5-2021

Proteomic Analysis of Mycobacteriophage CrimD

William Moeller

Proteomic Analysis of Mycobacteriophage CrimD

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Chemistry from
William & Mary

by

William Moeller

Accepted for Honors

John C. Poutsma

John C. Poutsma, Director

Rachel O'Brien

Rachel O'Brien

Kurt Williamson

Kurt Williamson

Williamsburg, VA
May 12, 2021

Abstract

Bacteriophages represent a large portion of the biomatter on our planet, and many of them have yet to be fully characterized. Here we discuss the proteomic analysis of a particular Bacteriophage, Mycobacteriophage CrimD. This phage was discovered on the Campus of William & Mary and has had its genome characterized. We took the next logical step of proteomic analysis.

In our analyses we made use high pressure liquid chromatography paired with linear ion trap mass spectrometry to analyze the proteome of CrimD at specific time points after the infection of its host, Mycobacterium smegmatis. Additionally, we used nanospray ionization with in-house produced analytical columns and emitters to analyze our samples. These techniques had been previously used in our lab to analyze different bacteriophages but required significant optimization in order to successfully analyze CrimD.

In our analyses we found that we are able to see different proteins being expressed in the different time samples. Consequently, we were able to assign each time point to represent a different phase in the replication cycle of CrimD, namely the lysogenic and Early and Late lytic phase of replication. We were also able to assign many proteins with unknown function to specific time points, opening the door for further characterization of these proteins and CrimD.

Table of Contents

Tables and Figures

Acknowledgements

I would firstly like to thank Prof. JC Poutsma whose guidance these last three years has been critical, especially in completing this project. I would also like to thank the Faculty and the Department of Chemistry at William & Mary who have had a large impact on me over my time at the College.

I would like to thank everyone who worked with me in the ion lab, especially Hao Qian, without whom this would not have been possible. I would also like to thank Anna Grace Towler and Sophie Messinger for their help on this project in addition to Henry Cardwell, Gwendylan Turner, and Alexis Brender A Brandis for being supportive throughout my time in lab.

I was finally like to thank the Charles Center and The National Institutes of Health for helping to fund this project.

Chapter 1. Introduction

1.1. Introduction to Bacteriophages

1.1.1 *General Introduction*

Bacteriophages, or phages, represent a significant portion of the biological material on our planet. A bacteriophage is, simply put, a virus that infects bacteria. The number of bacteriophages on the planet is estimated to be on the order of 10^{31} [1] and 10^{25} new bacteriophage infections take place every second.[2] The vast scale of bacteriophages begins to reveal the outsized impact that they have on microbial ecosystems. Phages play a role in regulating microbial ecosystems as diverse as the Pacific Ocean to a tuberculosis infection. Phages have long been of interest to science due not only to their outsized role in microbial life, but also their potential applications to humans. Phages have been used historically to treat diseases and have been proposed as potential alternatives to antibiotics for some diseases.[3,4] Additionally, nearly one in three bacteriophage proteins have no known homologs.[5] Viruses represent an enormous reservoir of genetic diversity, much of which has yet to be fully characterized. For example, when first identified, over half of SARS-CoV2 proteins were not fully characterized.[6] Study of viruses, and more specifically bacteriophages, is key to furthering understating of this enormously diverse and impactful group of biological entities.

At a basic level all viruses function the same way, they are obligate cellular parasites that hijack host cellular machinery to replicate their genomes and produce new viral particles. Bacteriophages do this by “injecting” their genetic material, most often dsDNA, through the host cell wall from where it interacts with host transcription/translation machinery beginning the infectious cycle. Many bacteriophages can replicate in two ways;

first by replicating their genome and producing new viral particles that lyse the host releasing the virions to infect new hosts. Secondly, Bacteriophages may insert their genome into the host's genome, from where it is replicated along with the host's genome. This combination of virus-host is termed a "lysogen". The first method of replication is termed "lytic" while the second is "lysogenic". All bacteriophages can undergo lytic replication, while only some can also undergo lysogenic replication. Those that can undergo both lytic and lysogenic replication are termed "temperate" phages. The lysogenic replication cycle allows a virus to hold off on releasing its progeny until it is under ideal conditions. [7]

1.1.2 *Introduction to the Phage Lab at W&M and CrimD*

In our experiments, we investigated a bacteriophage named CrimD that was first discovered in 2008 by Hilary Whelan as part of the William & Mary Phage Lab.[8] The Phage Lab at William & Mary is part of a program sponsored by the Howard Hughes Medical Institute. The Phage Lab involves a group of freshman searching around campus for phages that infect the nonpathogenic model of tuberculosis, *Mycobacterium smegmatis*. One of the phages is chosen to have its genome sequenced and full bioinformatics and gene mapping is performed as part of the lab. CrimD was the first phage that was fully sequenced as part of this project.[9] Since then, the collection of phages from this lab has grown dramatically. We are expanding on this project by taking the next logical step in characterizing these isolated phages: protein identification and differential expression.

1.1.3 *Introduction to Mycobacteriophage CrimD*

CrimD belongs to a family of viruses named *Caudovirales*, or tailed bacteriophages. *Caudovirales* are distinguished by their long "tail", they share a common origin, and their genetic information is in the form of dsDNA, under the Baltimore classification of virus, they are type 1.

They are the viruses one typically imagines when picturing a phage.[10] As it can infect mycobacteria, CrimD belongs to the mycobacteriophage group of viruses. Mycobacteriophages are known to often be able to infect multiple members of the mycobacterium genus. CrimD specifically belongs to a "cluster" of closely related phages called K1. [8,11] K1 phages are temperate phages.

Mycobacteriophage CrimD has 95 genes, numbered 1-96 (there is no gene 5).[11] The vast majority of CrimD genes are on the same strand of DNA, with only 3 being read on the opposite strand. Of the 95 genes, 91 have been assigned to Phamilies, that is, clusters of closely related genes across different phages. The remaining 4 are orphan genes with no known counterparts in any other phage.

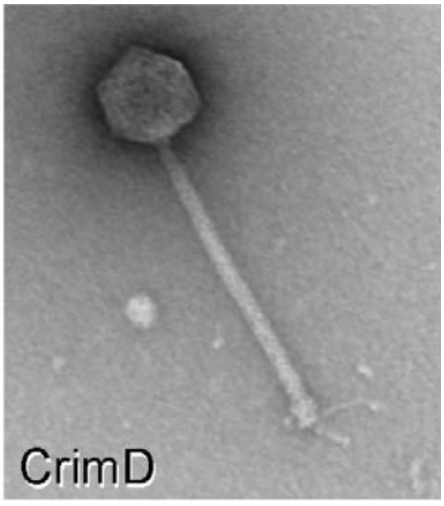

Figure 1.1- Scanning Electron microscope image of mycobacteriophage CrimD.[11]

1.2. *Introduction to Proteomics*

1.2.1 *General introduction to Proteomics*

The complete genome of an organism is not sufficient to fully understand what happens *in vivo*. Depending on the organism, some genes may not be expressed as proteins and/or may be modified in some way from the original genetic code. For this reason, we study the proteome, the whole collection proteins expressed in a cell.[12] By extension we will define proteomics for the purposes of this paper to mean the study of the proteome. Proteomics can take several different forms. A prominent example of this is full proteome discovery where the goal is to determine every protein expressed in a cell at a given time. This technique was only made possible by recent advances in technology such as the ultra-high resolution mass spectrometer(MS).[13] Full proteome discovery is often paired with multiple levels of separation, e.g., SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoreses followed by chromatography), in order to identify the lowest abundance proteins.[13]

Full proteome discovery can be contrasted with targeted proteomics. Targeted proteomics has expanded to encompass a wide variety of techniques but can be roughly thought of as any technique where specific proteins of interest are screened out of the sample. This generally requires greater knowledge of the sample beforehand, as it is not possible to screen for a protein that one does not know exists.[14] Targeted proteomics can be performed in such a way that only a specific protein of interest is analyzed, such as by using SIM (selected ion monitoring) in mass spectrometry, or alternatively can be performed after the primary analysis, where only specific proteins of interest are screened in a much larger sample.[14] The line between these techniques has, in recent

years, been blurred to some extent as it is now possible to perform targeted proteomics on such a variety of proteins as to analyze the whole proteome.[15] Many of these techniques rely heavily, if not entirely, on mass spectrometry, which has become a critical component of modern proteomics.

1.2.2 *Introduction to Mass Spec Based Proteomics*

Mass spectrometry-based proteomics is performed, as the name suggests, with a mass spectrometer (MS) as the detector. Proteins must first be separated, either by PAGE or by chromatographic means and then ionized prior to analysis in order to be identified.

Mass spectrometry allows for the proteins or peptides of interest to be identified by their overall mass to charge ratio (*m/z*) and the *m/z* of their fragments. This technique is termed tandem mass spectrometry. In tandem mass spectrometry, also called MS^n, the analyte is first isolated based on its mass to charge ratio then fragmented by some means and the fragment is isolated and can then, depending on the instrument, be fragmented again.[16]

1.2.3 *Introduction to Top-Down Proteomics*

Mass spectrometry-based proteomics can be roughly divided into two categories, bottom-up proteomics, and top-down proteomics. Bottom-up proteomics involves either the digestion of all the proteins in a sample, termed “shotgun” proteomics, or the digestion of targeted proteins from a gel or other separation followed by the analysis of the peptides. As this is the method that we use, it will be discussed in detail later[17]. In contrast, top-down proteomics involves mass spectrometric analysis of intact proteins usually following

some separation step - either by chromatography or by gel electrophoresis. Top-down proteomics can have many advantages, such as a superior ability to detect post-translational modifications and the preservation of some protein-protein interactions. The primary disadvantage is that very high-resolution MS is required. Top-down proteomics can also be far slower than bottom-up proteomics[18] Top-down proteomics also requires proteins to be fragmented in the gas phase, which can be difficult for many proteins.[17]

1.2.4 *Introduction to Bottom-up Proteomics*

In our lab, we have an LTQ linear ion trap mass spectrometer that does not have the resolution or mass range required for top-down studies. Rather, we use the bottom-up approach in which the protein(s) of interest are first digested into shorter peptides before mass spectrometric analysis. Digestion is accomplished using enzymes, such as trypsin, that cleave the proteins at predictable and consistent sites. For example, trypsin cleaves C-terminal to basic residues Lys and Arg. The resulting peptides are then analyzed and identified using analytical techniques, most commonly high-pressure liquid chromatography in combination with mass spectrometry. In the chromatography step, the peptides are typically separated based on polarity, and then analyzed by MS-MS. Ultimately, each peptide can be identified accurately because small peptides have predictable fragmentation patterns.[19] The experimental fragmentation pattern for each peptide is compared with a computationally predicted fragmentation pattern and a cross correlation “score” is assigned.[17] The proteins are then identified by comparing identified peptide sequences to known sequences from a database file. Mass spectrometry has increased in usefulness in recent years corresponding with the dramatic increase in the power of mass spectrometry.

