	$7.34e^{-05}$
negative	58996	$2.37e^{-05}$
positive	60173	$2.85e^{-06}$
positive	61431	$1.80e^{-05}$
positive	64008	$2.87e^{-05}$
negative	69950	$2.05e^{-05}$
positive	74161	$7.34e^{-05}$
positive	75823	$2.05e^{-05}$
negative	79046	$7.34e^{-05}$

Table A5. *Conserved repeat 5 in IPhane7 using MEME suite*

Strand	Position	P-value
negative	4222	$8.80e^{-05}$
negative	5659	$2.48e^{-05}$
negative	6073	$8.80e^{-05}$
negative	9568	$7.91e^{-06}$
negative	10839	$8.80e^{-05}$
negative	11714	$2.48e^{-05}$
positive	12044	$8.80e^{-05}$
positive	14402	$8.80e^{-05}$
positive	17166	$8.80e^{-05}$
negative	18963	$7.91e^{-06}$
negative	22154	$8.80e^{-05}$
negative	26703	$8.80e^{-05}$
positive	32429	$8.80e^{-05}$
negative	33411	$8.80e^{-05}$
negative	41580	$8.80e^{-05}$
positive	46949	$3.57e^{-05}$
positive	55349	$2.23e^{-07}$
positive	56221	$7.91e^{-06}$
positive	56903	$2.23e^{-07}$
positive	56980	$7.91e^{-06}$
positive	57104	$7.91e^{-06}$
positive	59636	$2.23e^{-07}$
positive	59834	$2.23e^{-07}$
positive	59987	$7.91e^{-06}$
positive	60064	$7.91e^{-06}$
positive	60250	$8.80e^{-05}$
negative	65243	$2.48e^{-05}$