CHARACTERIZATION OF DEFENSE MECHANISMS ACTINOBACTERIOPHAGES USE TO EVADE THEIR HOST BACTERIA

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 Science in Chemistry.

By

Hannah Bonogafsky

Director: Dr. Jamie Wallen
Associate Dean
College of Arts and Sciences
Chemistry and Physics Department

Committee Members: Dr. Maria Gainey, Chemistry and Physics
Brittania Bintz, Chemistry and Physics
Matt Burleson, Chemistry and Physics

April 2025

ACKNOWLEDGMENTS

I would like to thank Dr. Jamie Wallen for giving me the support, guidance, and encouragement needed to maintain motivation to keep my spirits up. I would like to thank those on my committee who have served in many roles. To Dr. Gainey for always generously giving me the materials I needed when I ran out, to Matt Burleson for giving me instrumentation help while always giving great advice, and to Brittania Bintz for helping keep positive and laugh. Thank you to the Chemistry and Physics department for the supplies needed for this research. Thank you to my friends and family who have helped me keep my spirits high through this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

CHARACTERIZATION OF DEFENSE MECHANISMS ACTINOBACTERIOPHAGES USE TO EVADE THEIR HOST BACTERIA

Hannah Bonogafsky

Western Carolina University (April 2025)

Director: Dr. Jamie Wallen

As a method to hide from their host and prevent degradation, bacteriophages (viruses that infect bacteria) will modify functional systems to evade detection. In literature, it has been found that a common modification mechanism employed by bacteriophages is adding a functional group to a DNA base such that restriction systems cannot recognize and cleave viral DNA. One example of these modifications include the addition of a methyl group to generate methylcytosine. Successful modification results in the host restriction enzyme mechanisms failing to recognize the foreign DNA, which allows the phage to thrive. Western Carolina University (WCU) is part of the Howard Hughes Medical Institute's (HHMI) Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program, and work from WCU undergraduate students suggests that there are novel DNA modification systems present in Actinobacteriophages (viruses that infect Actinobacteria). To help understand what modifications might be present, this thesis focuses on the characterization of a panel of viruses predicted to have modified genomes. Of particular interest is phage TinyTimothy, whose genome is resistant to digestion by a standard panel of enzymes and fails to provide PCR products using Q5® DNA polymerase. Analysis of the nucleotide content of TinyTimothy genomic DNA using liquid chromatography – mass spectrometry (LC-MS) reveals the four canonical nucleotides and an

additional unknown peak that elutes at ~18 minutes. Current efforts are focused on identifying the modification that generates this peak, as well as performing LC-MS on other viral genomes believed to be modified.

During infection of a bacterial cell, temperate bacteriophages can choose between either the lytic or lysogenic replication cycles. During the lytic cycle, the phage will produce new virions by replicating its DNA using host machinery. In the lysogenic cycle, rather than producing progeny, the phage will incorporate its genome into the host chromosome to remain dormant. Along with nucleotide modifications, viruses have also developed proteins known as immunity repressors that bind the phage genome to prevent transcription of lytic genes and allow the phage to remain dormant and hidden in the host. Our laboratory recently published a manuscript describing a novel repressor protein found in cluster A mycobacteriophages. A unique feature of this repressor is that it uses two domains to bind DNA as a monomer; most repressors described to date bind DNA as higher-ordered oligomers. A detailed bioinformatic analysis of this monomeric repressor revealed novel protein sequences in which the repressor is fused to other protein domains with a variety of predicted functions, thus giving reason that the repressor does more than simply silencing lytic genes. Repressor fusion proteins range in size from 300 to 1,900 amino acids, and the predicted functions of these fusions vary. A few of these repressor fusions are found in pathogenic bacteria, which provides a possible avenue for novel therapeutics. We have expressed and purified the repressor fusion protein from *Mycobacterium sp.* ENV421, which has a repressor domain fused to a domain of unknown function. Overall, our goal is to better understand the synergy between repressors and the diverse protein domains that make up these novel proteins.

CHAPTER ONE: INTRODUCTION

Actinobacteriophages (phages) are viruses that only infect bacteria from the phylum *Actinobacteria.* These phages are mostly composed of nucleic acid enclosed in a protein capsid. As genome sequencing technologies are becoming more advanced, fully annotated bacteriophage genomes have become more available to study. Since previous experiments working with these viruses have become the groundwork for molecular biology, it is important to continue studying how they evolve.[1] Western Carolina University is part of the Howard Hughes Medical Institute's (HHMI) SEA-PHAGES program in which students find and name novel phages and characterize them according to different properties found.[2]

Studying how bacteriophages evade their bacterial hosts has growing importance as it has been shown that phages can be used in therapy to treat diseases such as antibiotic-resistant *Mycobacterium tuberculosis* and *Mycobacterium abscessus* infections[3]. Given the dire consequences, the fear of antibiotic resistance is growing every day.

There are two stages of the replication cycle for a bacteriophage, the lytic and lysogenic cycles. In the lysogenic cycle, rather than producing progeny, the phage will incorporate its genome into the host chromosome to remain dormant. Temperate phages will stay dormant until entering the lytic cycle (if choosing to do so).[4] The bacteriophage will then use the host machinery to replicate its DNA, ultimately causing lysis of the host cell. In the lytic cycle, the phage will infect the bacterial host and immediately begin to replicate its DNA and produce new progeny of virions. Although lysogeny is good from an evolutionary standpoint, the phage evolves to control lysogeny because of the risk of the host remaining vulnerable to another round

of infection from closely related phages.[5] Phages that stay in the lysogenic cycle can replicate their genomes within the host DNA while ceasing to create new progeny.

Temperate phages contain a repressor protein that prevents the transcription of genes that activate the lytic cycle. Bacteriophages use repressor systems as a mechanism to evade their host bacteria, and special DNA recognition sites are required for repressor binding. In the case of some bacteriophages such as cluster A and CA, the repressor sites are found in operators and stoperators, which are found in promoter regions (operators) and in short intergenic regions (stoperators) of DNA.[6] Repressors provide phages homoimmunity to closely related bacteriophages, as the repressor protein can recognize and bind the operator/stoperator sites in the other phage's genome, thus down regulating their lytic cycle as well. In the model phage Lambda, the cI repressor binds to DNA as a dimer, which is common for most repressor proteins[4].

To date, only a few repressors from Actinobacteriophages have been biochemically characterized. Our laboratory recently published a manuscript describing a novel repressor protein found in cluster A mycobacteriophages[4]. TipsytheTRex is a cluster A2 mycobacteriophage that was discovered in a soil sample on Western Carolina University's campus[2]. Interest was sparked when it was discovered that TipsytheTRex shares 98% sequence identity with the well-studied repressor from bacteriophage L5. Our laboratory successfully determined the crystal structure of TipsytheTRex's repressor, revealing a novel fold[4]. This repressor system can be described as binding to DNA as a monomer using two domains. One is a helix-turn-helix (HTH) at the N-terminal end of the protein, while the C-terminal portion of the protein contains a novel fold called the stoperator domain. Both domains bind to DNA at

successive major grooves, and they are independent of one another. A linker that connects the two domains, known as the helix bridge, also contacts the backbone of DNA.

Our laboratory has gone on to determine the crystal structure of another immunity repressor from the cluster A1 phage Adahisdi. Our interest in this protein stems from the fact that although it has the same overall fold and domain architecture as TipsytheTRex (Figure 1), it recognizes a different DNA sequence and is therefore heteroimmune to cluster A2 bacteriophages.

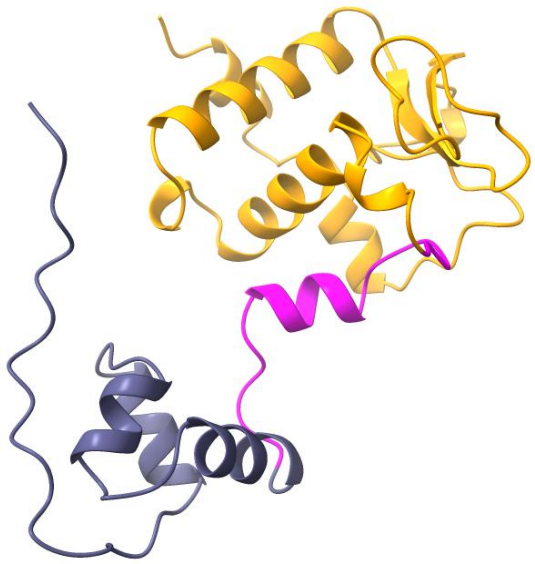

Figure 1: TipsytheTRex repressor. A cartoon representation of the repressor is colored by domain. The HTH is slate blue, the helix bridge is magenta and the stoperator is in orange.

Color coded in Figure 1 is the TipsytheTRex repressor. Interestingly, while the predicted structure of the Adahisdi repressor looks the same as the repressor for TipsytheTRex, it does not bind to the same DNA sequence. In the analysis of both the TipsytheTRex and Adahisdi repressor structures, a detailed bioinformatic analysis of Adahisdi's repressor was performed to identify all protein homologs using a position-specific iterative basic local alignment tool

(psiBLAST).[7] In our analysis we found that there were 1573 protein sequences that correspond to the ~180 residue long repressor. While this is fascinating, to our surprise we also found that there are much larger proteins that contain this repressor fused to other protein domains. These repressor fusions have a variety of predicted functions which support that the repressor does more than simply silence lytic genes.

Similar to how repressor systems allow bacteriophages to hide from their host, another mechanism of allowing bacteriophages to evade detection is through nucleotide modifications. DNA modifications do not cause a change in base pairing specificity, which allows for avoidance of the host.[8] 5-methylcytosine was found in *Mycobacterium tuberculosis* as the first identified modified nucleoside in 1925.[8] Later, 5-methylcytosine was found in bacteriophage Lambda.[8] To produce 5-methylcytosine a set of enzymes, called methyltransferases, are used by the host to modify specific DNA sequences, in which without this methylation the phage is unsuccessful at infecting the bacteria. Interestingly, it has been found that methyltransferases can be encoded via the phage or the host.[8]

Another example of a modified nucleotide in bacteriophages is in phage S-2L. Here the adenine is converted to 2-aminoadenine (known as Z).[9] After discovery of this modified nucleotide, there were questions about the mechanisms the bacteriophage used to convert adenine nucleotide fully into the Z nucleotide. It was discovered that there is a set of polymerases that are able to preferentially incorporate Z over the traditional A nucleotide during DNA replication.[9]This is a classic example of how bacteriophages modify their DNA to evade host detection via restriction enzymes.

TinyTimothy is an Actinobacteriophage that infects the bacterium *Microbacterium foliorum*. It was discovered at Western Carolina University by the SEA-PHAGES program and

contains 53 genes in total, with a genome length of 53932 base pairs.[2] TinyTimothy belongs to cluster EK2 along with 48 other bacteriophages. During routine characterization, we noticed two important observations were made: First, we were unable to amplify genes in the phage genome using traditional PCR and the high-fidelity Q5® DNA polymerase. When switching to a lower-fidelity Taq DNA polymerase a PCR product band of the expected size was observed, the results of which are shown in Figure 2.

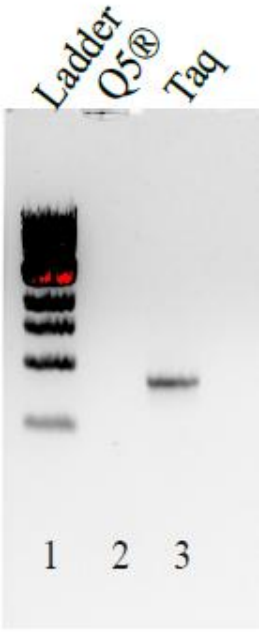

Figure 2: PCR amplification of TinyTimothy DNA using Q5® and Taq.

This provided clear evidence that something in the genome was likely inhibiting activity by Q5®. Second, some restriction enzymes predicted to cut TinyTimothy DNA did not digest the phage genome (Figure 3). Enzymes used initially for this assay, and their recognition sequences, are listed in Table 1. A virtual enzyme digest was performed with the enzyme panel to enable visualization of expected results if all enzymes properly cut TinyTimothy's DNA. Figure 3 shows the virtual and actual digest. In the experimental gel, pure DNA from TinyTimothy is

serving as a template for what an undigested sample of DNA should look like. EndoV is in an uncommonly used restriction enzyme that recognizes deoxyinosine and will be discussed in a later chapter. Out of the panel of enzymes used TinyTimothy DNA shown to be inhibited by 4 out of the 8 enzymes tested, as evidenced by the presence of a single band of high-molecular weight DNA.

Table 1: Restriction Enzyme List

Restriction Enzymes	Recognition Sequence Panel
BamHI	GGATCC
ClaI	ATCGAT
EcoRI	GAATTC
HaeIII	GGCC
HindIII	AAGCTT
NspI	RCATGY (Y=C or T, R= A or G)
SacII	CCGCGG

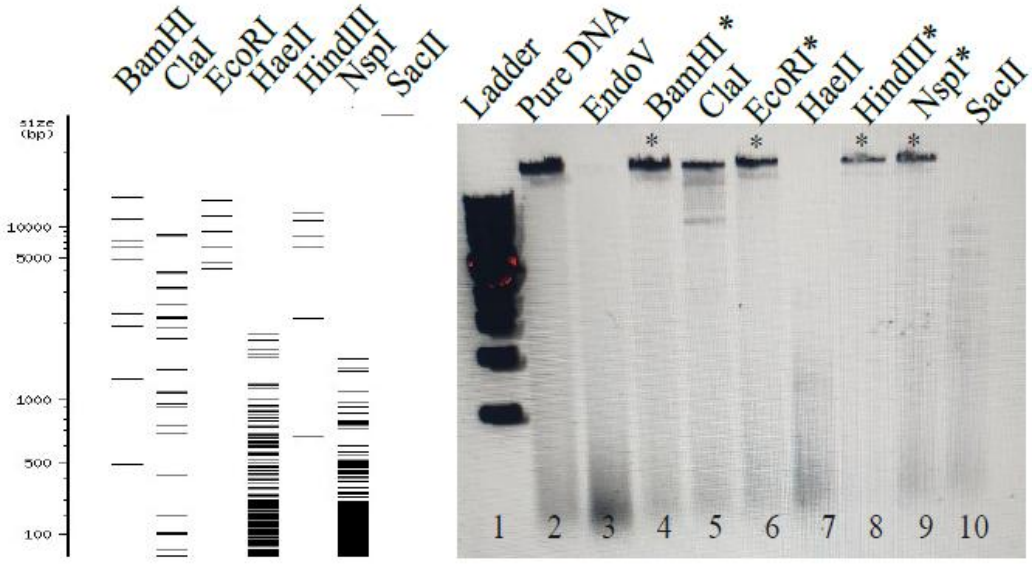

Figure 3: Enzymatic Digest on TinyTimothy's DNA. To the left is a virtual restriction enzyme simulation, the right image shows experimental results for restriction digest of TinyTimothy DNA using the same enzyme panel. Inhibited enzymes are notated with *.