1.3 Introduction to Mass Spectrometry

Mass spectrometers are powerful tools that allow analytes to be analyzed by their mass to charge ratio (*m/z*). The first MS was developed in 1912 by JJ Thomson, as part of his work on cathode ray tubes. Since then, the complexity and variety of MS has exploded. All MS, even Thomson's original, have the same components; an ionization source to produce gas phase ions out of the analyte, a mass analyzer that separates the ions based on *m/z*, and a detector. In proteomics studies the ionization sources are typically electrospray ionization (ESI), or matrix assisted laser desorption ionization (MALDI) or some variation on those techniques.[17,18]

1.3.1 Electrospray Ionization

While MALDI ionization offers several advantages to ESI, its primary limitation is that it does not couple with HPLC methods for high-throughput analyses. For this reason, ESI is the dominant source for both top-down and bottom-up proteomics studies.[17,18] The principle of ESI was first published in 1990 by John Fenn, and involves a sample in solution containing some adduct, such as protons from an acid, being eluted from a capillary into an electrode with a strong voltage potential applied (on the order of kV) opposite the instrument source, which is usually grounded. As the sample elutes, the voltage potential causes the sample to spray into a fine mist, which is then hit with an inert gas, which causes desolvation of the droplets. The charged particles will migrate due to the electric potential towards the inlet of the instrument. These particles are gas-phase ions and can then be analyzed by MS.[20,21] ESI has the distinct advantage of having a very high ionization efficiency, that is, the ratio of sample in the solution to sample that is ionized is high. ESI also allows for the consistent ionization of very large ions,[21] which

makes it ideal for proteomics as it allows one to confidently analyze the entire sample with minimal missed peptides or proteins.

1.3.2 *Nano-spray ionization*

A commonly used variation of ESI is nano spray ionization (NSI). NSI functions under the same basic principles as ESI, but the flow rates are much lower, typically 2-3 orders of magnitude lower, and no gas flow is used. The capillary inner diameter is much smaller, and comes to a very fine point, on the order of 5-10 microns, the tip of which is called the emitter. The very small emitter causes the spray droplets to be smaller than those formed by ESI, and they can be broken apart by electrostatic forces, hence the lack of a gas flow.[21,22] NSI has several distinct advantages over ESI, apart from the lack of a gas flow. These include a stronger resistance to contamination, that is, adducts of contaminants and analyte molecules are less likely to be seen. Additionally, NSI maintains a very high ionization efficiency, nearly 100%, and a very high portion of those ions can be trapped in the MS yielding a higher dynamic range.[22]

1.3.3 Linear Ion Trap Mass Spectrometers

It is also necessary to introduce linear ion trap mass spectrometry as it is the technique used to perform all of our analyses. A linear ion trap mass spectrometer (LIT) functions much in the same way as a quadrupole mass spectrometer, with some differences. A quadrupole mass spectrometer works by alternating the polarity of four rods creating a “saddle”, a stable region for ions of a certain *m/z*. The voltages and frequencies can be adjusted to change which *m/z* will remain stable within the four poles. The LIT works in a similar way with the addition of end caps which force the ions to move back and forth within the analyzer, “trapping” them.[23] LIT are also referred to as 2D ion

traps as the ions move only in 2D and to differentiate them from the 3D Paul traps. The primary advantage of using a LIT is that it allows for MS^n. MS^n Refers to the ability of a mass spectrometer to fragment the ions then analyze the fragments and then fragment them again to the nth degree of fragmentation and analysis. Theoretically this may be done an unlimited number of times, but in reality, some sample is lost with each level of fragmentation leading to lower and lower signals at higher levels.[23] There are a wide range of activation methods that can be used for fragmentation in LIT, but in our experiments, we use collision induced dissociation (CID), whereby the ions are impacted with a neutral collision gas, in our case helium, which causes the ion to fragment. There are various methods by which these fragments will be analyzed/prioritized, to be discussed below.

1.4 Data Analysis

1.4.1 Data Dependent MS

Two primary methods exist for data collection in bottom-up proteomics, Data dependent analysis (DDA) and data independent analysis (DIA). DIA involves fragmenting every ion seen in an initial scan. Because of the logistics of bottom-up proteomics, these scans must be performed over a limited mass range. This process is repeated until a full picture is gained. DDA works by only fragmenting specific ions within the initial scan, which allows for much faster sampling, at the expense of some trace ions.[24,25] We use DDA in all of our analyses. In DDA the ions to be fragmented can be chosen by a number of means, such as by picking the most abundant ions to fragment. This technique is heavily dependent on the quality of the chromatographic separation, as abundant ions can easily overwhelm the analysis if not properly separated. This can be

improved further by techniques such as dynamic exclusion where ions of a certain mass will not be fragmented again for a set period of time after having first been fragmented.[25] Dynamic exclusion paired with DDA and high-quality chromatography is a very powerful technique for proteomics analysis.

1.4.2 *Proteome Discoverer*

In our experiments we used the software Proteome Discoverer (Thermo Fisher Scientific) for our proteomic analyses. Proteome Discoverer uses FastA files as a "database" of proteins to search against. FastA files are commonly used for DNA code, but Proteome Discoverer uses amino acid sequences instead, and are composed of a simple list containing the names of proteins and the amino acid sequence using the one letter codes. Consequently, Proteome Discoverer will only determine the confidence of proteins that are listed in the FastA. This makes the analysis significantly faster and easier, but means that potential contaminants, or unknown proteins, might be missed and requires that the amino acid sequence of all the proteins be known in advance. Like all bottom-up proteomics software, Proteome Discoverer compares the experimental fragmentation spectra to computationally predicted[19] fragmentation patterns and then generates a confidence for the peptide, which is displayed as a color, with green being high, yellow medium and red low. The peptides are then compared to the amino acid sequence of the proteins in the FastA, which is then used to generate a "score" for the protein. The exact formula used to calculate the score is proprietary, but it takes into account a number of different factors. In addition to peptide confidence the program uses, among other factors which are not displayed, protein coverage and unique peptides in score calculations. Protein coverage is how much of a protein's amino acid sequence was

found in the sample and is expressed as a percent with 100% meaning that the entire amino acid sequence was found in the peptides in the sample. “Unique peptides” refers to peptide sequences that are found in the sample that are unique, or rare, to a certain protein. The presence of these peptides increases the score.

1.5 Proof of concept experiments with T7 and *E. coli*

1.5.1 Summary of Our Previous Proteomics work

Prior to my starting on this project, our lab was able to successfully perform proteomic analyses on bacteriophage infections. These analyses were performed using methods similar to the ones used on our analyses of CrimD, which will be discussed in Chapter 2. The original analyses were performed on *E. coli* infected with bacteriophage T7. T7 genes are either “early” or “late” and we were able to distinguish between these two types. These data show successful chromatography, and it should be noted that different proteins are seen at different times in the table. Because our lab set up has been previously shown to be successful, we believed that we could apply our set up, with significant optimization to a different system: Mycobacteriophage CrimD.

		Accession	Description	Score	Coverage	# Proteins	# UniquePeptides	# Peptides	# PSMs	# AAs	MW [kDa]	calc. pI
1		A0A140N6W0	Elongation factor Tu OS=Escherichia coli (strain B / BL21-...	14.37	44.92 %	1	0	16	21	394	43.3	5.45
2		A0A140NCI6	Elongation factor Tu OS=Escherichia coli (strain B / BL21-...	14.37	44.16 %	1	0	15	20	394	43.3	5.45
3		A0A140N7J1	50S ribosomal protein L2 OS=Escherichia coli (strain B / B...	5.06	25.27 %	1	0	9	10	273	29.8	10.93
4		A0A140N2Z9	30S ribosomal protein S9 OS=Escherichia coli (strain B / B...	4.51	32.31 %	1	2	5	6	130	14.8	10.95
5		A0A140NGG7	30S ribosomal protein S6 OS=Escherichia coli (strain B / B...	4.25	22.14 %	1	1	3	3	131	15.2	5.39
6		A0A140S563	50S ribosomal protein L7/L12 OS=Escherichia coli (strain...	4.03	35.54 %	1	1	4	21	121	12.3	4.65

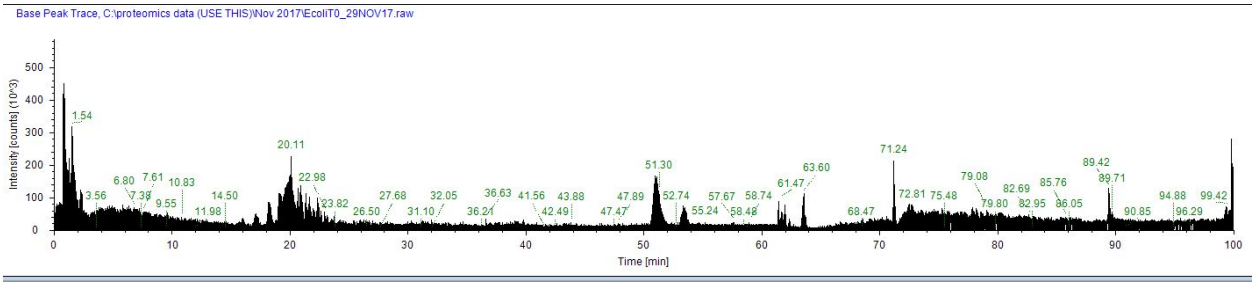

Figure 1.2 Partial output screen from Proteome Discoverer and chromatogram showing results from previous E. coli experiments. Note the tight peaks in the chromatogram, indicating good separation.

Protein	0 min	15 min	30 min	45 min	60 min
Phage shock protein C, PspC OS=Escherichia coli	x				
Protein 19.3 OS=Enterobacteria phage T7	x				
Gene 0.4 protein OS=Enterobacteria phage T7			x		
Protein 19.5 OS=Enterobacteria phage T7			x		
Spanin, outer lipoprotein subunit OS=Enterobacteria phage T7			x		
Bacterial RNA polymerase inhibitor OS=Enterobacteria phage T7				x	
Overcome classical restriction gp0.3 OS=Enterobacteria phage T7				x	
Phage shock operon rhodanese PspE OS=Escherichia coli				x	
Phage tail assembly protein T OS=Escherichia coli				x	
Phage tail protein I OS=Escherichia coli				x	
Protein 1.8 OS=Enterobacteria phage T7				x	
Protein 7.7 OS=Enterobacteria phage T7				x	
Lambda phage regulatory protein CIII OS=Escherichia coli					x
P2 phage tail completion R family protein OS=Escherichia coli					x
Phage shock protein B OS=Escherichia coli					x
Protein 4.1 OS=Enterobacteria phage T7					x
Tail tubular protein gp12 OS=Enterobacteria phage T7					x
Uncharacterized protein 1.1 OS=Enterobacteria phage T7	x	x			
Capsid assembly scaffolding protein OS=Enterobacteria phage T7	x				x
Phage holin, lambda family OS=Escherichia coli		x	x		

Phage shock protein PspD OS=Escherichia coli		x		x	
Putative bacteriophage protein OS=Escherichia coli		x		x	
Tail tubular protein gp11 OS=Enterobacteria phage T7			x	x	
Fusion protein 5.5/5.7 OS=Enterobacteria phage T7			x		x
Phage recombination protein Bet OS=Escherichia coli			x		x
Protein 5.3 OS=Enterobacteria phage T7			x		x
Protein suppressor of silencing OS=Enterobacteria phage T7			x		x
ExoO exonuclease VIII, ds DNA exonuclease OS=Escherichia coli				x	x
Inhibitor of dGTPase OS=Enterobacteria phage T7				x	x
Nucleotide kinase gp1.7 OS=Enterobacteria phage T7				x	x
Phage portal protein, PBSX family OS=Escherichia coli				x	x
Phage regulatory protein N OS=Escherichia coli				x	x
Protein 6.5 OS=Enterobacteria phage T7				x	x
Protein 7.3 OS=Enterobacteria phage T7				x	x
Protein 7 OS=Enterobacteria phage T7	x	x	x		
Protein 1.6 OS=Enterobacteria phage T7	x	x		x	
Protein 6.7 OS=Enterobacteria phage T7	x	x		x	
Protein 4.3 OS=Enterobacteria phage T7	x		x	x	
Phage minor tail protein G OS=Escherichia coli	x		x		x
Probable RecBCD inhibitor gp5.9 OS=Enterobacteria phage T7	x			x	x
Protein 3.8 OS=Enterobacteria phage T7	x			x	x

Single-stranded DNA-binding protein gp2.5 OS=Enterobacteria phage T7	x			x	x
Phage shock protein G OS=Escherichia coli		x		x	x
DNA ligase OS=Enterobacteria phage T7			x	x	x
Endolysin OS=Enterobacteria phage T7			x	x	x
Phage minor tail protein L OS=Escherichia coli			x	x	x
Phage shock protein A, PspA OS=Escherichia coli			x	x	x
Protein kinase 0.7 OS=Enterobacteria phage T7			x	x	x
Terminase, large subunit gp19 OS=Enterobacteria phage T7			x	x	x
Terminase, small subunit gp18 OS=Enterobacteria phage T7	x	x	x	x	
Phage tail tape measure protein, lambda family OS=Escherichia coli	x	x		x	x
Protein 0.6B OS=Enterobacteria phage T7	x	x		x	x
Phage-related tail fibre protein-like protein OS=Escherichia coli		x	x	x	x
Prophage minor tail Z family protein OS=Escherichia coli		x	x	x	x
DNA primase/helicase OS=Enterobacteria phage T7	x	x	x	x	x

Table 1.1 Partial results from earlier analyses showing identification of different proteins in bacteriophage T7 with x representing the protein being scored at that time point. The time points represent time since infection. This table shows the ability to identify viral proteins out of a vast quantity of bacterial proteins. We have improved on these results with our analyses.

1.5.2 Motivation for This Work

Mycobacteria are the causative agents of tuberculosis and Hansen's disease (leprosy) among other infections. Understanding the bacteriophages that infect these bacteria is critical to fully understanding these diseases. Research into mycobacteriophages has already yielded information about the bacteria leading to new

advances in gene induction and other technologies.[2] Having shown that we are able to determine the timing of gene expression in T7 phages after infection of *E. coli* samples, we took this as motivation for us to investigate the gene expression (proteome) of bacteriophage CrimD. We believe that proteomics is the optimal method for studying CrimD as it will allow us to see which proteins are actually expressed in vivo, as opposed to simply studying the genome of CrimD. Chapter 2 describes the experimental methods used in our study of CrimD proteomics and Chapter 3 describes our preliminary results on CrimD protein expression.

Chapter 2: Experimental Methods

2.1 General Workflow

We began our analyses of CrimD with infected mycobacteria samples prepared for us by Prof. Williamson in the Biology Department. We processed these samples by lysing the cells and then fractioning out the proteins. We digested the proteins into peptides and analyzed the samples using high pressure liquid chromatography (HPLC) coupled to MS. The data from these samples were analyzed using Proteome Discoverer as described above.

2.2 HPLC-MS protocols

2.2.1 Overview

In our experiments we used nanoflow high performance liquid chromatography (HPLC) combined with linear ion trap (LIT) mass spectrometry (MS) to perform all of our analyses. These methods were chosen for a combination of availability of resources and maximizing quality. Separation was based on reverse-phase chromatography. Our chromatographic separation was accomplished via gradient HPLC with our solvent gradient changing from a majority mixture of 98:2:0.1 high purity water: acetonitrile: formic acid, hereafter referred to as solvent A, to a majority of 98:2:0.1 acetonitrile: high purity water: formic acid, hereafter referred to as solvent B. Acetonitrile was chosen because it acts as a good non-polar solvent and is readily available, but has a sufficiently high dipole moment to dissolve some ionic or polar compounds. The formic acid was added to protonate the sample and is necessary for ionization using NSI. The mobile phase was transferred via 75 μm inner diameter x 355 μm outer diameter fused silica capillary produced by Polymicro. The autosampler used exclusively solvent A to perform the

sample injections. After each sample injection, we would do a repeat with a blank, usually high purity water, using an identical injection procedure. This was done in order to catch any peptides that were missed in the initial analysis. A further flush was then run in order to guarantee no carryover between injections. Over the course of my work on this project our protocols have changed dramatically.

2.2.2 Previous experimental protocols

2.2.2.1 Analytical columns and NSI emitters

When I began work on this project, all analyses were performed using an Eksigent nano-LC2D nano UHPLC paired with an Exigent autosampler. This HPLC is capable of 2D analysis and we initially used this functionality to transfer the analyte from the autosampler to a 3 cm C-18 guard column, in this case acting as a precolumn produced by Thermo Fisher. The guard column acts as a sort of filter to improve our signal. The column is a short, very non-polar column onto which the analyte is run. The sample is loaded onto the trap column at 100 nL/min flow rate and then a switching valve allows for backflow elution of the sample from the trap column onto the packed analytical column for separation. Use of the trap column has two benefits. First, our sample is compacted into a smaller “packet” for analyses. The smaller the packet that hits the analytical column, the better the separation will be. Secondly, very polar species, including most contaminants are not retained in the guard column, thus cleaning up the sample. Unfortunately, this also can cause a loss of the most polar peptides from the sample. As the ultimate goal is protein identification, losing some of the most polar peptides is an acceptable loss if it ensures that the contaminants are removed.

Our analytical columns were initially purchased commercially; an example includes the EASY-column produced by Thermo Scientific. These columns were made by placing a frit at one end of a glass capillary then filling the capillary with a bead that would act as the stationary phase. The frit is a semi-porous membrane that allows the mobile phase and any analyte contained within to pass through but does not allow the stationary phase to pass through. These columns, though effective, had wildly varying working lives, ranging from weeks to minutes. The inconsistency combined with the monetary cost of the commercial columns meant that their use was not sustainable.

The emitters used in our early experiments were fused silica capillary that had been pulled into the shape of a needle. The needles are typically around 1-2mm long with a final tip diameter of <5 um. Initially, we purchased these tips commercially from *New Objective*. These tips frequently had intermittent flow, that is, the flow during injections would cut out periodically, usually around every minute, before resuming. During the time when the spray was not functioning all signal was effectively lost. These emitters would also frequently clog. A clog manifests as a total lack of flow from the emitter and with a large spike in back pressure, usually on the order of around 2,000 psi, which would cause the pumps to exceed their maximum capacity and proceed to shut down. The needles clogged frequently, but there appears to be no pattern as to when or how the emitters clogged.[26]

Results from experiments that used the Eksigent autosampler and HPLC were inconsistent. Some sample runs would have what appeared to be excellent separation, while runs done with the same column less than 24 hours later would have very poor

separation, with most settings being held constant. I suspect that the cause of this was issues with the autosampler and flow settings of the nano-flow pumps.

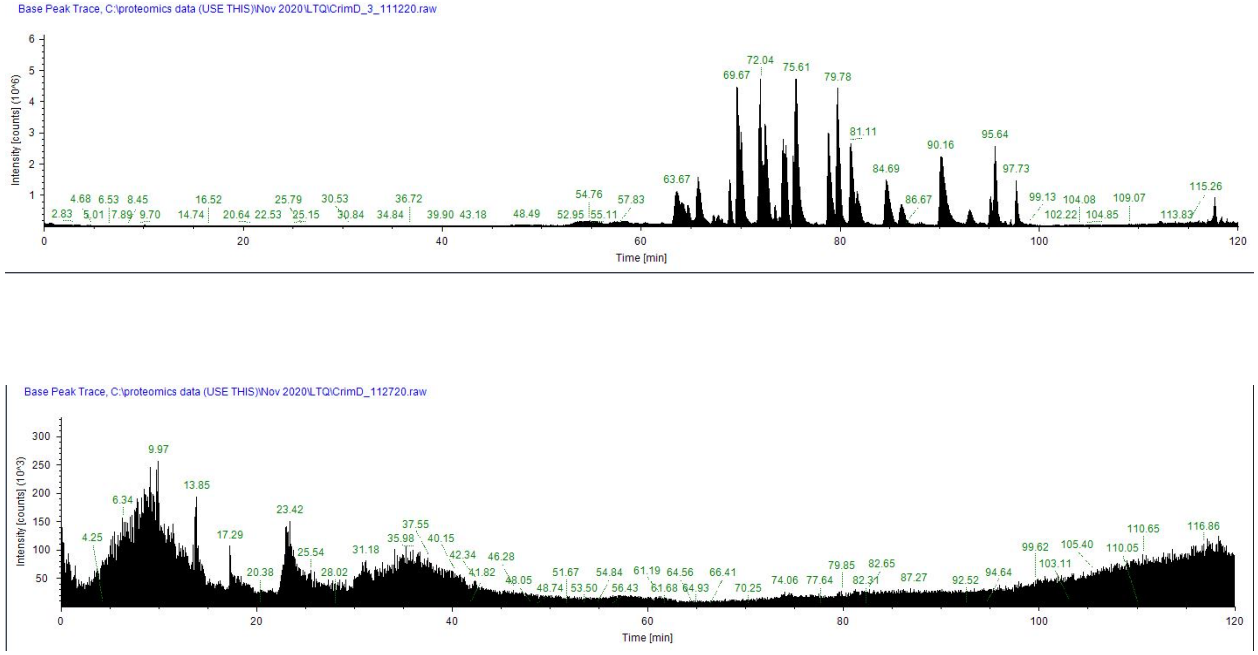

Figure 2.1 a and b 2 Chromatograms from around 2 weeks apart run with the same sample with identical settings, but different columns that were prepared at the same time as each other. Note the clean separation in a, contrasts with b where the separation is basically non-existent. The exact cause of the dramatic changes in quality remains unknown.