Given the results from both Q5® and restriction enzyme experiments, it appears that the DNA of TinyTimothy is modified.

The focus of this thesis is to characterize two different strategies used by Actinobacteriophages to hide from host bacteria during infection. The first method of characterization consists of analyzing the nucleotide composition of TinyTimothy as well as a small panel of other viruses, all of which infect *Microbacterium foliorum* and have shown unusual restriction digest patterns. Since TinyTimothy has been shown to produce PCR product using Q5® DNA polymerase coupled with the resistance of TinyTimothy DNA to a standard panel of enzymes, the analysis of nucleotide content of TinyTimothy's genomic DNA was performed via LC-MS. These results revealed the presence of the four canonical nucleotides and an additional unknown peak eluting at approximately ~18 minutes. This novel peak in is thought to be deoxyinosine (dI). If dI is present, the results of PCR amplification will be successful when using a polymerase specific to dI. Additionally, restriction digests using enzymes specific to dI should result in expected results if dI is present. Further investigation is needed to confirm this modification structurally. Furthermore, we have identified two other viruses that likely contain modified bases via results from LC-MS. Kors, a cluster EF bacteriophage, shows one unknown band in its chromatogram. Andromedas, a cluster EA2 bacteriophage, shows two unknown peaks. One peak matches the retention time that of the unknown peak obtained during LC-MS analysis for TinyTimothy, while the other matches the retention time for the unknown peak obtained for Kors.

The second method of characterization included comprehensive bioinformatic study of homologs of the Adahisdi repressor. Exploration consisted of, but is not limited to, assessment of homolog size, predicted function, the organisms they are found in, as well as the location of their

repressor domain in the fusion protein. Our results have shown that there are 30 proteins that are found to contain the repressor and are greater than or equal to 230 aa long. We have found that the repressor fusion proteins are located in 11 different types of bacteria and 5 different bacteriophages. We have found that some of their predicted functions include serpin, a pirin domain, a transcriptional MocR family regulator, an AAA domain- containing protein, or an ATP-binding protein. We were able to predict the structure of these fusion proteins which can help us better understand the correlation of their function. While there are still unknowns about these fusion proteins, we can use these results as a guide to further investigate each specific fusion repressor system.

CHAPTER TWO: DISCOVERY OF NUCLEOTIDE MODIFICATIONS

DNA Extraction

For all biochemical experiments, all bacteriophage DNA was derived from a viral stock which was concentrated to a high titer lysate. A polyethylene glycol virus precipitation allowed for virus purification before DNA extraction. Phenol chloroform DNA extraction yielded phage DNA concentrations ranging from 1000 ng/µl to 7000 ng/µl. DNA quantification was determined using the Thermofisher Nanodrop spectrophotometer with UV absorbance at 260 nm. Using phenol chloroform allowed for high yields of DNA required us to continue analyzing nucleotide composition analysis.

Initial Findings From LC-MS

We have shown that *Microbacterium foliorum* phage TinyTimothy has DNA is not recognized by standard restriction enzymes. Additionally, it was the discovery that TinyTimothy will not produce a amplification product when using high fidelity Q5® PCR polymerase (Figure 2). For further analysis, the nucleotide composition of TinyTimothy was further investigated.

Using New England Biolabs' (NEB) Nucleoside Digestion Mix, TinyTimothy's DNA was digested into single nucleosides so Liquid Chromatography-Mass Spectrometry (LC-MS) could be used for identification of individual nucleotides (nucleosides) which comprise the bacteriophage DNA.

After digestion of TinyTimothy's DNA, a total of 50 µl sample was injected into a silica-based, reversed-phase C18 column. The mobile phases used were 10 mM ammonium acetate pH 5.4 and methanol at a flow rate of 0.5 mL/ min. The detector was set at 260 nm to monitor nucleoside elution. Controls used were a deoxynucleotide triphosphate (dNTP) mix, as well as

each of the four individual nucleotides, both digested into nucleosides with the same process described above for phage DNA.

Shown in Figure 4 are the chromatographs for the dNTP control and for TinyTimothy DNA. Each peak is labeled according to expected elution times for individual nucleosides ran. The observed elution order is deoxycytidine (dC), deoxyguanosine (dG), deoxythymidine (dT), and deoxyadenosine (dA). Based off the dNTP controls, each peak in TinyTimothy's chromatogram is labeled accordingly. Interestingly, there were four peaks found that do not match any of the expected peaks for control dNTPs; however, peaks for all four canonical DNA nucleotides are present.

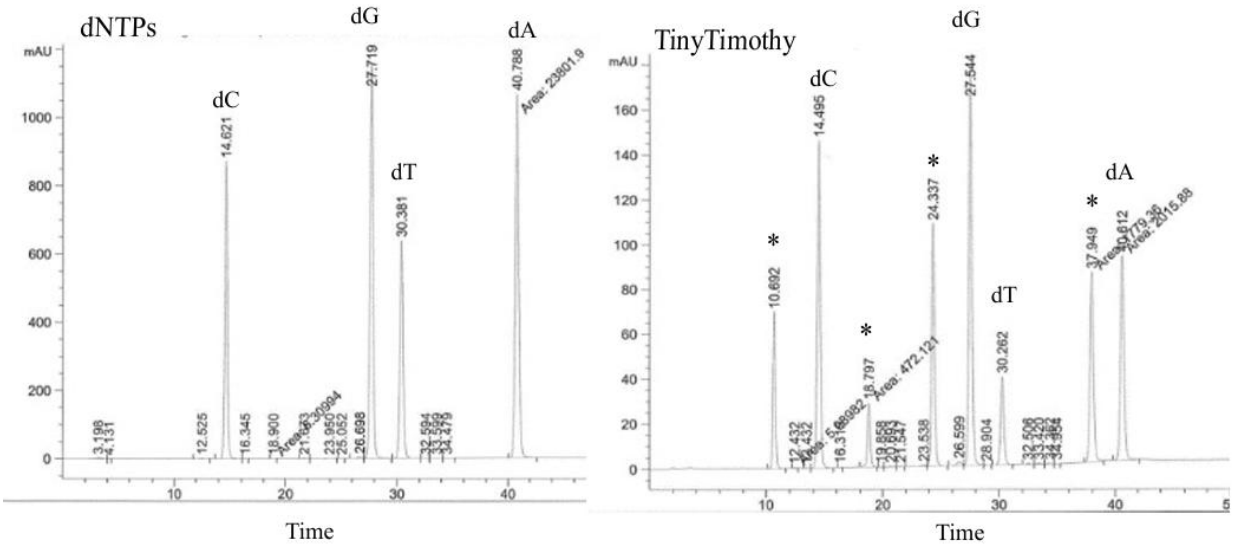

Figure 4: Initial LC-MS of dNTP controls and digested DNA from TinyTimothy. Labeled in the control in experimental chromatograms are the known nucleotides, with unknown notated in experimental chromatogram with *.

RNA Contamination

To identify unknown peaks present in TinyTimothy, we examined whether RNA contamination could be a source. A set of NTPs were purchased and processed exactly like the dNTPs and were then analyzed using LC-MS. Results digested for TinyTimothy DNA and RNA control results are shown in Figure 5. This run indicates that RNA is present in TinyTimothy

digests as three out of the four expected RNA nucleosides are identified in LC-MS data. Interestingly, there is still an unknown compound seen in the chromatogram. The retention times of each nucleoside are as follows: ~14 minutes for dC, ~26.5 minutes for dG, ~29.2 minutes for dT, ~ 39.5 minutes for dA, ~10.3 minutes for cytidine (C), ~13.7 minutes for uridine (U), ~23.1 minutes for guanosine (G), and ~36.5 minutes for adenosine (A).

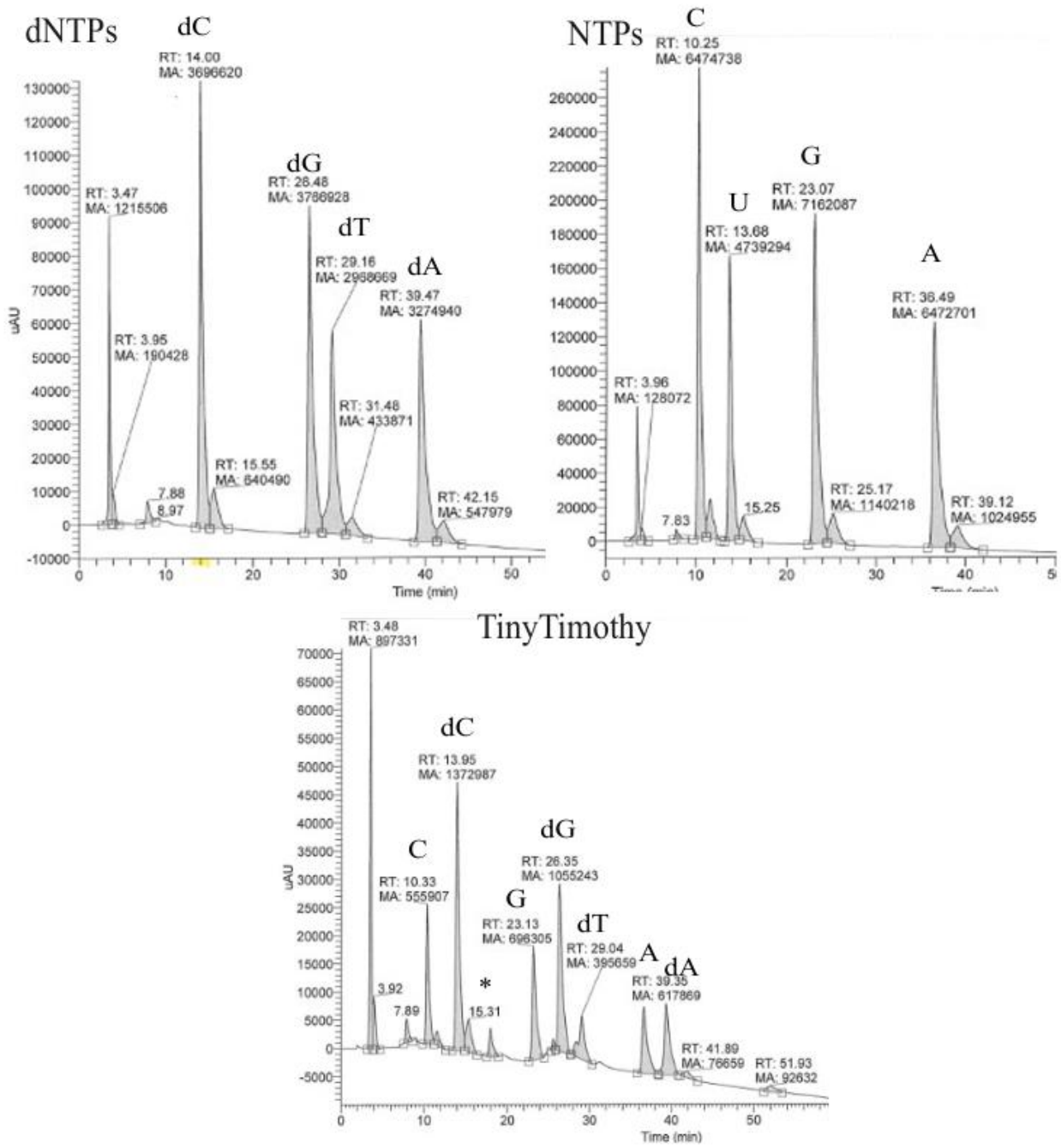

Figure 5: Second LC-MS that includes NTPs as a control, with unknowns notated with *.

To determine the best method for removal of RNA contamination, undigested DNA exposed to two enzymes including DNase treatment and RNase. These results are shown in Figure 6. Exposing DNase to TinyTimothy's DNA, DNA appears fully digested, as expected with DNase treatment. No digestion of TinyTimothy's DNA is evident after treatment with RNase.

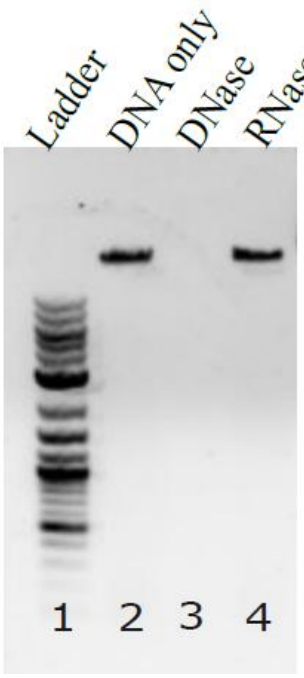

Figure 6: DNase and RNase treatment of TinyTimothy's DNA.

Contaminating RNA removal was accomplished by performing a second RNase treatment following DNA purification. The resulting DNA was desalted, then quantified using a Nanodrop spectrometer. Analysis of this DNA via LC-MS shows complete removal of RNA. Peaks for the four canonical nucleotides are evident as well as one other unknown species peak eluting between dC and dG. At this point in our analysis, we could not get our controls (dNTP) or digested TinyTimothy DNA to ionize via mass spectrometry, thus the lack of results shown of the samples.

Identifying the Unknown Nucleotide

Since the samples were not efficiently ionizing, both mobile phases were changed to water and methanol, both containing 0.1% formic acid. A note from this new method is that peak elution time is a little faster than what was observed when using the previous method. Additionally, the elution order of each peak is different. The order that each nucleotide elutes

corresponds to the mass obtained, thus known values of each nucleotide were compared to literature mass values and listed in (Table 2) along with the new retention times. The elution order is as follows: dC, dA, dG, dT.

Table 2: Masses and retention times of nucleosides from LC-MS.