2.2.2.2 Loading optimization

Over the course of my work on the project, we attempted several experiments without the guard column. In these experiments the initial signal in the mass spec was very high, but many less-abundant peptides were missed due to saturation with the more abundant peptides. By removing some of the very polar peptides using the guard column, we were better able to analyze less abundant peptides. In addition, the chromatographic separations without the guard column were invariably very poor.

The Eksigent autosampler initially used when I began work on this project had some significant drawbacks. We were ultimately unable to calibrate the flow meters to a

high accuracy. We incorrectly assumed that the lack of accuracy was largely inconsequential as the flows simply needed to be consistent relative to each other, which they were. What was made difficult by this lack of precision was loading the trap columns. The sample loop in the auto sampler had a set volume and transferring that volume onto the autosampler became a difficult balancing act between not loading it onto the column at all, versus running our sample too far on the trap column, which significantly hurt separation. The inconsistent flow rates also caused issues with injection, where multiple injections of a set volume were always the same as each other but were generally not proportional to the set volume. We never successfully loaded the guard column using the Eksigent with any consistency. The Shimadzu HPLC (see section 2.3) system was far more precise with its injections and flow rates, which proved crucial in optimizing our analyses.

2.3 The switch back to microflow-HPLC and ESI

A mechanical failure in the Eksigent autosampler required us to switch to a different HPLC instrument, a Shimadzu LC-20A, which is unfortunately not capable of flow rates on the order of magnitude required for nano-spray ionization. The typical flow rates for this instrument are on the order of microliters/min – mL/min, around 1-3 orders of magnitude higher than the nanoflow HPLC. These flow rates precluded the use of the nanoflow column, without use of additional plumbing. The primary downside of the higher flow was that orders of magnitude more solvent were needed when compared to the nanoflow system, which increases the cost and the environmental impact of our experiments. In addition, we lost the benefit of the aforementioned advantages of NSI. The microflow system did have some inherent benefits, though. The higher flow rate

reduced the risk of columns becoming clogged, which made it economical to purchase larger stainless-steel columns commercially that have a known and tested functionality. Our particular microflow system had the added benefit of integrating a UV-VIS detector and a column oven into the instrument set up. The UV-VIS detector is a form of absorption spectroscopy that allows us to visualize the sample in the column output separately from the MS. This spectroscopy dimension added a level of certainty to our data as we could verify peaks as being present in the MS and UV-VIS, and aided in trouble shooting, as a peak in one detector and not the other was a good indication of a potential issue. In our experiment we monitor absorbance at 254 nm. This number was chosen based on literature values that demonstrated a balance of good absorption of peptides with a low absorption of acetonitrile.[27] The column oven maintained a constant temperature of 40° C on the column. The increased temperature increases the rate at which the sample dissolves into the stationary phase, improving the separation. This system was used initially with an electrospray ionization source (ESI) with a heated probe. The heated source was used to help with the large volume of sheath gas that was required for this flow rate. We initially did not use the guard column arrangement for this instrument and instead injected directly onto the analytical column.

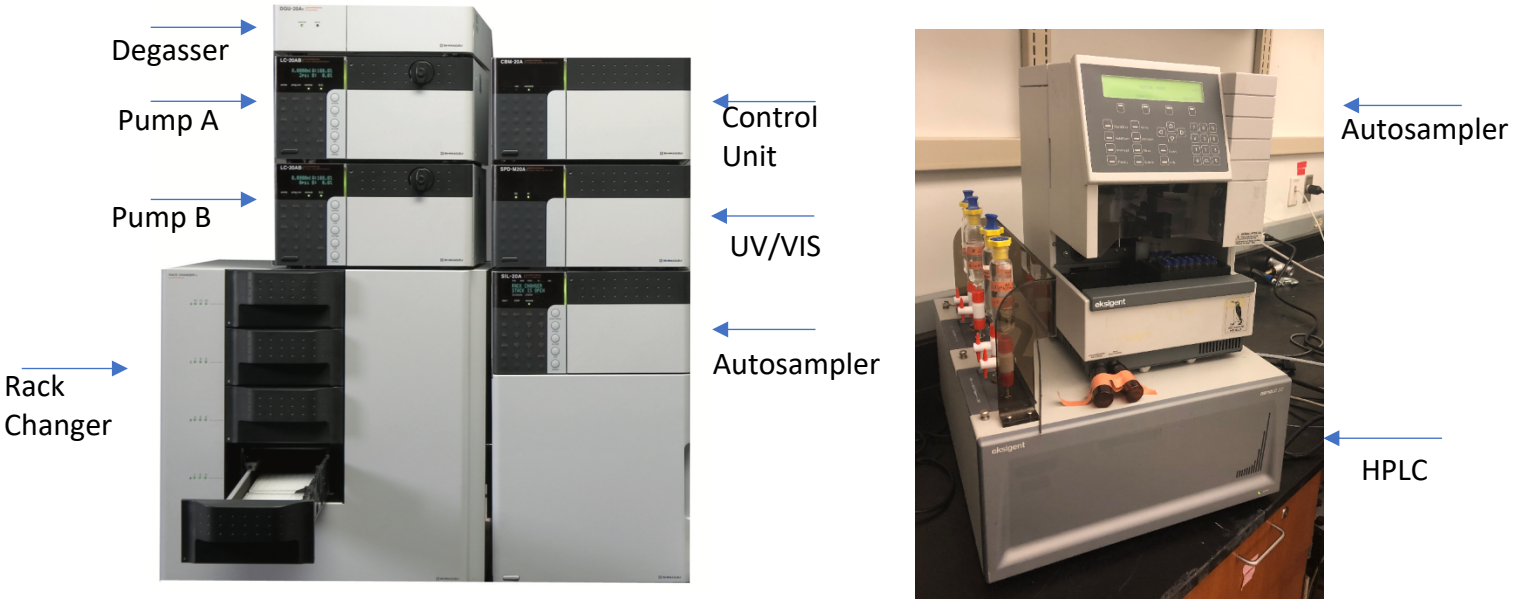

Figure 2.2- Diagrams of both instruments labeled with their different components. Left is the Shimazu LC20A system and the Eksigent nano LC2D is on the right. Not pictured in the Shimadzu diagram is the column oven, additionally the UV/Vis pictured is a different model, though they are superficially identical. The rack changer pictured is present, but not used in our set up. The Eksigent is not modular and hence all pumps/ control system are housed in a single unit.[28,29]

We attempted to perform some test analyses using the microflow system however, all of these experiments presented significant problems. We were able to see a high coverage of the proteins, that is, we could detect a majority of the peptides in a majority of the samples in our test solutions, but the confidence for these peptides and the protein scores were very low. Ultimately, the advantages of NSI required us to modify the Shimadzu plumbing system to allow for the generation of nL/min flow rates while still maintaining the ability to create specific solvent gradients. This change required the use of a flow “splitter”. This device, a model 620 produced by ASI, works by a similar method to an electric resistor. The flow is split into 2 directions and passed through two cartridges that limit the rate of flow through them, which allows for the generation of accurate flow rates in the nL/min range without using nanoflow pumps. It is important to note that the

total solvent used does not change with the splitter. The results from these experiments proved similar to the results we got when we used the nanoflow without the guard column.

2.4 Our current protocols

Over the course of our work, we have changed many of our protocols to optimize them better for this set up. We have faced many disappointments but have arrived at a point where we are able to perform our analyses consistently. We can reproducibly produce tips, columns and digest samples and successfully analyze them with the HPLC-MS.

2.4.1 Biology protocols

2.4.1.1 Cell lysis and protein extraction

Our samples were prepared by inoculating *Mycobacterium smegmatis* with the previously characterized phage, CrimD. The phages were allowed to grow in the sample for a specified amount of time and then were frozen at minus 80° Celsius to arrest growth. The times chosen were 30, 60 and 150 min post infection, and were chosen to help elucidate the timing of the genes as early or late. All of the previous steps were carried out by Professor Kurt Williamson, or by a student working in his lab. The samples were taken from the lab and lysed using the Pierce Mass Spec Sample Prep Kit for Cultured Cells produced by Thermo Fisher [30]. In this method the cells are first pelleted using a centrifuge then lysed by re-suspending the pellet with a detergent-based cell lysis buffer, with TritonX-100 acting as the detergent. The resuspension was incubated at 95^{o} C for 5 min. The nucleic acid components of the cell are sheared using a sonicator produced by Qsonica. This is done in order to reduce the viscosity of the sample. The solution is then centrifuged at 16000 g for 10 min, with the proteins largely remaining in the supernatant.

The concentration of proteins in the supernatant is confirmed by use of a Nanodrop produced by Thermo Scientific

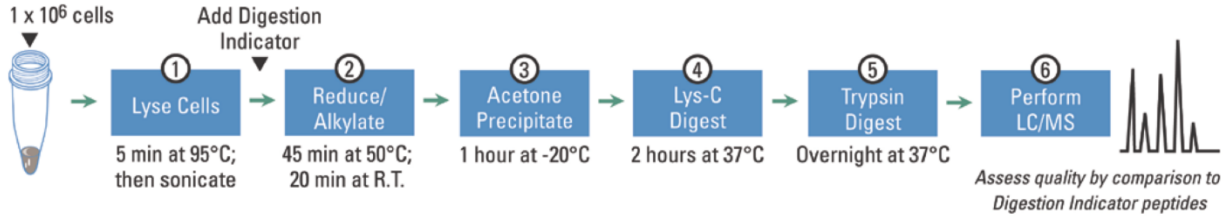

Figure 2.3. Flow-chart illustrating our digestion process from the protocol that was included with our commercial protein digestion kit [30]

2.4.1.2 Protein Digestion

Following protein extraction, the proteins are reduced and alkylated, by use of dithiothreitol (DTT) and iodoacetamide (IAA) respectively, to disrupt the tertiary structure of the proteins. DTT reduces the exposed disulfide bonds formed between cysteine residues that make up the tertiary structure, resulting in thiol groups being formed in place of the sulfide bonds. IAA then alkylates the thiol groups to prevent reformation of the disulfide bonds. This process includes incubating the proteins with DTT for 45 min at 50° C and then incubating with IAA for 20 min at room temperature. IAA is light sensitive, and this reaction is performed in the dark. The proteins are then precipitated out of the digestion buffer by mixing with acetone at -20° C overnight. The proteins are further purified by centrifugation and additional rinses with chilled acetone. The purified proteins are then re-suspended in a digestion buffer and digested first by Lys-C and then by trypsin. The digestion buffer, in combination with temperature, denatures the proteins, which allows enzymatic digestion to progress. Lys-C is a protease that cleaves primarily C-terminal to lysine residues, while trypsin is a protease that cleaves C-terminal to both lysine and arginine. The combination of enzymes is used as it has been found to have a

low number of missed cleavages, that is un-cleaved residues that act as targets of the digestion enzymes.[31] Lys-C digestion takes place at 37° C for 2 hours while the trypsin digestion takes place at 37° C overnight. Digestion is arrested by freezing at -80° C and proteins samples were stored at this temperature until ready for analysis. Before analysis the proteins were concentrated using a Savant DNA 120 SpeedVac Concentrator produced by Thermo Scientific then re-suspended in solvent A, as previously described, and frozen at -20° C in preparation for analysis.

2.4.2 HPLC parameters and gradient

The autosampler is set to inject a user-specified volume and then rinse the needle and purge the lines between each run. This, in addition to the blank runs, helps to minimize any carryover between runs. The primary runs and the blank runs had slightly different gradients. The microflow pumps we set to a total flow rate of 1 ml/min. After the splitter this becomes 1 µl/min (1000 nl/min). The gradient features 95% solvent A for 45 min followed by a ramp down to 55% solvent A for 45 min. This is then followed by a gentle ramp to 30% solvent A for 60 min followed by a sharp ramp to 5% solvent A for 20 min. The column is then washed for 10 min by quickly ramping back up to 95% solvent A. This results in a 180 min run total. The blank runs start with 95% solvent A for 5 min then have a quick ramp to 65% A for 10 min followed by a gentle ramp for 45 min to 30% A and a 5 min ramp to 5% A. The flush concludes with a 15 min wash for a total time of 75 min. The reasoning for the difference in gradients between the blank and sample runs was originally time: blanks were not originally expected to be worth investigating and thus were made shorter to speed up the analysis. When we discovered that the flush did regularly have high confidence proteins, we tried different gradients that all yielded

roughly equivalent results to the ones discussed above. Because of the apparent equivalence we reverted to the original gradients.

2.4.3 Analytical columns

In order to mitigate the cost of purchasing pre-made columns we began manufacturing our own packed analytical columns. We did this by producing a frit similar to that used commercially. Our frits were made in one of two ways. One method includes mixing a potassium silicate (Kasil) mixture produced by PQ Corporation with formamide and dipping a capillary directly into the solution. The capillary was then baked at 500 °C overnight. The resulting frits were often far too long to use and needed to be cut down the ideal length of around 1-3mm, which proved difficult as the frits are all but impossible to see except under a microscope and cutting the capillaries is, at best, an imperfect science. The second method of producing frits involves dropping the Kasil-formamide mixture onto a small piece of filter paper and then aggressively tapping the end of the capillary against the filter paper. Capillary action would then draw some of the mixture into the capillary and would also draw some of the filter paper into the capillary. The frit was then heated briefly, less than 10 s, with either a butane or propane torch. The frits made by this method were of much higher quality than the previous method, but the method had a very low yield. Occasionally as few as 1 in 3 capillaries would actually make a usable frit. Both of the previously described methods were adapted form methods originally developed by the University of Washington Proteomics Resource Center (UWPR).[32] The fritted capillaries were then “packed” using a method also developed by the UWPR. In this method the capillaries are placed in a “bomb”, which is a small brass

container that can be sealed except for two openings for allowing high pressure helium gas into and out of the bomb.

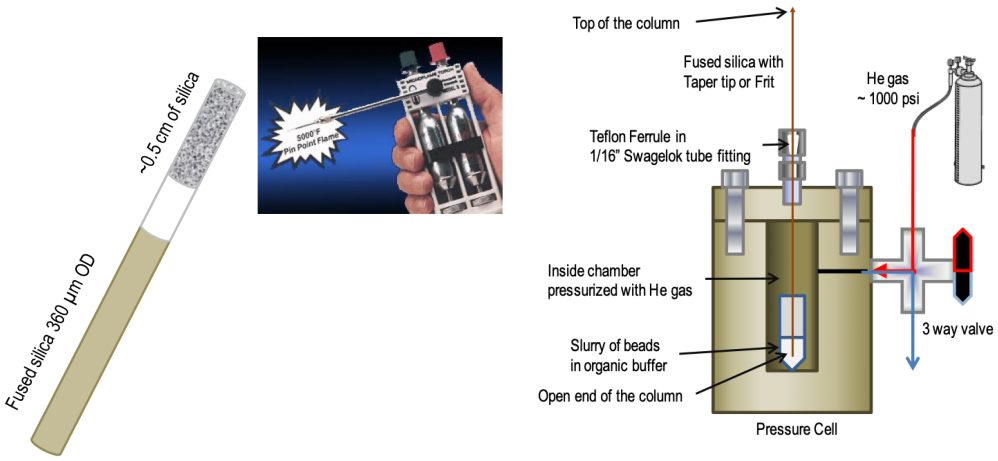

Figure 2.4 Figures from UWPR protocol for producing columns. Left shows the production of a frit and right shows the "bomb" for packing the columns. The flame pictured is different from that used in our protocols but is otherwise the same. The figure on the right closely matches our protocol. [32]

The capillary is placed in one inlet with the end of the capillary opposite the frit placed into an open container full of slurry containing the packing material. The other end is exposed to the air. The second inlet is used to vent in helium under high pressure, around 1,000 Psi. This forces the slurry through the capillary where the packing material is stopped by the frit, but the liquid and helium gas pass through. In our method we use methanol as the liquid in the slurry and our packing material is Zorbax 5-micron C-18 beads, produced by Agilent. Our method can be used, in theory, to produce analytical columns as long as 40 cm [32], but we have been unable to reproducibly make columns longer than ~15 cm. The reason behind this is that the capillaries can become "clogged" during the packing process, and any time after. The exact reasoning for this remains unknown[26], but in our experience longer columns clog more frequently. We have made

various attempts to rescue clogged columns, but to date all have failed, and clogged columns were discarded. The column is then placed onto the nanospray source, made by New Objective and is connected to an emitter.

2.4.4 Nanospray emitters

Due to the costs of the commercial emitters, we purchased a laser pipette puller, a model P-2000F produced by Sutter Instrument Company. This instrument allows us to produce our own emitters in-house at a significantly reduced cost. Initially, our success rate for producing working emitters was as low as 15%, from the initial number of capillaries prepared to the number that functioned as emitters, but we have successfully increased our skill to the point where we can reliably produce working tips consistently.

Figure 2.5 Sutter P2000 pipette puller. The black object in the center of the assembly is the housing for the laser. The mechanism on either side is the arms that hold and pull the capillary into the needle shape.[33]

Pulling our own emitters has allowed us to change our column preparation method as well. This pipette puller works by using a laser to heat up the center of the glass

capillary while applying pressure to either end. The causes the capillary to thin in the middle and eventually separate into two approximately equal emitters. We no longer need to rely on separate fritted columns as we can now produce emitters that have empty capillary of sufficient length behind them to be packed in the same way as a fritted column, thus creating a single packed column/emitter device.

While using the packed emitters dramatically increases the speed and success rate when producing new columns, this switch in protocol has the unfortunate side effect of reducing the working life of our columns. When the emitter clogs the entire column is now unusable. For most of the experiments described, we used columns with the attached emitters, but in an effort to produce longer columns some experiments were attempted with fritted columns. These ultimately failed, and all of the final analyses of the CrimD samples were performed using integrated columns and emitters. The tips and the columns are ultimately identical from one to the other, and the frit is not thought to have any impact on the sample so the two can be thought of as equivalent.

2.4.5 Mass spectrometer parameters

All experiments were performed on a Thermo LTQ XL linear ion trap mass spectrometer. In our experiments we used a spray voltage that was dependent on the needle and source, typical ranges are 1.75- 2.75 kV for the NSI. We used a mass range of 350-2000 *m/z*. We chose this range to eliminate as many contaminants as possible that have lower masses, while maximizing the number of peptides that would be observed. We used a DDA dependent scan system, with an isolation width of 2 *m/z*, where for each full scan the 4 most abundant ions in that scan would be fragmented. The scanning range in the dependent scans was increased to 110-2000 *m/z* as many

fragments had lower masses. They were fragmented by CID with a normalized collision energy of 35 (out of 100), an activation Q of 0.250, and an activation time of 30 ms. We used dynamic exclusion with masses being rejected after being seen three times within 60 seconds. Up to 200 masses can be excluded at any given time and they remain on the exclusion list for 180 seconds. Dynamic exclusion used a scanning width of +/- 1.5 *m/z*. These methods were developed based partially on methods given to us by the University of Arizona Proteomics Lab and were modified based on our specific experimental design.

2.4.6 Proteome Discoverer protocols

For our analyses in Proteome Discoverer, we use a workflow that was originally developed for use in bacteriophage T7 with *E. coli*. The workflow has been slightly modified to accommodate the CrimD analyses. The workflow is configured to work with our specific instrument. Our workflow uses a default minimum precursor mass of 350, and a default maximum of 5000. 350-5000 *m/z* range is larger than the mass range used on the MS in our experiments. This was left bigger as we did not see any harm in keeping the Proteome Discoverer range constant when the MS range was reduced. We worked under a tolerance of +/- 1.5 Da for the precursors and +/- 0.6 Da. These parameters are based off parameters given to us by the University of Arizona and have been slightly modified to optimize results. In our peptide fingerprinting we only looked for b and y fragment ions. This naming system is based on where in the peptide backbone the peptide is fragmented.[34] Ions of b- and y-type correspond to fragmentation on the peptide bond. The letters “b” and “y”” refers to which fragment retains the charge after peptide bond cleavage. We assumed a minimum peptide length of 6 and a maximum of 144.

Proteome discoverer works under the assumption of a single enzymatic digestion, for this reason we use trypsin as it also cleaves at the site preferred by Lys-C. We included 3 dynamic modifications to our peptides; dynamic in this instance means that this particular modification may or may not be present, while static means that modification is always present. These were N-terminal acetylation, C-terminal oxidation and methionine oxidation. We included one static modification: cysteine carbamidomethylation a result of the DTT and IAA processing. These were also based originally on the parameters provided to us by the University of Arizona proteomics lab.

Chapter 3. Results and discussion

I investigated the timing of the expression of gene products in CrimD to determine when each gene is active within the viral replication cycle. CrimD has 95 genes of which 21 have identified functions. We determined the time of expression, relative to the phage's life cycle, of some of these genes.