Nucleoside	Literature Mass Values (m/z)[10]	Mass Obtained from LC-MS [M+H] (m/z)	Retention Times (min)
dG	267	268.05	~24.2
dA	251	252.06	~19.4
dC	227	228.05	~9.3
dT	242	243.05	~24.5

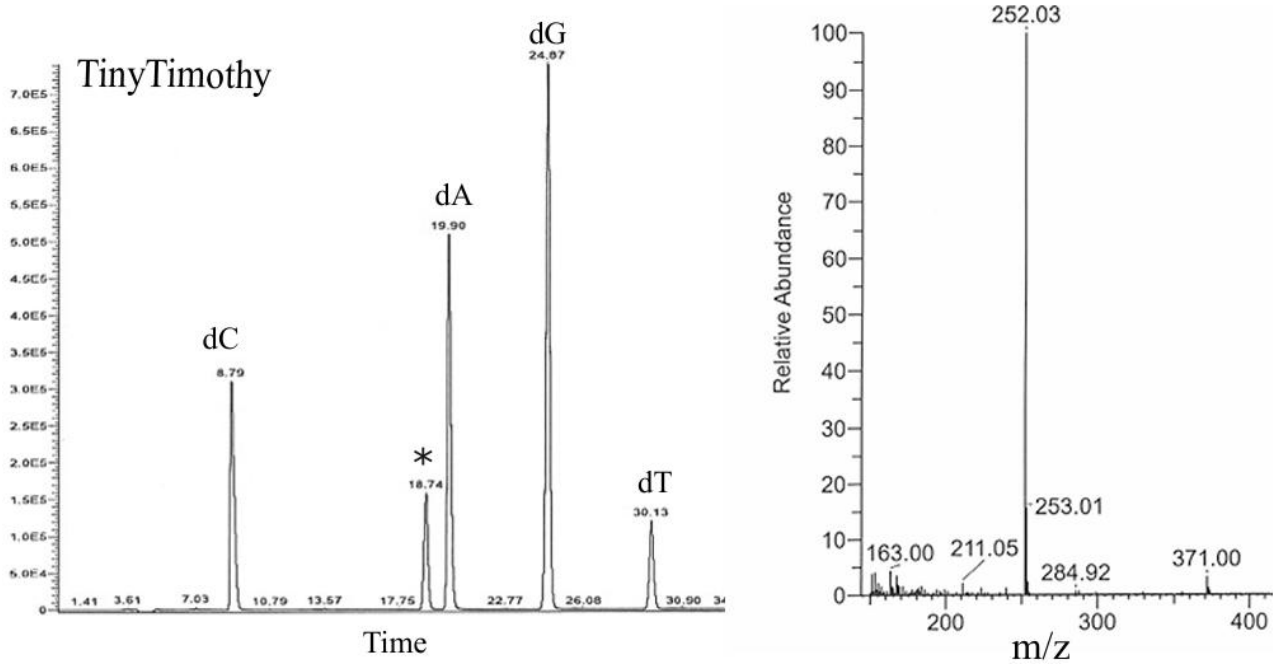

Figure 7: LC-MS Results from Modified method of TinyTimothy DNA including mass of unknown (peak notated with *).

Next, to increase the signal of the unknown peak, the injection volume was increased from 50 µl to 100 µl. Figure 7 shows the results of the unknown compound through LC-MS. The unknown peak elutes at ~18 minutes and has a mass of 252 m/z. Also demonstrated are the new elution times and order of the nucleotides.

The literature[11] mass value reported for deoxyinosine matches the mass we obtained in LC-MS data for our unknown peak. To determine if inosine is the modified nucleotide in TinyTimothy DNA, the restriction enzyme EndoV, which cleaves inosine was purchased. Additionally, we also reamplified TinyTimothy genomic DNA using the enzyme Q5U®, which reads uracil and inosine-containing templates. Recall the panel of restriction enzymes test and their cleavage sites are shown back in table 1. We also added the enzyme EndoV which cleaves at dI, therefore if TinyTimothy DNA is modified with dI, EndoV should digest its DNA. Recall the gel shown in Figure 3 of the restriction digest of TinyTimothy DNA using the enzyme panel and EndoV. In agreement with the mass observed in the LC-MS data, the enzyme EndoV does indeed cleave TinyTimothy DNA.

Figure 8 provides the results of the PCR reaction with Q5U®. As observed previously, Q5® fails to amplify TinyTimothy DNA. However, Q5U® readily forms a PCR product using TinyTimothy DNA. Although both of these results support that the modification in TinyTimothy could be dI, a purchased standard of dI had a retention time of ~ 24.8 minutes shown in Figure 9, which does not match with the unknown in TinyTimothy with a retention time of ~ 18 minutes (Figure 7).

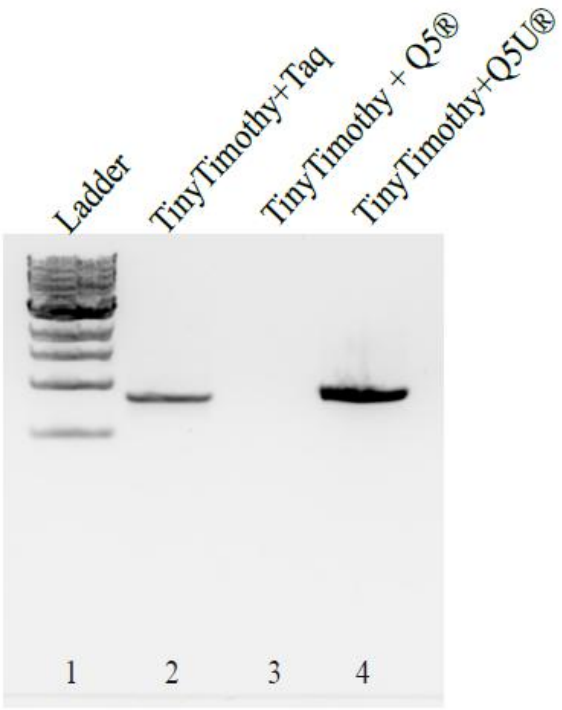

Figure 8: PCR results for amplification of TinyTimothy DNA using Q5U® which recognizes deoxyinosine containing templates.

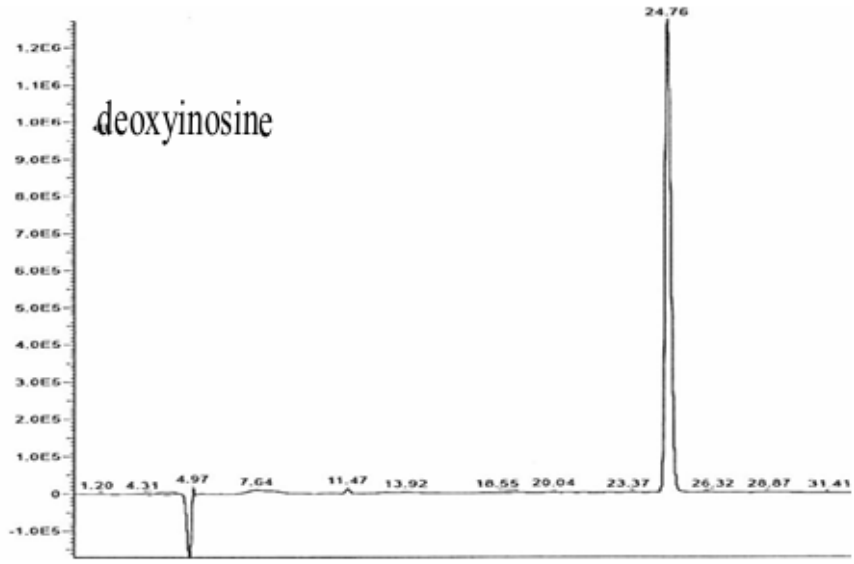

Figure 9: Deoxyinosine standard chromatogram

Panel of Viruses

Using LC-MS, two other viruses were analyzed for modified nucleotides, including Andromedas and Kors are cluster EA2 and EF bacteriophages, respectively. Both bacteriophages were found at Western Carolina University, and both are isolated from the host *Microbacterium foliorum*. Restriction digest profiles for these viruses also suggest that they have modified nucleotides. Each virus was tested to see if it produced a PCR product using Q5®, (Figure 10). While Andromedas shows that there is inhibition with Q5® but produces a strong product with Taq, (lanes 2 and 3) Kors produces a product with both polymerases, (lanes 6 and 7). TinyTimothy's results remain the same as the initial testing (Figure 2) in which demonstrates a product with Taq but no product with Q5®.

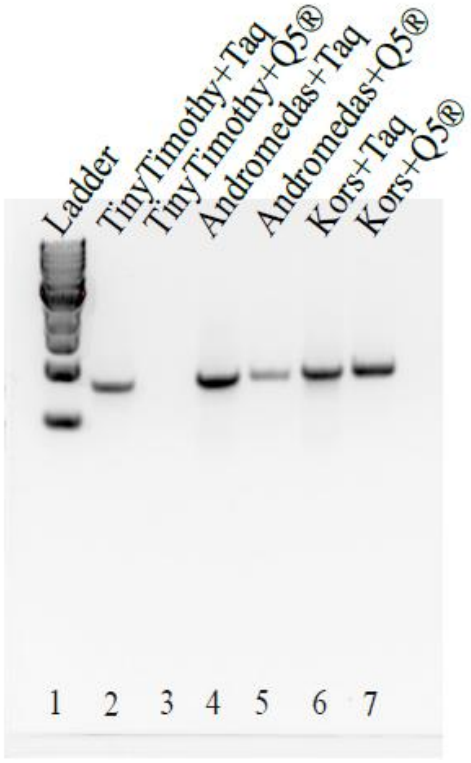

Figure 10: PCR products using a gene from TinyTimothy, Andromedas, and Kors using a high and low fidelity polymerase.

Continuing with exploration of modified nucleotides, the DNA of both of these phages were extracted and digested in the same manner as DNA from TinyTimothy, and the LC-MS results are shown in Figure 11. Kors and Andromedas both seem to contain all the canonical nucleotides; however, they each have extra bands. In Kors this is eluted after the dA and is eluting as a double peak. Andromedas has two extra bands in which one is similar to TinyTimothy's modification, eluting after dC and the other similar to Kors which is eluting after dA.

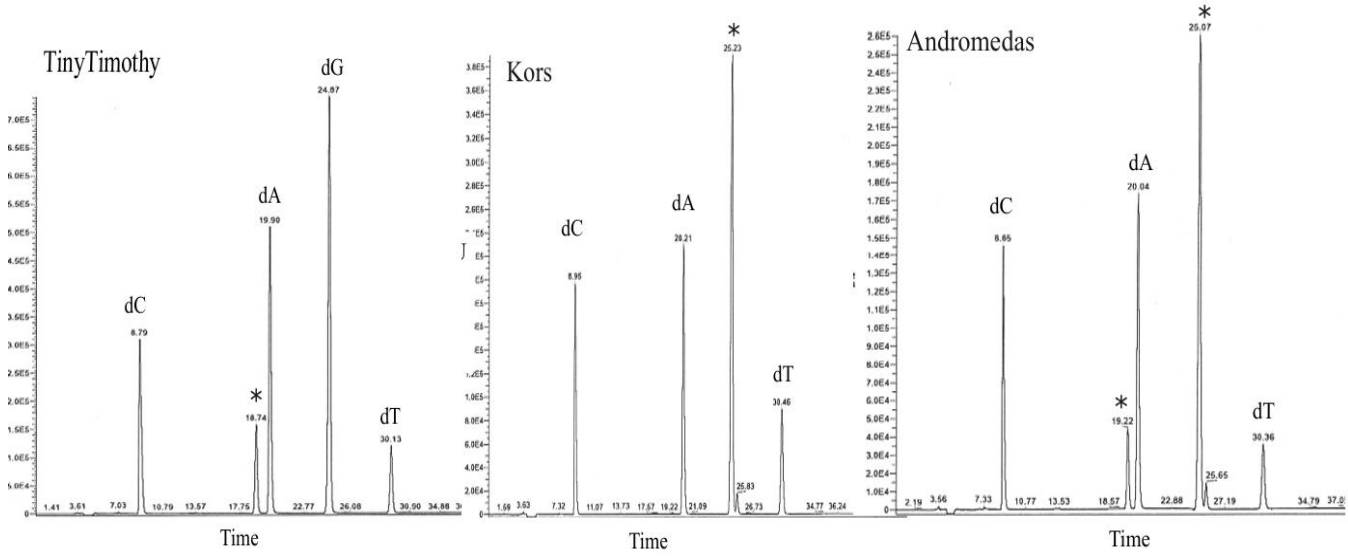

Figure 11: Viral panel genome assessment using LC-MS. Unknown marked with *.

Discussion

Discovering a modified nucleotide is not novel, however it is still intriguing to think about why bacteriophages are modifying their DNA. At the start of this experimental process, it seemed like the polymerases being used were not functioning correctly because TinyTimothy would not produce an amplification product when using Q5®. When Taq polymerase was able to support amplification, it then occurred that there could be something in the DNA that was

inhibiting DNA amplification when using the Q5® enzyme. Continuing through with the restriction enzyme digest panel, it was interesting to see how many enzymes did not cut DNA from TinyTimothy. All inhibited enzymes that are seen seem to have a common splicing point at guanine or adenine.

A challenge when analyzing TinyTimothy's DNA using LC-MS was the RNA contamination. Even after standard RNase treatment after phage precipitation, an additional treatment is needed to completely remove the RNA that derives from the bacterial host. It is interesting that three of the four nucleosides of RNA were found in the samples, with the exception of uracil. This is expected since the literature shows we found that other researchers were running into this situation as well[12]; however, they did not report on how the contamination was eliminated.

The RNase digest that was run on purified DNA of TinyTimothy (Figure 6), we know that there is not RNA in the genome of TinyTimothy. If there were then the results of RNase treatment (lane 4) would look similar to the results from the DNase treatment (lane 3). This study allowed us to optimize a method that would eradicate RNA that was co-extracted with DNA. The trials of this consisted of tweaking the amount of DNA digested, the length of time it was allowed to incubate, and the ratio of RNase to amount of DNA.

During the process of eliminating RNA contamination, we also optimized the LC-MS method for ionization of our samples. Now that the samples were ionizing, we realized we needed to increase the signal of the unknown peak from TinyTimothy DNA such that analysis via mass spectrometry would be successful. Once the injection volume of the sample was increased, we saw an increase in peak size on the chromatogram that allowed for accurate mass detection.

More analysis is to be done on discovering what exactly the modification in TinyTimothy DNA is. The mass (Figure 7), the digestion with EndoV, and the PCR results with Q5U® (Figure 8), strongly suggest that the modified nucleotide is dI; however, LC-MS data on a deoxyinosine standard shows the correct mass, but an elution time that differs from the unknown peak in TinyTimothy, shown in Figure 9. While we are consistently getting an elution time of ~18 minutes for the unknown peak in TinyTimothy, when dI is run alone it is eluting at a time of ~24.8 minutes. We purchased the nucleoside of dI, therefore we did not digest it using the same digestion kit as was used with the other DNA samples. Further work is needed to resolve this discrepancy.

To expand on this work, we decided to run LC-MS on DNA from Andromedas and Kors, two viruses isolated from the same bacterial host as TinyTimothy that also show evidence of a modified genome. In Figure 11, we see that LC-MS data from Andromedas it looks very similar to data from TinyTimothy. Also, with Andromedas PCR results look similar to the PCR products from TinyTimothy (Figure 10); however, the product is not fully inhibited. Kors on the other hand, yields PCR products using both Taq and Q5® polymerases, but LC-MS data indicates a potential modification. In Andromedas, two potential modifications are shown, one of which is similar to TinyTimothy's modification (figure 11). The other peak that Andromedas contains is similar to that of Kors. Andromedas and Kors are also missing the dG elution peak. Due to the intensity of the peaks, we thought at first their unknown ~ 25 minutes was dG, however dG elutes ~24.5 minutes, therefore it cannot be dG. To obtain more confirmation about these two peaks, we will have to obtain the masses of each one. These potential modifications are also interesting because Andromedas is from cluster EA, Kors from EF, and TinyTimothy EK. While they are from different clusters, the clusters are similarly related. Due to time constraints, we did

not analyze additional viruses via LC-MS, and additional work is needed for both Andromedas and Kors to identify the potential modified nucleotides.

Future Directions

There are many more avenues to continue this research on modified nucleotides. To start, confirmation of what is the modified nucleotide is needed that is in TinyTimothy. To do this we would need to figure out the structure of the peak at ~18 minutes. Discovering the structure can be achieved in a few different ways, one of which is using LC-MS/MS with electrospray ionization (ESI). ESI is a soft ionization technique which ionizes the intact molecule. Tandem mass spectrometry or MS/MS can be used to break down molecules and they can be analyzed, thus obtaining a structure. Electrospray ionization is great to use with polar samples because of their ability to ionize by accepting or donating protons. Once there is a method developed for ESI for TinyTimothy DNA, this can then be used on the panel of bacteriophages that we have analyzed via LC-MS. We believe that TinyTimothy has a novel modification that needs confirmation. More questions that will be answered include is the modification actually deoxyinosine or something else? Is the modification something that has not been characterized before? Is it a purine or pyrimidine? What is the likeness of the other phages modifications to TinyTimothy?

An exhaustive panel of bacteriophages should also be analyzed to see if they also contain modified nucleotides. In this study two other phages were examined, it would be interesting to see if multiple members of a cluster contains these modifications. Are particular modifications restricted to certain clusters? Are modifications restricted to bacteriophages with certain hosts? How different are the modifications from various different categories of bacteriophages? Before starting analysis via LC-MS, a comprehensive panel of restriction enzymes should be tested.

Additionally, it should be determined whether bacteriophages will produce a amplification products with different polymerase enzymes.

After identifying what the modification in TinyTimothy's DNA, the mechanism by which the modifications arise should be elucidated. Does TinyTimothy use enzymes of its own or from its bacterial host to induce this modification? Is this modification derived from another canonical nucleotide? TinyTimothy does not seem to fully replace a nucleotide, given it contains all four canonical nucleosides shown via LC-MS, however does it partially replace one? These questions can start to be answered by delving into bioinformaticanalysis of TinyTimothy's genome. To begin, we would need to perform genome annotation followed by identification of gene sequence This can then be input into a function prediction database, such as HHpred. To determine if a gene has a function that would produce an enzyme that would modify its DNA. As mentioned in the introduction chapter, phage S-2L contains a modification that completely replaces dA in its genome. Pezo et al. show that there is a protein N6-succino-2-amino-2'-deoxyadenylate synthase that makes S-2L's Z nucleotide[9]. One more tool that would be beneficial would be is to input TinyTimothy's protein sequences for each gene into psiBLAST[7], running the program until no new iterations are found. These results would produce distant relatives of the sequences given which would enable us to look for repetitive functions that might suggest DNA modification. Furthermore, a study could be conducted to see what would happen to the phage's virulence if identified modification enzymes were removed. Would TinyTimothy still infect its host, or would the host cleave off the DNA via restriction enzymes? If we could take this enzyme and introduce it into another bacteriophage that contains canonical DNA, would it produce this modification as well?

Once the enzyme is discovered, creating phylogenetic trees could reveal whether this modification originated from the bacteriophage or the bacteria host. To achieve this, we would need to search through databases with the enzyme sequence, looking for other organisms containing this enzyme. Once these homologs are located, using strict algorithms, a phylogenetic tree can then be created which will show the roots of where each organism came from and how closely related they are to each other.

CHAPTER THREE RESULTS OF BIOINFORMATIC ANALYSIS

Overall Results Of Protein Size

On 01/22/25 the stoperator domain from the Adahisdi immunity repressor was analyzed through NCBI's psiBLAST. PsiBLAST performs a protein-protein sequence search to find proteins with high similarity to with the input sequence.[7] Using this as a basis, it uses a position specific search algorithm to identify position specific amino acid substitution scores, thus determining closely related proteins based on residues at this position.[7] With each iteration, new sequences found are highlighted. The Adahisdi stoperator search took 9 iterations before no new sequences were identified, and the final search produced a total of 1573 protein sequences that were within the correct e-value threshold of 0.005. The repressor found in cluster A2 TipsytheTRex and cluster A1 Adahisdi are ~180aa long; therefore, to find repressors with additional domains, we selected proteins larger than 230aa.Within this search, we identified 30 sequences that were greater than 230 amino acids long, which suggested that additional protein domains may be fused to the repressor protein. Table 3 provides results including the length of each protein, accession numbers, predicted function, and the organism that contains the protein. From our results in Table 3, there seems to be a lack of variation of function if the protein is less than ~400aa. The predicted functions include DNA-binding protein, a protein with function unknown (hypothetical protein or domain of unknown function (DUF)), or a repressor protein. Proteins longer than 400aa long have the predicted functions of a serpin, a pirin domain, a transcriptional MocR family regulator, a AAA domain- containing protein, or an ATP-binding protein.

Table 3: Discovered Repressor Fusion Domains From psiBLAST

Accession Number	***Bacteria***	***Length (aa)***	***Predicted Function***
WP_346537388	*Micromonospora sp. DPT*	233	sigma factor-like helix-turn-helix DNA-binding protein
GAA2655495	*Streptomyces spororaveus*	234	hypothetical protein
WP_280840815	*Micromonospora sp. A200*	235	helix-turn-helix domain-containing protein
WP_266766066	*Streptomyces sp. NBC_00638*	242	DUF6192 family protein
WP_190092309	*Streptomyces melanogenes*	249	hypothetical protein
WP_216699470	*Actinotalea ferrariae*	252	hypothetical protein
WP_319431842	*Mycobacterium sp. RTGN5*	261	hypothetical protein
WP_233209289	*Mycobacterium sp. ENV421*	262	hypothetical protein
WP_235677880	*Mycolicibacterium sarraceniae*	262	hypothetical protein
WP_327204034	*Rhodococcus pyridinivorans*	275	hypothetical protein
WP_012877027	*Xylanimonas cellulosilytica*	277	hypothetical protein
WP_291278509	*Galactobacter sp.*	281	hypothetical protein
WP_231748648	*Mycobacterium sp. M26*	285	hypothetical protein
WP_367880880	*Rhodococcus pyridinivorans*	319	hypothetical protein
WP_207207318	*Xylanimonas protaetiae*	331	hypothetical protein
AHK31009	*Rhodococcus opacus PD630*	333	Repressor-like immunity protein
WP_083613183	*Mycolicibacterium mageritense*	422	serpin family protein
MCW2843333	*Nocardioides sp.*	445	Pirin domain protein
WP_190267212	*Gordonia hankookensis*	602	serpin family protein
RZS62629	*Xylanimonas ulmi*	603	DNA-binding transcriptional MocR family regulator
ACZ32369	*Xylanimonas cellulosilytica DSM 15894*	611	Putative transcriptional regulator, GntR family
WP_246228823	*Mycolicibacterium psychrotolerans*	648	hypothetical protein
WP_234792637	*Mycolicibacterium fortuitum*	1516	AAA domain-containing protein
OBJ98398	*Mycolicibacterium fortuitum*	1875	ATP-binding protein IstB
	Bacteriophage		

AMW64332	*Mycobacterium* phage ChipMunk	231	hypothetical protein
WXX09781	*Mycobacterium* phage MS619	232	transcriptional repressor
AOQ29445	*Mycobacterium* phage Bigfoot	236	Immunity repressor
XEN16495	*Mycobacterium* phage PhesterPhotato	229	hetero-immunity repressor
UGL61796	*Mycobacterium* phage Grub	252	Immunity repressor
YP_005087008	*Rhod*ococcus phage RGL3	268	Transcriptional repressor

Types of Organisms Meta Analysis

Table 3 highlights that these proteins are found in both bacteria and bacteriophages. One important thing to note though is that it has not been established whether the proteins are bacterial in origin, or if they are found in prophages. Interestingly, of the 30 proteins selected for further analysis 6 are from bacteriophages and 24 are from bacteria. The bacteria identified include: *Actinotalea ferrariae, Galactobacter sp., Gordonia hankookensis, Micromonospora sp.* strains: *DPT & A200, Mycobacterium sp.* strains*: ENV421, M26, & RTGN5, Mycolicibacterium* species*: fortuitum, mageritense, psychrotolerans, & sarraceniae, Nocardioides sp., Rhodococcus* species*:opacus PD630 & pyridinivorans, Streptomyces* species*: melanogenes, sp. NBC_00638, & spororaveus, and Xylanimonas* species*: protaetiae, ulmi, cellulosilytica & cellulosilytica DSM 15894.* To learn more about what types of bacteria these proteins are coming from, peer reviewed literature was consulted to fill in this unknown.

Actinotalea ferrariae is Gram-stain-positive and aerobic. It is rod shaped and also non-motile. This bacterium was originally found in an iron mine, a discovery that emended the

description of the genus *Actinotalea.*[13] At present, it does not seem to be pathogenic, nor have any bacteriophages been identified that use it as a host.

Galactobacter sp. are Gram-stain-positive and aerobic. They are rod-shaped and non-motile. It was found originally in raw cow's milk in Germany.[14] To date, it is non-pathogenic and does not have any recorded bacteriophages that use it as a host.

Gordonia hankookensis sp. is a genus of aerobic, non-motile, Gram-stain-positive bacteria that are rod/coccus shape. Originally found in a soil sample near a wastewater treatment plant, *Gordonia hankookensis sp*. have not been found to be pathogenic.[15] There 3378 phages that infect *Gordonia,* 9 different species, and 16 different strains.[16] Clusters that these phages are part of include A, C, D, and Singletons. Noted from literature, there is a common immunity repressor found in these often temperate phages.[15] From *Gordonia hankookensis sp*. specially, it does not have a characterized bacteriophage as its pathogen.

Micromonospora sp. are a genus of Gram-stain-positive, aerobic bacteria, and differentiates into both mycelia and spores.[17] Species of *Micromonospora* have roles in nitrogen fixation, biocontrol biofuel production, and are being studied as drug candidates.[18] Both strains DPT and *A200* do not appear to have a known bacteriophage as their pathogen, nor have they been found to be pathogenic.

Mycobacterium are rod shaped and Gram-stain-positive, aerobic bacteria. Two of the most common pathogens from this genus are *M. tuberculosis* and *M. leprae*, causing tuberculosis and leprosy, respectively. There are many bacteriophages that have been found to use *Mycobacterium* as their host. Recorded to date are 10 species, 24 strains, and 14255 different bacteriophages.[16] The results given from the repressor fusion proteins found are from *Mycobacterium sp*, with the sp meaning that the species is not fully described. The strain

ENV421 has been shown to use chemicals as a source of energy, some including methyl tert-butyl ether, tert-butyl alcohol, and propane.[19] Strain *M26* was isolated from a sample from a respiratory tract infection.[20] Strain *RTGN5* has been found in a root nodule from *Alnus glutinosa*, an Alder tree which has been studied for the use of bioremediation.[21]

Mycolicibacterium genus are well-known bacteria that are pathogenic to humans and animals. It is worth noting that *Mycolicibacterium* is a sub species from *Mycobacterium*. *Mycolicibacterium* are Gram-stain-positive and are rod shaped.[22] *M. fortuitum* has been shown to have a natural antibiotic resistance which makes it such a virulent pathogen.[23] To date there are no bacteriophages identified that use this species as a host, however *M. smegmatis* phage KVT1 has been shown to target and grow using *M. fortuitum* as its host.[24] *M. mageritense* is highly pathogenic and has been found in pleural fluids from a patient infected with pleurisy.[25] There are no known bacteriophages that infect this species of bacteria. *M. psychrotolerans* does not have a recorded pathogen nor bacteriophage. The first record of it was that it was isolated from pond water that was near a uranium mine.[26] *M. sarraceniae* has been found from a water sample that was obtained from a pitcher plant.[27] It is not a recorded pathogen, nor does it have a bacteriophage.

Nocardioides is a Gram-stain-positive bacteria that is aerobic and irregularly rod-shaped. *Nocardioides* has been isolated from industrial wastewater and contaminated soil and has been examined for properties that degrade environmental pollutants.[28] Specifically, *Nocardioides sp*. were found in crude oil samples and are thought to be able to breakdown oil.[28] There have been bacteriophages identified that infect *Nocardiodes.*[29] There are also no reported cases suggesting that it is pathogenic.

Rhodococcus is a Gram-stain-positive bacteria that is an aerobic and non-motile coccobacillus. *Rhodococcus* is best known as being an animal pathogen and contains 192 known phages that infect genus.[16,30] *Rhodococcus opacus PD630* has three total characterized phages which use it as a host bacterium, though none have yet been clustered or sequenced. *Rhodococcus pyridinivorans*, used in the biodegradation of plastics, does not serve as host for any characterized phages.[31]

Streptomyces is a Gram-stain-positive bacteria that is aerobic and complex.[32] This type of bacteria has been found to be useful in producing many different types of antibiotics and pharmaceutical products.[33] Streptomyces as a genus has been associated with 1642 phages that have been characterized to date.[16] *Streptomyces sporoaveus* has been found to produce an antibiotic against two types of bacteria and six types of fungi.[34] Currently there are no characterized bacteriophages that are associated with the species of *Streptomyces* in which the protein repressor fusions have been found.

Xylanimonas is a Gram-stain-positive bacteria that is aerobic and coccoid or rod shaped. Both *X. cellulosilytic*a and *X.ulmi* were found in decayed trees.[35] *X. protaetiae* has been found in the gut of insect larva.[36] None of these species have been found to be pathogenic, nor has a bacteriophage been found using it as a host.

Four of the repressor fusion proteins are from bacteriophages that infect the host *Mycobacterium smegmatis mc²155,* which include phages ChipMunk, Bigfoot, PhesterPhotato, and Grub. Each of these have been characterized as temperate phages belonging to various in clusters. Chipmunk is from cluster A2, Bigfoot is from A1, PhesterPhotato is from F1, and Grub is from A3[16]. Phage RGL3 infects the host *Rhodococcus globerulus Rglo35* and has been verified to infect *Rhodococcus erythropolis Rery29.*[16] RGL3 is a temperate phage from cluster CA. From

the psiBLAST results, phage MS619 is classified as a *Mycobacterium phage*, however there is no reported information classifying this phage except that it is from the phylum *Uroviricota*, class *Caudoviricetes.*[37]

Location of Repressor Domain and Function of Other Domains

For each protein found, the location of the repressor domain was identified using Max Planck Institute's bioinformatics toolkit, HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred).[38] The results from this are found in Table 4. For 9 of the proteins, the repressor is found on the N-terminus, while the repressor is found at the C-terminus of 21 of the repressor fusion proteins. Whether the repressor is on the N or C-terminus, it does not seem to have any impact on the function of the domain attached to it.