3.1 Identified proteins from CrimD

3.1.1 Scored Proteins

In total, 75 proteins were successfully identified for the time-values across several analyses. For the purposes of this discussion successfully identified will be defined as having a score of greater than zero in Proteome Discoverer. Due to time constraints, complete statistics have not yet been performed on this data and these values do not have confidence intervals associated with them. More than 20 sample runs were analyzed, but 2 failed to have any proteins with a score. A sample run consists of a primary run where the sample is injected or a flush run immediately after. The flush, or blank, consists of a similar protocol as the sample run with the exception that instead of sample, water was injected.

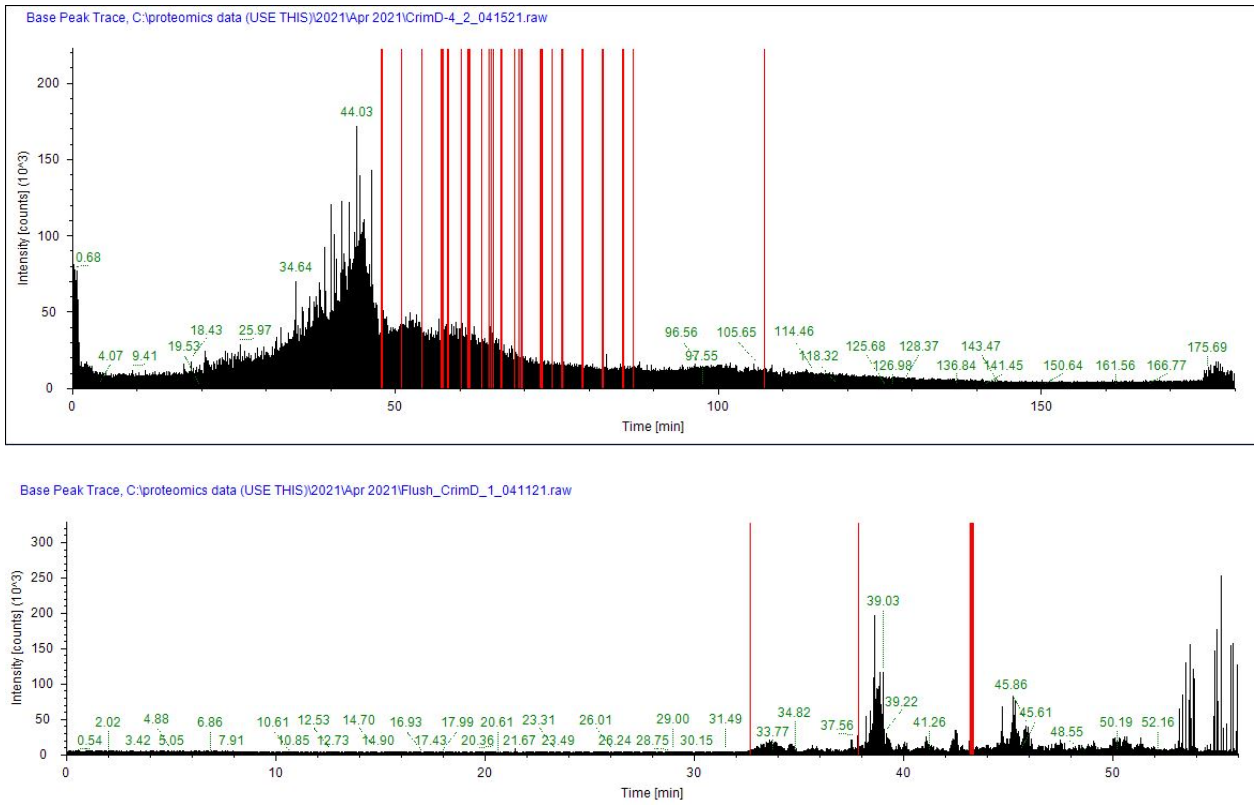

Figure 3.1. Chromatograms from a T30, above, and T150 run, below. The T30 run is a sample run and the T150 is a blank. The red lines represent the presence of a peptide, in this case the most confident peptide from the highest scored protein. Note the tight band in the T150 compared to the more spread out T30. The gaps in the T30 (from 50 min to 90 min) are due to the dynamic exclusion.

It is strongly suggested from our chromatographic data that we are loading too many peptides onto our column, i.e., the concentration of the sample is far too high. This is believed to be the cause of the poor-quality chromatograms during our sample runs but the high concentration of peptides results in high protein scores. The flush runs analyze residual peptides, and therefore have far better chromatograms, but have slightly lower scores due to the lower peptide concentrations. Since we are analyzing both sets of data, we feel confident in the assignment of the proteins from the different time points despite having inefficient chromatographic separation in the primary runs. We are currently

optimizing our sample injection protocols in order to avoid overloading the column but due to time constraints those data are not in this document.

Proteins are either listed as a number, i.e., 24, which represents a protein whose function has yet to be identified, or as the hypothetical function of the protein. Several proteins have the same function, for example the minor tail subunit, for the purposes of clarity we labeled these proteins as a or b.

Protein	T30%	T60 %	T150%	Protein	T30%	T60 %	T150%
HNH endonuclease	9%	25%	0%	CRIMD_48	9%	50%	0%
CRIMD_95	0%	0%	0%	CRIMD_47	0%	0%	0%
CRIMD_94	9%	25%	0%	CRIMD_46	9%	25%	0%
CRIMD_93	0%	25%	0%	CRIMD_45	9%	0%	0%
CRIMD_92	0%	0%	0%	CRIMD_44	0%	0%	0%
CRIMD_91	0%	0%	0%	immunity repressor	0%	0%	40%
CRIMD_90	36%	75%	40%	CRIMD_42	0%	0%	0%
CRIMD_89	18%	0%	20%	integrase	18%	50%	20%
CRIMD_88	9%	0%	0%	CRIMD_40	0%	0%	0%
CRIMD_87	0%	0%	0%	CRIMD_39	9%	25%	20%
CRIMD_86	0%	25%	0%	CRIMD_38	18%	25%	0%
RtcB	18%	50%	0%	CRIMD_37	0%	0%	0%
CRIMD_84	0%	0%	0%	CRIMD_36	18%	25%	40%
CRIMD_83	9%	0%	0%	CRIMD_35	18%	50%	40%
CRIMD_82	9%	25%	0%	CRIMD_34	9%	0%	0%
CRIMD_81	9%	0%	0%	CRIMD_33	36%	25%	0%
CRIMD_80	9%	25%	0%	CRIMD_32	18%	25%	0%
CRIMD_79	0%	0%	0%	lysin B	36%	50%	20%
CRIMD_78	9%	25%	0%	lysin A	0%	25%	80%
CRIMD_77	0%	25%	0%	CRIMD_29	9%	75%	40%
CRIMD_76	0%	25%	0%	CRIMD_28	18%	25%	0%
peptidase	18%	0%	0%	minor tail subunit_a	18%	0%	0%
CRIMD_74	64%	75%	40%	CRIMD_26	0%	0%	20%
CRIMD_73	9%	25%	0%	minor tail subunit_b	0%	0%	20%
RusA	0%	50%	0%	CRIMD_24	9%	50%	0%
DNA primase/helicase	18%	100%	20%	minor tail subunit_c	18%	0%	20%
CRIMD_70	27%	25%	40%	minor tail subunit_d	36%	25%	60%

CRIMD_69	0%	0%	0%	tapemeasure protein	55%	75%	100%
NrdH	0%	25%	20%	tail assembly chaperone_a	0%	25%	0%
CRIMD_67	27%	25%	40%	tail assembly chaperone_b	0%	25%	0%
CRIMD_66	0%	0%	0%	major tail subunit	27%	25%	60%
CRIMD_65	0%	0%	0%	tail terminator	0%	25%	0%
CRIMD_64	9%	0%	20%	CRIMD_16	0%	0%	20%
CRIMD_63	0%	0%	0%	head-to-tail stopper	0%	0%	0%
CRIMD_62	0%	25%	0%	head-to-tail adaptor	18%	25%	0%
CRIMD_61	0%	25%	0%	major capsid protein	18%	50%	80%
CRIMD_60	9%	0%	0%	scaffolding protein	0%	25%	20%
CRIMD_59	9%	0%	40%	CRIMD_11	9%	0%	0%
CRIMD_58	45%	25%	0%	capsid maturation protease	73%	50%	60%
DnaQ	0%	0%	20%	portal protein	9%	0%	20%
CRIMD_56	0%	0%	0%	terminase	18%	0%	20%
CRIMD_55	0%	0%	0%	CRIMD_7	0%	0%	0%
CRIMD_54	0%	25%	0%	CRIMD_6	0%	0%	0%
CRIMD_53	0%	0%	0%	CRIMD_4	0%	0%	0%
CRIMD_52	0%	0%	0%	CRIMD_3	9%	0%	0%
CRIMD_51	0%	0%	20%	CRIMD_2	9%	50%	0%
WhiB	0%	0%	20%	CRIMD_1	0%	0%	0%
CRIMD_49	9%	25%	80%				

Table 3.1 showing proteins found in the three time intervals. The percent refers to what percent of runs had the given protein, CrimD_xx denotes protein without known function, and the number represents the gene number as it is listed in the genome, posted on PhageDB.org

Table 3.1 shows all of the proteins found in our analyses. The major capsid protein was the highest scored protein in the majority of the T150 runs and was scored much lower in the T30 and T60 runs. The Capsid Maturation protease/ MuF like Minor capsid subunit (CMP/MMCS) is a protein whose gene belongs to a phamily, a group of related

phage genes, which serve one of those two functions. CMP/MMCS was the highest scored protein in a majority (6/11) of the T30 runs and one of the T150 runs. The T150 and T60 runs had consistently higher scores on average for the most abundant proteins, this was mostly due to the blanks from the T30 runs having very low scores overall. The T150 and T60 runs consistently had higher scores in the blank run than in the sample runs.

3.2 Discussion

In analyzing these data, it is important to note that these samples contain millions of cells and potentially billions of proteins. The replication cycle for the phages infecting these cells is not perfectly uniform. At any given moment some cells might be uninfected, while others are in the lytic phase, though efforts were made to minimize this. The data are collected from the whole culture and therefore any variations in cell growth will be seen in the data. We believe that this is the reason that some proteins that are typically involved in the lytic replication cycle are being seen, albeit with low scores and not consistently, in the T30 runs.

Capsid maturation proteins are proteins that cleave immature capsid proteins to yield mature capsid proteins, while MuF-like proteins are a type of toxin.[35] MuF proteins are often included in the capsid and injected into the cell with the viral DNA, but their exact purpose is not fully understood.[35] One of the proposed functions of MuF toxins is to inhibit cellular growth in the surrounding bacteria. Because CMP/MMCP is incorporated into the capsid, its higher scores at the T150 time sample are also unsurprising.