Table 4: HHpred Results of Repressor Fusion Domains

Organism	**Repressor Domain**	**PDB ID**	**Description**
***Mycobacterium* phage ChipMunk**	C	N/A	
***Mycobacterium* phage MS619**	C	N/A	
Micromonospora sp. DPT	N	N/A	
Streptomyces spororaveus	C	N/A	
Micromonospora sp. A200	N	N/A	
***Mycobacterium* phage Bigfoot**	C	N/A	
Streptomyces sp. NBC_00638	N	N/A	
***Mycobacterium* phage PhesterPhotato**	C	2GM5	An activated, truncated gamma-delta resolvase tetramer
Streptomyces melanogenes	C	2OA4	Northeast Structural Genomic Consortium Target SiR5
Actinotalea ferrariae	N	N/A	

***Mycobacterium* phage Grub**	N	N/A	
Mycobacterium sp. RTGN5	C	N/A	
Mycobacterium sp. ENV421	C	N/A	
Mycolicibacterium sarraceniae	C	N/A	
***Rhodococcus* phage RGL3**	N	N/A	
Rhodococcus pyridinivorans	C	N/A	
Xylanimonas cellulosilytica	C	N/A	
Galactobacter sp.	C	N/A	
Mycobacterium sp. M26	N	N/A	
Rhodococcus pyridinivorans	C	N/A	
Xylanimonas protaetiae	C	N/A	
Rhodococcus opacus PD630	C	N/A	
Mycolicibacterium mageritense	C	1JMO	Heparin Cofactor II-S195A Thrombin Complex
Nocardioides sp.	C	6D0P	Quercetin 2,3-dioxygenase from Acinetobacter baumannii
Gordonia hankookensis	C	4AFX	Reactive loop cleaved ZPI in I2 space group
Xylanimonas ulmi	N	3EZ1	Putative aminotransferase (MocR family) (YP_604413.1) from DEINOCOCCUS GEOTHERMALIS
Xylanimonas cellulosilytica DSM 15894	N	3AOW	Pyrococcus horikoshii kynurenine aminotransferase in complex with AKG
Mycolicibacterium psychrotolerans	C	1JMO	Heparin Cofactor II-S195A Thrombin Complex
Mycolicibacterium fortuitum	C	3 domains found: 4B3F, 8FTK, 3R3P	Lghmbp2 helicase, Chaetomium thermophilum SETX, Homing Endonuclease I-Bth0305I Catalytic Domain
Mycolicibacterium fortuitum	C	3 domains found: 4B3F, 8FTK, 3R3P	Lghmbp2 helicase, Chaetomium thermophilum SETX, Homing Endonuclease I-Bth0305I Catalytic Domain

Ten different protein domains have been identified to be fused to the repressor domain (Table 4 correlating with their predicted function in Table 3). To gain better insight into what more the repressors do or why they would need another domain, a brief synopsis of the uncommon domains found has been provided. These domains are as follows: a serpin protein domain, a pirin domain, a MocR domain, an AAA domain- containing protein, and an ATP-binding protein.

In short a serpin protein is a serine protease inhibitor.[39] Since their initial identification as serpins, they have been shown to perform a variety of different functions such as transporter of proteins such as small molecules like hormones, or they target other classes of proteases.[40] The structure of serpins consists of three β-sheets, eight to nine α- helices, and a flexible loop on the top of the backbone called a reactive center loop. This loop, which is used for interactions with target proteases that to produce permanent inhibition of a protease, is eventually cleaved. This unique feature of serpin proteins is called a "suicide" mechanism where once they are used as an inhibitor they go through an intensive amount of change that renders them useless. Along with cleavage of their loop, they also go through intense distortion where some of the protein is also unfolded.[39] In viral serpins, it has been found that they will have a smaller structure in order to keep the genome smaller.[39] In humans, serpins will regulate hemostasis, a process that stops bleeding, angiogenesis, tissue remodeling, or inflammation.[41] Some of the commonly known serpins are called antithrombin, heparin cofactor II, PAI-1, protease nexin-1, and Alpha 2 antiplasmin.[41] In viruses, serpins have a slightly different function. Their job is to repress the host inflammatory responses and to increase the infection, such that when the serpin was taken out of the viruses, virulence decreased.[40] Interestingly, two of the repressor fusion domains contain a serpin that is a heparin cofactor II, with the PDB ID 1JMO found in Table 4. These are

from *Mycolicibacterium mageritense* and *Gordonia hankookensis*. The literature associated with this specific protein recorded in PDB explains how the heparin cofactor II interacts with thrombin causing protease inhibition and intensive conformational changes.[42]

In the cupin family, pirins have a bicupin fold that binds iron ions. Pirins have a unique two cross linked β-barrel domain.[43] Highly conserved amongst these proteins is a metal binding site in the N-terminal barrel, thus relating to its function. Pirins have been described as nuclear proteins suggested to be transcription cofactors.[43] It is thought that these proteins play an essential role in biological processes and since they are highly conserved amongst bacteria, plants, and mammals.[44] Even though pirins are understudied, they have various functions in each organism within which they are found. Human pirins are transcriptional co-regulators of transcription factors whereas bacterial pirins serve a role in the synthesis of antibiotics.[44,45] Interestingly, increased pirin expression is commonly found as a response to oxidative stress, thus leading to cell death.[46] Another feature that has been found in pirins is the ability to facilitate nuclear factor κB, which is a cellular regulator that signals for inflammatory responses, in DNA binding when the center of it contains iron[46]. In the fusion domain from the *Nocardioides sp.*, a pirin protein has been fused to the repressor. Its PDB ID is 6D0P and is described as "1.88 Angstrom Resolution Crystal Structure of Quercetin 2,3-dioxygenase from *Acinetobacter baumannii*". There is no literature associated with this specific structure as of yet.

MocR is a subfamily of GntR regulators which are characterized by their 2 domains that include the N-terminal domain as a winged-helix-turn-helix domain and a second C-terminal domain used for oligomerization and effector binding.[47] MocR differs with its C terminus domain as an aspartate aminotransferase. Often, MocR is notated as GabR, which is the best characterized regulator from the GntR family deriving from *Bacillus subtilis.*[48] The N-terminus

of MocR allows for DNA binding, while the C-terminus binds to a ligand, thus causing a conformational change of the whole protein.[49] Aminotransferases are catalysts to the production of amino acids and use pyridoxal-5'-phosphate as a coenzyme.[50] The protein associated with the MocR is from *Xylanimonas ulmi*. Similarly, *Xylanimonas cellulosilytica* DSM 15894 is in the GntR regulator family. HHpred results for both proteins did not show which specific PDB ID that the repressor associated with, however we can tell from the psiBLAST results that there are both domains attached. There are 23 proteins that have 100% of our queried domain that appear conserved via HHpred results. Two examples of specific proteins in the PDB include 3EZ1, putative aminotransferase (MocR family) (YP_604413.1) from *Deinococcus geothermalis* and 3AOW, a *Pyrococcus horikoshii* kynurenine aminotransferase in complex with AKG.

An AAA domain containing protein is in the structural class of P-loop NTPases, which is a DNA binding domain containing an αβα core. This core contains Walker A and Walker B motifs commonly found in helicase enzymes.[50] Hydrolysis is driven by coordination of Mg^{2+} and water[51]. Some AAA domains contain arginine residues within their core that interact with ATP during the catalytic cycle.[51] AAA domains can be found in various organisms and systems. Some of these include a system involving a sliding clamp aiding in the DNA polymerase preventing it from falling off at the replication fork, or the loading of AAA domains to unfold proteins such as a ubiquitin-tagged protein.[51] Along with DNA binding, some AAA domains are also regulators for RNA polymerases.[51] Another common feature of AAA domains is that when substrate binding, they have been shown to stimulate ATPase activity.[51]

In the same family as AAA domain containing proteins, IstB is a homolog that bends target DNA clamping around it as a dimer. The name in part means insertion sequence that is a DNA transposon found in bacteria.[52] As a whole, DNA binding and hydrolysis support IstB so it

can bind to DNA and form its dimer.[52] In our research, we have found two types of ATP binding proteins, both from *Mycolicibacterium fortuitum*. One protein's function was specifically found to be "AAA domain-containing protein" and the other is called "ATP-binding protein IstB". HHpred results (Table 4) shows fusion proteins all contain the same domains with three domains other than the repressor. The other two domains that are a part of these two proteins seem to be correlated with the AAA family of proteins as they are identified as an endonuclease catalytic domain (PDB ID: 3R3P) and a hydrolase domain (PDB ID: 4B3F) respectively. The main domain associated with AAA and ATP binding is under the ID 8FTK.[53]

Structural Analysis

Based on the crystal structure of the TipsytheTRex immunity repressor, the typical fold of a repressor contains an N-terminal, helix-turn-helix DNA binding domain (HTH), a helical bridge, and the C-terminal stoperator domain.[4] Both the HTH and the stoperator bind to DNA, with the helix bridge linking the two domains. FASTA sequences from psiBLAST were input into Alphafold, and all 30 structures have been predicted and color coded based on the repressor domain and the fusion domain, shown in Figure 12.[54] A few of the structures will be discussed to highlight interesting features.

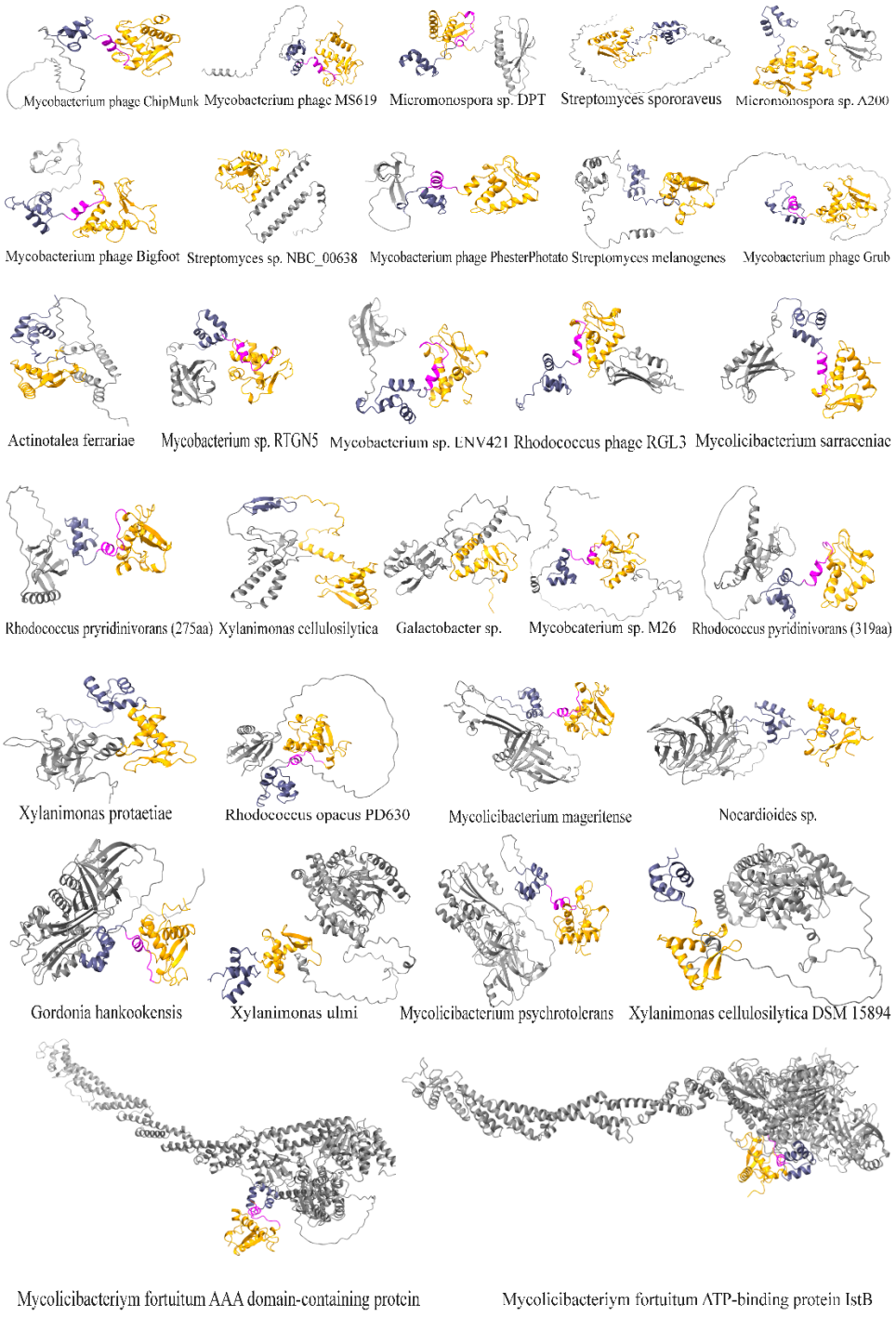

Figure 12: Thirty predicted structures of the repressor fusions. Labeled in grey is the fusion domain, in blue is the HTH DNA binding motif, in magenta is the helical bridge, and in orange is the stoperator domain.

To start, it is worth noting that some of the structures are not shown to contain a helical bridge. This could be because Alphafold either did not accurately predict the structure because it is a singular helix or simply because it does not contain one. It is difficult to judge which is the case, and it will therefore not be mentioned in this analysis.

Streptomyces sp. NBC_00638 is a 242 aa long protein which does not have a predicted function. As seen in the model shown in Figure 13, it only contains a stoperator domain and a domain with unpredicted function. Since there is no HTH, there is only one domain from the repressor that can contribute to DNA binding. The stoperator domain of the repressor is on the N-terminus, which is interesting because there were only 9 predicted proteins that contain the repressor on the N-terminus. On the other hand, the *Galactobacter sp*. protein also has an unpredicted function of the fusion domain, which lacks the HTH domain. In this protein, the stoperator domain is present on the C-terminal end of the protein. Even though the full function of the protein found in *Galactobacter sp*. is unknown, observation of the structure suggests it is more complex than the protein found in *Streptomyces sp. NBC_00638.*

Xylanimonas cellulosilytica fusion protein does not have a predicted function, is 277 residues long, and contains a C-terminal repressor. An interesting feature of this structure is that when analyzing the domains of the repressor, the HTH sequence is conserved with Adahisdi, however, the Alphafold prediction contains a helix with β-sheets.

Of the predicted functions, there are three serpins amongst these structures. They are from *Mycolicibacterium mageritense, Gordonia hankookensis,* and *Mycolicibacterium psychrotolerans*. These three all contain the repressor on the C-terminus and their sizes are 422aa, 602aa, 648aa, respectively. Each of these three structures conserves the β-sheets, α-helices, and the reactive loop of the serpin domain.

Both of the proteins from *Mycolicibacterium fortuitum* are over 1000aa long. These proteins contain 3 domains other than the C-terminal repressor domain, and they conserve a tightly coiled α-helix that compacts the domains.

Discussion

This study attempts to determine what the repressors do other than simply silence lytic gene expression. Our goal is to better understand the synergy between repressors and the diverse protein domains that make up these novel proteins.

Discovering mechanisms used by bacteriophages to evade host detection systems has been of great importance. A great example of this is the monomeric repressor found in TipsytheTRex.[4] While this structure has been characterized as a DNA binding immunity repressor, we have discovered a great number of protein homologs of diverse function that are larger than the typical ~180aa long repressor. We conducted a psiBLAST search which identified 30 proteins larger than 230 aa long, ten of which contained a domain with functions we have investigated deeper.

The results obtained in Table 3 show that the majority of the fusions are found in bacteria with few identified from bacteriophages. This is interesting, because the immunity repressor that we study is found in bacteriophages. This suggests that bacteria may have hijacked this phage protein to diversify the function of their protein repertoire. Alternatively, the repressor domain gene may have originated in bacteria, later transferred to bacteriophages genome. As stated earlier, we cannot rule out the possibility that these fusion proteins are in prophages, and this will need to be further investigated. The fusions found in bacteriophages were all characterized as immunity repressors, and they did not have a significant secondary domain attached to them. From the various types of bacteria that were identified, all were gram-stain-positive bacteria and

primarily rod shaped. The repressor and the fusion domains from bacteria are working together to serve a purpose, whether that be DNA binding or a different function, there is more to be learned. Bacteriophages identified as having repressor fusion proteins belong to clusters A, F, and CA. Discovering repressors from bacteriophages in clusters F and C is fascinating because, to date, repressors have only been validated from bacteriophages belonging to the clusters A, G, K, I, N, P, and Q.[4]

DNA binding is a common theme with these repressors and with some of the fusions. For example, MocR proteins, have a winged HTH motif on their N-terminus. AAA and ATP binding proteins, also bind to DNA and have DNA clamping features as well. What has not been discovered from this research is how bacteria use repressor proteins with two different DNA binding domains.

There are also a few functions other than DNA binding that are associated with the fusion repressor proteins. Fusion proteins containing serpins are protease inhibitors, have been identified exclusively in bacteria. It would be interesting to see if the bacteria use these proteins mechanism to target proteases in bacteriophages. Proteases found in bacteriophages can be found during phage assembly during the transition from procapsid to a mature capsid.[2] Pirins are unique because they bind to metals as nuclear proteins localized in the cytoplasm. They also induce cell death when exposed to oxidative stress. What the functions of these are in combination with the repressors is still unknown.

Lastly, by visualizing the structures of all the repressor fusion domains, we can compare these proteins of known structure. Observing the structures, it appears that even if there is not currently a predicted function, many include more than just the repressor. Commonly seen among additional domains is a linker that connects it to the repressor. The two ATP binding

proteins are not seen to have any loops but rather a coiled α-helix. It would be interesting to see if this coil brings the two domains together such that they will function as one unit. In order to elucidate the actual structures of each repressor fusion, obtaining the crystal structure of each protein would be required. It would also be beneficial to see how these proteins bind to DNA.

Future Directions

Bioinformatics has been helpful in discovering other repressor proteins and the domains that they are fused with. Now that we have a predicted structure of all proteins identified with HHpred, more work could be done biochemically characterizing them to further study their function. To do this, we could take the sequence given from the psiBLAST results and express and purify the protein. If it is soluble, we can then use it for additional testing. As preliminary data for future directions, we have already started to biochemically characterize the repressor fusion protein from *Mycobacterium sp. ENV421* and *Mycolicibacterium mageritense*, which has a repressor domain fused to a domain of unknown function and a serpin, respectively. Using the FASTA sequence given from psiBLAST, we were able to order codon optimized genes which then were expressed in *E. coli* Bl21 cells. From here, we were able to purify the protein using a Nickel affinity column and then ran the product of this on a SDS-PAGE gel to see if it was soluble. We found that *Mycobacterium sp. ENV421* was pure and soluble (Figure 14, lane 5), whereas *Mycolicibacterium mageritense* was not soluble (Figure 14, lane 8). Obtaining a soluble protein means that the process of producing a crystal structure that can be initiated. Protein crystallization is a method used to create structured lattices for complex macromolecules, allowing for greater stability and visualization of the protein structure.

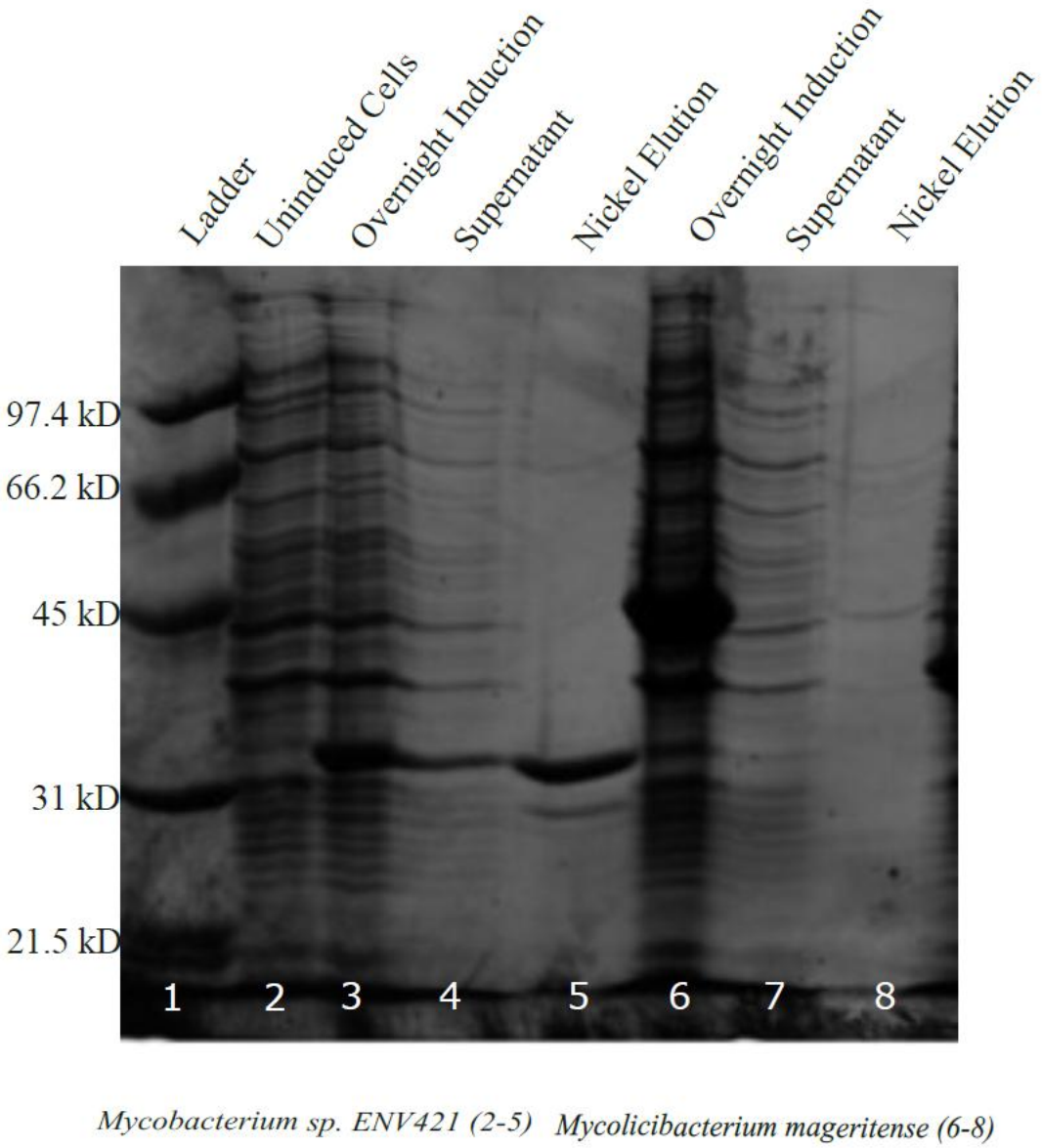

Figure 13: SDS-PAGE analysis of the repressor proteins from *Mycobacterium sp. ENV421* and *Mycolicibacterium mageritense*.

For domains with known functions, once we have expressed the protein, we can begin to study how the domain functions with the repressor attached to it. Does it have a target? Is the extra function necessary for the survival of the bacteria or the phage?

After obtaining a soluble protein, we can start to search for its purpose in relation to bacteria. Most of the repressor fusions found were from bacteria so this would be interesting to

characterize. Using bioinformatics, we can also search for the closest related bacteriophage that these repressor fusions belong to. From here we can see if the related repressor fusion binds to the DNA of the closely related bacteriophage. Furthermore, would the protein bind to the DNA of TipsytheTRex, from which the repressor was first characterized? If it does not, does the extra domain have a correlation to its DNA binding properties?

Many of these fusions have a domain that is characterized as having unknown function. Doing more research, we could attempt to see what function this domain serves. Does it impact DNA binding? What relationship does it have pertaining to silencing genes? For the bacteriophage fusions, if this extra domain was removed or mutated, would it affect its ability to infect the host? Does the location of the repressor matter when it is attached to the fusion domain? We could splice off the domain to see if silencing lytic genes continues, and we could see if any other function was lost during this process. On the other hand, we could keep the domain and splice the repressor end to see what would happen.

Another question to ask with the monomeric repressor is where did it come from? Is the origin from bacteria or phages? We can use the sequences of these repressor fusions to do bioinformatic research on where the repressor originated. By taking the sequences found in psiBLAST we can use strict algorithms to create a phylogenetic tree that will show the roots of where each organism came from and how closely related, they are to each other by predicting a common ancestor. As preliminary data we have used a basic algorithm using unipro UGENE, a unified bioinformatics toolkit, to create a phylogenetic tree.[55] Using the FASTA sequences obtained from psiBLAST we used the align with ClustalW feature, from there selecting the build phylogenetic tree function. Using the default parameters, it produced a tree including all thirty of the repressor fusion proteins, (Figure 14). This tree gives us the basis of how all of the proteins

are connected, however it does not fully connect to a common ancestor, therefore a deeper investigation needs to be conducted.

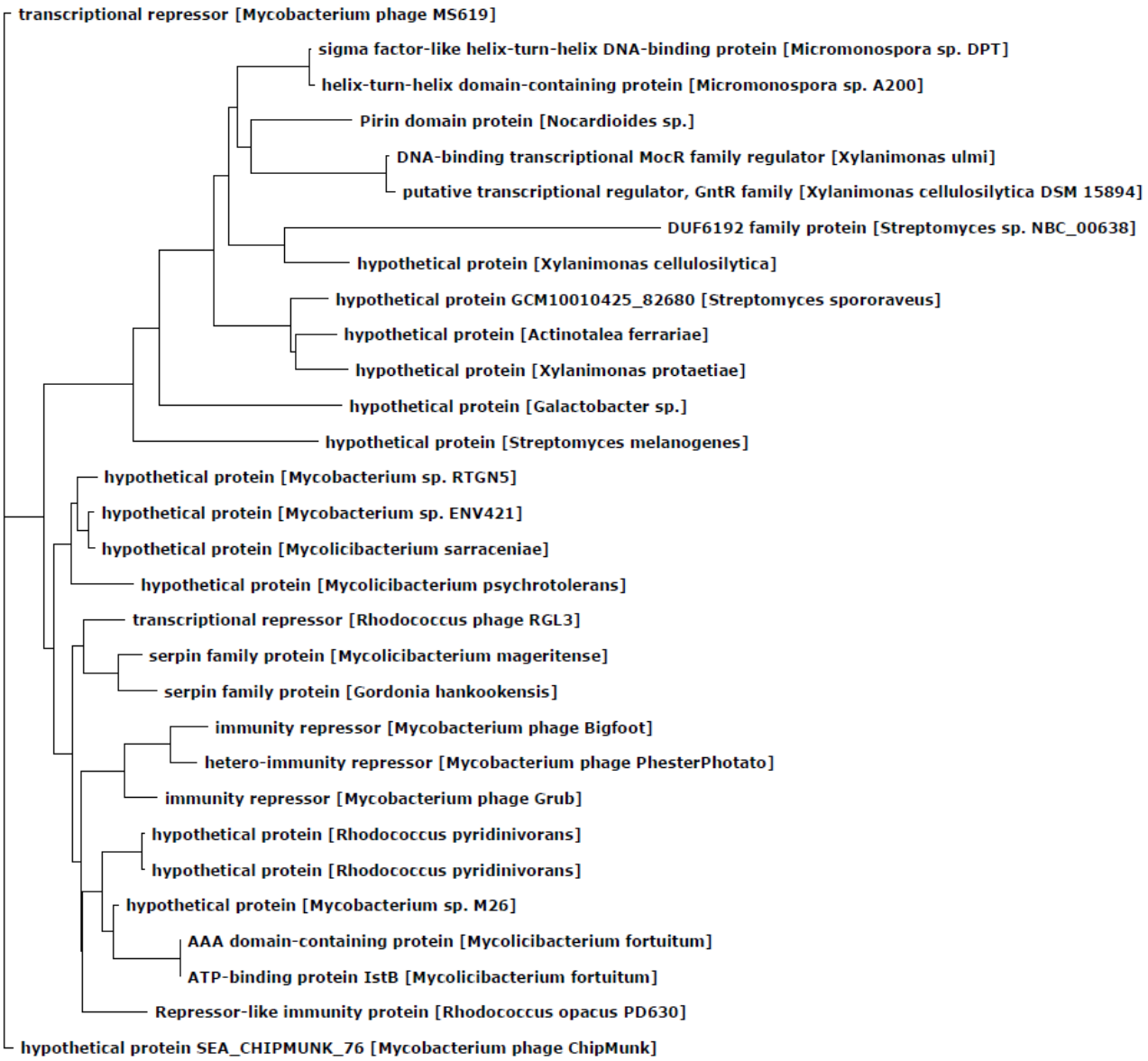

Figure 14: Phylogenic tree using repressor fusion protein sequences derived from psiBLAST.

CHAPTER FOUR: MATERIALS AND METHODS

Nucleotide Modification Discovery

Preparation of Cells

Microbacterium foliorum NRRL B-24224 cells obtained from a -80 ºC freezer stock were streaked onto peptone yeast calcium agar (PYCa) plates which contain 4.5 mM $CaCl_2$ and 0.1% dextrose. A single colony was picked and added to 50mL PYCa media containing 4.5 mM $CaCl_2$ and 0.1% dextrose in a baffled flask. The culture was grown with shaking at 30 ºC for 72 hours to reach saturation.