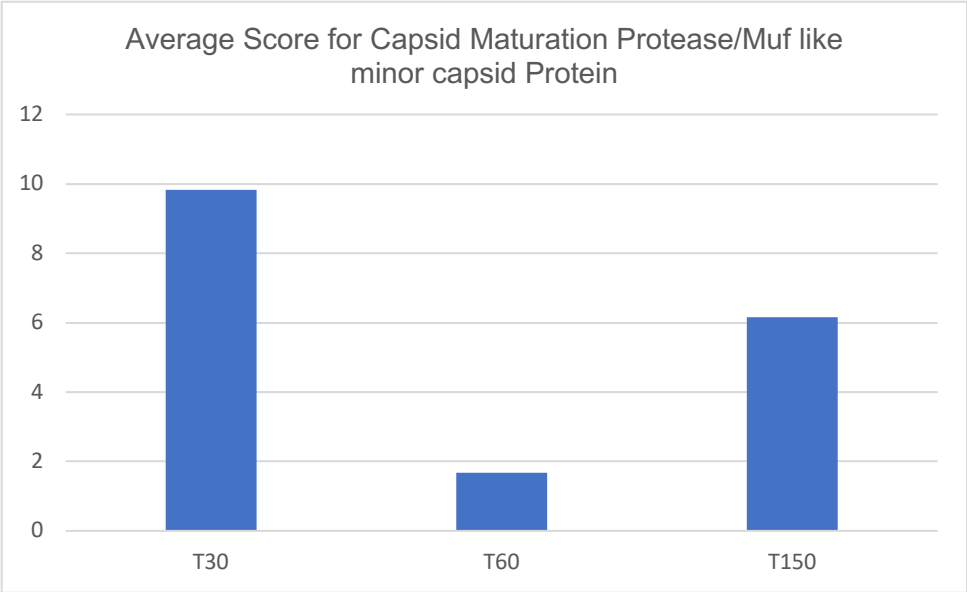

Figure 3.2 Graph comparing the average score of the CMP/MMCP in the different time points. Runs that had at least one scored protein where CMP/MMCP was not scored were counted as a 0.

We also suspect protein Unknown 74 to be involved in the lysogenic replication cycle as it is also present early in the viral replication cycle. Integrase, which is responsible for integrating the phage genome into the host genome, is also related to lysogenic replication. Another protein of note is the RtcB protein found in the T30 and T60 runs. This protein functions to circularize RNA molecules by linking the 5' and 3' ends together. This functionality plays a role in the synthesis of tRNAs.[36] Its presence in the earlier time sample is consistent with its functionality and we suspect it is involved in the lytic replication cycle.

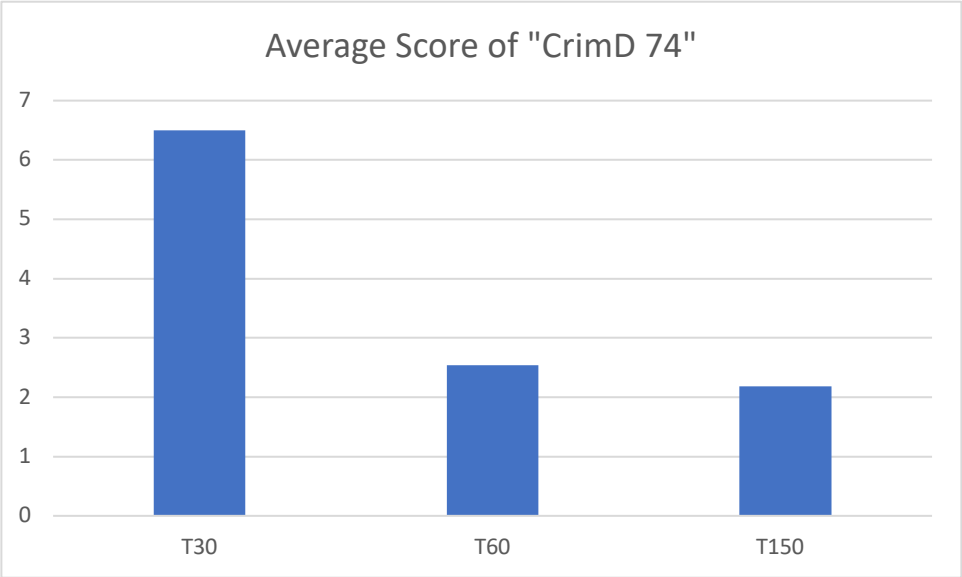

Figure 3.3 Graph comparing the average score of the CrimD 74 in the different time points. Runs that had at least one scored protein where CrimD 74 was not scored were counted as a 0. CrimD 74 is suspected to be involved in the lysogenic replication cycle based on its higher score in the T30 runs. Of note though, CrimD 74 was present in a higher portion of T60 runs than T30 runs (75% vs 64%).

The presence of the major capsid protein and Lysin A in the T150 samples strongly suggest that the phage has entered the late lytic part of its replication cycle. This would suggest that the proteins that are predominantly in the T150 samples are primarily involved in the later lytic replication cycle. The major capsid protein was the dominant protein in the T150 samples and was also scored far lower, when scored at all in the other runs. Major capsid protein is, of course, the major component of the viral capsid and during lytic replication it would become one of the most produced proteins in the cell in preparation for cell lysis.

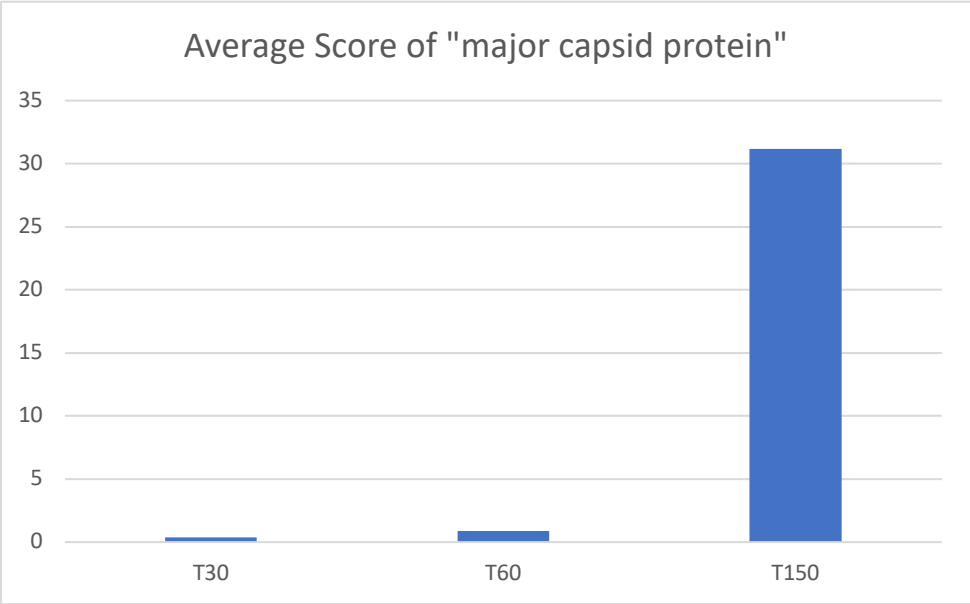

Figure 3.4 Graph comparing the average score of the major capsid protein in the different time points. Runs that had at least one scored protein where major capsid protein was not scored were counted as a 0. The major capsid protein is involved in the lysogenic replication cycle and this graph suggests that the vast majority of cells are in the lytic replication cycle at the T150 time point while far fewer are lytic at T30 or T60.

Based on the proteins observed in the T60 time point it is less obvious which phase of phage replication a majority of the cells are in. The high scores of the viral DNA helicase in the T60 time point, in the graph below, combined with its ubiquity, it was present in every T60 run, suggests that the virus has begun to replcate its genome. Viral helicases are a very common feature of DNA viruses and some RNA viruses. They play important roles in regulating the viral replication cycle, and significant diversity exists between different viral helicases, though many do share a common origin.[37]

CrimD DNA primase/helicase is likely involved in the replication of the viral genome. Because the lysis proteins and capsid proteins are not as highly scored or widely present in T60 as opposed to T150 we believe that the CrimD is in the early lytic replication phase during the T60 time sample. During this phase the virus predominantly

replicates its genome while lysing the host genome. We believe that any unknown proteins predominantly present in the T60 time samples, such as CrimD 90 or 48, are proteins involved in DNA replication

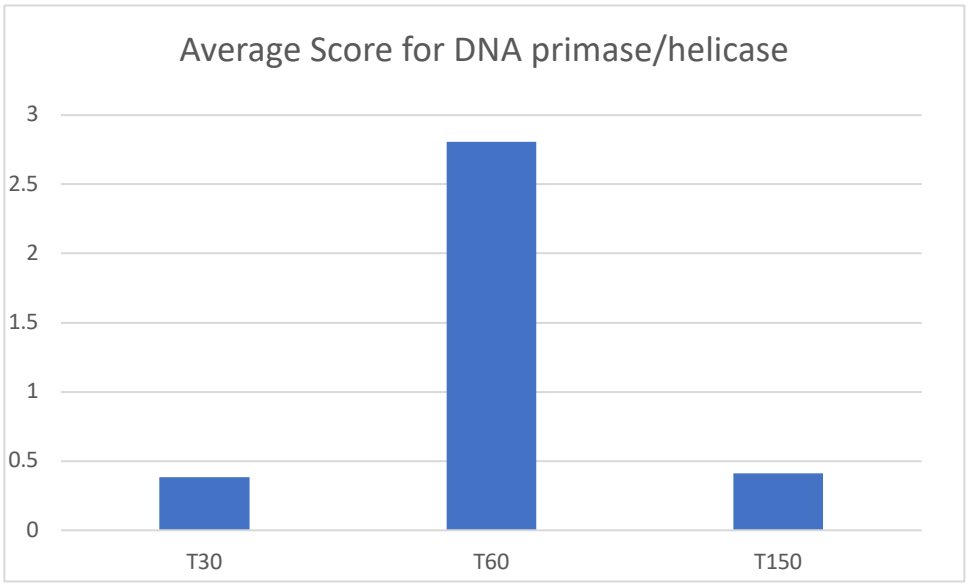

Figure 3.5 Graph comparing the average score of DNA helicase in the different time points. Runs that had at least one scored protein where DNA helicase was not scored were counted as a 0. The DNA primase/helicase is involved in viral genome replication. We believe that its presence predominantly in T60 suggests that CrimD is in the early lytic phase of its replication cycle.

3.3 Conclusions

Here we have described the proteomics analysis of mycobacteriophage CrimD. CrimD was previously characterized as part of the William & Mary Phage Lab. We took the next logical step of full proteomic analysis. Although we have been able to perform proteomic analyses on bacteriophages previously, our set up required much optimization to be able to function sufficiently well to perform these analyses. We analyzed three time points in the CrimD life cycle measured from time of infection. These time points correspond to Early, Middle and Late genes respectively. We identified 75 bacteriophage proteins and determined in which time points each of those proteins was expressed. We

believe that the next step in the work is bioinformatics to begin to assign functions to some of the unknown proteins and further characterize mycobacteriophage CrimD.