Virus Preparation

TinyTimothy, Kors, and Andromedas were found at Western Carolina University as a part of the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program. After extraction, the virus was plaque purified, sequenced, and recorded in GenBank under accession numbers MK878904 (TinyTimothy) and MH590606 (Andromedas). Kors has yet to be deposited to GenBank. The sequencing facility used for Kors and Andromedas was Pittsburgh Bacteriophage Institute using Illumina shotgun sequencing methods. The sequencing facility used for TinyTimothy was the Western Carolina University Forensic Science Sequencing facility using Illumina shotgun sequencing methods.

To generate high titer virus stocks, 50 µL serial dilutions of each virus were mixed with 0.5 mL of cells, 4.5 mL PYCa, 1M $CaCl_2$ and 40% dextrose top agar, and then poured on PYCa plates. Plates were incubated for 24-48 hours at 30 ºC. Webbed plates were flooded with 8 mL of Phage Buffer (10 mM Tris, 10mM $MgSO_4$, 68 mM NaCl, 1mM $CaCl_2$), for ~24 hours at 4 ºC, and lysates were harvested by filtration using 0.22 µm PES syringe filters. For viral DNA

extraction, 8 mLs of high titer virus stock was treated with 4 mLs of phage precipitation solution (30% w/v Polyethylene Glycol 8000 and 3.3 M NaCl). This was allowed to incubate overnight at 4 ºC on a nutator for maximum yield. A high-speed centrifuge spin at 10,000xg for 20 minutes at 4ºC pelleted the virus, and the supernatant was removed, leaving the pellet. The phage pellet was suspended in 0.5 mL phage buffer.

Removal of non-phage DNA and RNA was achieved by adding 500 µL of the resuspended phage with 6.25 µL 1 M $MgCl_2$, 0.5 µL DNAse and 1 µL RNAse A to the tube then vortexing and incubating for 30 minutes at room temperature. 20 µL 0.5 M EDTA, 2.5 µL ProteinaseK, and 25 µL 10% SDS was added and incubated at 55ºC for an hour after vortexing vigorously. In a fume hood, 500 µL of phenol:choloroform:isoamyl solution was added to the tube, the solution was mixed, then centrifuged for 5 minutes at 13,000 rpm. The top aqueous layer was transfered to a new tube. For ethanol precipitation of phage DNA, 1 mL of 95% ethanol and 50µL of 1M sodium acetate solution was added to the aqueous solution, inverted to mix, then placed on ice for 5 minutes. The sample was centrifuged for 10 minutes at 13,000 rpm, and the supernatant removed. The resulting DNA pellet was washed with 500 µL of 70% ethanol, then centrifuged for 5 minutes at 13,000 rpm. The DNA pellet was allowed to air dry, then was resuspended in 25 µL of cloning water. The amount of purified DNA was quantified using a Thermofisher Nanodrop spectrophotometer.

Restriction Enzyme Digest

For each restriction enzyme digest panel, between 0.5-1 µg of DNA was digested with 1µL of the corresponding restriction enzyme in the appropriate NEB buffer at 37ºC for 30 minutes. Products were then run on an 0.8% agarose gel in 1X TAE buffer at 100 volts for 1 hour. A 1kb ladder, as well as undigested pure DNA, were included with each gel.

Virtual Restriction Enzyme Digest

To create virtual restriction enzyme digestions, the DNA sequence of each phage was uploaded to NEB's REBsites website (https://tools.neb.com/REBsites/index.php), and parameters were set to digest the phage DNA with the restriction enzymes BamHI, ClaI, EcoRI, HaeIII,HindIII, NspI, SacII, and SacI.

DNA Preparation for HPLC

After DNA extraction, a mixture of 1 µL RNase, 5 µL purified DNA, and 10 µL Phage Buffer was incubated 37 ºC for 3 hours. NEB's Monarch Spin PCR & DNA Cleanup Kit was used to clean and desalt the DNA for further analysis. Purified DNA was quantified using a Thermo Nanodrop. The Nucleoside Digestion Mix kit from New England Biolabs was used to digest genomic DNA into nucleosides. The included protocol from the kit was followed, scaling up three times for all reagents and including 5 µg of DNA for each reaction.

PCR

Genomic DNA template for PCR testing was derived from boiling 50 uL of fresh high titer phage stock. NEB kits and protocols were used to prepare each PCR. The polymerases used for PCR were: OneTaq® 2X Master Mix, Q5© Hot Start High-Fidelity 2X Master Mix, and Q5U® Hot Start High-Fidelity DNA Polymerase. Primers were purchased from Integrated DNA Technologies (IDT). TinyTimothy's primer set was designed to amplify gene 2, Andromedas' primer was from gene 46, and Kors from gene 51. Once PCR products were complete, each sample was loaded on a 0.8% agarose gel prepared in 1X TAE buffer. The gels were run at 100 V for 1 h at room temperature in 1X TAE buffer.

LC-MS

After nucleotides were digested into nucleosides, they were injected onto an Agilent 1100 liquid chromatography system consisting of a degasser, quaternary pump, autosampler, column compartment and diode array detector. The column used is a Waters T3 C18, 100 angstrom, 3 micron particle size with dimensions of 4.6 x 150 mm. The mobile phases used for the LC were Mobile Phase A: Water + 0.1% Formic Acid, Mobile Phase B: Methanol + 0.1% Formic Acid. LC method parameters were as follows: Begin at 98% Mobile Phase A and ramp up to 25% Mobile Phase B over 60 minutes. The injection volume: 100 µL, with the flow rate at 0.5 mL/min. The detector wavelength was set at 260 nm, and the column oven was ambient (25ºC). The mass spectrometer parameters were as follows: ESI-positive polarity with a full scan type in centroid mode with a mass range of 100-800 m/z, spray voltage of 4.6 kV, sheath gas pressure 40 arb units, aux gas pressure 12 arb units, and the capillary temperature set to 350 ºC.

Bioinformatics

Repressor Fusions Discovery

psiBLAST (NCBI, 2004)[56] [7]was used to discover fusion proteins containing the adahisdi stoperator sequence (>Adahisdi gp71 MRTTREQLPRLSLEVIEALKATGETEADIARMYGVTPQAVSWHVHTYGGKLTARQVIRR EYPFKVPEPLSQCTPHKRLRDHGEYIATRGKGMKEYKLKRLRSFYRMLRENNWVVEFD PNIPPIPGVSKRGGWAYRERQESDEDLLIRVNEYTTLSEIGRHHIWRFPSVEP, highlighted in yellow is the stoperator sequence). Default parameters were used except increasing the output to 20000 maximum sequences. Iterations were run until no new sequences were found. Proteins larger than 230 residues were selected and their sequences downloaded in FASTA format. These

sequences were loaded into UCSF ChimeraX version 1.9 (UCSF ChimeraX, 2023)[57] for structure prediction using the program AlphaFold[54] found at (https://alphafold.ebi.ac.uk/). To use AlphaFold the following steps were taken: tools, structure prediction, alphafold, paste desired sequence, click fetch. This produced a graphic of each protein structure used.

Through Max Planck Institute's bioinformatics toolkit, HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred)[58] was used with default parameters to locate the repressor portion of the protein, as well as predicted functions of the fusion proteins. The sequence of each protein was the input from downloading the FASTA files retrieved from the psiBLAST iterations. The function of the protein corresponds with an ID that is found in RCSB's Protein Data Bank (https://www.rcsb.org/)[59].

Biochemical Characteristics of Repressor Fusion

Codon-optimized genes were purchased from Twist Biosciences for the repressor fusion proteins characterized. The following are the descriptions of the proteins purchased >WP_083613183.1 serpin family protein [*Mycolicibacterium mageritense*] and >WP_233209289.1 hypothetical protein [*Mycobacterium sp. ENV421*]. These sequences were obtained through the psiBLAST experiment described above. These proteins were expressed by transforming both plasmids into *E.coli* BL21 (DE3) cells. Cells were grown in 50 mL LB at 37 ºC with shaking at 225 RPM to an OD600 = ~0.5, chilled on ice, then induced overnight at 16 ºC with 1 mM IPTG. Cells were harvested by centrifugation at 3,000 RPM for 30 minutes, and the cell pellets were resuspended in 35 mL of Nickel Buffer A (50mM Tris 8.0, 0.5 M NaCl, 10% glycerol and 1mM DTT), and stored at -80 ºC.

For protein purification, cells were thawed on ice, and 0.1 mg/mL lysozyme was added at 4 ºC for 30 minutes. Cells were lysed via sonication, then spun at 17,000 RPM for 45 minutes.

The supernatant was passed over a gravity nickel column equilibrated in Nickel buffer A. This was passed over two times, then columns washed with Nickel buffer A, then Nickel buffer A with 40mM imidazole. Protein was eluted with Nickel buffer A plus 250 mM imidazole. Protein solubility and identification of expression of the *Mycobacterium sp, ENV421* protein was analyzed via SDS-PAGE.

For large-scale protein expression, 3 L of *E.coli* cells were used, with expression and harvesting the same as described above for the small-scale cultures. Cells were resuspended in ~100 mL of Nickel buffer A, lysed by sonication, then spun at 17,000 RPM for 45 minutes. The resulting supernatant was passed over a 5 mL BioRad Nickel column equilibrated in Nickel Buffer A, washed with 10mM of imidazole, followed by a 40mM imidazole wash. The protein was eluted using a 40-250 mM imidazole gradient over 200mL, with a collection of 3 mL fractions. Fractions were run on a 15% SDS-PAGE gel to identify the repressor fusion protein. Pure fractions were pooled and dialyzed overnight in 2L of buffer (20mM Tris 8.0, 0.5 M NaCl, 0.5 mM EDTA, 5% glycerol and 0.5mM DTT). The protein was concentrated using a spin concentrator, then quantified using a Nanodrop spectrometer. The protein was aliquoted into 30µL aliquots, flash frozen in liquid nitrogen, and stored at -80ºC.

Phylogenetic Tree

Using default parameters with the bioinformatics program Unipro UGENE: a unified bioinformatics toolkit (Unipro UGENE, 2012)[55], we input all 30 of the FASTA sequences of the repressor fusion proteins into the multiple alignment feature, aligning using the ClustalW imbedded into UGENE. After alignment, the feature “build tree” was selected and run using default parameters to create the phylogenetic tree.

REFERENCES

(1) Allan Campbell. The Future of Bacteriophage Biology. *Nat Rev Genet* **2003**, *4* (6), 471–477.

(2) Russell, D. A.; Hatfull, G. F. PhagesDB: The Actinobacteriophage Database. *Bioinformatics* **2017**, *33* (5), 784–786. https://doi.org/10.1093/bioinformatics/btw711.

(3) Hatfull, G. F. Mycobacteriophages: From Petri Dish to Patient. *PLoS Pathog* **2022**, *18* (7), 1–25. https://doi.org/10.1371/journal.ppat.1010602.

(4) McGinnis, R. J.; Brambley, C. A.; Stamey, B.; Green, W. C.; Gragg, K. N.; Cafferty, E. R.; Terwilliger, T. C.; Hammel, M.; Hollis, T. J.; Miller, J. M.; Gainey, M. D.; Wallen, J. R. A Monomeric Mycobacteriophage Immunity Repressor Utilizes Two Domains to Recognize an Asymmetric DNA Sequence. *Nat Commun* **2022**, *13* (1). https://doi.org/10.1038/s41467-022-31678-6.

(5) Mavrich, T. N.; Hatfull, G. F. Evolution of Superinfection Immunity in Cluster A Mycobacteriophages. *mBio* **2019**, *10* (3). https://doi.org/10.1128/mBio.00971-19.

(6) Brown, K. L.; Sarkis, G. J.; Wadsworth, C.; Hatfull, G. F. Transcriptional Silencing by the Mycobacteriophage L5 Repressor. *EMBO J* **1997**, *16* (19), 5914–5921. https://doi.org/10.1093/EMBOJ/16.19.5914.

(7) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res* **1997**, *25* (17), 3389–3402. https://doi.org/10.1093/NAR/25.17.3389.

(8) Weigele, P.; Raleigh, E. A. Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. *Chem Rev* **2016**, *116* (20), 12655–12687. https://doi.org/10.1021/acs.chemrev.6b00114.

(9) Pezo, V.; Jaziri, F.; Bourguignon, P. Y.; Louis, D.; Jacobs-Sera, D.; Rozenski, J.; Pochet, S.; Herdewijn, P.; Hatfull, G. F.; Kaminski, P. A.; Marliere, P. Noncanonical DNA Polymerization by Aminoadenine-Based Siphoviruses. *Science (1979)* **2021**, *372* (6541), 520–524. https://doi.org/10.1126/science.abe6542.

(10) *Detection of Modified Bases in Bacteriophage Genomic DNA*.

(11) Crippen, C. S.; Lee, Y.-J.; Hutinet, G.; Shajahan, A.; Sacher, J. C.; Azadi, P.; de Crécy-Lagard, V.; Weigele, P. R.; Szymanski, C. M. Deoxyinosine and 7-Deaza-2-Deoxyguanosine as Carriers of Genetic Information in the DNA of Campylobacter Viruses . *J Virol* **2019**, *93* (23). https://doi.org/10.1128/jvi.01111-19.

(12) Borges, A. L.; Radkov, A.; Thuy-Boun, P. S. A Workflow to Isolate Phage DNA and Identify Nucleosides by HPLC and Mass Spectrometry. **2022**. https://doi.org/10.57844/ARCADIA-1EY9-J808.

(13) Li, Y.; Chen, F.; Dong, K.; Wang, G. Actinotalea Ferrariae Sp. Nov., Isolated from an Iron Mine, and Emended Description of the Genus Actinotalea. *Int J Syst Evol Microbiol* **2013**, *63* (PART9), 3398–3403. https://doi.org/10.1099/IJS.0.048512-0/CITE/REFWORKS.

(14) Hahne, J.; Isele, D.; Von Heilborn, D. H.; Czaja-Hasse, L.; Huttel, B.; Lipski, A. Galactobacter Caseinivorans Gen. Nov., Sp. Nov. and Galactobacter Valiniphilus Sp. Nov., Two Novel Species of the Family Micrococcaceae, Isolated from High Bacterial Count Raw Cow's Milk. *Int J Syst Evol Microbiol* **2019**, *69* (9), 2862–2869. https://doi.org/10.1099/IJSEM.0.003570/CITE/REFWORKS.

(15) Pope, W. H.; Mavrich, T. N.; Garlena, R. A.; Guerrero-Bustamante, C. A.; Jacobs-Sera, D.; Montgomery, M. T.; Russell, D. A.; Warner, M. H.; Hatfull, G. F. Bacteriophages of Gordonia Spp. Display a Spectrum of Diversity and Genetic Relationships. *mBio* **2017**, *8* (4), e01069-17. https://doi.org/10.1128/MBIO.01069-17.

(16) *The Actinobacteriophage Database| Hosts*. https://phagesdb.org/hosts/ (accessed 2025-03-14).

(17) Hirsch, A. M.; Valdés, M. Micromonospora: An Important Microbe for Biomedicine and Potentially for Biocontrol and Biofuels. *Soil Biol Biochem* **2010**, *42* (4), 536–542. https://doi.org/10.1016/J.SOILBIO.2009.11.023.

(18) Yan, S.; Zeng, M.; Wang, H.; Zhang, H. Micromonospora: A Prolific Source of Bioactive Secondary Metabolites with Therapeutic Potential. *J Med Chem* **2022**, *65* (13), 8735–8771. https://doi.org/10.1021/ACS.JMEDCHEM.2C00626/ASSET/IMAGES/MEDIUM/JM2C00626_0047.GIF.

(19) Masuda, H.; McClay, K.; Steffan, R. J.; Zylstra, G. J. Characterization of Three Propane-inducible Oxygenases in Mycobacterium Sp. Strain ENV421. *Lett Appl Microbiol* **2012**, *55* (3), 175–181. https://doi.org/10.1111/J.1472-765X.2012.03290.X.

(20) Phelippeau, M.; Asmar, S.; Osman, D. A.; Sassi, M.; Robert, C.; Michelle, C.; Musso, D.; Drancourt, M. “Mycobacterium Massilipolynesiensis” Sp. Nov., a Rapidly-Growing Mycobacterium of Medical Interest Related to Mycobacterium Phlei. *Sci Rep* **2017**, *7*, 40443. https://doi.org/10.1038/SREP40443.

(21) Thompson, R. M.; Fox, E. M.; Del Carmen Montero-Calasanz, M. Draft Genome Sequences of Five Mycobacterium Strains, Isolated from Alnus Glutinosa Root Nodules. **2024**. https://doi.org/10.1128/mra.01132-23.

(22) Gupta, R. S.; Lo, B.; Son, J. Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Front Microbiol* **2018**, *9* (FEB), 67. https://doi.org/10.3389/FMICB.2018.00067/FULL.

(23) Morgado, S.; Ramos, N. de V.; Freitas, F.; da Fonseca, É. L.; Vicente, A. C. Mycolicibacterium Fortuitum Genomic Epidemiology, Resistome and Virulome. *Mem Inst Oswaldo Cruz* **2022**, *116*, e210247. https://doi.org/10.1590/0074-02760210247.

(24) Nayak, T.; Kakkar, A.; Singh, R. K.; Jaiswal, L. K.; Singh, A. K.; Temple, L.; Gupta, A. Isolation and Characterization of a Novel Mycobacteriophage Kashi-VT1 Infecting Mycobacterium Species. *Front Cell Infect Microbiol* **2023**, *13*, 1173894. https://doi.org/10.3389/FCIMB.2023.1173894/BIBTEX.

(25) Niitsu, T.; Kuge, T.; Fukushima, K.; Matsumoto, Y.; Abe, Y.; Okamoto, M.; Haduki, K.; Saito, H.; Nitta, T.; Kawano, A.; Matsuki, T.; Motooka, D.; Tsujino, K.; Miki, K.; Nakamura, S.; Kida, H.; Kumanogoh, A. Pleural Effusion Caused by Mycolicibacterium Mageritense in an Immunocompetent Host: A Case Report. *Front Med (Lausanne)* **2021**, *8*, 797171. https://doi.org/10.3389/FMED.2021.797171/BIBTEX.

(26) Trujillo, M. E.; Velázquez, E.; Kroppenstedt, R. M.; Schumann, P.; Rivas, R.; Mateos, P. F.; Martínez-Molina, E. Mycobacterium Psychrotolerans Sp. Nov., Isolated from Pond Water near a Uranium Mine. *Int J Syst Evol Microbiol* **2004**, *54* (5), 1459–1463. https://doi.org/10.1099/IJS.0.02938-0/CITE/REFWORKS.

(27) Tran, P. M.; Dahl, J. L. Mycobacterium Sarraceniae Sp. Nov. and Mycobacterium Helvum Sp. Nov., Isolated from the Pitcher Plant Sarracenia Purpurea. *Int J Syst Evol Microbiol* **2016**, *66* (11), 4480–4485. https://doi.org/10.1099/IJSEM.0.001377/CITE/REFWORKS.

(28) Ma, Y.; Wang, J.; Liu, Y.; Wang, X.; Zhang, B.; Zhang, W.; Chen, T.; Liu, G.; Xue, L.; Cui, X. Nocardioides: "Specialists" for Hard-to-Degrade Pollutants in the Environment. *Molecules* **2023**, *28* (21), 7433. https://doi.org/10.3390/MOLECULES28217433.

(29) Pulverer, G.; Schütt-Gerowitt, H.; Schaal, K. P. Bacteriophages of Nocardia Asteroides. *Med Microbiol Immunol* **1975**, *161* (2), 113–122. https://doi.org/10.1007/BF02121752.

(30) Chatterjee, A.; DeLorenzo, D. M.; Carr, R.; Moon, T. S. Bioconversion of Renewable Feedstocks by Rhodococcus Opacus. *Curr Opin Biotechnol* **2020**, *64*, 10–16. https://doi.org/10.1016/j.copbio.2019.08.013.

(31) Guo, W.; Duan, J.; Shi, Z.; Yu, X.; Shao, Z. Biodegradation of PET by the Membrane-Anchored PET Esterase from the Marine Bacterium Rhodococcus Pyridinivorans P23. *Communications Biology 2023 6:1* **2023**, *6* (1), 1–13. https://doi.org/10.1038/s42003-023-05470-1.

(32) *Streptomyces | Antibiotic Production, Soil Microbe & Actinomycete | Britannica*. https://www.britannica.com/science/Streptomyces (accessed 2025-03-14).

(33) El-Naggar, N. E. A. Streptomyces-Based Cell Factories for Production of Biomolecules and Bioactive Metabolites. *Microbial Cell Factories Engineering for Production of Biomolecules* **2021**, 183–234. https://doi.org/10.1016/B978-0-12-821477-0.00011-8.

(34) Chang, P. C.; Liu, S. C.; Ho, M. C.; Huang, T. W.; Huang, C. H. A Soil-Isolated Streptomyces Spororaveus Species Produces a High-Molecular-Weight Antibiotic AF1

against Fungi and Gram-Positive Bacteria. *Antibiotics* **2022**, *11* (5), 679. https://doi.org/10.3390/ANTIBIOTICS11050679/S1.

(35) Rivas, R.; Sánchez, M.; Trujillo, M. E.; Zurdo-Piñeiro, J. L.; Mateos, P. F.; Martínez-Molina, E.; Velázquez, E. Xylanimonas Cellulosilytica Gen. Nov., Sp. Nov., a Xylanolytic Bacterium Isolated from a Decayed Tree (Ulmus Nigra). *Int J Syst Evol Microbiol* **2003**, *53* (1), 99–103. https://doi.org/10.1099/IJS.0.02207-0/CITE/REFWORKS.

(36) Heo, J.; Kim, S. J.; Kim, M. A.; Tamura, T.; Saitou, S.; Hamada, M.; Kim, J. S.; Hong, S. B.; Kwon, S. W. Lactococcus Protaetiae Sp. Nov. and Xylanimonas Protaetiae Sp. Nov., Isolated from Gut of Larvae of Protaetia Brevitarsis Seulensis. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* **2020**, *113* (7), 1009–1021. https://doi.org/10.1007/S10482-020-01413-6/TABLES/4.

(37) Labra, O. R.; Montiel-Garcia, D.; Reddy, V. S.; Goodrum, F. Virus World Database (VWdb), an API-Enabled Database of Virus Taxonomy. *J Virol* **2023**, *97* (8). https://doi.org/10.1128/JVI.00620-23.

(38) Gabler, F.; Nam, S. Z.; Till, S.; Mirdita, M.; Steinegger, M.; Söding, J.; Lupas, A. N.; Alva, V. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **2020**, *72* (1), e108. https://doi.org/10.1002/CPBI.108.

(39) Law, R. H. P.; Zhang, Q.; McGowan, S.; Buckle, A. M.; Silverman, G. A.; Wong, W.; Rosado, C. J.; Langendorf, C. G.; Pike, R. N.; Bird, P. I.; Whisstock, J. C. An Overview of the Serpin Superfamily. *Genome Biol* **2006**, *7* (5), 1–11. https://doi.org/10.1186/GB-2006-7-5-216/FIGURES/3.

(40) Maas, C.; de Maat, S. Therapeutic SERPINs: Improving on Nature. *Front Cardiovasc Med* **2021**, *8*, 648349. https://doi.org/10.3389/FCVM.2021.648349.

(41) Bouton, M.; Geiger, M.; Sheffield, W. P.; Irving, J. A.; Lomas, D. A.; Song, S.; Satyanarayanan, R. S.; Zhang, L.; McFadden, G.; Lucas, A. R. The Under-appreciated World of the Serpin Family of Serine Proteinase Inhibitors. *EMBO Mol Med* **2023**, *15* (6), e17144. https://doi.org/10.15252/EMMM.202217144.

(42) Baglin, T. P.; Carrell, R. W.; Church, F. C.; Esmon, C. T.; Huntington, J. A. Crystal Structures of Native and Thrombin-Complexed Heparin Cofactor II Reveal a Multistep Allosteric Mechanism. *Proc Natl Acad Sci U S A* **2002**, *99* (17), 11079. https://doi.org/10.1073/PNAS.162232399.

(43) Pang, H.; Bartlam, M.; Zeng, Q.; Miyatake, H.; Hisano, T.; Miki, K.; Wong, L. L.; Gao, G. F.; Rao, Z. Crystal Structure of Human Pirin: AN IRON-BINDING NUCLEAR PROTEIN AND TRANSCRIPTION COFACTOR. *Journal of Biological Chemistry* **2004**, *279* (2), 1491–1498. https://doi.org/10.1074/JBC.M310022200.

(44) Xu, G.; Shi, J.; Qiao, J.; Liao, P.; Yong, B.; Zhong, K. Genome-Wide Identification and Characterization of the Pirin Gene Family in Nicotiana Benthamiana. *Genes 2025, Vol. 16, Page 121* **2025**, *16* (2), 121. https://doi.org/10.3390/GENES16020121.

(45) Wendler, W. M. F.; Kremmer, E.; Förster, R.; Winnacker, E. L. Identification of Pirin, a Novel Highly Conserved Nuclear Protein. *Journal of Biological Chemistry* **1997**, *272* (13), 8482–8489. https://doi.org/10.1074/jbc.272.13.8482.

(46) Liu, F.; Rehmani, I.; Esaki, S.; Fu, R.; Chen, L.; De Serrano, V.; Liu, A. Pirin Is an Iron-Dependent Redox Regulator of NF-KB. *Proc Natl Acad Sci U S A* **2013**, *110* (24), 9722–9727. https://doi.org/10.1073/PNAS.1221743110/-/DCSUPPLEMENTAL/PNAS.201221743SI.PDF.

(47) Milano, T.; Angelaccio, S.; Tramonti, A.; Di Salvo, M. L.; Contestabile, R.; Pascarella, S. A Bioinformatics Analysis Reveals a Group of MocR Bacterial Transcriptional Regulators Linked to a Family of Genes Coding for Membrane Proteins. *Biochem Res Int* **2016**, *2016* (1), 4360285. https://doi.org/10.1155/2016/4360285.

(48) Hermann, L.; Dempwolff, F.; Steinchen, W.; Freibert, S. A.; Smits, S. H. J.; Seubert, A.; Bremer, E. The MocR/GabR Ectoine and Hydroxyectoine Catabolism Regulator EnuR: Inducer and DNA Binding. *Front Microbiol* **2021**, *12*, 764731. https://doi.org/10.3389/FMICB.2021.764731/FULL.

(49) Okuda, K.; Ito, T.; Goto, M.; Takenaka, T.; Hemmi, H.; Yoshimura, T. Domain Characterization of Bacillus Subtilis GabR, a Pyridoxal 5′-Phosphate-Dependent Transcriptional Regulator. *The Journal of Biochemistry* **2015**, *158* (3), 225–234. https://doi.org/10.1093/JB/MVV040.

(50) Snider, J.; Thibault, G.; Houry, W. A. The AAA+ Superfamily of Functionally Diverse Proteins. *Genome Biol* **2008**, *9* (4), 216. https://doi.org/10.1186/GB-2008-9-4-216.

(51) Khan, Y. A.; White, K. I.; Brunger, A. T. The AAA+ Superfamily: A Review of the Structural and Mechanistic Principles of These Molecular Machines. *Crit Rev Biochem Mol Biol* **2022**, *57* (2), 156–187. https://doi.org/10.1080/10409238.2021.1979460.

(52) de la Gándara, Á.; Spínola-Amilibia, M.; Araújo-Bazán, L.; Núñez-Ramírez, R.; Berger, J. M.; Arias-Palomo, E. Molecular Basis for Transposase Activation by a Dedicated AAA+ ATPase. *Nature 2024 630:8018* **2024**, *630* (8018), 1003–1011. https://doi.org/10.1038/s41586-024-07550-6.

(53) Appel, C. D.; Bermek, O.; Dandey, V. P.; Wood, M.; Viverette, E.; Williams, J. G.; Bouvette, J.; Riccio, A. A.; Krahn, J. M.; Borgnia, M. J.; Williams, R. S. Sen1

Architecture: RNA-DNA Hybrid Resolution, Autoregulation, and Insights into SETX Inactivation in AOA2. *Mol Cell* **2023**, *83* (20), 3692-3706.e5. https://doi.org/10.1016/J.MOLCEL.2023.09.024.

(54) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature 2021 596:7873* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(55) Okonechnikov, K.; Golosova, O.; Fursov, M.; Varlamov, A.; Vaskin, Y.; Efremov, I.; German Grehov, O. G.; Kandrov, D.; Rasputin, K.; Syabro, M.; Tleukenov, T. Unipro UGENE: A Unified Bioinformatics Toolkit. *Bioinformatics* **2012**, *28* (8), 1166–1167. https://doi.org/10.1093/BIOINFORMATICS/BTS091.

(56) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J Mol Biol* **1990**, *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

(57) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Tools for Structure Building and Analysis. *Protein Science* **2023**, *32* (11), e4792. https://doi.org/10.1002/PRO.4792.

(58) Gabler, F.; Nam, S. Z.; Till, S.; Mirdita, M.; Steinegger, M.; Söding, J.; Lupas, A. N.; Alva, V. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **2020**, *72* (1), e108. https://doi.org/10.1002/CPBI.108.

(59) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28* (1), 235–242. https://doi.org/10.1093/NAR/28.1.235.