References

(1) Wommack, K. E.; Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiol. Mol. Biol. Rev.* **2000**, *64* (1), 69–114. https://doi.org/10.1128/mmbr.64.1.69-114.2000.

(2) Pedulla, M. L.; Ford, M. E.; Houtz, J. M.; Karthikeyan, T.; Wadsworth, C.; Lewis, J. A.; Jacobs-Sera, D.; Falbo, J.; Gross, J.; Pannunzio, N. R.; Brucker, W.; Kumar, V.; Kandasamy, J.; Keenan, L.; Bardarov, S.; Kriakov, J.; Lawrence, J. G.; Jacobs, W. R.; Hendrix, R. W.; Hatfull, G. F. Origins of Highly Mosaic Mycobacteriophage Genomes. *Cell* **2003**, *113* (2), 171–182. https://doi.org/10.1016/S0092-8674(03)00233-2.

(3) Broxmeyer, L.; Sosnowska, D.; Miltner, E.; Chacoón, O.; Wagner, D.; Mc Garvey, J.; Barletta, R. G.; Bermudez, L. E. Killing of Mycobacterium Avium and Mycobacterium Tuberculosis by a Mycobacteriophage Delivered by a Nonvirulent Mycobacterium: A Model for Phage Therapy of Intracellular Bacterial Pathogens. *J. Infect. Dis.* **2002**, *186* (8), 1155–1160. https://doi.org/10.1086/343812.

(4) Verity, R.; Okell, L. C.; Dorigatti, I.; Winskill, P.; Whittaker, C.; Imai, N.; Cuomo-Dannenburg, G.; Thompson, H.; Walker, P. G. T.; Fu, H.; Dighe, A.; Griffin, J. T.; Baguelin, M.; Bhatia, S.; Boonyasiri, A.; Cori, A.; Cucunubá, Z.; FitzJohn, R.; Gaythorpe, K.; Green, W.; Hamlet, A.; Hinsley, W.; Laydon, D.; Nedjati-Gilani, G.; Riley, S.; van Elsland, S.; Volz, E.; Wang, H.; Wang, Y.; Xi, X.; Donnelly, C. A.; Ghani, A. C.; Ferguson, N. M. Estimates of the Severity of Coronavirus Disease 2019: A Model-Based Analysis. *Lancet Infect. Dis.* **2020**, *20* (6), 669–677. https://doi.org/10.1016/S1473-3099(20)30243-7.

(5) Yin, Y.; Fischer, D. Identification and Investigation of ORFans in the Viral World. *BMC Genomics* **2008**, *9*, 1–10. https://doi.org/10.1186/1471-2164-9-24.

(6) Yoshimoto, F. K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. *Protein J.* **2020**, *39* (3), 198–216. https://doi.org/10.1007/s10930-020-09901-4.

(7) Acheson, N. *Fundamentals of Molecular Virology*, 2nd ed.; Wiley, 2011.

(8) Mycobacterium phage CrimD https://phagesdb.org/phages/CrimD/.

(9) Science Education Alliance : The Phage Lab https://hhmi.wm.edu/phage_lab.html.

(10) Maniloff, J.; Ackermann, H. W. Taxonomy of Bacterial Viruses: Establishment of Tailed Virus Genera and the Order Caudovirales. *Arch. Virol.* **1998**, *143* (10), 2051–2063. https://doi.org/10.1007/s007050050442.

(11) Pope, W. H.; Ferreira, C. M.; Jacobs-Sera, D.; Benjamin, R. C.; Davis, A. J.; DeJong, R. J.; Elgin, S. C. R.; Guilfoile, F. R.; Forsyth, M. H.; Harris, A. D.; Harvey, S. E.; Hughes, L. E.; Hynes, P. M.; Jackson, A. S.; Jalal, M. D.; MacMurray, E. A.; Manley, C. M.; McDonough, M. J.; Mosier, J. L.; Osterbann, L. J.; Rabinowitz, H. S.; Rhyan, C. N.; Russell, D. A.; Saha, M. S.; Shaffer, C. D.; Simon, S. E.; Sims, E. F.; Tovar, I. G.; Weisser, E. G.; Wertz, J. T.; Weston-Hafer, K. A.; Williamson, K. E.; Zhang, B.; Cresawn, S. G.; Jain, P.; Piuri, M.; Jacobs, W. R.; Hendrix, R. W.; Hatfull, G. F. Cluster k Mycobacteriophages: Insights into the Evolutionary Origins of Mycobacteriophage Tm4. *PLoS One* **2011**, *6* (10), 1–22. https://doi.org/10.1371/journal.pone.0026750.

(12) Banks, R. E.; Dunn, M. J.; Hochstrasser, D. F.; Sanchez, J. C.; Blackstock, W.;

Pappin, D. J.; Selby, P. J. Proteomics: New Perspectives, New Biomedical Opportunities. *Lancet* **2000**, *356* (9243), 1749–1756. https://doi.org/10.1016/S0140-6736(00)03214-1.
(13) de Godoy, L. M. F.; Olsen, J. V.; de Souza, G. A.; Li, G.; Mortensen, P.; Mann, M. Status of Complete Proteome Analysis by Mass Spectrometry: SILAC Labeled Yeast as a Model System. *Genome Biol.* **2006**, *7* (6), 1–15. https://doi.org/10.1186/gb-2006-7-6-r50.
(14) Borràs, E.; Sabidó, E. What Is Targeted Proteomics? A Concise Revision of Targeted Acquisition and Targeted Data Analysis in Mass Spectrometry. *Proteomics* **2017**, *17* (17–18), 17–18. https://doi.org/10.1002/pmic.201700180.
(15) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. Full Dynamic Range Proteome Analysis of S. Cerevisiae by Targeted Proteomics. *Cell* **2009**, *138* (4), 795–806. https://doi.org/10.1016/j.cell.2009.05.051.
(16) Skeie, J. M.; Roybal, C. N.; Mahajan, V. B. Proteomic Insight into the Molecular Function of the Vitreous. *PLoS One* **2015**, *10* (5), 1–19. https://doi.org/10.1371/journal.pone.0127567.
(17) Zhang, Y.; Fonslow, B.; Shan, B.; Baek, M.-C.; Yates III, J. Protein Analysis by Shotgun Proteomics. *Chem. Rev.* **2013**, *113*, 2343–2394. https://doi.org/10.1002/9781118970195.ch1.
(18) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519. https://doi.org/10.1146/annurev-anchem-071015-041550.
(19) Frank, A. M. Predicting Intensity Ranks of Peptide Fragment Ions. *J Proteome Res* **2009**, *8* (5), 2226–2240. https://doi.org/10.1021/pr800677f.Predicting.
(20) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F. Electrospray Ionization-Principles and Practice. *Mass Spectromety Rev.* **1990**, *9*, 37–70.
(21) Wilm, M. Principles of Electrospray Ionization. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009407. https://doi.org/10.1074/mcp.M111.009407.
(22) Juraschek, R.; Dulcks, T.; Karas, M. Nanoelectrospray — More Than Just A. *J. Am. Soc. Mass Spectrmetry* **1999**, *0305* (98), 300–308.
(23) Douglas, D. J.; Frank, A. J.; Mao, D. Linear Ion Traps in Mass Spectrometry. *Mass Spectrom. Rev.* **2005**, *24* (1), 1–29. https://doi.org/10.1002/mas.20004.
(24) Guo, J.; Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **2020**, *92* (12), 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.
(25) Guan, S.; Taylor, P. P.; Han, Z.; Moran, M. F.; Ma, B. Data Dependent-Independent Acquisition (DDIA) Proteomics. *J. Proteome Res.* **2020**, *19* (8), 3230–3237. https://doi.org/10.1021/acs.jproteome.0c00186.
(26) Noga, M.; Sucharski, F.; Suder, P.; Silberring, J. A Practical Guide to Nano-LC Troubleshooting. *J. Sep. Sci.* **2007**, *30* (14), 2179–2189. https://doi.org/10.1002/jssc.200700225.
(27) Antosiewicz, J. M.; Shugar, D. UV–Vis Spectroscopy of Tyrosine Side-Groups in Studies of Protein Structure. Part 2: Selected Applications. *Biophys. Rev.* **2016**, *8* (2), 163–177. https://doi.org/10.1007/s12551-016-0197-7.
(28) *Shimadzu LC20 Series*; Shimazu.

(29) Moeller, W. *The Eksigent in the Afternoon*; USA, 2021.
(30) INSTRUCTIONS PierceTM Mass Spec Sample Prep Kit for Cultured Cells https://assets.fishersci.com/TFS-Assets/LSG/manuals/MAN0011864_Pierce_MassSpecSamplePrep_CulturedCells_UG.pdf.
(31) Glatter, T.; Ludwig, C.; Ahrné, E.; Aebersold, R.; Heck, A. J. R.; Schmidt, A. Large-Scale Quantitative Assessment of Different in-Solution Protein Digestion Protocols Reveals Superior Cleavage Efficiency of Tandem Lys-C/Trypsin Proteolysis over Trypsin Digestion. *J. Proteome Res.* **2012**, *11* (11), 5145–5156. https://doi.org/10.1021/pr300273g.
(32) UWPR. Packing Capillary Columns and Pre-Columns (Traps). UWPR 2019.
(33) Sutter P2000 https://www.sutter.com/productLG/P-2000_lg.jpg.
(34) Wysocki, V. H.; Resing, K. A.; Zhang, Q.; Cheng, G. Mass Spectrometry of Peptides and Proteins. *Methods* **2005**, *35* (3 SPEC.ISS.), 211–222. https://doi.org/10.1016/j.ymeth.2004.08.013.
(35) Jamet, A.; Touchon, M.; Ribeiro-Gonçalves, B.; Carriço, J. A.; Charbit, A.; Nassif, X.; Ramirez, M.; Rocha, E. P. C. A Widespread Family of Polymorphic Toxins Encoded by Temperate Phages. *BMC Biol.* **2017**, *15* (1), 1–12. https://doi.org/10.1186/s12915-017-0415-1.
(36) Chakravarty, A. K.; Subbotin, R.; Chait, B. T.; Shuman, S. RNA Ligase RtcB Splices 3′-Phosphate and 5′-OH Ends via Covalent RtcB-(Histidinyl)-GMP and Polynucleotide-(3′)Pp(5′)G Intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (16), 6072–6077. https://doi.org/10.1073/pnas.1201207109.
(37) Frick, D. N.; Lam, A. M. I. Understanding Helicases as a Means of Virus Control D. *Curr Pharm Des* **2006**, *12* (11), 1315–1338.